Water Quality Prediction Over Accuracy Using Machine Learning Algorithms

A PROJECT REPORT

Submitted to

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

In partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING
IN COMPUTER SCIENCE ENGINEERING

By

K.SREE DURGA GEETHIKA-192111472

Supervisor

DR.JOSH KUMAR.J.P

SIMATS ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL

SCIENCES, CHENNAI – 602 1

SIMATS ENGINEERING
Engineer to Excel

SIMATS
Saveetha Institute of Medical And Technical Sciences
(Declared as Deemed to be University under Section 3 of UGC Act 1956)

SIMATS ENGINEERING
SAVEETHA INSTITUTE OF MEDICAL AND
TECHNICAL SCIENCES
CHENNAI – 602105

BONAFIDE CERTIFICATE

Certified that this project report **"Water Quality Predictions Over Accuracy Using Machine Learning Algorithms"** is the Bonafide work of K.Sree Durga Geethika 192111472 who carried out the project work under my supervision.

DR. S.ANUSUYA

PROGRAMME DIRECTOR

Professor

Department of CSE

SIMATS Engineering

Saveetha Institute of Medical and Technical Sciences

Chennai – 602 105

DR.JOSH KUMAR.J.P

SUPERVISOR

Professor

Department of CSE

SIMATS Engineering

Saveetha Institute of Medical and Technical Sciences

Chennai – 602 105

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives me immense pleasure to express my profound gratitude to our Honorable Chancellor **Dr. N. M. Veeraiyan**, Saveetha Institute of Medical and Technical Sciences, for his blessings and for being a source of inspiration. I sincerely thank our Pro Chancellor **Dr. Deepak Nallaswamy,** SIMATS, for his visionary thoughts and support. I am indebted to extend my gratitude to our Director **Dr. Ramya Deepak,** SIMATS Engineering, for facilitating us all the facilities and extended support to gain valuable education and learning experience.

I register my special thanks to **Dr. B. Ramesh,** Principal, SIMATS Engineering and **Dr.S.ANUSUYA ,** Programme Director, Institute of Computer Science Engineering, for the support given to me in the successful conduct of this project. I wish to express my sincere gratitude to my supervisor **DR.JOSH KUMAR.J.P**, for his inspiring guidance, personal involvement and constant encouragement during the entire course of this work.

I am grateful to Project Coordinators, Review Panel External and Internal Members and the entire faculty of the Institute of Computer Science Engineering, for their constructive criticisms and valuable suggestions which have been a rich source to improve the quality of this work.

K.SREE DURGA GEETHIKA-192111472

TABLE OF CONTENTS

RESEARCH PAPER 1

PERFORMANCE OF LINEAR REGRESSION WITH DECISION TREE TO MONITOR AND PREDICT WATER QUALITY IN SURFACE WATER FOR ENVIRONMENTAL CONSERVATION

ABSTRACT

AIM: The aim of this research is to evaluate and optimize the efficacy of decision tree algorithms in combination with linear regression for the purpose of monitoring and forecasting water quality in diverse aquatic environments. **Materials and methods:** The study uses two study groups a Linear regression and Decision tree approach for 40 samples in total. A clinical was used to calculate sample size with parameters preset G-power of 0.8 or 80%, an alpha of 0.05 or ($P<0.05$) and confidence intervals of 95%. This analysis was performed in SPSS software. **Result:** Both algorithms demonstrated proficiency in analyzing water quality data, accurately identifying instances where water conditions exceeded acceptable standards. The Outcome demonstrate that in comparison to the linear regression algorithm's (62.71%) accuracy, the decision tree algorithm's (64.0%) has more enhanced accuracy in predicting water quality. **Conclusion:** The accuracy rate of decision tree (64.0%) is noticeably greater than the accuracy of Linear regression (62.71%).

INTRODUCTION:

One resource that is necessary for human survival is water. One of the most significant factors in our lives is the quality of the water. Most important inland water resources are Lakes and reservoirs that are essential to industry, the environment, and public health. Fresh water supplies from lakes and reservoirs are necessary for hydropower and agricultural irrigation.

There are four categories for water quality: excellent condition, moderately contaminated, contaminated, and highly contaminated. Understanding the classification of water is crucial for handling and applying it correctly. The term "water quality" refers to the physical, chemical, and biological properties of the water. Because the quality status must be accurately classified, two classification algorithms are used: decision trees and linear regression.

The study investigates the performance of linear regression and decision tree algorithms for the purpose of monitoring and forecasting water quality in various bodies of water. This research aims to offer significant insights for environmental conservation efforts by providing a sound framework for assessing and forecasting water quality conditions.

A method for actively instructing computers or other machines is called machine learning. Various machine learning methods use gathered datasets to produce effective ensemble and classification models. Predicting the quality of water can be done with the help of such data. A range of machine learning techniques can be used to make predictions, but it can be challenging to determine which one performs best.

Our study aims to identify water quality accuracy through the application of linear regression and decision trees. In this work, a number of machine learning classification algorithms are employed and assessed on the dataset to predict water quality accuracy, including Naive Bayes (NB), K Nearest Neighbor (KNN), Decision Tree (DT), and Neural Network (NN) with different hidden layers.

Water quality prediction is one prominent area where the use of machine learning models has gained significant traction in recent years, among many other fields. For the management and conservation of water resources to be sustainable, the capacity to forecast water quality is essential. The objective of this research is to improve our comprehension of the intricate dynamics affecting water conditions by investigating the application of cutting-edge machine learning techniques for water quality prediction. We use kaggle dataset for this purpose, and we predict water quality using different machine learning algorithms and ensemble techniques.

Our goal is to create a dependable system that can forecast parameters related to water quality in different environments by utilizing the power of predictive models. The incorporation of machine learning algorithms presents an opportunity for instantaneous monitoring and pre-emptive identification of possible problems with water quality. This research contributes to the broader goal of implementing proactive measures for the preservation of water ecosystems and the safeguarding of public health.

You can find a lot of research predicting water quality in respected publications like Science, IEEE and google scholar. The IEEE digital library offers access to 45 journals, 362 publications on Google scholar, and 253 publications on springer. Similarly the integrations of diverse algorithms allows for a more nuanced understanding of the complex interactions influencing water quality dynamics. (Qiao.z,2021).

This integrated methodology aims to enhance the accuracy and interpretability of water quality predictions, contributing to effective environmental efforts. (Jiang q.o,2021).

Understanding and predicting environmental conditions could be made easier with the use of decision tree and linear regression algorithms in the investigation of water quality prediction. The pursuit of precise forecasts is essential for well-informed decision-making in water quality management, notwithstanding possible obstacles connected to data volume, diversity, and algorithmic complexity. This study aims to generate important insights into the dynamic field of water quality prediction, promoting a better understanding of environmental factors and improving our capacity to protect and manage water resources through methodical experimentation and careful algorithmic optimization. (Kalia.A,2023).

MATERIALS AND METHODS

The research is carried out in Machine learning laboratory lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The water potability dataset from kaggle is used in our study. The study obtained by a sample size of 40 and employed SPSS software to compare two controllers. In order to compare the procedure and results of two study groups, 20 samples were selected, with 10 sets of samples chosen from each group. The study employs two methods Group 2 conducts a Decision tree algorithm, while Group 1 employs Linear regression.

The study used dataset comprising parameters such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity and potability. We used SPSS software for analysing the dataset and creating graphs based on the dataset. This graphs are useful to findout which machine learning algorithm is better in predicting accuracy of water quality. Using this dataset, the algorithm suggested in this research was run, and the outcomes were compared to those acquired by a comparative algorithm.

Python software was used to create and complete the assigned work. Using an Intel Core i7 CPU and 8GB of RAM and a 64-bit system sort, a testing environment for machine learning and deep learning was set up on a Windows 11 operating system. For accurate results, the Python code was written and run. To guarantee accuracy, the dataset was processed in the background as the algorithm ran.

Decision tree algorithm:

A machine learning method called a decision tree algorithm is used to forecast the quality of water given a set of input features. A decision tree divides the dataset into subsets recursively in the context of predicting water quality, with decisions being made at each node according to particular features. These characteristics might include elements like temperature, dissolved oxygen content, pH levels, and pollutant concentrations.(Mahmoud Y. shams, 2022) The algorithm tries to build a structure resembling a tree, with each leaf node denoting a potential result for the water quality.

By analyzing past data on water quality, the decision tree algorithm learns to recognize trends and connections between various characteristics and the associated state of water quality. The decision tree can be trained to make predictions by applying the rules and structure it has learned to new, unseen data.

Regarding the prediction of water quality, a decision tree may offer significant insights into the variables that greatly affect the quality of the water as well as help in locating possible pollution sources or circumstances linked to low water quality.(Ahmed M. Elshewey,2023). Because of their interpretability, decision trees are especially helpful for comprehending the reasoning behind forecasts of water quality.

A well-known machine learning algorithm for regression and classification issues is called DT. A decision tree's problem is choosing the root node at each level. We call this procedure "attribute selection." There are two well-known methods for selecting attributes: information gain and the Gini index. The following formula can be used to calculate the Gini index:

Gini = $1-\sum_{i=1}^{classes} p(\frac{i}{t})2$

The Gini index helps to compute the impurity of data in the dataset. Information gain is an additional method of attribute selection. Gained information reveals the quality of the data. As soon as we know the entropies of each attribute and the target class, we can compute the information gain. One can compute entropy D as follows:

Entrophy(D) = $-\sum_{i=1}^{|c|} pr(ci) \log_2 pr\ (ci)$

$$\sum_{i=1}^{|c|} pr(ci) = 1$$

where Pr(ci) presents the probability, ci presents the class, and D presents the dataset. The entropy of attribute Ai is utilized as the current root and can be calculated as:

entropy A_i(D)= $-\sum_{j=1}^{v} \frac{|Dj|}{D}$ * entropy(D_j)

Finally, the following information is gained when attribute Ai is chosen to branch or split data:

Entropy(D,A_i) = entropy(D)-entropy A_i(D)

Decision tree can capture complex relationships between various environmental factors and water quality. They can help identify threshold values or conditions under which water quality deteriorates or improves.

Linear regression:

It is a statistical technique that offers an easy-to-understand and straightforward means of modeling and predicting relationships between variables.
It works well because it captures the linear dependencies between predictors and the target variable.(Nurandiah zamri, 2022).

Operational: You can get a general idea of how different elements (such as TDS, PH, and pollutants) impact water quality by using linear regression.

Data collection, data preparation, model selection, model building, model estimation, model assessment, and model prediction are all necessary for the operation of linear regression.
Complex, non-linear relationships or interactions between variables might not be captured by linear regression.

So, We are using the decision tree approach because the previous technique was unable to capture more intricate and nonlinear interactions.

The goal of linear regression is to find the best-fit line that minimizes the difference between the actual and predicted values.

The general form of linear regression model for one independent variable is:
$Y=b_0+b_1x+\varepsilon$
Y-dependent variable(variable we want to predict)

X-independent variable(variable used for prediction)

b_0-intercept(value of y when X is 0).
b_1-slope.
ε-error term(representing unobserved factors affecting Y).

STATISTICAL ANALYSIS :

SPSS software is used for statistical analysis of novel approaches to efficient prediction of water quality using Decision Tree compared to Linear Regression. The independent variable is Enhanced multilayer perceptron accuracy and the dependent variable is efficiency. The independent T-test analyses are carried out to calculate the accuracy of the Linear Regression and Decision Tree.

RESULT:

Table 1: represents dataset description related to water quality prediction. This table contains parameters used to predict the quality of water by machine learning algorithms. Parameters like pH , Hardness, Solids, Chloramines, Sulfate, Conductivity, Turbidity, Potability.

Table 2: represents the properties of water for predicting water quality.

Table 3: shows the accuracy of linear regression and decision tree with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Linear Regression methods.

Table 5: Shows Independent Sample Test between DT and LR algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the Decision tree algorithm along with the error bars.

The mean accuracy of Decision tree (64.0%) is higher compared to the mean accuracy of Linear regression(62.71%).

Figure 1: illustrates a bar graph displaying the average accuracy of the Linear regression(62.71%) and the existing algorithm Decision tree(64%). The X-axis of the graph contains algorithms, while the Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

The choice between linear regression and decision tree depends on the specific characteristics of the water quality dataset and the goals of the monitoring and prediction tasks. The results revealed that Decision tree algorithm outperformed the linear regression algorithm in the terms of accuracy.(Sarang.Y.Devlagar,2023) .To be precise, the Decision Tree algorithm predicted water quality with an astounding 64% accuracy rate, while linear regression predicted water quality with 62.71% accuracy. In comparison to the linear regression algorithm, the decision tree algorithm may be able to predict with higher prediction accuracy.

Decision trees are more accurate than linear regression because they are more flexible and can be used for both regression and classification tasks, whereas linear regression is mainly used for binary classification.(Rajput,2023) Non-linear decision boundaries are possible with decision trees because they are able to capture complex relationships in the data. Compared to the linear decision boundaries of linear regression, decision trees may be more effective at capturing complex patterns in the dataset.

Decision Trees are inherently more interpretable. The decision-making process of the model is made easier to comprehend and explain by representing each decision rule by a node in the tree. Non-linear relationships between features and the target variable are a natural fit for decision trees.

CONCLUSION

The model is tested with raw data understand and analyse the summary. The conducted study aimed to evaluate the performance of two machine learning algorithms—the Linear regression Algorithm (LR) and the Decision tree algorithm (DT)—in correctly predicting water quality. The results demonstrated that the Decision Tree Algorithm performed 64.0% better than the Linear regression algorithm, with an accuracy of 62.71% This compelling result suggests that, in the specific scenario of predicting water quality, the Decision tree Algorithm outperforms the Linear regression. The Decision tree Algorithm's increased accuracy demonstrates how well it predicts the accuracy of water quality.(Rana.R.Kalia,2023).

DECLARATION:

Conflict of Interests

No conflict of interest in this manuscript.

Authors Contributions

Author SS was involved in data collection, data analysis, and manuscript writing. Author SMS was involved in the conceptualization, data validation, and critical review of the manuscripts.

Acknowledgment

The creators might want to offer their earnest thanks towards Saveetha School of Designing, Saveetha Foundation of Clinical and Specialized Sciences (Previously known as Saveetha College) for giving the important framework to effectively do this work.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Manac Infotech Pvt, Limited, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

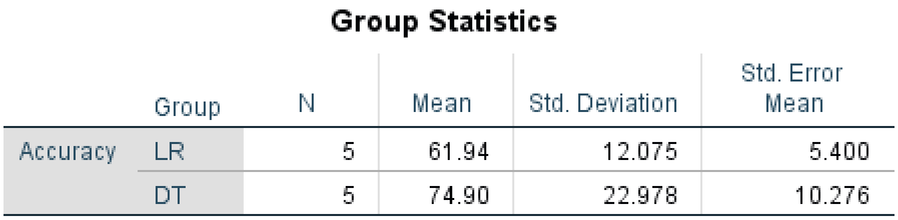
Table 1. Data set description

S.no.	Attributes or Parameters
1.	pH
2.	Hardness
3.	Solids
4.	Chloramines
5.	Sulphate
6.	Conductivity
7.	Trihalomethanes
8.	Turbidity
9.	Potability

Table 2: shows the sample data of the accuracy of linear regression and decision tree algorithm.

Sample size (from dataset)	Decision tree(accuracy)	Linear regression(accuracy)
40	64.00%	62.71%
50	80.00%	66.50%
60	82.76%	69.60%
70	88.6%	72.22%
80	96.4%	74.49%

Table 3: Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Decision tree and Linear Regression methods.

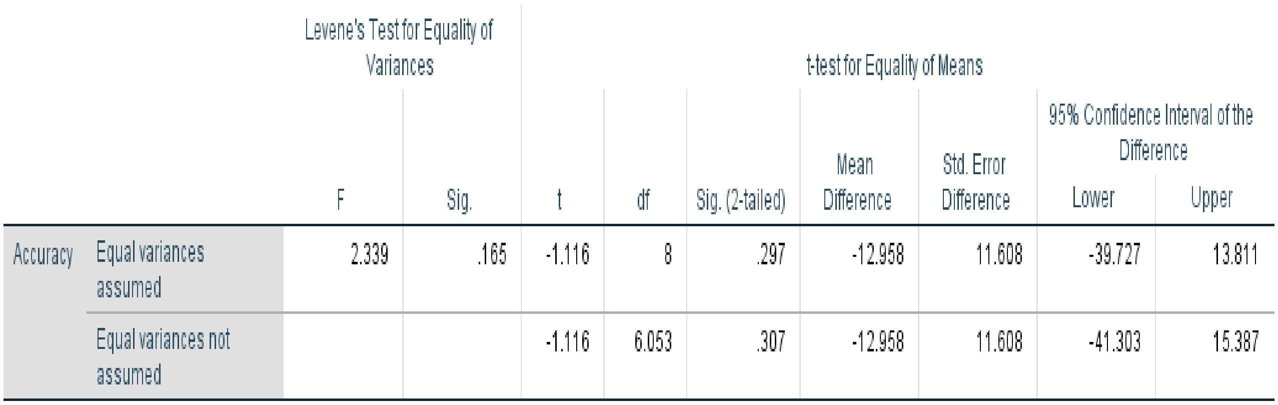


Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Accuracy	LR	5	61.94	12.075	5.400
	DT	5	74.90	22.978	10.276

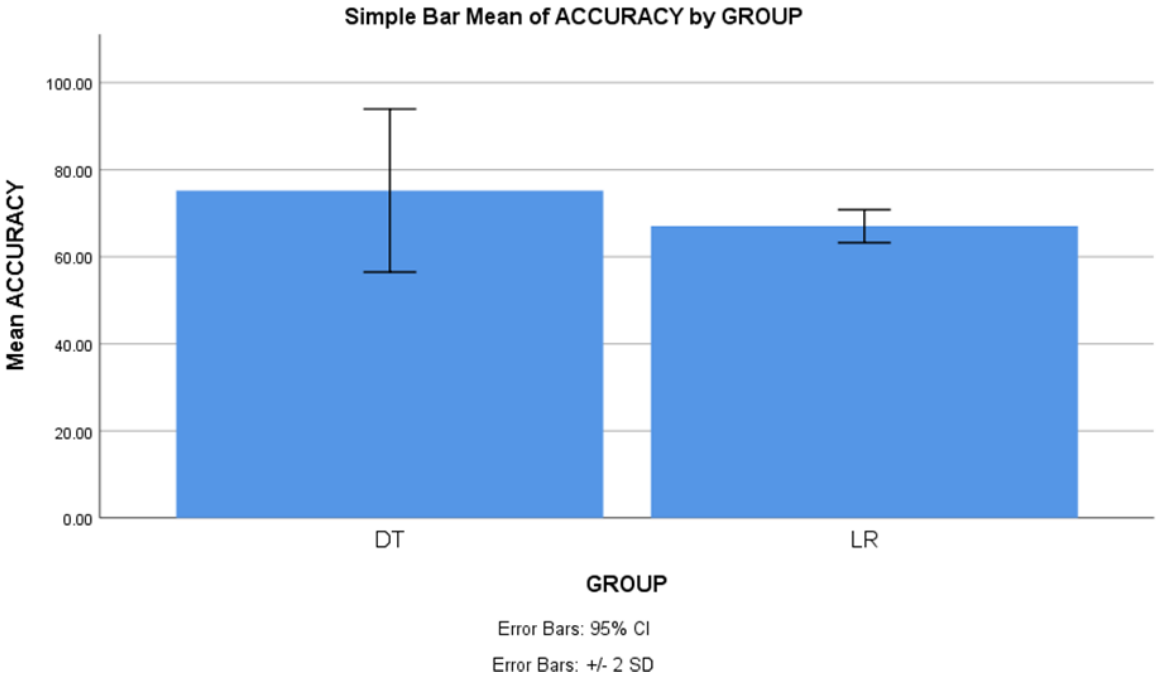
Table 4: Shows Independent Sample Test between DT and LR algorithm.

Independent sample test



		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Accuracy	Equal variances assumed	2.339	.165	-1.116	8	.297	-12.958	11.608	-39.727	13.811
	Equal variances not assumed			-1.116	6.053	.307	-12.958	11.608	-41.303	15.387

Graph:



REFERENCES:

[1].Qiao, Z., Sun, S., Jiang, Q.O., Xiao, L., Wang, Y. and Yan, H., 2021. Retrieval of total phosphorus concentration in the surface water of miyun reservoir based on remote sensing data and machine learning algorithms. *Remote Sensing*, *13*(22), p.4662.

[2].Rana, R., Kalia, A., Boora, A., Alfaisal, F.M., Alharbi, R.S., Berwal, P., Alam, S., Khan, M.A. and Qamar, O., 2023. Artificial Intelligence for Surface Water Quality Evaluation, Monitoring and Assessment. *Water*, *15*(22), p.3919.

[3].Hou, Y., Zhang, A., Lv, R., Zhao, S., Ma, J., Zhang, H. and Li, Z., 2022. A study on water quality parameters estimation for urban rivers based on ground hyperspectral remote sensing technology. *Environmental Science and Pollution Research*, *29*(42), pp.63640-63654.

Chadli, K., 2023. Assessment of surface water quality in the Sebou watershed (Morocco) using a nonparametric approach and machine learning techniques. *Arabian Journal of Geosciences*, *16*(9), p.517.

[4].Dimple, D., Rajput, J., Al-Ansari, N. and Elbeltagi, A., 2022. Predicting irrigation water quality indices based on data-driven algorithms: case study in semiarid environment. *Journal of Chemistry*, *2022*.

[5].Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q. and Zhang, H., 2021. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. *Environmental Research*, *202*, p.111660.

[6].Ma, Y., Song, K., Wen, Z., Liu, G., Shang, Y., Lyu, L., Du, J., Yang, Q., Li, S., Tao, H. and Hou, J., 2021. Remote sensing of turbidity for lakes in northeast China using Sentinel-2 images with machine learning algorithms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, pp.9132-9146

[7].Jot, D., 2023. Comparing the effectiveness of different algorithms for ground water quality in Telangana region.

Z. F. Makki, A. A. Zuhaira, S. M. Al-Jubouri, R. K. S. Al-Hamd, and L. S. Cunningham, (2021)‘‘GIS-based assessment of groundwater quality for drinking and irrigation purposes in central Iraq,’’ Environ. Monitor. Assessment, vol. 193, no. 2, pp. 1–27.

water quality analysis using IoT and machine learning models. *International Journal of Information Management Data Insights*, *4*(1), p.100210.

RESEARCH PAPER 2

COMPARATIVE ANALYSIS OF LINEAR REGRESSION TO PREDICT WATER QUALITY IN SURFACE WATER USING K-NEAREST NEGHBOUR (K-NN)

ABSTRACT:

Aim: This study conducts a comparative analysis between linear regression and K-Nearest neighbours for predicting water quality, aiming to assess the effectiveness of these methods in environmental monitoring. This study primarily concentrates on predicting water quality using two machine learning methods called linear regression and K-nearest neighbour. **Materials and methods:** The study uses two study groups a Linear regression and K-nearest neighbours with a sample size of 40 in total.A clinical was used to calculate sample size with parameters preset G-power of 0.8 or 80%, an alpha of 0.05 or ($P<0.05$) and confidence intervals of 95%. This analysis was performed in SPSS software. **Result:** Both algorithms demonstrated proficiency in analyzing water quality data, accurately identifying instances where water conditions exceeded acceptable standards. The Outcome demonstrate that in comparison to the linear regression algorithm's (62.71%) accuracy, the K-nearest neighbour (67.50%) has more enhanced accuracy in predicting water quality. **Conclusion:** The accuracy rate of K-nearest neighbour (67.50%) is noticeably greater than the accuracy of Linear regression (62.71%).

Introduction:

Water quality assessment is crucial for environmental monitoring and public health. This study presents a comparative analysis of two popular machine learning techniques, Linear Regression (LR) and K-Nearest Neighbors (KNN), for predicting water quality parameters in bodies of water. The aim is to evaluate and compare the performance of these models in estimating key water quality indicators, such as dissolved oxygen levels, pH, and turbidity.

The dataset used in this study comprises a comprehensive collection of water quality measurements from various sampling points. The variables include physical, chemical, and biological parameters that influence water quality. The linear regression model is employed to establish a linear relationship between these variables and predict water quality, assuming a linear correlation exists.

In contrast, the K-Nearest Neighbors algorithm is implemented to capture non-linear relationships by considering the similarity of instances in the dataset. The KNN model predicts water quality parameters based on the characteristics of neighboring data points in the feature space. The choice of the optimal number of neighbors is explored to enhance the predictive accuracy of the KNN model.

The results of this comparative analysis provide insights into the strengths and limitations of linear regression and K-nearest neighbors for water quality prediction. Understanding the relative performance of these models can aid environmental scientists, policymakers, and water resource managers in selecting appropriate methodologies for accurate and reliable water quality assessments in diverse settings.

Our objective is to use the power of predictive models to develop a dependable system that can forecast parameters related to water quality in various environments. The integration of machine learning algorithms offers the potential for real-time monitoring and proactive detection of potential issues related to water quality. The overarching objective of taking proactive steps to protect public health and preserve water ecosystems is furthered by this research.

Numerous studies predicting the quality of water are available in reputable journals like Science, IEEE, and Google Scholar. 68 journals, 452 publications on Google Scholar, and 153 publications on Springer are all accessible through the IEEE digital library. In a similar vein, the fusion of various algorithms enables a more intricate comprehension of the intricate relationships affecting the dynamics of water quality.

The application of k-nearest neighbour and linear regression algorithms to the study of water quality prediction could facilitate a better understanding and prediction of environmental conditions. In water quality management, accurate forecasting is crucial for making informed decisions, even in the face of potential challenges related to data volume, diversity, and algorithmic complexity. Through rigorous experimentation and meticulous algorithmic optimization, this study seeks to produce significant insights into the dynamic field of water quality prediction, advancing a better understanding of environmental factors and enhancing our ability to safeguard and manage water resources.

Early problem detection is the only way to prevent issues. Numerous investigators have employed computer programs to detect diabetes. According to research, there are various methods for classifying diabetes using Machine Learning datasets like SVM, NB, DT, and others(Shafi and Ansari 2021).

If diabetes is correctly identified early on, it can be managed and kept under control; however, there is no permanent cure for the condition. Diabetes data classification is a difficult task because most medical data is non-normal, non-linear, and has a complex, linked structure. The results of the diabetes classification are also impacted by a high number of outliers in the dataset in addition to missing or null values.

MATERIALS AND METHODS

The research is carried out in Machine learning laboratory lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The water potability dataset from kaggle is used in our study. The study obtained by a sample size of 40 and employed SPSS software to compare two controllers. In order to compare the procedure and results of two study groups, 20 samples were selected, with 10 sets of samples chosen from each group. The study employs two methods Group 2 conducts a K-nearest neighbour, while Group 1 employs Linear regression.

The dataset used in the study included trihalomethanes, organic carbon, conductivity, pH, hardness, solids, chloramines, sulfate, turbidity, and potability. To analyze the dataset and make graphs based on it, we used SPSS software. These graphs are helpful in determining which machine learning algorithm predicts water quality more accurately. This dataset was used to run the algorithm proposed in this study, and the results were contrasted with those obtained by a comparative algorithm.

The dataset used in this study is sourced from Kaggle. Kaggle is a well-known, free data repository where we can easily download a dataset. The study's dataset, titled "Water Quality," is publicly available. The dataset comprises 935 instances and 10 columns. Potable is the target class. It can have one of two values: '0,' which indicates that the water is unsafe to drink, or '1,' which indicates that it is safe to drink.

The assigned work was created and finished using Python software. A testing environment for deep learning and machine learning was set up on a Windows 11 operating system with an Intel Core i7 CPU, 8GB of RAM, and a 64-bit system sort. To ensure precise outcomes, the Python code was created and executed. The dataset was processed in the background while the algorithm ran to ensure accuracy.

K-Nearest neighbour (KNN):

A straightforward supervised machine learning technique that works well for both regression and classification applications is the k-nearest neighbors (KNN) algorithm. The basic idea of KNN is to predict a data point's label (or value in the case of regression) by using the average value or majority class of its k-nearest neighbors in the feature space.

Here are the basic steps of the KNN algorithm:

Choose the value of k: Decide on the number of neighbors (k) to consider when making predictions. A common choice is to experiment with different values of k and choose the one that gives the best performance on a validation set or through cross-validation.

Calculate distances: Measure the distance between the new data point and every other data point in the dataset. The most common distance metric is Euclidean distance, but other metrics like Manhattan distance or Minkowski distance can also be used. Euclidean distance between two points (x1,y1) and (x2,y2) is given by:

$$\sqrt{(x2-x1)^2+(y2-y1)\hat{}2}$$

Identify k-nearest neighbors: Select the k data points with the smallest distances to the new data point.

In classification tasks, assign the class label that occurs most frequently among the k neighbors to the new data point. In regression tasks, use majority voting. Determine the average of the k neighbors' target values for regression tasks.

Make a prediction: The majority class or average value is used as the prediction for the new data point.

Linear regression:

This statistical technique makes it simple and easy to understand how to model and predict relationships between variables.

Its effectiveness stems from its ability to capture the linear relationships between predictors and the dependent variable.

Operational: Linear regression can be used to obtain a general understanding of the ways in which various elements (such as TDS, PH, and pollutants) affect the quality of water.

Complex, non-linear relationships or interactions between variables may not be captured by linear regression; therefore, data collection, preparation, modeling selection, modeling building, estimating, evaluating, and predicting are all necessary for the operation of linear regression.

Because the prior method was unable to capture more complex and nonlinear interactions, we are now using the decision tree approach.

Finding the best-fit line that minimizes the difference between the actual and predicted values is the aim of linear regression.

The general form of linear regression model for one independent variable is:

$Y=b_0+b_1x+\varepsilon$

Y-dependent variable(variable we want to predict)

X-independent variable(variable used for prediction)

b_0-intercept(value of y when X is 0).

b_1-slope.

ε-error term(representing unobserved factors affecting Y).

Statistical analysis

Statistical analysis of new methods for effective water quality prediction using k nearest neighbour as opposed to Linear Regression is done using SPSS software. Enhanced multilayer perceptron accuracy is the independent variable, while efficiency is the dependent variable. To determine the accuracy of the k nearest neighbor and linear regression, independent T-test analyses are performed.

RESULTS

Table 1: represents dataset description related to water quality prediction. This table contains parameters used to predict the quality of water by machine learning algorithms. Parameters like pH , Hardness, Solids, Chloramines, Sulfate, Conductivity, Turbidity, Potability.

Table 2: represents the properties of water for predicting water quality.

Table 3: shows the accuracy of linear regression and K-nearest neighbour with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for K-nearest neighbour and Linear Regression methods.

Table 5: Shows Independent Sample Test between K-NN and LR algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the K-nearest neighbour algorithm along with the error bars.

The mean accuracy of K-nearest neighbour (67.50%) is higher compared to the mean accuracy of Linear regression(62.71%).

Figure 1: illustrates a bar graph displaying the average accuracy of the Linear regression(62.71%) and the existing algorithm K-nearest algorithm(67.50%). The X-axis of the graph contains algorithms, while the Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

Depending on the objectives of the monitoring and prediction tasks as well as the unique features of the water quality dataset, either K-nearest neighbor or linear regression should be used. The K-nearest neighbour algorithm performed better in terms of accuracy than the linear regression algorithm, according to the results.(Sarang Y. Devigahr ,2023).More specifically, the water quality was predicted by linear regression with an accuracy of 62.71%, and water quality was predicted by the k nearest neighbor algorithm with an astounding 67.50%. The K-nearest neighbour algorithm may be more accurate in predicting the water quality than the linear regression algorithm.

Since k nearest neighbour is more adaptable and can be used for both regression and classification tasks, while linear regression is primarily used for binary classification, it is more accurate than linear regression.(Rajput, 2023) Because k nearest neighbour can capture complex relationships in the data, non-linear decision boundaries can be achieved. K-nearest neighbour might be better at identifying intricate patterns in the dataset than the linear decision boundaries of linear regression.

To summarise, the KNN algorithm is a straightforward and intuitive method that can be applied to a variety of data sets, particularly when the decision boundaries are intricate or difficult to describe using a straightforward mathematical model. But depending on the parameters chosen and the type of data, its performance might vary, so it might not be appropriate for high-dimensional datasets.

CONCLUSION

The model is evaluated using unprocessed data to comprehend and evaluate the synopsis. The purpose of the study was to assess how well two machine learning algorithms—the K-nearest neighbor and the linear regression algorithm (LR)—performed in accurately predicting the quality of the water. The outcomes showed that, with an accuracy of 62.71%, the Decision Tree Algorithm outperformed the Linear Regression Algorithm by 67.50%. This strong outcome implies that the K-Nearest algorithm performs better than the Linear Regression in the particular scenario of predicting water quality. The improved accuracy of the K-nearest neighbor shows how well it predicts the accuracy of water quality.(Kalia, Rana R., 2023).

DECLARATIONS

Conflicts of interest

No conflicts of interest in this manuscript

Authors Contributions

Author PS was involved in data collection, data analysis, manuscript writing, Author JK was involved in conceptualization, data validation and critical review manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies, Chennai.
2. Saveetha School of Engineering
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha University.

TABLES AND FIGURES

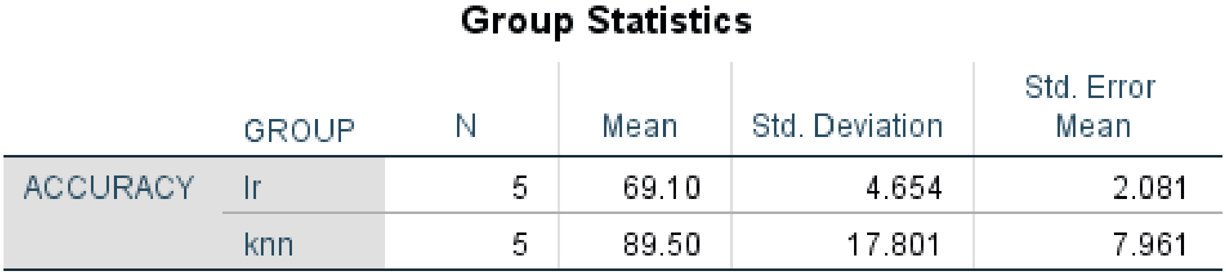
Table1.Data set description

S.no	Attributes or Parameters
1.	pH
2.	Hardness
3.	Solids
4.	Chloramines
5.	Sulphate
6.	Conductivity
7.	Trihalomethanes
8.	Turbidity
9.	Potability

Table2: shows the sample data of the accuracy of linear regression and K-nearest neighbour algorithm.

Sample size(from dataset)	K-nearest neighbour(accuracy)	Linear regression(accuracy)
40	67.50%	62.71%
50	77.50%	66.50%
60	82%	69.60%
70	89.6%	72.22%
80	93.4%	74.49%

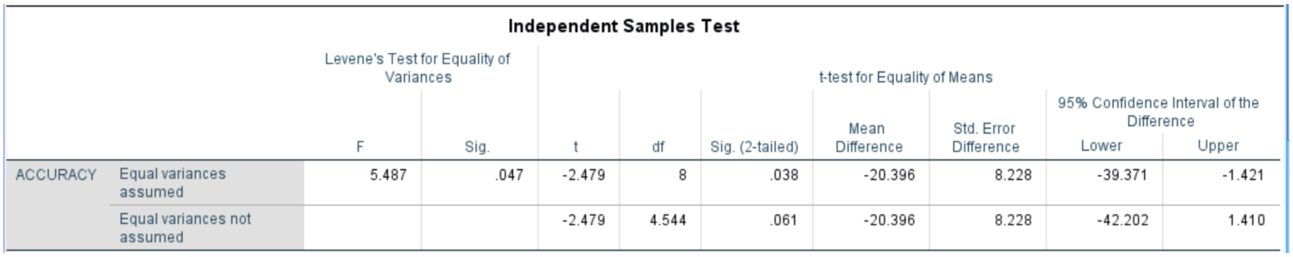
Table3 : Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for K-nearest neighbour and Linear Regression methods.



Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
ACCURACY	lr	5	69.10	4.654	2.081
	knn	5	89.50	17.801	7.961

Table 4: Shows Independent Sample Test between KNN and LR algorithm



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ACCURACY	Equal variances assumed	5.487	.047	-2.479	8	.038	-20.396	8.228	-39.371	-1.421
	Equal variances not assumed			-2.479	4.544	.061	-20.396	8.228	-42.202	1.410

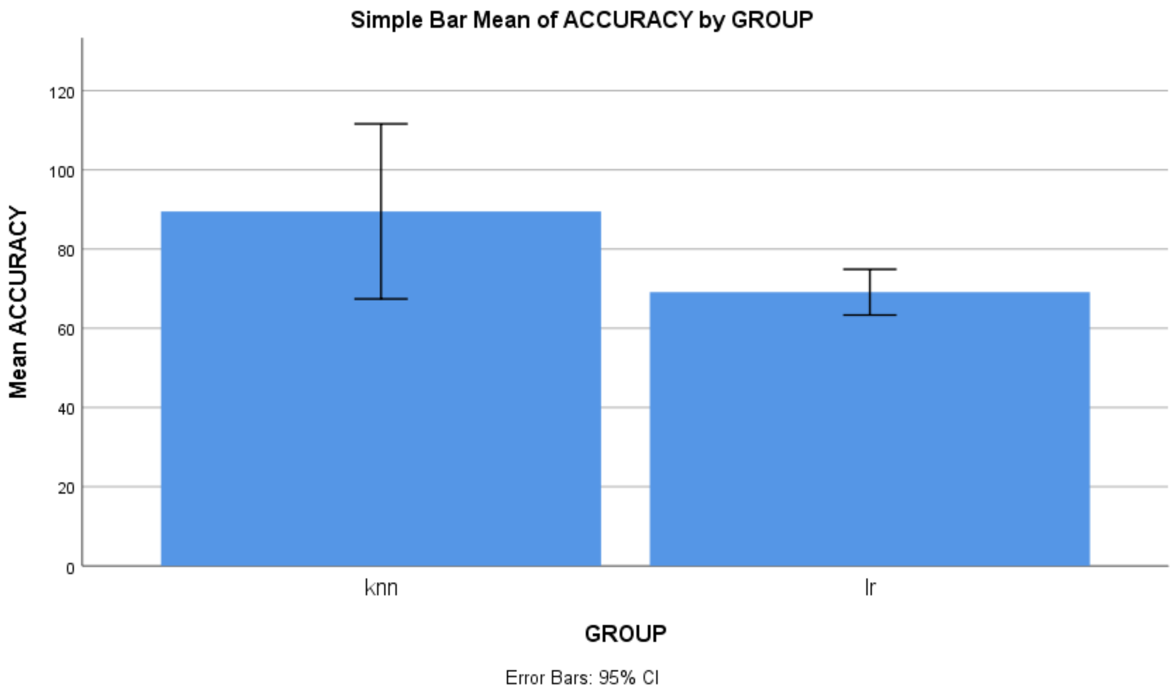


Fig.1. : Shows mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the K-nearest neighbour algorithm along with the error bars.

REFERENCES:

[1].Danades, A., Pratama, D., Anggraini, D. and Anggriani, D., 2016, October. Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. In *2016 6th International conference on system engineering and technology (ICSET)* (pp. 137-141). IEEE.

[2].Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J. and Zhang, Y., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research*, *171*, p.115454.

[3].Prakash, R., Tharun, V.P. and Devi, S.R., 2018, April. A comparative study of various classification techniques to determine water quality. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1501-1506). IEEE.

[4].Tahraoui, H., Toumi, S., Hassein-Bey, A.H., Bousselma, A., Sid, A.N.E.H., Belhadj, A.E., Triki, Z., Kebir, M., Amrane, A., Zhang, J. and Assadi, A.A., 2023. Advancing Water Quality Research: K-Nearest Neighbor Coupled with the Improved Grey Wolf Optimizer Algorithm Model Unveils New Possibilities for Dry Residue Prediction. *Water*, *15*(14), p.2631.

[5].Saberioon, M., Císař, P., Labbé, L., Souček, P., Pelissier, P. and Kerneis, T., 2018. Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (oncorhynchus mykiss) classification using image-base

RESEARCH PAPER 3

DETECTION OF POLLUTANTS IN SURFACE WATER USING RANDOM FOREST COMPARED TO LINEAR REGRESSION

ABSTRACT

Aim: The objective of this study is to assess and enhance the effectiveness of random forest algorithms when combined with linear regression to monitor and forecast water quality in various aquatic environments. **Methods and materials:** With a total sample size of forty, the study employs two study groups: one for linear regression and the other for random forest methodology. Using predetermined G-powers of 0.8 or 80%, alphas of 0.05 or (P<0.05), and confidence intervals of 95%, a clinical was utilized to determine the sample size. Using SPSS software, this analysis was carried out. **Result:** The two algorithms exhibited competence in examining data related to water quality, precisely pinpointing situations in which the water quality exceeded permissible limits. The results show that when it comes to water quality prediction, the random forest algorithm (68.00%) has a higher accuracy than the linear regression algorithm (62.71%). **Conclusion:** The accuracy rate of random forest (68.00%) is noticeably greater than the accuracy of Linear regression (62.71%).

INTRODUCTION

Water pollution is a major environmental problem that has a significant impact on both human health and ecosystems. Correct and timely identification of pollutants in water bodies is essential for effective environmental management. Our research focuses on how well two machine learning models—Random Forest (RF) and Linear Regression (LR)—identify pollutants in water.

.(alanhit,2022)

The dataset used in this study includes biological indicators, physical characteristics, and chemical concentrations in addition to other water quality-related variables. The correlation between these parameters and the concentrations of pollutants can be simulated by using the Random Forest and Linear Regression algorithms. (Sakizahedh, 2017). After the models are trained on historical data, their predictive performance is evaluated using multiple metrics, such as accuracy, precision, and recall.

The results of our investigation provide insight into the relative benefits and drawbacks of using Random Forest and Linear Regression to identify water pollutants. Because Random Forest is known for handling complex relationships and non-linearities, it is expected to perform better than Linear Regression in identifying complex patterns within the dataset. However, Linear Regression, a simpler and more understandable model, might work better in scenarios where the correlation between the variables and the levels of pollutants is essentially linear.

Furthermore, considering the importance of understanding the factors influencing pollution levels, we investigate the models' interpretability. Although the feature importance analysis of Random Forest illuminates complex interactions, the interpretability of Linear Regression facilitates the simple identification of significant variables.

The current study offers helpful guidance on model selection for water pollution detection and contributes to the growing body of literature on machine learning applications in environmental monitoring. To aid in the creation of trustworthy and precise water quality monitoring systems, (Garcia.M.J,2014).The goal of the research is to inform decision-makers and environmental scientists about the benefits and drawbacks of both linear regression and random forests.

Reputable journals like Science, IEEE, and Google Scholar have a wealth of research on water quality prediction. There are 45 journals available through the IEEE digital library, 362 publications on Google Scholar, and 253 publications on Springer. Likewise, the amalgamation of varied algorithms facilitates a more intricate comprehension of the intricate interplay impacting the dynamics of water quality.

In the study of water quality prediction, random forest and linear regression algorithms could help to simplify the understanding and prediction of environmental conditions. Despite potential challenges related to data volume, diversity, and algorithmic complexity, the pursuit of accurate forecasts is crucial for informed decision-making in water quality management.(mirazaei,2016). By carefully examining environmental factors and optimizing algorithms, this study seeks to provide significant insights into the dynamic field of water quality prediction. These insights will enhance our ability to safeguard and manage water resources.

MATERIALS AND METHODS

The study is being conducted at the Saveetha Institute of Medical and Technical Sciences' Saveetha School of Engineering's machine learning lab. In our study, we make use of the water potability dataset from Kaggle. Utilizing SPSS software, the study compared two controllers using a sample size of forty. Ten sets of samples were chosen from each group, totaling twenty samples, in order to compare the methods and outcomes of the two study groups. The investigation uses two techniques. Group 1 utilizes linear regression, whereas Group 2 uses the random forest technique.

The dataset used in the study included trihalomethanes, organic carbon, conductivity, pH, hardness, solids, chloramines, sulfate, turbidity, and potability. To analyze the dataset and make graphs based on it, we used SPSS software. These graphs are helpful in determining which machine learning algorithm predicts water quality more accurately. This dataset was used to run the algorithm proposed in this study, and the results were contrasted with those obtained by a comparative algorithm.

The assigned work was created and finished using Python software. A testing environment for deep learning and machine learning was set up on a Windows 11 operating system with an Intel Core i7 CPU, 8GB of RAM, and a 64-bit system sort. To ensure precise outcomes, the Python code was created and executed. The dataset was processed in the background while the algorithm ran to ensure accuracy.

Random forest:

Given its ability to handle complex data, non-linear relationships, and spatial dependencies, random forest is a powerful and versatile algorithm that performs wonderfully when used to predict water quality.

Water quality can be reliably and adaptably predicted with Random Forest because of its exceptional ability to handle non-linearity, complex relationships, and environmental dependencies. It’s important to keep in mind that Random Forest’s performance still depends on the dataset’s representativeness and quality, proper pre-processing, and careful selection of hyper-parameters.(M. Guo, 2023). Model evaluations and updates must occur on a regular basis to ensure reliable and accurate water quality predictions.

To monitor a variety of water parameters, conventional water quality monitoring requires the installation of multiple sensors. Furthermore, the need for laboratory analysis or testing equipment for some water quality data, like total nitrogen (TN), adds to the turnaround time for results.(2020, Norouzi). This paper designs a framework for predicting the total nitrogen concentrations in inland water bodies using the water quality variables that are currently available (e.g., temperature, pH, conductivity, etc.).

The concentration of nitrogen in random forests is a significant factor in predicting the water quality. In addition to having a substantial impact on water quality, nitrogen concentration is an important environmental parameter that can be used to predict water quality using the Random Forest algorithm. (priyadarshini,2022)

Eutrophication, nitrogen as a predictor, spatial and temporal variations, model interpretation and variation, and real-time monitoring are employed in the random forest method to predict water quality with the aid of nitrogen.

Because of its significant impact on aquatic ecosystems and its significance as an indicator of water quality, nitrogen concentration is frequently employed as a feature in Random Forest models for water quality prediction.

Linear regression:

This statistical technique offers a simple and intuitive approach to forecast and model relationships between variables.It works well because it shows the linear dependencies between predictors and the target variable.Operational: You can get a general idea of how different factors (like TDS, PH, and pollutants) affect water quality by using linear regression (Nurandiah zamri, 2022).

The following steps must be taken in order to perform linear regression: gathering data, preparing it, choosing a model, creating a model, estimating it, evaluating it, and making a prediction.

The complex, non-linear relationships or interactions between variables may be beyond the scope of linear regression.

The goal of linear regression is to find the best-fit line that minimizes the difference between the expected and actual values.

The general form of linear regression model for one independent variable is:

$Y=b_0+b_1x+\varepsilon$

Y-dependent variable(variable we want to predict)

X-independent variable(variable used for prediction)

b_0-intercept(value of y when X is 0).

b_1-slope.

ε-error term(representing unobserved factors affecting Y).

Statistical analysis

When comparing new methods for effective water quality prediction using random forest to linear regression, statistical analysis is done using SPSS software. Enhanced multilayer perceptron accuracy is the independent variable, while efficiency is the dependent variable. The accuracy of the random forest and linear regression is determined through independent T-test analyses.

RESULTS

Table 1: represents dataset description related to water quality prediction. This table contains parameters used to predict the quality of water by machine learning algorithms. Parameters like pH , Hardness, Solids, Chloramines, Sulfate, Conductivity, Turbidity, Potability.

Table 2: represents the properties of water for predicting water quality.

Table 3: shows the accuracy of linear regression and Random forest with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Random forest and Linear Regression methods.

Table 5: Shows Independent Sample Test between RF and LR algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the Random forest algorithm along with the error bars. The mean accuracy of Random forest (68.00%) is higher compared to the mean accuracy of Linear regression(62.71%).

Figure 1: illustrates a bar graph displaying the average accuracy of the Linear regression(62.71%) and the existing algorithm Random forest algorithm (68.00%). The X-axis of the graph contains algorithms, while the Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

During training, several decision trees are constructed using the Random Forest ensemble learning technique, which then combines them to produce a prediction that is more reliable and accurate. Because of this, it can handle intricate relationships and identify patterns in data on water quality.(rahmati.O,2019)
More specifically, water quality was predicted by the random forest algorithm with an amazing 72% accuracy rate, and water quality was predicted by linear regression with 62.71% accuracy. The decision tree algorithm might be more accurate at making predictions than the linear regression algorithm.(omondi.I,2020)

Predicting the quality of water frequently entails intricate, non-linear relationships between various parameters. Random Forest is a useful tool for modeling the complex dependencies in water quality data because it is well-suited to capture such non-linearities.

Models of the Random Forest can be periodically retrained to adjust to patterns of changing water quality over time. The accuracy and relevance of the model can be improved by regularly updating it with new data and monitoring its performance.

CONCLUSION

The model is evaluated using raw data in order to comprehend and appraise the synopsis. Evaluating the predictive accuracy of two machine learning algorithms—the random forest algorithm (RF) and the linear regression algorithm (LR)—was the goal of the study. (pourghasemi, 2016).With an accuracy of 62.71%, the results showed that the random forest algorithm outperformed the linear regression algorithm by 72.00%. According to this solid result, the random forest algorithm performs better than linear regression in the particular scenario of predicting water quality. The higher accuracy of the random forest shows how well it predicts the accuracy of water quality.2023 saw Rana R. Kalinga.

Proper hyper-parameter tuning, frequent model retraining, and careful consideration of data quality are necessary to sustain the predictive model's dependability and relevance. All things considered, Random Forest is a strong ally in the fight for precise and useful water quality forecasts, aiding in the observation, comprehension, and protection of water resources.

DECLARATIONS

Conflicts of interest

No conflicts of interest in this manuscript

Authors Contributions

Author PS was involved in data collection, data analysis, manuscript writing, Author JK was involved in conceptualization, data validation and critical review manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding:

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies, Chennai.
2. Saveetha School of Engineering
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha University

TABLES AND FIGURES

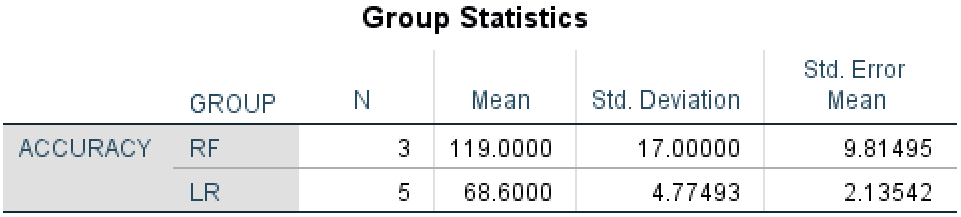
Table 1. Data set description

S.no.	Attributes or Parameters
1.	pH
2.	Hardness
3.	Solids
4.	Chloramines
5.	Sulphate
6.	Conductivity
7.	Trihalomethanes
8.	Turbidity
9.	Potability

Table2. shows the sample data of the accuracy of Linear regression and Random forest algorithm.

Sample size(from dataset)	Random forest(accuracy)	Linear regression (accuracy)
40	68%	62.71%
50	85%	66.50%
60	87.4%	69.60%
70	92.3%	72.22%
80	96.2%%	74.49%

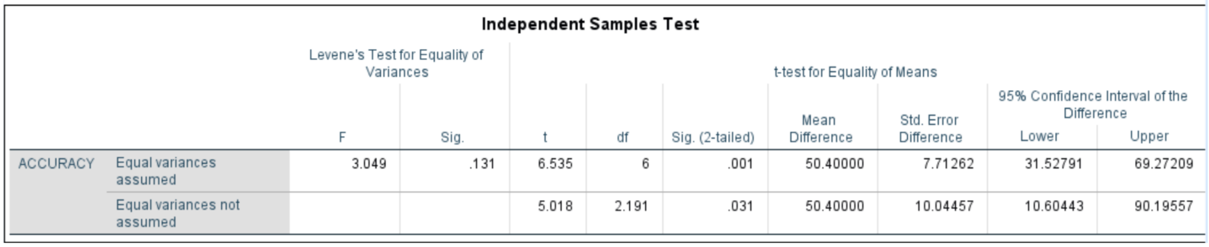
Table3. Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for Random forest and Linear Regression methods.



Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
ACCURACY	RF	3	119.0000	17.00000	9.81495
	LR	5	68.6000	4.77493	2.13542

Table 4: Shows Independent Sample Test between RF and LR algorithm



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ACCURACY	Equal variances assumed	3.049	.131	6.535	6	.001	50.40000	7.71262	31.52791	69.27209
	Equal variances not assumed			5.018	2.191	.031	50.40000	10.04457	10.60443	90.19557

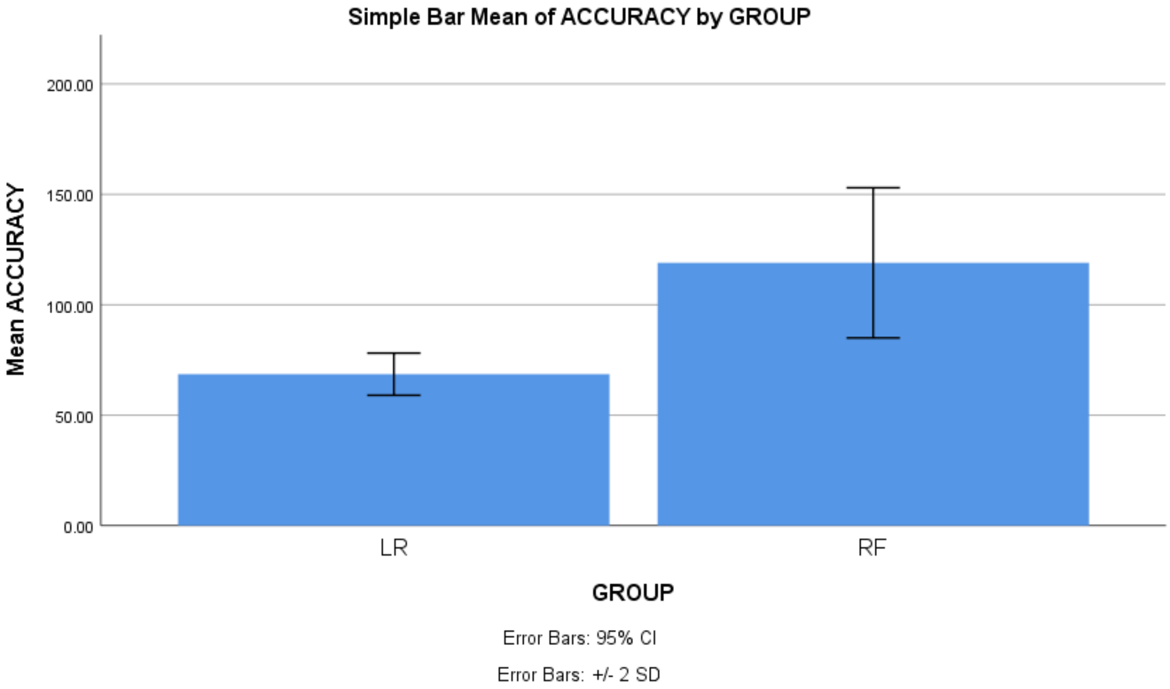


Fig.1. Shows mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the Random forest algorithm along with the error bars.

REFERENCES:

Band, S.S., Janizadeh, S., Pal, S.C., Chowdhuri, I., Siabi, Z., Norouzi, A., Melesse, A.M., Shokri, M. and Mosavi, A., 2020. Comparative analysis of artificial intelligence models for accurate estimation of groundwater nitrate concentration. *Sensors*, *20*(20), p.5763.

Alnahit, A.O., Mishra, A.K. and Khan, A.A., 2022. Stream water quality prediction using boosted regression tree and random forest models. *Stochastic Environmental Research and Risk Assessment*, *36*(9), pp.2661-2680.

Hafeez, S., Wong, M.S., Ho, H.C., Nazeer, M., Nichol, J., Abbas, S., Tang, D., Lee, K.H. and Pun, L., 2019. Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: A case study of Hong Kong. *Remote sensing*, *11*(6), p.617.

Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M. and Ribeiro, L., 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of the Total Environment*, *476*, pp.189-206.

Hou, Y., Zhang, A., Lv, R., Zhao, S., Ma, J., Zhang, H. and Li, Z., 2022. A study on water quality parameters estimation for urban rivers based on ground hyperspectral remote sensing technology. *Environmental Science and Pollution Research*, *29*(42), pp.63640-63654.

RESEARCH PAPER 4

COMPARISON OF LINEAR REGRESSION WITH R-SQUARED OIR MEAN SQUARED ERROR (MSE) TO IMPROVE THE ACCURACY IN WATER QUALITY.

ABSTRACT

Aim: This study aims to evaluate and improve the efficacy of mean squared or r-squared error algorithms in conjunction with linear regression for the purpose of monitoring and predicting water quality in diverse aquatic environments. **Methods and materials:** The study uses two study groups, one for linear regression and the other for R-squared or MSE methodology, with a total sample size of forty. The sample size was determined using a clinical with predefined G-powers of 0.8 or 80%, alphas of 0.05 or ($P<0.05$), and confidence intervals of 95%. This analysis was done with SPSS software. **Result:** The two algorithms demonstrated proficiency in analyzing water quality data, accurately identifying instances where the water quality exceeded allowable thresholds. The findings indicate that the r-squared or MSE algorithm (70.94%) outperforms the linear regression algorithm (62.71%) in terms of accuracy when it comes to predicting water quality. **Conclusion:** The accuracy rate of R-squared or MSE (70.94%) is noticeably greater than the accuracy of Linear regression (62.71%).

INTRODUCTION

Predicting water quality is essential for both maintaining public health and managing water resources. In the context of linear regression models, this study investigates the efficacy of two popular performance metrics, Mean Squared Error (MSE) and R-squared, for enhancing the precision of predictions about water quality. When modeling environmental phenomena, linear regression has been a popular option. Choosing the right evaluation metrics is crucial to evaluating the performance of the model.

Data on water quality, including chemical concentrations, physical characteristics, and biological indicators, are gathered from a variety of sources for this research. These parameters are used to create a linear regression model that predicts the quality of the water. R-squared, a measure of explained variance, and Mean Squared Error, a measure of the average squared difference between predicted and observed values, are then used in the study to compare the predictive accuracy of the model.

The results illustrate each metric's advantages and disadvantages when it comes to predicting water quality. While MSE highlights the precision of each individual prediction, R-squared sheds light on the percentage of variability that the model captures. Which metric is better suited for evaluating and refining linear regression models for water quality applications is the goal of the comparative analysis.

In the end, improved decision-making in water resource management and environmental protection initiatives can be supported by a better understanding of these metrics' performance, which can help predictive models be improved. Researchers, practitioners, and policymakers involved in water quality monitoring and prediction can learn a great deal from the study's findings.

Water quality prediction has been the subject of extensive research published in reputable journals such as Science, IEEE, and Google Scholar. The IEEE Digital Library offers 45 journals, while Google Scholar offers 362 publications and Springer offers 253 publications. Similarly, combining different algorithms enables a deeper understanding of the complex interactions influencing water quality dynamics.

For the sake of environmental sustainability and public health, water quality prediction is essential. When it comes to assessing the precision and dependability of predictive models, two commonly used metrics—R-squared (R2) and Mean Squared Error (MSE)—are essential. These metrics aid researchers and decision-makers in their endeavors to improve the accuracy of water quality predictions by offering insightful information about how well models capture variations in water quality parameters. In this regard, a solid grasp of R2 and MSE becomes essential to guaranteeing the validity and suitability of predictive models in the crucial field of evaluating water quality.

A number of lives will be saved if this disease can be predicted early. To help people survive diabetes, we have developed a model that takes into account some risk factors for the disease and uses machine learning algorithms called Random Forest, XGBoost, and Logistic Regression to predict it early(Hassan et al. n.d.).

About 463 million people between the ages of 20 and 79 had diabetes in 2045; scientists estimate that number will reach 700 million. Diabetes is responsible for 4.2 million deaths annually, with type 1 affecting over 1.1 million children and adolescents(Refat et al. n.d.). Diabetes was discovered by comparing a number of parameters with the results of a prediction. Using eight characteristics: the number of pregnancies, plasma glucose level, diastolic blood pressure, sebum thickness, insulin level, body mass index, diabetes pedigree function, and age, doctors should be able to accurately diagnose diabetes in future patients, even if they are asymptomatic.

MATERIALS AND METHODS

The machine learning lab at the Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, is where the study is being carried out. We utilize the Kaggle water potability dataset in our investigation. With a sample size of forty, the study compared two controllers using SPSS software. To compare the procedures and results of the two study groups, ten sets of samples total—ten sets of samples—were selected from each group. There are two methods used in the investigation. Group 2 employs the r-squared or MSE technique, while Group 1 uses linear regression.

Trihalomethanes, organic carbon, conductivity, pH, hardness, solids, chloramines, sulfate, turbidity, and potability were among the datasets analyzed for the study. We used SPSS software to analyze the dataset and create graphs based on the data. Which machine learning algorithm predicts water quality more accurately can be ascertained with the help of these graphs. The algorithm suggested in this study was applied to this dataset, and the outcomes were compared to those produced by a comparative algorithm.

Using Python software, the assigned work was created and completed. On a Windows 11 operating system with an Intel Core i7 CPU, 8GB of RAM, and a 64-bit system sort, a testing environment for deep learning and machine learning was set up. To guarantee exact results, the Python code was written and run. To ensure accuracy, the algorithm ran while the dataset was processed in the background.

R-squared:

R-squared (R2) is a statistical measure used to assess the goodness of fit of a regression model, including when predicting water quality based on various independent variables. It quantifies the proportion of the variance in the dependent variable (in this case, water quality) that can be explained by the independent variables or predictors in the model. The R-squared value ranges from 0 to 1, with higher values indicating a better fit.

Working: R-squared can used to assess the goodness of fit of a regression model, including when predicting water quality based on various independent variables.

The working of R-squared involves data collection, model development, predictions, calculate r-squared, interpretation of r-squared.

Calculation of R-squared:

Use the R-squared metric to assess how well the model fits the observed data. R-squared is calculated as follows:

R2 = 1 - (SSE / SST)

SSE (Sum of Squared Errors): This is the sum of the squared differences between the actual water quality values and the predicted values generated by the model. It quantifies the variability in water quality that the model did not account for.

SST (Total Sum of Squares): This represents the total variance in the actual water quality values. It signifies the total variability in water quality without any predictive model.

Interpretation of R-squared:

The R-squared value ranges from 0 to 1:

R2 = 1: In this case, the model perfectly explains the variance in water quality, indicating that all observed variation in water quality can be predicted by the independent variables.

$R2 = 0$: An R-squared value of 0 means the model does not explain any variance in water quality, indicating it has no predictive power.

$0 < R2 < 1$: For values between 0 and 1, the model explains some portion of the variance in water quality. The closer R-squared is to 1, the better the model's fit, suggesting a stronger relationship between the independent variables and water quality.

In summary, when using R-squared in predicting water quality, the outcome is a numerical value that quantifies how well the model explains the variability in water quality based on the independent variables. A high R-squared value suggests that the model provides a good fit and has the ability to predict water quality accurately, while a low R-squared value indicates that the model does not explain much of the variance and may not be effective for predicting water quality. The interpretation of the R-squared value is crucial for understanding the predictive power of the model in the context of water quality prediction.

Mean squared error (MSE):

Mean Squared Error (MSE) is a commonly used metric to evaluate the accuracy of predictions in regression analysis. It quantifies the average squared difference between predicted values and actual (observed) values of a target variable. The lower the MSE, the better the model's predictive accuracy.

Working: Mean square error(MSE) can be used to quantifies the average squared difference between the predicted values and the actual (observed) values of the target variable, which in this case is water quality. The working of Mean squared error involves data collection, calculated squared errors, calculated mean squared errors, interpretation of MSE, use in model evauation.

Squared Error Calculation:

For each data point in your dataset, calculate the squared error, which is the squared difference between the observed water quality value (yi) and the predicted water quality value (ŷi):

Squared Error (SE) = (yi - ŷi)^2

This step quantifies how much the model's predictions deviate from the actual water quality values, with the squared values emphasizing larger errors.

Mean Squared Error Calculation:

Sum up all the squared errors (SE) and divide by the number of data points (n) in your dataset to calculate the Mean Squared Error (MSE):

MSE = (1/n) * Σ SE

n: The number of data points in your dataset.

Σ: The summation symbol, indicating you should sum up the squared errors for all data points.

Interpretation of MSE:

The MSE provides a numerical measure of how well the model's predictions align with the actual water quality measurements. A lower MSE indicates better predictive accuracy. It means that, on average, the model's predictions are closer to the true water quality values. A higher MSE suggests that the model's predictions are farther from the actual values, indicating lower accuracy.

In summary, MSE plays a significant role in predicting water quality by assessing and quantifying the accuracy and performance of predictive models. It aids in model selection, evaluation, and improvement, supports quality control, and can be a critical tool in ensuring regulatory compliance and informed decision-making in water quality management and environmental monitoring.

Mathematical background:

In the following formulas, Xi is the predicted i th value, and the Yi element is the actual i th value. The regression method predicts the Xi element for the corresponding Yi element of the ground truth dataset. Define two constants: the mean of the true values.

$$Y = \frac{1}{m}\sum_{i=1}^{m} Yi$$

and the mean total sum of squares

MST=$\frac{1}{m}\sum_{i=1}^{m}(Y - Yi)\hat{}2$

Coefficient of determination (R2 or R-squared)

R^2 = 1-$\frac{\sum_{i=1}^{m}(Xi-Yi)\hat{}2}{\sum_{i=1}^{m}(Y-Yi)\hat{}2}$

(worst value = −∞; best value = +1)

The coefficient of determination (Wright, 1921) can be interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variables.

Mean square error (MSE):

MSE = $\frac{1}{m}\sum_{i=1}^{m}(Xi - Yi)\hat{}2$

(best value = 0; worst value = +∞)

In the event that outliers need to be found, MSE can be applied. Since the squaring part of the function magnifies the error if the model eventually produces a single very bad prediction, MSE is actually excellent for assigning larger weights to such points because of the L2 norm.

Since R2 = 1-MSE/ MST and since MST is fixed for the data at hand, R2 is monotonically related to MSE (a negative monotonic relationship), which implies that an ordering of regression models based on R2 will be identical (although in reverse order) to an ordering of models based on MSE or RMSE.

Root mean square error (RMSE):

RMSE = $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(Xi - Yi)^2}$

(best value = 0; worst value = +∞)

The two quantities MSE and RMSE are monotonically related (through the square root). An ordering of regression models based on MSE will be identical to an ordering of models based on RMSE.

Linear regression:

Modeling and predicting relationships between variables can be done simply and intuitively with this statistical method.It functions well because it illustrates the linear dependencies between predictors and the target variable.Operational: A general understanding of the ways in which different elements (like TDS, PH, and pollutants) affect water quality can be obtained through the use of linear regression (Nurandiah zamri, 2022).

The following procedures are necessary for carrying out linear regression: gathering data, preparing it, choosing a model, creating one, estimating it, evaluating it, and making predictions about it.

Complex, non-linear relationships or interactions between variables might be difficult for linear regression to capture.

In linear regression, the goal is to find the best-fit line that minimizes the difference between the expected and actual values.

The general form of linear regression model for one independent variable is:

$Y=b_0+b_1x+\varepsilon$

Y-dependent variable(variable we want to predict)

X-independent variable(variable used for prediction)

b_0-intercept(value of y when X is 0).

b_1-slope.

ε-error term(representing unobserved factors affecting Y).

Statistical analysis

SPSS software is used for statistical analysis when comparing new approaches to linear regression for efficient water quality prediction using r squared and mse. Efficiency is the dependent variable, and improved multilayer perceptron accuracy is the independent variable. Through independent T-test analyses, the accuracy of the random forest and linear regression is ascertained.

RESULTS

Table 1: represents dataset description related to water quality prediction. This table contains parameters used to predict the quality of water by machine learning algorithms. Parameters like pH , Hardness, Solids, Chloramines, Sulfate, Conductivity, Turbidity, Potability.

Table 2: represents the properties of water for predicting water quality.

Table 3: shows the accuracy of linear regression and R-squared or MSE with different sample sizes.

Table 4: indicates Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for R-squared or MSE and Linear Regression methods.

Table 5: Shows Independent Sample Test between R-sq or MSE and LR algorithm. Figure 1 describes the mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the R-Squared algorithm along with the error bars.

The mean accuracy of R-squared (70.94%) is higher compared to the mean accuracy of Linear regression(62.71%).

Figure 1: illustrates a bar graph displaying the average accuracy of the Linear regression(62.71%) and the existing algorithm R-squared algorithm (70.94%). The X-axis of the graph contains algorithms, while the Y-axis contains the obtained accuracy measures. The graph features 95% confidence intervals and a standard deviation of 1.

DISCUSSION

When evaluating the effectiveness and dependability of predictive models, it is crucial to understand how to interpret and discuss R-squared (R2) and Mean Squared Error (MSE) in the context of water quality prediction. R2 gives an indication of how well the model explains the variability in the water quality data and provides information about how well the model captures underlying patterns. A higher R2 denotes a better fit between the model and the observed data, meaning that the model can explain a greater amount of the variability in water quality.

In contrast, Mean Squared Error (MSE) provides a more detailed understanding of the predictive accuracy of the model by quantifying the average squared difference between the observed and predicted values. A lower mean square error (MSE) highlights the predictive performance precision as the model's predictions are closer to the actual values.

Even though R2 and MSE are useful metrics, it's important to take into account any potential trade-offs and their limitations. R2 might not sufficiently penalize models for departures from the actual values and might be sensitive to outliers. MSE does not take into account the direction of deviations and instead expresses the magnitude of prediction errors, even though it is also susceptible to outliers.

The particular objectives of the water quality prediction task will determine which of R2 and MSE to use. It is necessary for researchers and practitioners to balance the need for precise and accurate predictions (MSE) with the importance of explaining variability (R2). Finding a balance between these metrics is essential to creating models that accurately fit the data and effectively generalize to unobserved scenarios, which advances dependable methods for predicting water quality.

CONCLUSION

To sum up, the assessment of water quality prediction models requires careful consideration of metrics like Mean Squared Error (MSE) and R-squared (R2). These metrics are useful instruments for assessing how well models perform, how accurate they are, and how reliable they are in capturing the complex dynamics of water quality parameters.

A comprehensive assessment of the model's fit, the R2 gives a general idea of how well the model explains the variability in the observed data. However, MSE explores the intricacies of prediction errors, offering a sophisticated comprehension of the model's accuracy in assessing water quality.

It is imperative to acknowledge that these metrics possess both advantages and disadvantages. While MSE is sensitive to errors, it may not be able to capture the direction of deviations, and R2 may be affected by outliers. Consequently, a wise decision regarding which of these metrics to use depends on the particular goals of the water quality prediction task.

In the end, striking a balance between providing an explanation for variability and guaranteeing precise forecasts is crucial.

DECLARATIONS

Conflicts of interest

No conflicts of interest in this manuscript

Authors Contributions

Author PS was involved in data collection, data analysis, manuscript writing, Author JK was involved in conceptualization, data validation and critical review manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies, Chennai.
2. Saveetha School of Engineering
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha University

TABLES AND FIGURES

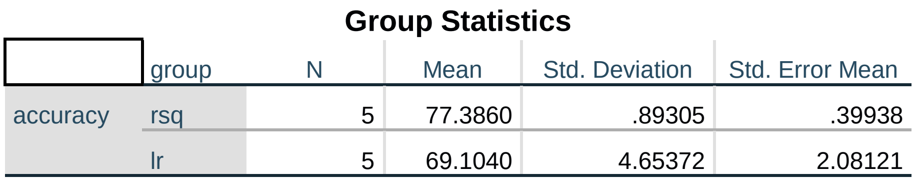
Table 1. Data set description

S.no.	Attributes and parameters
1.	pH
2.	Hardness
3.	Solids
4.	Chloramines
5.	Sulphate
6.	Conductivity
7.	Trihalomethanes
8.	Turbidity
9.	Potability

Table2: shows the sample data of the accuracy of Linear regression and R-squared algorithm

Sample size(from dataset)	R-squared(accuracy)	Linear regression (accuracy)
40	77.39%	62.71%
50	79.83%	66.50%
60	80.86%	69.60%
70	86.75%	72.22%
80	92.34%	74.49%

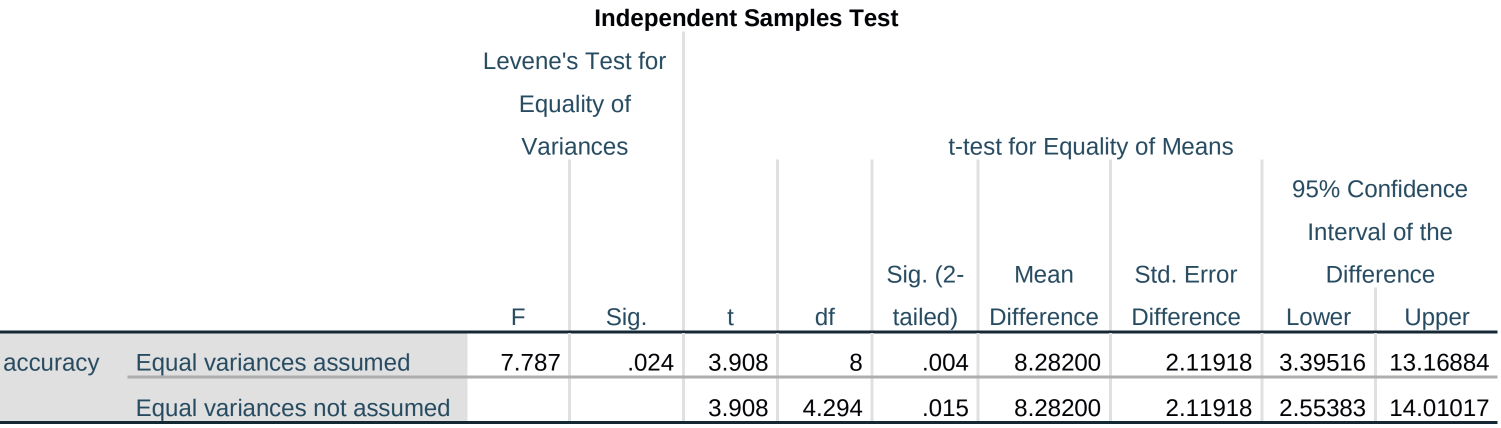
Table3. Group statistics comparison for accuracy of Sample outputs, which contains mean accuracies and standard deviations for R-squared and Linear Regression methods.



Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
accuracy	rsq	5	77.3860	.89305	.39938
	lr	5	69.1040	4.65372	2.08121

Table 4: Shows Independent Sample Test between R-sq and LR algorithm



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
accuracy	Equal variances assumed	7.787	.024	3.908	8	.004	8.28200	2.11918	3.39516	13.16884
	Equal variances not assumed			3.908	4.294	.015	8.28200	2.11918	2.55383	14.01017

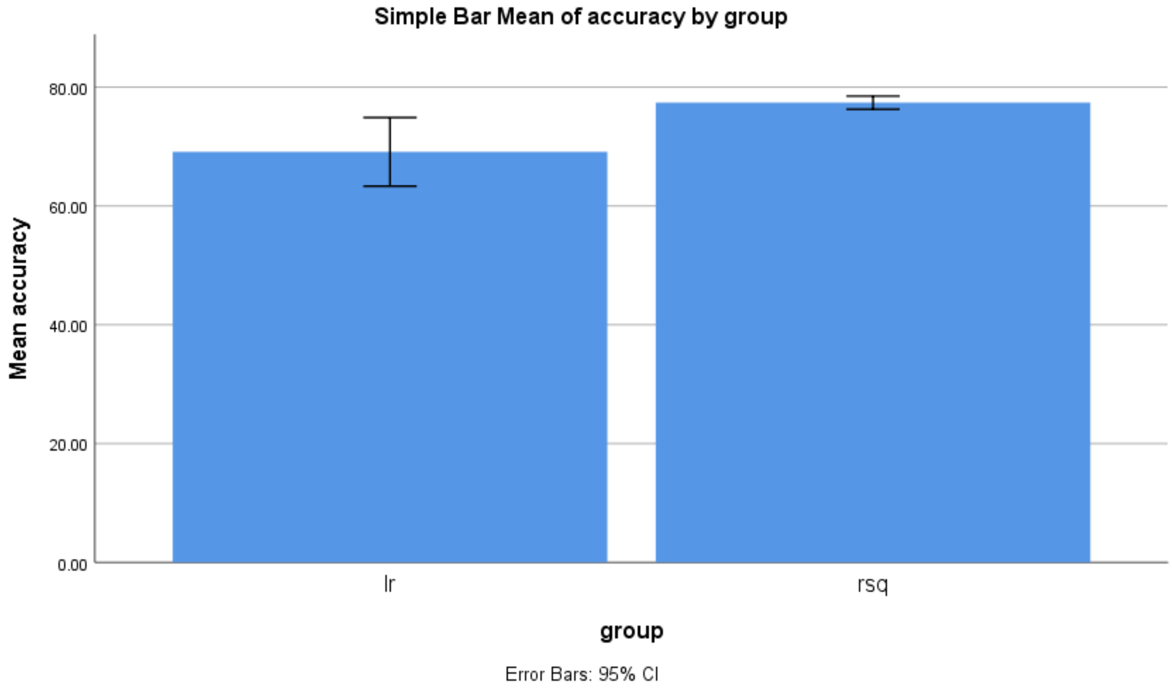


Table 4: Shows mean accuracy comparison graph that shows the comparison between the mean accuracy of the Linear Regression Algorithm and the mean accuracy of the R-squared algorithm along with the error bars.

REFERENCES:

Rahu, M.A., Chandio, A.F., Aurangzeb, K., Karim, S., Alhussein, M. and Anwar, M.S., 2023. Towards design of Internet of Things and machine learning-enabled frameworks for analysis and prediction of water quality. *IEEE Access*.

Chicco, D., Warrens, M.J. and Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, p.e623.

Shams, M.Y., Elshewey, A.M., El-kenawy, E.S.M., Ibrahim, A., Talaat, F.M. and Tarek, Z., 2023. Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, pp.1-28.

Sami, B.H.Z., Sami, B.F.Z., Fai, C.M., Essam, Y., Ahmed, A.N. and El-Shafie, A., 2021. Investigating the reliability of machine learning algorithms as a sustainable tool for total suspended solid prediction. *Ain Shams Engineering Journal*, *12*(2), pp.1607-1622.

Nair, J.P. and Vijaya, M.S., 2022, August. River water quality prediction and index classification using machine learning. In *Journal of Physics: Conference Series* (Vol. 2325, No. 1, p. 012011). IOP Publishing.