Ann. For. Res. 65(1): 6732-6744, 2022
ISSN: 18448135, 20652445

ANNALS OF FOREST RESEARCH
https://www.e-afr.org/

ESTIMATING QUALITY OF WELL WATER USING MACHINE LEARNING MODELS – A CASE STUDY FROM INDIA

Gowri Ganesh N.S [1], Venkata Vara Prasad.D [2]

[1] Department of Artificial Intelligence and DataScience, Saveetha Engineering College, Thandalam,
Chennai- 602105, India
Email ID: gowriganeshns@saveetha.ac.in

[2]Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakam-603110, Chennai, India.
Email I: dvvprasad@ssn.edu.in

Abstract

Water quality is continuously deteriorating with the release of unprocessed industrial effluents, sewage, and wastewater from the households, agriculture runoff, and untreated wastewater has contaminated the water bodies like rivers, lakes, and ponds which in turn affects the groundwater. The quality of water is being affected by several parameters such as pollution, acid rain, and other chemicals from agriculture runoff which include fertilizers and pesticides which make the water toxic. The quality of water that is being taken has a direct effect on the health of a living organism, the consumption of impure water causes various water-borne diseases like cholera, diarrhea and affects child mortality. To overcome these problems, in this project we are going to predict the water quality using various machine learning(ML) algorithms. The training phase includes the usage of various models such as Logistic Regressor (LR), Random Forest(RF), Extra Tree, Decision Tree(DT), Support Vector Machine (SVM), and XG Boost. The models were evaluated and the results of five machine learning models were compared. Out of these five models, Random Forest performed best with prediction accuracy of 98% and precision of 97%.

Keywords: Prediction; Chengalpattu Well Water; Machine Learning; Random Forest; Extra Tree Regressor.

1. Introduction

Water covers 71% of Earth's surface and is vital for all known forms of life. Recently, safe drinking water is very scarce and also being polluted by urbanization and industrialization. With the rapid development of the economy and accelerated urbanization, water pollution has become more and more serious. Water quality has a direct impact on public health and the environment. Consuming impure water has adverse effects on health. Hence, it is important to check the quality of water before consumption. The contamination of groundwater mainly occurs in areas with dense populations. To reduce the amount of contamination that is happening in various water bodies it is essential to assess different aspects of water quality. Predicting water quality parameters a few steps ahead can be beneficial to reduce the number of water-borne diseases which are occurring at large. Therefore, it is necessary to predict the quality of water before consuming it. The current

methods that are developed require a lot of labor and are time- consuming need an alternative automated technology to predict the quality of water.

The proposed work aims to perform water quality prediction of well water data in the Chengalpattu district of Tamil Nadu. And perform analysis and comparison between machine learning and time series algorithms and predict the maximum accuracy. And, to develop an identical process in Auto ML and Auto DL to have a better observation of the results. The results help to get an overview of the standard of water quality and the degree of contamination. Several machine learning models have been built to predict the water quality to date, but they weren't accurate enough. The parameters considered were also not sufficient. They were'nt able to handle the imbalanced and multidimensional datasets. Hence, this work employs machine learning to meet the requirements that the previously used models could'nt achieve. Due to its nonlinear nature, the prediction of water quality becomes a difficult task. But the application of various machine learning techniques has been becoming a powerful source for prediction. These techniques employ historical data of the water quality for the training of machine learning algorithms and help in predicting their future behavior. The Machine Learning models such as Random Forest, Decision Tree, XG

Boost, Logistic Regression and Extra tree regressor are used for prediction. The parameters considered are turbidity, phosphate, nitrate, iron, pH, chloride, and sodium, total dissolved salts (TDS) and chemical oxygen demand (COD). These models are able to handle complex and nonlinear type, large datasets. They are apt to make predictions, when the number of parameters considered is large and on time series data. Based on the predicted values from these models, the accuracy of the machine learning models and time series models is analyzed and compared to find the most suitable model.

2. Literature survey

Ahmed et al. explored 15 supervised machine learning algorithms such as random forest, gradient boosting algorithm, SVMs, stochastic gradient descent, ridge regression, lasso regression, multiple linear regression, logistic regression, polynomial regression, elastic net regression, neural net / Multi-Layer Perceptrons (MLP), Gaussian naive Bayes, K nearest neighbour, decision tree, and bagging classifier to predict the water quality of Rawal Water Lake. Regression algorithms were used to estimate the water quality index and classification algorithms are used to estimate water quality class. These models are analyzed on four input parameters, temperature, turbidity, pH, and total dissolved solids. The data set was collected from the Pakistan Council of Research in Water Resources (PCRWR). To evaluate the accuracy of the regression model the various parameters that are considered are mean absolute error, mean square error, root means squared error and R Squared error. And for classification metrics used are accuracy, precision, recall, F1 score. While analyzing gradient boosting and polynomial was found to be efficient regression algorithm for predicting WQI whereas MLP performed better in predicting WQC. [1]

Prasad et al. proposed a work that uses a machine learning algorithm to identify the water quality of the lake, water samples were taken from Korattur Lake and were tested using various machine learning algorithms such as SVM, Logistic Regression, Random Forest, Decision Tree, and Naive Bayesian. The observations were based on accuracy, Precision, and execution time. Random Forest was observed as the suitable model amongst these algorithms with maximum accuracy of 93% and least execution time. [2]

SasoDzeroski et al. addressed the problem of inferring chemical parameters of river water quality from biological ones. They applied the machine learning algorithm which uses Regression tree induction to the biological and chemical data on the water quality of Slovenian rivers. The data about Slovenian rivers come from the Hydrometeorological Institute of Slovenia performs water quality monitoring for most Slovenian rivers and maintains a database of water quality samples. The data covers the six years from 1990 to 1995. The physical and chemical samples include the measured values of sixteen different parameters such as biological oxygendemand(BOD), chlorine concentration (Cl), CO2 concentration, electrical conductivity, chemical oxygen demand(K2Cr2O7 and KMnO4), concentrations of ammonia(NH4), NO2, NO3 and dissolved oxygen (O2), alkalinity(pH), PO4, oxygen saturation, SiO2, water temperature, and total hardness.[3, 4]

Amir HamzehHaghiab et al. investigated the performance of artificial intelligence techniques including artificial neural network (ANN), group method of data handling (GMDH), and support vector machine (SVM) for predicting water quality components of Tireh River located in the southwest of Iran with DO, BOD, pH, COD, K, Na, EC, Temperature, Mg as parameters. Different types of transfer and kernel functions were tested to develop ANN and SVM. The results of the models based on the DDR index, it was found that the lowest DDR value was related to the performance of the SVM model. The structure of SVM showed that the best accuracy was related to the RBD as the kernel function. Results of ANN indicated that its accuracy is acceptable for practical purposes. [5]

Yafra khan et al. predicted the quality of the natural water resources like lakes, rivers, streams, estuaries. They developed a water quality prediction model using Artificial Neural Network (ANN) and time-series analysis. The data was obtained from the United States Geological Survey (USGS) of the year 2014. The data includes the measurements of four parameters as Chlorophyll, Specific Conductance, Dis-solved oxygen, and Turbidity. They evaluated the performance of their model with Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE), and

Regression Analysis. Their model proves to be a reliable one with the prediction accuracy indicating much- improved results with the lowest MSE and the best Regression value for Specific Conductance (0.99). [6]

Lu H et al. proposed hybrid decision tree-based machine learning models for short-term water quality prediction. The basic models used for these two hybrid models are extreme gradient boosting (XGBoost) and random forest (RF) and an advanced data denoising technique - complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) was used. The CEEMDAN used in the two models was utilized to decompose the raw data with large fluctuations so that the performance of XGBoost and RF can be better. Water samples were collected from the Gales Creek site of Tualatin River in Oregon, the USA the data was collected from May 1st to July 20th, 2019, and divides the raw data into training sets and test sets according to the ratio of 9:1. Two models are used to predict water temperature,
dissolved oxygen, pH value, specific conductance, turbidity, and FDOM, and the prediction results are compared with other benchmark models such as LSTM, RF, and XG Boost. The proposed models outperformed the other benchmark models. [7]

Fitore Muharemi et al. proposed the solution to some challenges when dealing with water quality time series data. They used machine learning and deep learning models such as logistic regression, linear discriminant analysis, support vector machines (SVM), artificial neural network (ANN), deep neural network (DNN), recurrent neural network (RNN), and long shortterm memory (LSTM). The performance evaluation is conducted using the F-score metric. They collected the data from a public water company located in Germany. The parameters they chose are as follows: Time, Water Temperature, Turbidity, Chlorine (cl), pH, Chlorine dioxide, Redox, Electric conductivity, and flow rate. [8, 9]

Prasad et al. proposed deep learning models such as Artificial Neural Network, Recurrent Neural Network, and long-short term memory were used in this paper. These models were tested on water samples of Korattur Lake. And these models were evaluated based on accuracy, precision, and execution time. LSTM was observed as the most efficient algorithm with an accuracy of 94 %, highest precision, and least execution time. [10]

Singha S et al. used extreme gradient boosting, random forest, and artificial neural networks to perform a comparison with the deep learning model to predict the quality of groundwater. A total of 226 groundwater samples were collected from an agriculturally concentrated area Arang of Raipur district, Chattisgarh, India. And entropy weight-based groundwater quality index was computed by measuring numerous physicochemical parameters. All the models were evaluated based on the five-performance metrics such as MSE, MAE, RMSE and MAPE, and R2. DL algorithm outperformed other models with RMSE = 1.254. Relatively higher prediction performance is observed in the XGBoost model. The order of the model's performance is DL>XGBOOST>RF as per R2 values.[11]

Archana Solanki et al. predicted the water quality challenges in the reservoir. The water quality is predicated on a continuous water quality dataset from the Chaskaman reservoir, with pH, dissolved

oxygen, and turbidity as parameters using supervised learning such as ANN techniques. Further analysis using the advanced predictive technique, such as deep learning and Performance analysis was done using metrics such as Mean squared error and mean absolute error. This system can be implemented on a system to continuously monitor the quality of the water. It can be helpful to monitor the quality of water in any uncertain condition.[12]

P.Varalakshmi et al. used the Naıve Bayesian model to predict the water quality. The climatic conditions and the environmental impacts were considered to decide if the water is suitable for drinking purpose. They considered the water quality parameters such as Chloride, Nitrate, Nitrogen, TDS, pH and Hardness to predict the water quality. Thus, this model is designed for assessing the water quality with respect to drinking water standards and calculating the posterior probability.[13]

Xiang Yunrong et al. proposed a machine learning model LS-SVM which predicts the water quality of the Liuxi River in Guangzhou. Their model combined the least squares support vector machine (LS-SVM) with particle swarm optimization (PSO) to overcome the shortcomings in the traditional BP algorithms. They also compared the models such as SVRPSO, BPNN, ARIMA, and GML. Their parameters include DO and COD. Their algorithm is simple to implement and effective and is inexpensive in terms of memory and time required. This approach provides solutions with better quality within a reasonable time limit. [14]

Hamid ZareAbyaneh et al. the efficiency of ANN and MLR models was investigated in the prediction of two major water quality parameters, BOD and COD, in Ekbatan wastewater treatment plant, Tehran, Iran. The performance of the models was evaluated using the coefficient of correlation (r) and root mean square error statistics (RMSE). The results indicate that the ANN model with minimum input parameters, temperature (T), pH, total suspended solids, and total suspended could be successfully used for predicting BOD and COD concentrations. This result suggests that the use of more input parameters will not necessarily lead to improvements in predicted results, but the type of input parameters is more important than its number.[15]

3. Methodology

This work's focal point is on predicting the quality of well water in the Chengalpattu district of Tamil Nadu. First, the dataset is collected and pre-processed to clean the data and for feature selection. After the feature selection, the classes are assigned to the data based on the values of the parameters. Once the classes have been assigned, the data set is split into training and testing data. The Machine Learning models such as Random Forest, Decision Tree, Logistic Regression, Extra Tree and XG Boost. The data is then tested and evaluated to find the accuracy of prediction, precision, and thus the quality of the water. The accuracy obtained from all the models is compared and analyzed to find the best suitable model. Drinking water quality pursues the accuracy of prediction results and the stability of prediction error fluctuation. The conventional models used

for prediction had many drawbacks. The seasonal effects on parameters were not considered. They could not identify the factors that make it unfit for drinking. The under and over-fitting of predictions were not handled. Water quality affected by meteorological and hydrological factors was not predicted accurately. So, our work was proposed with models that can handle large and complex datasets of nonlinear type. They are suitable for making predictions on datasets with huge number of parameters and on time series data. It also performs well with unstructured and semi structured data. The output acquired is more informative than any other algorithm. The architecture of the proposed system is shown in Figure 1.

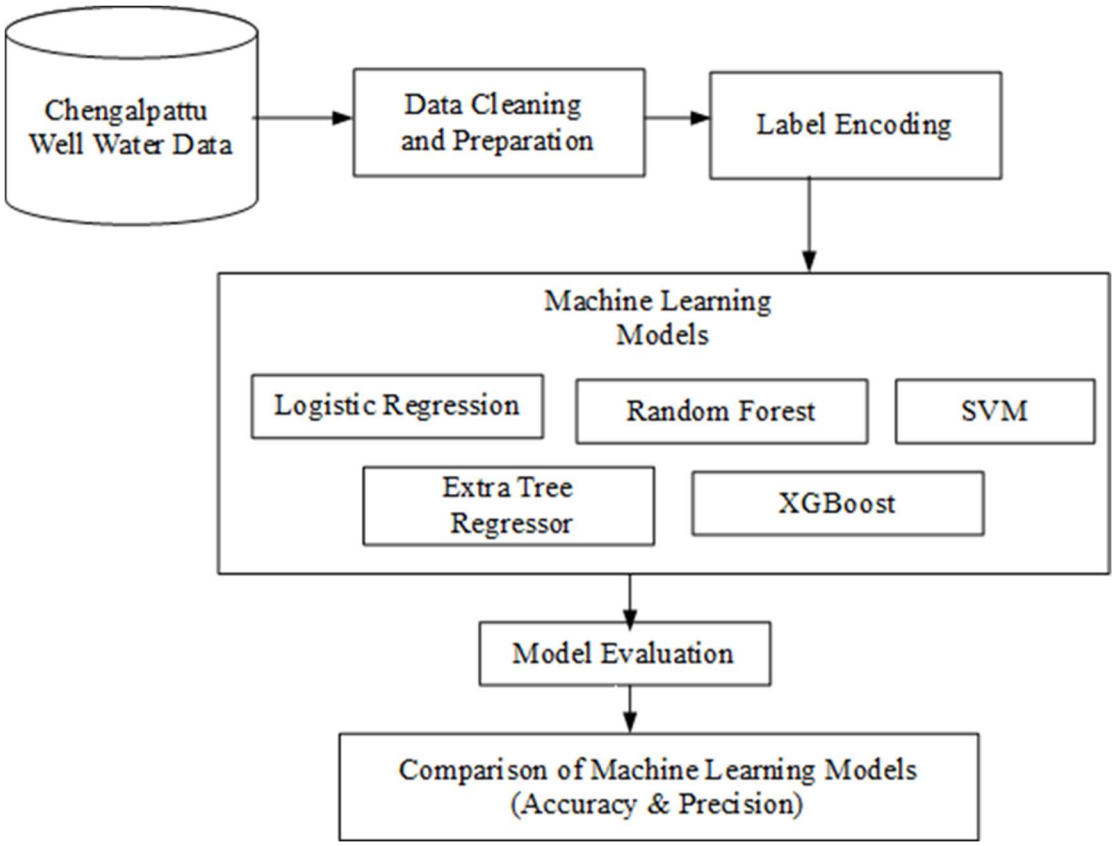

Figure 1. Methodology for Water Quality Prediction

The first step is the data pre-processing, which involves cleaning the data and feature selection. Data cleaning is the process of detecting and removing corrupt or inaccurate records from the dataset. Feature selection refers to selecting the most important attributes which contribute most to the prediction variable (Class) because having irrelevant features in our data can decrease the accuracy of the models and makes the model learn based on irrelevant features. The next step is assigning classes to the data by considering the values of all the parameters. The water data is classified based on WQI and assigned class names as excellent, good, poor and very poor. Once the classes have been assigned for all the data, the dataset is split into training and testing sets.

Among the available data, 80% of the data is used for training and the remaining 20% is used for testing. The models used for training include the machine learning models such as Random Forest, Logistic Regression, Decision Tree, XG Boost and Extra tree regressor. The classification algorithms were evaluated based on precision and accuracy. Based on the outcomes obtained from models, the results are analysed to find the best suitable model for water quality prediction.

3.1. Dataset Collection

The dataset was collected from various wells in the Chengalpattu district of Tamil Nadu. It consists of water data for over 28 consecutive years (1992 to 2020). The dataset consists of about 10,248 records and consists of 13 chemical parameters such as TDS- Total Dissolved Solids, NO2+NO3-Nitrate, Ca- Calcium, Mg - Magnesium, Na - Sodium, K - Potassium, Cl -Chloride, So4 - Sulfate, F - Fluoride, CO3 + HCO3 (Alkalinity), pH- power of hydrogen, HAR_Total-Total Hardness and SAR - Sodium Adsorption Ratio. The sample dataset is shown in Table 1.

Table 1: Sample Dataset

TDS	NO2+NO3	Ca	Mg	Na	K	Cl	SO4	F	HCO3	pH_GEN	HAR_Total	SAR
276	11	32	9	53	2	53	17	0.26	122	7.7	115	2.13243
175	6	22	10	23	1	28	34	0.23	61	7.8	95	1.020864
404	5	52	15	58	20	53	77	0.47	214	7.9	190	1.823238
350	0	40	34	46	2	53	41	0.48	268	8.1	240	1.292377
1483	24	72	61	322	84	390	240	0.42	415	7.8	430	6.749407
1432	26	56	61	311	86	432	240	0.51	262	8	390	6.843795
535	2	116	1	92	2	64	29	0.64	445	7.5	295	2.335613
283	1	58	10	37	2	43	19	0.44	220	7.6	185	1.180493
320	3	46	17	44	11	25	36	0.62	232	8.5	185	1.408163
968	7	96	97	117	4	411	72	0.35	244	8.4	640	2.013829
903	7	92	58	150	9	248	144	0.48	342	7.8	470	3.015375
1170	26	124	72	193	4	319	120	0.47	445	7.9	605	3.411239
625	1	40	88	64	2	209	96	0.48	244	8.1	460	1.295335
213	4	42	23	5	1	7	17	0.54	201	8	200	0.154006

1120	3	64	4	368	9	330	48	0.28	567	8.2	175	12.06053
533	6	100	28	55	1	184	41	0.42	195	7.5	365	1.252687
322	6	42	19	51	4	32	19	0.95	256	7.7	185	1.639986
220	4	34	9	28	2	39	35	0.3	110	8	120	1.103259
269	11	28	18	35	3	60	40	0.16	73	8.2	145	1.269004
1074	14	44	146	133	9	347	113	0.05	439	7.7	710	2.170387
350	1	16	46	55	2	57	20	0.05	299	8	230	1.580279
1215	3	22	69	299	59	372	168	0.09	342	8.3	340	7.065974
501	3	84	27	64	9	121	9	0.13	275	8.4	320	1.554541
248	4	60	17	7	2	39	8	0.02	98	8.5	220	0.205441

3.2. Water Quality Index

The step involved in calculating the water quality index is elaborated in the work [2]. The chemical parameters are TDS- Total Dissolved Solids, NO2+NO3- Nitrate, Ca- Calcium, Mg - Magnesium, Na - Sodium, K - Potassium, Cl - Chloride, So4 - Sulfate, F - Fluoride, CO3 + HCO3 (Alkalinity), pH- power of hydrogen, HAR_Total - Total Hardness and SAR - Sodium Adsorption Ratio. The permissible limit and desirable limits for each chemical parameter is given in the Table 1. The minimum and maximum value for each parameter which was obtained from the data is also tabulated. Based on the degree of impact each chemical parameter could causes on the water, the weights are assigned to the parameters. The weights are assigned from 1 to 5, 5 being the high degree of impact on water as given in Table 2.

Table 2: Influential Chemical Parameters on Water Quality.

Chemical Parameter	Permissible Limit	Acceptable / Desirable Limit	Minimum Value in the data	Maximum Value in the data	Assigned Weights
TDS- Total Dissolved Solids	5000	2000	2000	5000	5
NO2+NO3 -	100	45	45	100	3
Ca - Calcium	200	75	75	200	4
Mg - Magnesium	100	30	30	100	3
Na - Sodium	2	2	2	2	5

K - Potassium	11	0.4	0.4	11	3
Cl -Chloride	1000	250	250	1000	5
So4 - Sulfate	400	200	200	400	3
F - Fluoride	1 - 1.5	1 - 1.5	1	1.5	2
CO3 + HCO3 (Alkalinity)	600	200	200	600	3
pH- power of hydrogen	6.5-8.5	6.5-8.5	6.5	8.5	4
HAR_Total - Total Hardness	300	300	300	300	2
SAR - Sodium Adsorption Ratio	10	1	1	10	1
				Total	**47**

3.3. Dataset Description

The Table 3 shows the details about the distribution of Chengalpattu Well Water data into excellent, good, poor, and very poor water.

Table 3: Dataset Description

Dataset	No. of Records	No. of Parameters	No. of Classes	Class Distribution
Chengalpattu Well Water Data	10248	13	4	Excellent - 1420 Good Water- 2201 Poor Water- 4299 Very Poor Water- 2328 Not Suitable for drinking - 0

The categorization of drinking water status was done based on the water quality index level as given in the Table 4.

Table 4: Categorization of drinking water based on WQI

Water Quality Index Level	Water Quality Status
0-50	Excellent
50-100	Good
100-200	Poor

200-300	Very Poor
>300	Not suitable for drinking

3.4. Data Cleaning and Feature Selection

Data cleaning is the process of preparing data for classification and prediction by removing the data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. It also includes removing the record with one or more missing data in the dataset. Because training the data with irrelevant data produces inaccurate results. Feature selection refers to selecting the most important attributes which contribute most to the prediction variable (Class) because having irrelevant features in your data can decrease theaccuracy of the prediction.

3.5. Dataset Splitting

Before training the machine learning model it is necessary to split the data into training and testing sets. After splitting the data, the model is trained and tested with a certain part of the data to measure the accuracy of the model's performance. The data was split in the ratio of 4:1 for training and testing respectively. For multi-class classification, out of 10,000 samples, 8000 samples were used for training and 2000 for testing.

4. Methods

The machine learning algorithms such as Random Forest (RF), Extra Tree Regressor, XGBoost, Logistic Regression (LR) and SVM and are used for training and testing. The machine learning models are applied for multi-class data. The working of these machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) are explained in the work [10]. The Extra Tree Regression and XGBoost are detailed as follows.

4.1. Extra Tree Regression

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. Extremely Randomized Trees, or Extra Trees for short, is an ensemble machine learning algorithm. Specifically, it is ensemble of decision trees and is related to other ensembles of decision trees algorithms such as bootstrap aggregation (bagging) and random forest. The Extra Trees algorithm works by creating a large number of unpruned decision trees from the training dataset. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification.

- Regression: Predictions made by averaging predictions from decision trees.
- Classification: Predictions made by majority voting from decision trees.

4.2. XG Boost

XGBoost stands for eXtreme Gradient Boosting. XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. The name xgboost, though, refers to the engineering goal to push the limit of computations resources for boosted tree algorithms.

5. Results

The machine learning models such as Logistic Regression, Random Forest, Extra Tree Regressor, Support Vector Machine (SVM) and XGBoost were trained and evaluated using test data. The accuracy of the machine learning models and precision was tabulated in the Figure 3. The accuracy obtained with Logistic Regression was found to be 90% and the precision was found to be 93%. The accuracy obtained with Random Forest was found to be 98% and the precision was found to be 97%. The accuracy obtained with Extra Tree Regressor was found to be 93% and the precision was found to be 97%. The accuracy obtained with SVM was found to be 97% and the precision was found to be 96%. The accuracy obtained with XGBoost was found to be 97% and the precision was found to be 97%.

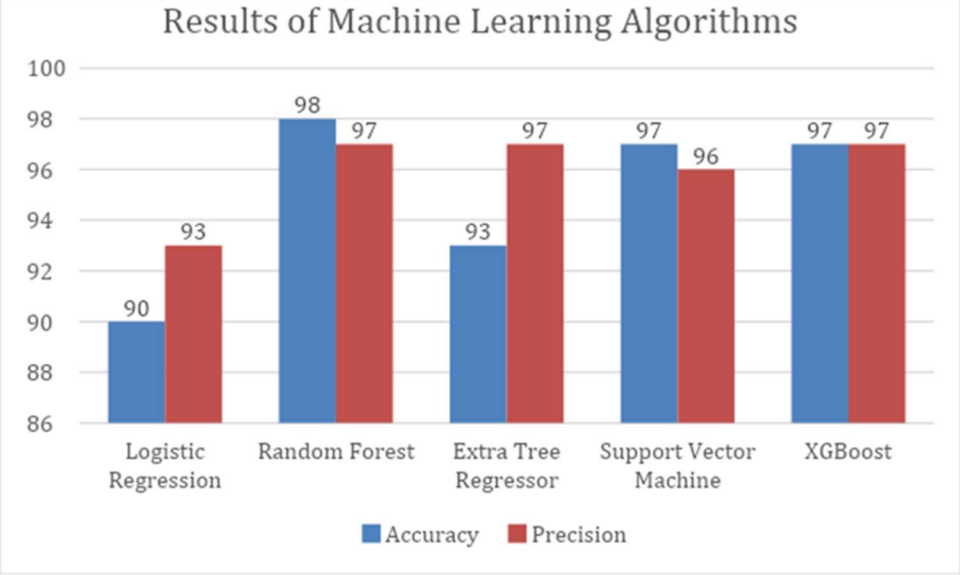

Figure 3. Accuracy & Precision of Machine Learning Algorithms

7. Conclusion

Water is an inevitable resource for all living organisms and has got a serious significance in checking the quality due to the pollution caused by the influence of various external factors like industrial effluents, acid rain, dumping of degradable and non-degradable wastes, etc., and therefore it is necessary to check the quality of water before consumption. Hence, the work focuses on predicting the water quality of well water in the Chengalpattu district of Tamil Nadu. The models used for Training and Testing include Machine Learning models such as Random Forest

(RF), Support Vector Machine (SVM), Extra Tree Regressor (ET), Logistic Regression (LR) and XG Boost binary classification and multi-class classification. The machine learning models produced an average accuracy of around 96.8%. Out of the five machine learning algorithms, Random Forest was found to be the best suitable algorithm for this work since it produces the highest precision of 98%. The random forest also has the highest precision compared to all the other machine learning algorithms. So, the random forest can be used for the prediction of water quality given the attributes. The work can be extended byreducing the over fitting of data and the data set can be trained using hybrid models of machine learning and deep learning. Since deep learning algorithms can handle large amounts of data, the hybrid model might be more efficient as it can produce high accuracy as well as can handle large data sets.

9. References

[1]. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. Water, 11(11), 2210.

[2]. Prasad, V. V. D., Venkataramana, L. Y., Kumar, P. S., Prasannamedha, G., Soumya, K., &Poornema, A. J. (2020). Water quality analysis in a lake using deep learning methodology: prediction and validation. International Journalof EnvironmentalAnalytical Chemistry., DoI: https://doi.org/10.1080/03067319.2020.1801665

[3]. Chen, E., & He, X. J. (2019). Crude oil price prediction with decision tree based regression approach. Journal of International Technology and Information Management, 27(4), 2-16.

[4]. Džeroski, S., Demšar, D., &Grbović, J. (2000). Predicting chemical parameters of river water quality from bioindicator data. Applied Intelligence, 13(1), 7-17.

[5]. Haghiabi, A. H., Nasrolahi, A. H., &Parsaie, A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal, 53(1), 3-13.

[6]. Khan, Y., & See, C. S. (2016, April). Predicting and analyzing water quality using Machine Learning: a comprehensive model. In 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT) (pp. 1-6). IEEE.

[7]. Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere, 249, 126169.

[8]. Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication, 3(3), 294-307.

[9]. Poornima, S., &Pushpalatha, M. (2019). Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units. Atmosphere, 10(11), 668.

[10]. Prasad D., Vara V., Venkataramana L.Y., Kumar P.S., Prasannamedha G., Soumya K., and Poornema A. (2021). Prediction on water quality of a lake in Chennai, India using machine learning algorithms. Desalination And Water Treatment, vol. 218:44–51.

[11]. Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., & Kumar, S. (2021). Prediction of groundwater quality using efficient machine learning technique. Chemosphere, 276, 130265.
[12]. Solanki, A., Agrawal, H., &Khare, K. (2015). Predictive analysis of water quality parameters using deep learning. International Journal of Computer Applications, 125(9), 0975-8887.
[13]. Varalakshmi, P., Vandhana, S., &Vishali, S. (2017, January). Prediction of water quality using Naive Bayesian algorithm. In 2016 Eighth International Conference on Advanced Computing (ICoAC) (pp. 224-229). IEEE.
[14]. Xiang, Y., & Jiang, L. (2009, January). Water quality prediction using LS-SVM and particle swarm optimization. In 2009 Second International Workshop on Knowledge Discovery and Data Mining (pp. 900- 904). IEEE.
[15]. ZareAbyaneh, H. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering, 12(1), 1-8.