Turkish Journal of Engineering

https://dergipark.org.tr/en/pub/tuje

e-ISSN 2587-1366

Application of machine learning algorithms in the investigation of groundwater quality parameters over YSR district, India

Jagadish Kumar Mogaraju[*1]

[1]*Indira Gandhi National Open University, Life Sciences, Kadapa, India*

Keywords
Groundwater
Machine learning
Support Vector Machines
Classification Trees
Random Forest

Research Article
DOI: 10.31127/tuje.1032314

Received: 04.12.2021
Accepted: 02.03.2022
Published: 01.04.2022

Abstract

Human life sustained for decades due to the availability of basic needs, and freshwater is one of them. However, groundwater quality is constantly under pressure. This can be attributed to anthropogenic activities not limited to urban areas but to rural zones. Machine learning methods like linear discriminant analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbour (KNN), Support Vector Machines (SVM) and, Random Forest (RF) models were used to analyse groundwater quality variables. The mean accuracy of each classifier was calculated, and the obtained mean accuracies were 77.5% (LDA), 87% (CART), 96% (KNN), 93.5% (SVM) and 96% (RF). RF and KNN models were selected as optimal models with higher accuracy. This study made it apparent that machine learning algorithms can estimate and predict water quality variables with significant accuracy. In this study, the observations and variables were compared with the water quality index and drinking water limits provided by the Bureau of Indian Standards. The water quality index for each observation was calculated. If at least four variables have a higher value than prescribed limits, it was assigned a value of 1; if more than four variables reported higher values, it was assigned a value of 2.

1. Introduction

Machine learning seeks to predict an outcome by extracting patterns from big datasets, usually in the form of an algorithm [1]. Machine learning is an advanced tool to understand groundwater quality variables over a study area [2]. Machine learning tools were used in the planning of several irrigation projects all over the world [3]. Mining is prevalent in the study area and is considered an essential economic source [4]. The irrigation, salinization, ion exchange, carbon dissolution and weathering processes can affect the groundwater quality. Some of them can be due to anthropogenic activities [5]. This region's deterioration of groundwater quality is mainly due to overexploitation and contamination [6]. Agriculture has been an essential economic source in India, and 60% of the populace depend on it for livelihood [7]. The quality of groundwater can be due to the percolation and infiltration of pollutant-laden rainwater; however, domestic and agricultural activities get involved in some places [8]. Crystallines, shale, limestone and quartzite are some litho units that can affect groundwater quality [9-10]. Approximately 91000 hectares of land in the study area are irrigated by water from the local canals. One thousand three hundred sixty-eight minor irrigation tanks irrigate 47000 hectares [9]. The rise in the groundwater level was 2.11m in alluvium, 2.50m in limestones, 3.82m in shales, 5.35m in crystallines and 7.32 in quartzites [9]. The groundwater quality contamination due to nitrates and pesticides was studied using machine learning models like Extreme Gradient Boosting (XGB), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) and were evaluated for their accuracy [11]. Support Vector Regression (SVR), Artificial Neural Network (ANN), Random Forest (RF), and Adaptive Boosting (Adaboost) models were used to forecast and evaluate water quality indexes [3]. The ensemble models of RF and Boosted Regression Trees (BRT) were investigated with Multivariate Discriminative Analysis (MDA) [12]. SVM, MDA and BRT models were used in installing a framework for evaluating nitrate contamination in groundwater [13]. Artificial Intelligence techniques like SVM, Naïve Bayes classifier and Particle Swarm Optimization (PSO) were used to predict the water quality index [14]. ANN models

[*] **Corresponding Author**

[*](jagadishmogaraju@gmail.com) ORCID ID 0000-0002-6461-8614

Cite this article

Mogaraju, J. K. (2023). Application of machine learning algorithms in the investigation of groundwater quality parameters over YSR district, India. Turkish Journal of Engineering, 7(1), 64-72

were compared with GIS tools to delineate the potential zones of groundwater in Ethiopia [15]. SVM, Multivariate Adaptive Regression Splines (MARS), k-nearest neighbour (KNN), ANN, RF, BRT, penalized discriminant analysis (PDA), flexible discriminant analysis (FDA), quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA) models were compared in combination with GIS tools to assess groundwater quality in Iran [16]. RF, C5.0 and MARS models integrated with GIS were used in potential groundwater mapping [17]. The collected feature vectors were subjected to machine learning-based feature selection to determine the best feature sets for predicting soil water content [18]. Determining the soil surface humidity in vegetated areas is problematic; hence, polarimetric decomposition models and machine learning-based regression models were used to solve the problem [19].

Machine Learning has been one of the most quickly evolving areas in Artificial Intelligence research. Decision trees are highly effective and straightforward to understand. On the other hand, individual trees can be susceptible to slight data changes. Even greater prediction can be accomplished by using this variability to develop many trees from the same data [20]. Several Machine Learning algorithms have acquired popularity in part owing to their transparency. The Decision Tree algorithm, also known as the Classification and Regression Trees (CART) algorithm, is one of them. The CART algorithm is a classification algorithm used to construct a decision tree using Gini's impurity index.

The main purpose of this study is to investigate and report the prediction accuracies of the variables of groundwater quality using machine learning algorithms. The primary contribution of this study is to extend the role of machine learning in understanding the subsurface hydrology parameters and their characteristics on spatial domain.

2. Method

2.1. Data collection

The datasets used in this study were collected from the Central Ground Water Board (CGWB), Government of India. These datasets can be accessed at [21-22]. The shapefile used in producing location map of the study area was downloaded from the Website of Geodata, The University of Texas at Austin [23]. The location map of the study area is revealed in Figure 1.

The water samples were collected from 56 places, and a Physico-chemical analysis was performed. The water analysis was focused on Bicarbonates (HCO_3), Calcium (Ca), Chloride (Cl), Electric Conductivity (EC), Fluoride (F), Magnesium (Mg), Sodium (Na), Nitrates (NO_3), pH, Residual Sodium (RSC), SAR, SO_4, Total Hardness (TH) and Total Alkalinity. These variables were compared with the drinking water limits proposed by the Bureau of Indian Standards (BIS) (IS 10500:2012). The values of every variable were compared with the BIS limits and then assigned labels like 'N' (Normal) and 'H' (High). If at a given observation, the number of 'H's are ≤ 4, then it was assigned a value of '1' (Manageable), and if the number of 'H's are >4, then a value of '2' (High) was assigned. In order to pass the data onto the machine learning tools, the factor levels were kept to a minimum of 2. Unfortunately, some of the data is unavailable. Instead of opting for data imputation techniques, Inverse Distance Weighting was used to compensate for the missing values. 'R' version 4.1.1 with packages like 'caret,' 'mlbench,'randomForest' and other packages that were integrated with caret packages were used in this analysis. The data was split into 'training' and 'test' data. 80% of the observations (rows) were used as training data, and the remaining 20% was used as test or validation data. One linear algorithm (LDA), two non-linear algorithms (CART & KNN) and advanced algorithms (SVM & RF) were used. These algorithms were run using 10-fold cross-validation. Accuracy and Kappa values were obtained.

This work was framed for classification. Hence, accuracy was mainly considered to select the appropriate model instead of R^2 and RMSE values. Mean accuracy post comparing five models were considered to select the appropriate model. The skill of the selected model was estimated using test data. A confusion matrix was prepared using both training and test or validation data. Fourteen numerical variables and a 1-factor variable with two levels, i.e., '1' and '2', were used. The descriptive statistics are represented in table 1.

2.2. Study area

Waterlogging is caused by the intensive use of surface water in irrigation project command areas. The increased use of groundwater for agriculture, industry, and home purposes produces ongoing depletion of water levels, well drying, and water quality issues [24]. Thus, water resource management is necessary to protect aquifers and ensure they continue to provide water at a reasonable cost. In the drought-prone Cuddapah district, integrated geological, hydrological (surface and groundwater), and geochemical elements have been researched to develop and manage water resources. Crystallites, quartzites, shales, and limestones are the primary lithological units. Canal water irrigates about 91 000 acres of land in the Cuddapah area. In addition, 1368 minor irrigation tanks irrigate a registered ayacut of roughly 47 000 ha [9].

In the entire district, 503 spring channels originating from rivers/streams have been identified, with the ability to irrigate around 8700 acres. In quartzites, 5.35 m (crystallines), 3.82m (Shales), 2.50m (Limestone), and 2.11m in alluvium, the average seasonal rise in groundwater level is 7.32 m, 5.35 m(crystallines,) 3.82 m (shales), 2.50m in limestones, and 2.11m in alluvium [9]. Large amounts of groundwater are accessible in mining sites, which can be used and managed appropriately by the irrigation department/cultivators. According to groundwater assessment studies, the district has 584 million m^3 of groundwater accessible for future irrigation [8]. According to chemical analysis, the groundwater quality in various rock units is within legal limits for irrigation and residential use; however, specific conductance, chloride, and fluoride levels are high in a few spots. This could be due to untreated effluents, a faulty drainage system, or fertilizer application [25].

Table 1. Descriptive statistics

		Valid	Mean	Std. Deviation	Minimum	Maximum
HCO3	1	32	230.650	91.415	28.418	391.530
HCO3	2	24	441.555	203.365	213.751	972.903
Ca	1	32	39.455	19.418	7.865	112.000
Ca	2	24	79.923	30.369	41.270	162.424
Cl	1	32	142.035	81.228	19.883	350.685
Cl	2	24	526.148	812.553	50.012	4220.785
EC	1	32	960.865	399.591	137.111	1662.483
EC	2	24	2763.739	3219.268	1064.999	17435.179
F	1	32	0.497	0.253	0.056	0.983
F	2	24	0.897	0.502	0.370	2.147
Mg	1	32	29.314	11.706	3.985	51.672
Mg	2	24	61.970	33.765	25.642	172.373
Na	1	32	117.287	72.153	12.436	292.843
Na	2	24	427.485	711.630	96.625	3645.152
NO3	1	32	26.125	20.753	5.435	116.000
NO3	2	24	74.934	109.238	0.776	558.252
pH	1	32	3.432	1.381	0.604	7.410
pH	2	24	5.467	1.310	3.046	7.945
RSC	1	32	1.415	0.989	0.159	3.621
RSC	2	24	4.197	2.690	1.000	12.858
SAR	1	32	2.134	1.240	0.257	4.522
SAR	2	24	6.603	8.958	1.418	45.873
SO4	1	32	64.895	33.530	6.859	131.419
SO4	2	24	206.106	299.020	53.558	1564.692
TH	1	32	219.010	85.925	36.014	371.911
TH	2	24	454.290	209.380	212.140	1073.733
TA	1	32	190.081	74.262	23.293	320.926
TA	2	24	372.755	185.069	196.160	961.798

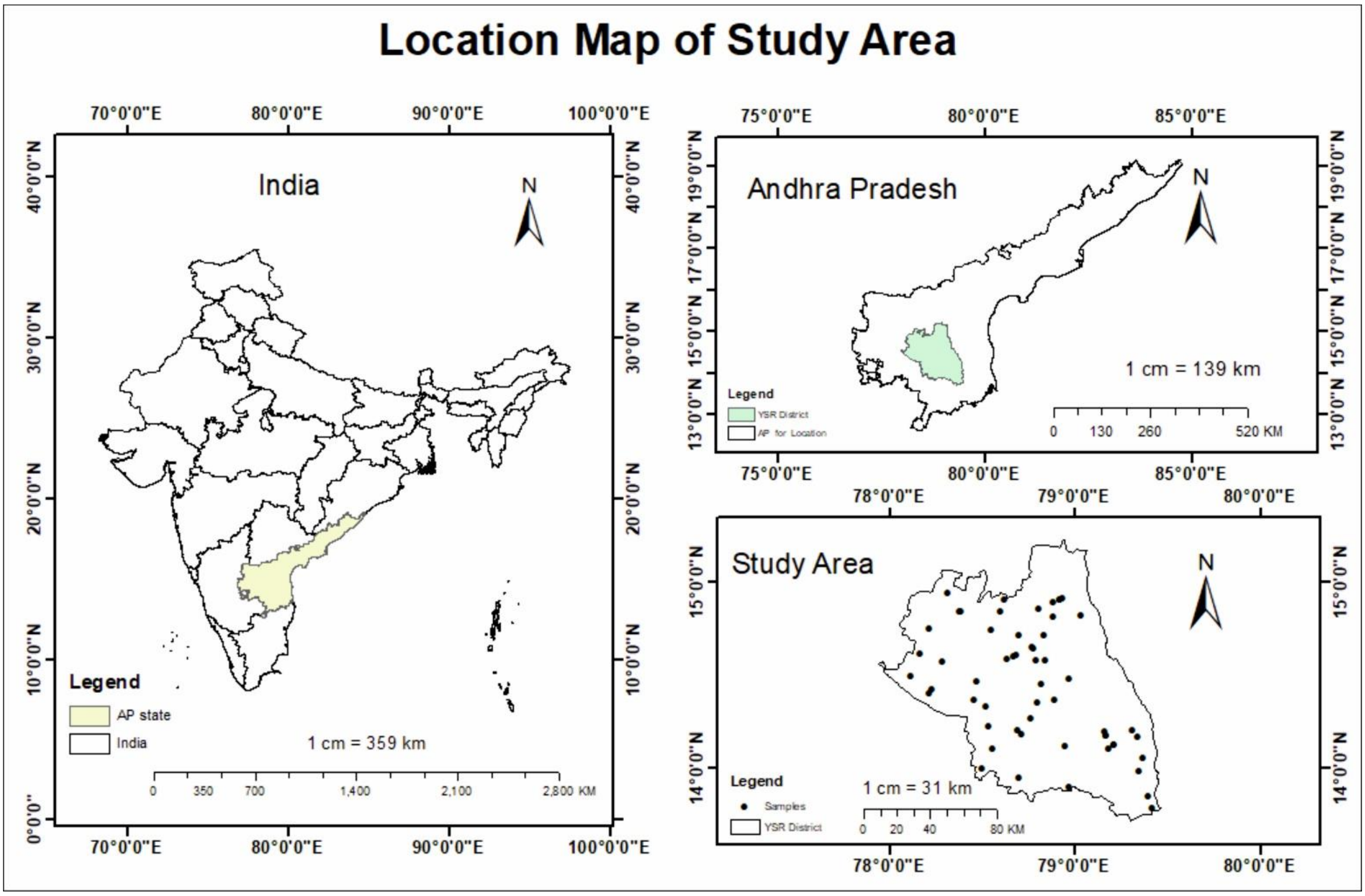

Figure 1. Location map of the study area

2.3. Methodology

The detailed methodology is shown in figure 2. In the pre-processing step, linear discriminant analysis (LDA) is the most often used dimensionality reduction technique. The target is to project a dataset onto a lower-dimensional space with excellent class-separability to minimize overfitting ("dimensionality's curse") and reduce computational costs. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation approaches for dimensionality reduction. PCA is an "unsupervised" technique. It ignores class labels and aims to find the directions (known as principal components) that maximize a dataset's variance. Unlike PCA, LDA is "supervised," which means it calculates the directions ("linear discriminants") that will represent the axes that maximize the separation between several classes. Although it may appear logical that LDA is superior to PCA for multi-class classification tasks with known class labels, this is not necessarily the case.

Data classification can be done in a variety of ways. Two widely used data categorization and dimensionality reduction techniques are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). When the within-class frequencies are unequal, and their performances have been tested on randomly produced test data, it is handled using Linear Discriminant Analysis. This approach maximizes the ratio of between-class variation to within-class variance in each given data set, ensuring maximum separability. The classification challenge in speech recognition is tackled with the help of Linear Discriminant Analysis.

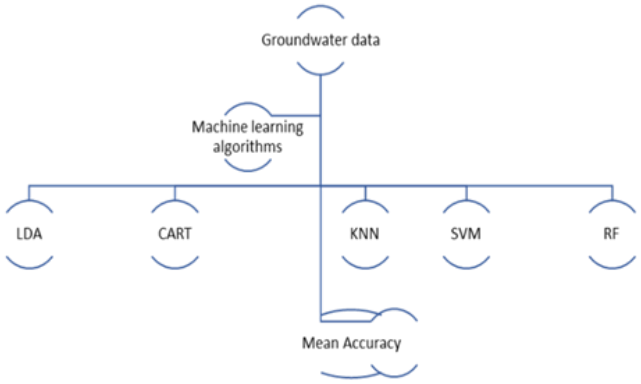

Figure 2. Methodology

2.3.1. Linear Discriminant Analysis (LDA)

In comparison to Principal Components Analysis, we opted to design an algorithm for LDA in the hopes of delivering better classification. The most significant distinction between LDA and PCA is that PCA focuses on feature classification, whereas LDA focuses on data classification. When PCA transforms data sets to a different space, the shape and position of the original data sets change. In contrast, LDA seeks to provide more class separability and create a decision zone between the given classes. This strategy also aids in a better understanding of the feature data distribution [26].

$$S_B = \sum_{i=1}^{c} Ni(mi - m)(mi - m^T)$$

Where 'S_B' is the between-class matrix, 'm' is the overall mean, 'm_i' is the sample mean, and 'N_i' is the size of the respective classes.

Discriminant analysis is a method for distinguishing between two or more groups that meet the study's requirements. Linear combinations of discriminating variables that measure qualities on which the groups are predicted to differ are generated, resulting in a model extrapolating to the rest of the region. Linear discriminant analysis (LDA) has previously been used in New Zealand hydrological studies to distinguish between groups of rivers at base flow conditions [27]. Discriminant analysis has also been used to investigate the distribution of nitrate in groundwater [28]. However, only limited studies have used LDA to discriminate between zones of varying redox status. It is a straightforward machine learning algorithm with a wide range of applications.

2.3.2 Classification and Regression Trees (CART)

The Classification and Regression Trees (CART) algorithm is a decision tree classification technique that uses Gini's impurity index as a splitting condition [29]. CART is a binary tree created by continuously splitting each node into two child nodes. Statistician Leo Breiman coined the term to characterize Decision Tree algorithms to solve classification or regression predictive modeling problems. A Decision Tree is a technique for predictive analysis. The decision tree is the predictive model used here. It is used to go from observations about an item represented by branches to the item's target value, represented by leaves. Decision trees are popular machine learning approaches due to their readability and simplicity. The nodes in the decision tree are divided into sub nodes based on an attribute's threshold value. The CART algorithm uses the Gini Index criterion to find the best homogeneity for the sub nodes. The root node is used as the training set, and the best attribute and the threshold value are used to divide it into two parts. In addition, the subsets are divided using the same rationale. This process is continuously repeated until the tree has the last pure sub-set or the maximum number of leaves conceivable in that growing tree. Tree Pruning is another name for this [30].

$$GI = \sum_{i=0}^{c} Pi(1 - Pi)$$

'GI' is the Gini Index, and 'P' is the estimated output.

The Supervised Learning category includes the k-Nearest Neighbour method used for classification and regression. It is a flexible approach that may also fill in missing values and resample datasets.

2.3.3 K-Nearest Neighbour (KNN)

As the name suggests, the K-Nearest Neighbour algorithm uses k-Nearest Neighbours or Data points to forecast the class or continuous value for a new Datapoint. The nearest neighbours are the data points with the shortest distance in feature space from our new data point. Moreover, k is the number of data points we consider in our method implementation. As a result, while utilizing the KNN method, the distance metric and the K value are two key factors. The most often used distance measure is Euclidean distance. We can also employ Hamming, Manhattan, and Minkowski distances depending on needs. It considers all of the data points in the training dataset when predicting class/constant value for a new data point. Instead of learning and storing weights, the entire training dataset is saved in memory. As a result, the whole training dataset represents the KNN model [31].

There is overfitting of data/high variance at low K levels. As a result, the test error is significant while the training error is low. Because the nearest neighbour to that point is that point itself, the error is always zero in train data when K=1. As a result, with smaller K values, test error is considerable even while training error is minimal. This is referred to as overfitting. The test error decreases when the value for K is increased. However, after a specific K value, bias/underfitting occurs, and test error increases. So, we may say that the test data error is high at first (due to variation). It drops and stabilizes, and with a higher K value, it rises again (due to bias). When the test error stabilizes and is low, the K value is optimal. We can choose K=8 for our KNN algorithm implementation based on the error curve [32].

It classifies data into a category that is quite similar to the new data [33]. Distance-based approaches are often employed to solve data categorization problems. The k-nearest neighbour classification technique is one of the most extensively used distance-based algorithms (k-NN). This classification compares the distances between the test sample and the training samples to get the final classification result. The conventional k-NN classifier works well with numerical data.

$$R^* \leq R_{knn} \leq R^* \text{ (2-MR*/M-1)}$$

Where R^* is the Bayes error rate, R_{knn} is the k-NN error rate, and M is the number of classes.

2.3.4 Support Vector Machines (SVM)

SVMs (Support Vector Machines) are a novel machine learning technique based on Statistical Learning Theory (Vapnik-Chervonenkis or VC-theory). For the estimate of dependencies and predictive learning from finite data sets, VC-theory has a solid mathematical foundation. SVM is dependent on the Structural Risk Minimisation principle, which aims to reduce both empirical risk and model complexity while maintaining good generalizability.

SVMs (supervised vector machines) are supervised machine-learning algorithms used in classification and regression models. SVMs are more powerful than regression models, but they work best with limited datasets. First, every data point is plotted in an n-dimensional space, with n equalling the number of characteristics. Then a hyperplane is created to divide (classify or sort) the clusters physically. This approach uses the hyperplane to maximize the distance (or margin) between classes while ignoring outliers. When linear separation is not achievable, kernels alter data to make it more separable [34]. SVM (support vector machines) is a supervised learning algorithm that may be used to solve classification and regression problems such as support vector classification (SVC) and support vector regression (SVR) (SVR). However, it is only used for small datasets because processing them takes too long.

$$\text{SVM}= [\tfrac{1}{n} \textstyle\sum_{i=1}^{n} \max(0,1) - yi(w^T xi - b))] + \lambda \parallel w \parallel^2$$

Where ‘w’ is the average vector, x_i is a p-dimensional real vector, and ‘b’ is the boundary.

2.3.5 Random Forests (RF)

RF is a supervised ML algorithm commonly used to solve classification and regression problems. It creates decision trees from various samples, using the majority vote for classification and the average for regression. One of the essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and discrete variables, as in regression and classification. It outperforms the competition when it comes to categorization difficulties [35]. Bagging is a random forest ensemble approach. Bagging takes a random sample of data from the complete set and puts it in a virtual bag. As a result, row sampling is used to substitute the samples (Bootstrap Samples) provided by the Original Data in each model. Row sampling with replacement is known as the "bootstrap" step [36]. Because random forests are built from subsets of data, and the final output is based on average or majority rating, overfitting is avoided. It is, on the whole, slower. Random forest selects data at random, forms a decision tree, and averages the results. It does not rely on any formulas. Bagging, or bootstrap aggregation, is used by the Random Forest classifier to create an ensemble of classification and regression tree (CART)-like classifiers [37].

$$Ni_j = w_jC_j - W_{left(j)}C_{left(j)} - W_{right(j)}C_{right(j)}$$

Where ‘Ni_j’ is the importance of node j, w_j is the weighted number of samples reaching node j, C_j is the impurity value of the node j, $left_{(j)}$ is the child node from left split on node j and $right_{(j)}$ is the child node from right split on node j.

3. Results

3.1.1 Statistical metrics

The dataset was passed onto machine learning algorithms like LDA, CART, KNN, SVM and RF. The number of resamples employed was 10. The mean accuracy of each classifier was calculated, and the obtained mean accuracies were 77.5% (LDA), 87% (CART), 96% (KNN), 93.5% (SVM) and 96% (RF). RF and KNN models were selected as optimal models with higher accuracy and represented as dot plots (Table 2 & Figure 3).

$$\text{Average Accuracy} = \frac{1}{|D|}\sum_{1}^{|D|}\frac{\mathrm{xor}(yi,y\ \ )}{|L|}$$

|D| the number of samples and |L| the number of labels, and y_i is the actual label, y^ the predicted label.

Response y_i and covariates x_i for i=1...n, and Loss function is L. The NIR rate of a model f is the average loss of f over all combinations of yi and xi is given as

$$\text{NIR} = \frac{1}{n^2}\ \sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{L}(\text{yi}, \text{f(xj)})$$

Table 2. Mean accuracy of each classifier

Classifier	Mean Accuracy
LDA	77.5%
CART	87%
KNN	96%
SVM	93.5%
RF	96%

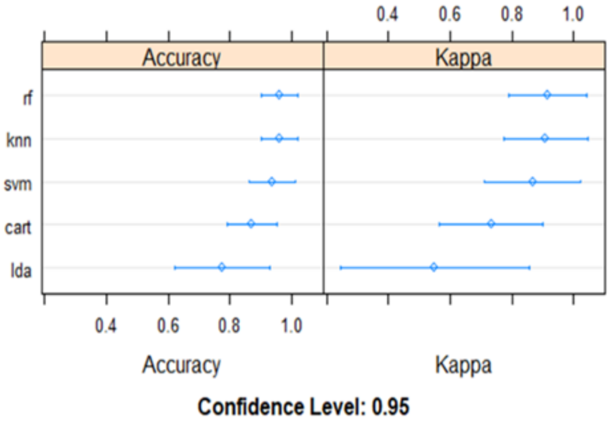

Figure 3. Dot plot

The original dataset was split into training and test data. The type of random forest used in this model is classification with 500 trees. The number of variables attempted at each split is equal to 3. The Out of Bag (OOB) estimate of error rate for training data from all the variables is 12.5%. Confusion matrix using training dataset was represented in a table. The prediction was made using the rf model produced using training data over factor variable (WQ_2LN). The prediction accuracy obtained was 100% with 24 observations ('1') and 16 observations ('2'). This accuracy was 100% obtained at 95% CI within 0.9119 and 1. The No Information Rate (NIR) was 0.6, and the p-value of accuracy is greater than NIR (1.337e-09). The positive class obtained was '1' with a sensitivity and specificity of 1. The positive and negative predictive value was equal to 1. The prevalence and detection rate were 0.6. The detection prevalence is 0.6 with a balanced accuracy of 1.0. The prediction was made using the rf model passed over test data. The accuracy obtained was 100% at 95% CI within 0.7941 and 1. The NIR value is 0.5, and the p-value of accuracy was greater than NIR (1.526e-05). The specificity and sensitivity were 1, with positive and negative prediction values equal one. The positive class obtained was '1'. The variable importance observed was in the order of EC > Cl > NO_3 > Mg > pH > TA > TH > Ca > HCO_3 > Na > SAR > SO_4 > RSC >F. The mean decrease Gini values reflected EC (3.583) as a top player in determining the accuracy. The confusion matrices of training and test data are shown in Tables 3, 4, 5 and 6. Variable importance and mean Gini decrease is shown in Table 7 and 8.

The original dataset was split into training and test data. To avoid confusion, the factor variable (WQ_2LN) was assigned labels 'Yes' and 'No.' The method used in this process was repeated cross-validation with 10-fold and three repeats. The accuracy obtained using training data was 92.5% at k=5 (Table 9). The confusion matrix resulted in an accuracy of 100% at 95% CI with 0.7151 and 1. The NIR value was 0.634, and the p-value of accuracy is more significant than NIR (0.00693). The value of sensitivity and specificity is 1. The positive and negative prediction value obtained was equal to 1. The prevalence value is 0.3636, and the detection rate was 0.3636. The balanced accuracy obtained was 1. The positive class obtained was 'No.' The test data was allowed to run on this model, and a confusion matrix was obtained. The accuracy obtained with test data was 100% at 95% CI with 0.7151 and 1. The NIR value obtained was 0.6364 with a p-value of accuracy greater than NIR (0.00693). The sensitivity, specificity, positive and negative predictive value was 1. The prevalence and detection rate were equal to 0.3636. The balanced accuracy obtained was equal to 1. The positive class is 'No.' For further analysis, the training data was subjected to tuneLength of 20 with 'center' and 'scale' pre-processing (Table 10 & 11).

Table 3. Confusion Matrix (Training Dataset)

	Reference	
Prediction	**1**	**2**
1	24	0
2	0	16

Table 4. RF-Training Dataset statistics

Accuracy	100%
95% CI	(0.9119, 1)
No Information Rate	60%
P-Value [Acc > NIR]	1.337e-09

Table 5. Confusion Matrix (Test Data)

	Reference	
Prediction	1	2
1	8	0
2	0	8

Table 6. RF-Test Dataset statistics

Accuracy	100%
95% CI	(0.7941, 1)
No Information Rate	50%
P-Value [Acc > NIR]	1.526e-05

Table 7. Variable importance

	Overall
HCO_3	1.02559
Ca	1.306406
Cl	2.005265
EC	3.583079
F	0.324502
Mg	1.783165
Na	0.703416
NO3	1.877962
pH	1.538661
RSC	0.616339
SAR	0.662882
SO4	0.657861
TH	1.307576
TA	1.411797

Table 8. Mean decrease Gini

	Mean DecreaseGini
HCO_3	1.0255904
Ca	1.3064055
Cl	2.0052647
EC	3.5830793
F	0.3245019
Mg	1.7831649
Na	0.7034163
NO_3	1.8779617
pH	1.5386611
RSC	0.6163392
SAR	0.6628815
SO4	0.6578606
TH	1.3075757
TA	1.4117973

Table 9. Accuracy of KNN classifier (Training data)

k	Accuracy	Kappa
5	0.925	0.840909
7	0.908333	0.807576
9	0.895	0.771212
11	0.863333	0.706061
13	0.876667	0.742424
15	0.841667	0.659091
17	0.843333	0.660606
19	0.843333	0.660606
21	0.803333	0.578788
23	0.796667	0.563636

Table 10. Confusion Matrix (Test data)

	Reference	
Prediction	No	Yes
No	4	0
Yes	0	7

Table 11. Statistics (KNN)

Accuracy	100%
95% CI	(0.7151, 1)
No Information Rate	63%
P-Value [Acc > NIR]	0.00693

4. Discussion

Machine learning (ML) tools are used in several studies across domains with a high accuracy rate. This study assumes that ML tools can predict the values of the groundwater quality variables with both accuracy and prediction. This work compared five machine learning algorithms under classification mode. Two of the five algorithms provided higher accuracy in predicting the groundwater quality variables. The data was split into training and test data, and their respective accuracies were good. The groundwater surveys are always expensive, and ML tools can predict accurate values for the points that are unknown or yet to be explored. This study can be extended to surface water quality parameters and propagation of the pollutants. The area selected for this study is not conducive to regular groundwater surveys due to topographic inconvenience. It calls for a need to use ML algorithms.

5. Conclusion

As the prediction and modeling are always based on the data availability, it is often buoyant in most areas for several reasons. Artificial intelligence, Machine learning and Geostatistics can help us in filling the gap in hydrological research. Though interpolation serves this immediate purpose, whatever is left in prediction studies can be easily satisfied with ML tools. There is a need to explore many aspects of groundwater in this area, and it is expected that machine learning can be added to the methodologies. The deep learning model backed with the neural networks are used in understanding several aspects of groundwater and the pressure of population on it. Since groundwater is being over exploited for obvious reasons, this study might aid the researchers in developing integrated machine learning and AI models in saving water for present and future generations.

The second and subsequent lines of each bibliography should be indented 0.5 cm inward as shown in this text.

Thesis should be written as Master's Thesis or Doctoral Thesis in the reference list.

Acknowledgement

The author would like to acknowledge Central Groundwater Board, Government of India for making data available.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Aytaç, E. (2020). Unsupervised learning approach in defining the similarity of catchments: Hydrological response unit-based k-means clustering, a demonstration on Western Black Sea Region of Turkey. International Soil and Water Conservation Research, 8(3), 321–331. https://doi.org/10.1016/j.iswcr.2020.05.002
2. Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., & Kumar, S. (2021). Prediction of groundwater quality

using efficient machine learning technique. Chemosphere, 276. https://doi.org/10.1016/j.chemosphere.2021.130265

3. Bilali, A., Taleb, A., & Brouziyne, Y. (2021). Groundwater quality forecasting using machine learning algorithms for irrigation purposes. Agricultural Water Management, 245. https://doi.org/10.1016/j.agwat.2020.106625
4. Yenugu, S. R., Vangala, S., & Badri, S. (2020a). Groundwater quality evaluation using GIS and water quality index in and around inactive mines, Southwestern parts of Cuddapah basin, Andhra Pradesh, South India. HydroResearch, 3, 146–157. https://doi.org/10.1016/j.hydres.2020.11.001
5. Brindha, K., Pavelic, P., Sotoukee, T., Douangsavanh, S., & Elango, L. (2017). Geochemical Characteristics and Groundwater Quality in the Vientiane Plain, Laos. Exposure and Health, 9(2), 89–104. https://doi.org/10.1007/s12403-016-0224-8
6. Reddy, B. M., V.Sunitha, M.Prasad, Reddy, Y. S., & Reddy, M. R. (2019). Evaluation of groundwater suitability for domestic and agricultural utility in semi-arid region of Anantapur, Andhra Pradesh State, South India. Groundwater for Sustainable Development, 9, 100262. https://doi.org/10.1016/j.gsd.2019.100262
7. Datta, P. S., & Tyagi, S. K. (1996). Major Ion Chemistry of Groundwater in Delhi Area: Chemical Weathering Processes and Groundwater Flow Regime. Journal of Geological Society of India (Online Archive from Vol 1 to Vol 78), 47(2), 179–188.
8. Raju, N. J. (2007). Hydrogeochemical parameters for assessment of groundwater quality in the upper Gunjanaeru River basin, Cuddapah District, Andhra Pradesh, South India. Environmental Geology, 52(6), 1067–1074. https://doi.org/10.1007/s00254-006-0546-0
9. Ramakrishna Reddy, M., Janardhana Raju, N., Venkatarami Reddy, Y., & Reddy, T. V. K. (2000). Water resources development and management in the Cuddapah district, India. Environmental Geology, 39(3), 342–352. https://doi.org/10.1007/s002540050013
10. Sreedevi, P. D. (2004a). Groundwater Quality of Pageru River Basin, Cuddapah District, Andhra Pradesh. Journal of Geological Society of India (Online Archive from Vol 1 to Vol 78), 64(5), 619–636.
11. Bedi, S., Samal, A., Ray, C., & Snow, D. (2020). Comparative evaluation of machine learning models for groundwater quality assessment. Environmental Monitoring and Assessment, 192(12), 776. https://doi.org/10.1007/s10661-020-08695-3
12. Mosavi, A., Hosseini, F. S., Choubin, B., Abdolshahnejad, M., Gharechaee, H., Lahijanzadeh, A., & Dineva, A. A. (2020). Susceptibility Prediction of Groundwater Hardness Using Ensemble Machine Learning Models. Water, 12(10), 2770. https://doi.org/10.3390/w12102770
13. Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., & Pradhan, B. (2018). A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Science of The Total Environment, 644, 954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054
14. Agrawal, P., Sinha, A., Kumar, S., Agarwal, A., Banerjee, A., Villuri, V. G. K., … Pasupuleti, S. (2021). Exploring Artificial Intelligence Techniques for Groundwater Quality Assessment. Water, 13(9), 1172. https://doi.org/10.3390/w13091172
15. Tamiru, H., & Wagari, M. (2021). Comparison of ANN model and GIS tools for delineation of groundwater potential zones, Fincha Catchment, Abay Basin, Ethiopia. Geocarto International, 0(0), 1–19. https://doi.org/10.1080/10106049.2021.1946171
16. Naghibi, S. A., Pourghasemi, H. R., & Abbaspour, K. (2018). A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. Theoretical and Applied Climatology, 131(3), 967–984. https://doi.org/10.1007/s00704-016-2022-4
17. Golkarian, A., Naghibi, S. A., Kalantar, B., & Pradhan, B. (2018). Groundwater potential mapping using C5.0, random forest, and multivariate adaptive regression spline models in GIS. Environmental Monitoring and Assessment, 190(3), 149. https://doi.org/10.1007/s10661-018-6507-8
18. Acar, E., & Özerdem, M. S. (2020). On a yearly basis prediction of soil water content utilizing sar data: A machine learning and feature selection approach. Turkish Journal of Electrical Engineering & Computer Sciences, 28(4), 2316–2330. Retrieved from https://online-journals.tubitak.gov.tr/publishedManuscriptDetails.htm?id=27563
19. Acar, E., Ozerdem, M. S., & Ustundag, B. B. (2019). Machine Learning based Regression Model for Prediction of Soil Surface Humidity over Moderately Vegetated Fields. 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), 1–4. 8820461 https://doi.org/10.1109/AgroGeoinformatics.2019.
20. Al-Adhaileh, M. H., & Alsaade, F. W. (2021). Modelling and prediction of water quality by using artificial intelligence. Sustain., 13. https://doi.org/10.3390/su13084259
21. https://indiawris.gov.in/wris/#/GWQuality
22. http://cgwb.gov.in/GW-data-access.html
23. Districts, India, 2016—University of Texas Libraries GeoData. (n.d.). Retrieved November 21, 2021, from https://geodata.lib.utexas.edu/catalog/stanford-sh819zz8121
24. Yenugu, S. R., Vangala, S., & Badri, S. (2020b). Monitoring of groundwater quality for drinking purposes using the WQI method and its health implications around inactive mines in Vemula-Vempalli region, Kadapa District, South India. Applied Water Science, 10(8), 202. https://doi.org/10.1007/s13201-020-01284-2
25. Sreedevi, P. D. (2004b). Groundwater quality of Pageru River basin, Cuddapah District, Andhra Pradesh. Journal of Geological Society of India, 64.
26. Castro, C. L., & Braga, A. P. (2013). Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. IEEE Transactions on Neural Networks and Learning

Systems, 24. https://doi.org/10.1109/TNNLS.2013.2246188

27. Collins, R., & Jerkins, A. (1996). The impact of agriculture land use on stream chemistry in the middle Hills of the Himalayas, Nepal. Journal of Hydrology, 185. https://doi.org/10.1016/0022-1694(95)03008-5
28. Ako, A. A., Eyong, G. E. T., Shimada, J., Koike, K., Hosono, T., Ichiyanagi, K., … Roger, N. N. (2014). Nitrate contamination of groundwater in two areas of the Cameroon Volcanic Line (Banana Plain and Mount Cameroon area). Applied Water Science, 4(2), 99–113. https://doi.org/10.1007/s13201-013-0134-x
29. Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing, 135. https://doi.org/10.1016/J.NEUCOM.2013.05.059
30. Ajmera, T. K., & Goyal, M. K. (2012). Development of stage discharge rating curve using model tree and neural networks: An application to Peachtree Creek in Atlanta. *Expert Systems with Applications*, 39. https://doi.org/10.1016/j.eswa.2011.11.101
31. Zhou, Z. H., & Liu, X. Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 18. https://doi.org/10.1109/TKDE.2006.17
32. Zhang, C., Tang, Y., Xu, X., & Kiely, G. (2011). Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Applied Geochemistry, 26.
33. Cunningham, P., & Delany, S. J. (2021). k-Nearest Neighbour Classifiers—A Tutorial. Conference Papers. https://doi.org/10.1145/3459665
34. Celestino, A. E. M., Cruz, D. A. M., Sánchez, E. M. O., & Reyes, F. G. (n.d.). Groundwater Quality Assessment: An Improved Approach to K-Means Clustering, Principal Component Analysis and Spatial Analysis: A Case Study. Retrieved from https://core.ac.uk/display/156977871
35. Biau, G. (2012). Analysis of a Random Forests Model. Journal of Machine Learning Research, 13(38), 1063–1095. Retrieved from http://jmlr.org/papers/v13/biau12a.html
36. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random Forests. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (pp. 587–604). New York, NY: Springer. https://doi.org/10.1007/978-0-387-84858-7_15
37. Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. Pattern Recognition Letters, 27(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011