



Modeling Groundwater Spring Potential of Selected Geographical Area Using Machine Learning Algorithms

S. Husen^(✉), S. Khamitkar, P. Bhalchandra, S. Lohkhande, and P. Tamsekar

School of Computational Sciences, S.R.T.M.University, Nanded, Maharashtra 431606, India
husen09@gmail.com, s_khamitkar@yahoo.com, srtmun.parag@gmail.com,
lokhande_sana@rediffmail.com, pritam.tamsekar@gmail.com

Abstract. The major objective of this paper was to produce groundwater spring potential maps for the selected geographical area using machine learning models. Total seven ML algorithms, viz Logistic Regression, Random Forest, SVM, Gradient Boosting Classifier, XG Boost, KNN, and Decision Tree were deployed. Further modeling was done using six hydrological-geological aspects that control the site of springs during the course of this research. Finally, groundwater spring potential was modeled and planning using fusion of these technologies.

Keywords: Water management · GIS · AHP · DSS · ML

1 Introduction

Groundwater is supposed to be one of the mainly priceless natural treasures. Due to a number of factors such as steady temperature, extensive accessibility, incomplete susceptibility to contamination, short growth cost, and lack of consistency, this treasure is shrinking down [1]. Further, the speedy raise in human populace has tremendously raised the requirement for groundwater provisions for drinking, agricultural, and industrial reasons [2]. Thus, the demand and supply equation is severely damaged leading to massive scarcity. To cater demands for clean groundwater rose in recent years, the demarcation of groundwater spring potential area develop into an important solution. The implementation of this solution needs resolve safety, and good management structure [3]. This is addressed by the creation of groundwater maps. A large number of researchers used integrated GIS, RS, and geo-statistics techniques for groundwater planning [4]. A lot of studies have also examined groundwater viable using the probabilistic model [5]. Additionally, a lot of current research studies use GIS methods build in with frequency ratio (FR), logistic regression (LR), models [6]. After careful investigations of the highlighted references and contemporary research, our work has used Logistic Regression, Random Forest, SVM, Gradient Boosting Classifier, XG Boost, KNN, and Decision Tree methods, included in a GIS to forecast groundwater spring locations for selected the study area [7]. Machine learning (ML) algorithms is a quickly rising area of prognostic modeling, where it is difficult to recognizes structure in complex, frequently

nonlinear data and generates correct prognostic models [8]. The ML approaches showed better authority for deciding complex relationships as ML approaches are not limited to the conventional hypothesis which is usually used along with predictable and parametric approaches. On the backdrop of these discussions, the main objective of this study was to predict groundwater spring potential maps using Logistic Regression, Random Forest, SVM, Gradient Boosting Classifier, XG Boost, KNN, and Decision Tree methods. The selected geographical area is the Kalmnoori taluka, Maharashtra state, India. The result of this research will afford a methodology to expand SPM that can be used by not only government departments, but also by the private sector for groundwater examination, estimation, and fortification.

2 Methodology

Kalmnoori taluka is taken as the study area. The Geographical location of Kalmnoori taluka lies within the coordinate of latitudes between 19.67° North latitude and 77.33° east longitudes. It is in the Hingoli district, Maharashtra state, India and its area is approximately 51.76 km^2 . It has an average elevation of 480 m. The Kayadhu is the main river which flows from the study area. The study area is taken out from the top sheet collected from the certified government agency, and it is geo-referenced. In order to generate the DSS for recognition of the Artificial Water Recharge Site (AWRS) the following process has been adopted.

2.1 Spatial Data Collection

Remote sensing data of CartoDEM of 2.5 m resolution and 3.5 m resolution satellite data of IRS P5 LISS-III, which are taken from the Bhuvan portal [9]. Geomorphology and Lineament data is taken out from the WMS layer of the Bhuvan portal. The particulars of the hydrological soil group map are taken from the National Bureau of Soil Survey.

2.2 Formulation of Criteria for Artificial Water Recharge Site

For the selection of Artificial Water Recharge Site, we have selected the below listed six parameters namely Geomorphology, Slope, Drainage density, Lineament density, LU/LC, and Soil. Geomorphology must be pediment-piedplain complex, anthropogenic terrain and water body should be good. The slope required is 0° – 5° . Drainage density is 0 to $3.5/\text{km}^2$. Lineament density is 0 – 1.5 km^2 . LU/LC must be a cropland, water body or scrubland. The soil must be sand or loam.

2.3 Generation of Criteria Map Using GIS

For AWRS the formulation of criteria is completed by using parameters like Geomorphology, soil, slope, LULC, lineament density, and drainage density. After that reclassification process is done for the same unit of all layers. And finally, all vector layers are converted into the raster layer.

2.4 Deriving the Weights Using AHP

Literature reviews, expert advises and from local field experience for the comparative significance to assigning weight age for every parameter and classes. The aim of using an analytic hierarchy process (AHP) is to recognize the ideal option and also conclude a position of the alternatives when all the decision criteria are measured concurrently [10].

2.5 Weighted Overlay Analysis

After giving the weightage of every major parameter has been determined, the weightage for the associate class of major parameters have been assigned. “Weighted Overlay” method is used for overlaying all the layers according to their importance. “Weighted Overlay” is a sub method of Spatial Analyst Tools in ArcGIS. Further, ranking is given from 1 to 5 according to their weight. Where 5 represents excellent prospects and 1 represents a poor prospect of groundwater. Lastly according to importance and rank of layers its entity features are given to each layer [11].

2.6 Data Set

The data set was extracted from the inventory map of artificial groundwater zones in such a manner that it was balanced. The Random Point Selection function of ArcGIS was used for data preparation. The data set consists of 500 samples which consist of 7 parameters namely Geomorphology, soil, slope, LULC, lineament density, drainage density, and result map. Among this, the result map is the target variable and the remaining 6 parameters are the response variable. The sample points used to build the data set from the study area is shown in Fig. 1a.

2.7 Exporting Data in Python

The data set is imported in Python and data exploration is done. In Python, we have done one hot encoding on data. The dataset consists of continuous variable in order to apply classification algorithms the data should be in categorical form, therefore One hot encoding is used to transform the continues variable to categorical. Finally, the dataset is split into training (70%) and testing (30%) dataset.

After the above analysis, efforts are taken for training the Model on Data Set. Following classification algorithms are used to train the model on a dataset.

KNN: The Fig. 1b shows the map of the Artificial Ground Water Recharge Site (AGWRS) predicted by K-Nearest Neighbor (KNN) model of the machine learning algorithm. In this map classes of artificial groundwater recharge site such as very poor, poor, moderated, good, very good are shown along with theses classes Artificial Ground Water Recharge Point predicted by KNN algorithm are also shown. In k-Nearest Neighbors where the k is the number of observations are tuned to 1, 3, 5, 7, 9 and found that $k = 5$ gives the utmost results. By applying k-NN with the value $k = 5$, the confusion matrix for the applied model was created. Total 105 artificial groundwater recharge site (AGRW) points were correctly classified as good followed by 21 as moderate, and 1,

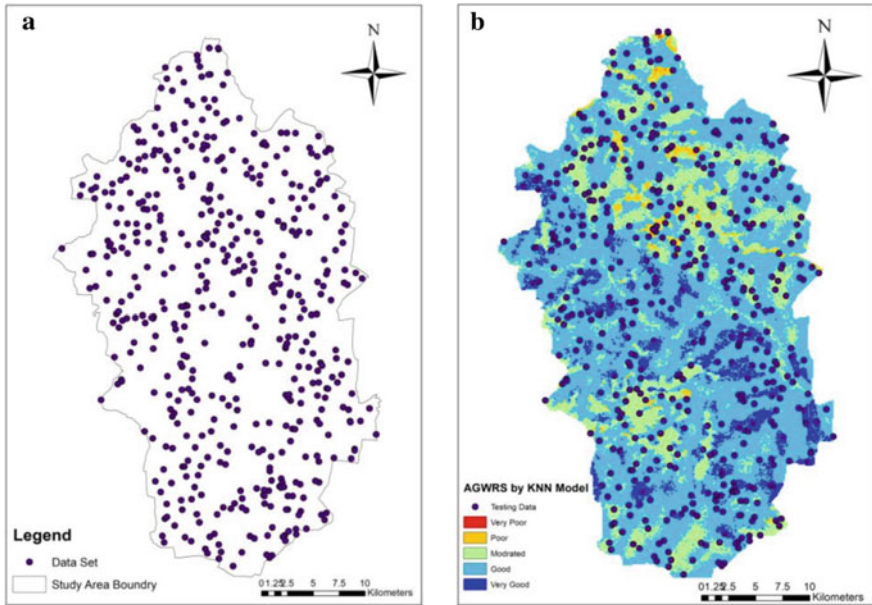


Fig. 1 a Data set, b AGWRS by KNN model

4, 3 as very poor, poor, and very good, respectively. The misclassifications for each class were 3, 1, 9, 13, and 5 for very poor, poor, moderate, good, and very good class, respectively. The overall accuracy of the model is 0.81% and the error rate is 0.19%.

Random Forest: The Fig. 2a shows the map of the artificial groundwater recharge site (AGWRS) predicted by the Random Forest (RF) model of the machine learning algorithm. This map shows the artificial groundwater recharge point predicted by the RF model and classes of artificial groundwater recharge sites such as very poor, poor, moderated, good, very good. The model is giving relatively good prediction results. This model gave 100% accuracy while predicting very poor class. In poor class, only 1 instance is misclassified as moderate. In moderate, out of 27 instances, 20 were correctly classified as moderate and 2, 7 were misclassified as good and poor, respectively. Out of 165 instances, 108 were classified as good and 2, 9, 5 were misclassified as very poor, moderate, and very good, respectively. In class very good, 3 instances were classified correctly and 3 were misclassified as good. Random forest algorithm has given 0.82 accuracy.

Support Vector Machine: Artificial groundwater recharge site (AGWRS) predicted by Support Vector Machine (SVM) model of machine learning algorithm map is shown in Fig. 2b. This map shows the artificial groundwater recharge point that is testing data and classes of artificial groundwater recharge site such as very poor, poor, moderated, good, very good. In parameter tuning, we have used a linear kernel. Using this approach, the trained model achieved good accuracy, while predicting class very poor. In poor class, 3 instances were correctly classified as poor class and 1 was misclassified as moderate. In class good, out of 121 instances, 108 were correctly classified as good and 1, 7, 5

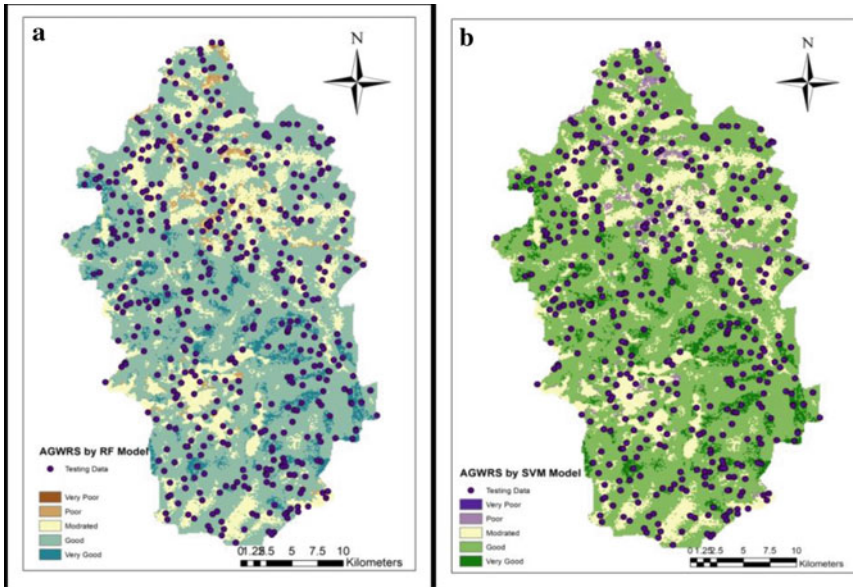


Fig. 2 a AGWRS by RF model, b AGWRS by SVM model

instances were misclassified as very poor, moderate, and very good, respectively. In class very good, 3 instances were classified correctly and 3 misclassified as good. Accuracy of SVM algorithm is 0.83.

XG Boost: In Fig. 3a shows the map of the artificial groundwater recharge site (AGWRS) predicted by the XG Boost (XGB) model of machine learning algorithm. The xg boost’s model is a linear combination of decision trees. To make a prediction xg boost calculates predictions of individual trees and adds them. The trained model achieved better accuracy while predicting class very poor. In poor class, 3 instances were correctly classified as poor class and 1 is misclassified as moderate. In moderate, 21 instances out of 32 were correctly classified as moderate and 2 and 9 were misclassified as poor and good. In good class, out of 121 instances, 106 were correctly classified as good and 2, 8, 5 instances were misclassified as very poor, moderate, and very good, respectively. In class very good, 3 instances classified correctly and 3 misclassified as good. Accuracy of XG Boost algorithm is of 0.81.

Logistic Regression: The Fig. 3a shows the map of the Artificial Ground Water Recharge Site (AGWRS) predicted by the Logistic Regression (LR) model of machine learning algorithm. In this map classes of artificial ground water recharge site such as very poor, poor, moderated, good, very good, along with these classes Artificial Ground Water Recharge Point data is shown on map which is predicted by LR algorithm. The related confusion matrix of the trained logistic regression model has achieved better accuracy while predicting class very poor. In poor class, 1 instance was correctly classified as poor class and 1 was misclassified as moderate. In moderate, 11 instances out of 20 were correctly classified as moderate and 2, 4, and 3 were misclassified as very poor, poor, and good, respectively. In class good, out of 142 instances, 115 were correctly

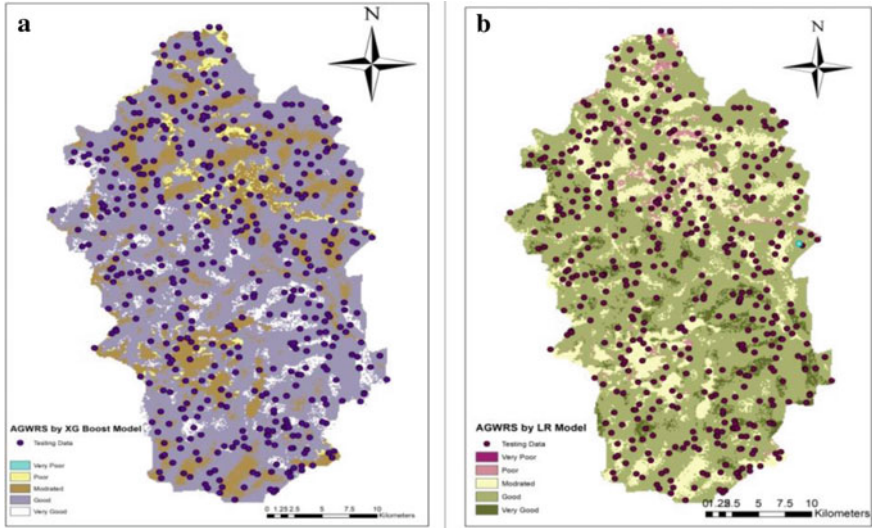


Fig. 3 a AGWRS by XG Boost model, b AGWRS by LR model

classified as good and 1, 18, and 8 instances were misclassified as very poor, moderate, and very good, respectively. Overall accuracy of LR algorithm is 0.77.

Gradient Boosting Classifier: In Fig. 4a shows the map of the Artificial Ground Water Recharge Site (AGWRS) predicted by the Gradient Booster Classifier (GBC) model of machine learning algorithm. This map shows that, the artificial groundwater recharge point is testing data and classes of artificial groundwater recharge site such as very poor, poor, moderated, good, very good. The related confusion matrix of the trained gradient boosting classifier model was obtained. While predicting class very poor out of 3 instances 2 were correctly classified as very poor and 1 instance was misclassified as good. In poor class, 2 instances were correctly classified as poor and 2 were misclassified as moderate. In moderate, 18 instances out of 29 were correctly classified as moderate and 2, 9 were misclassified as poor and good, respectively. In class good, out of 121 instances, 104 were correctly classified as good and 2, 10, and 5 instances were misclassified as very poor, moderate, and very good, respectively. Overall accuracy of GBC algorithm is 0.78.

Decision Tree: The artificial groundwater recharge site (AGWRS) predicted by the Decision Tree (DT) model of machine learning algorithm map is shown in Fig. 4b. This map shows that, the artificial groundwater recharge point is testing data and classes of artificial groundwater recharge site such as very poor, poor, moderated, good, very good. While predicting class very poor out of 2 instances 2 were correctly classified as very poor. In poor class, 4 instances were correctly classified as poor and 1 was misclassified as moderate. In moderate, 22 instances out of 37 were correctly classified as moderate and 1, 14 were misclassified as poor and good, respectively. In class good, out of 114 instances, 100 were correctly classified as good and 2, 7, and 5 instances were misclassified as very poor, moderate, and very good, respectively. Overall accuracy of Decision tree algorithm is 0.78.

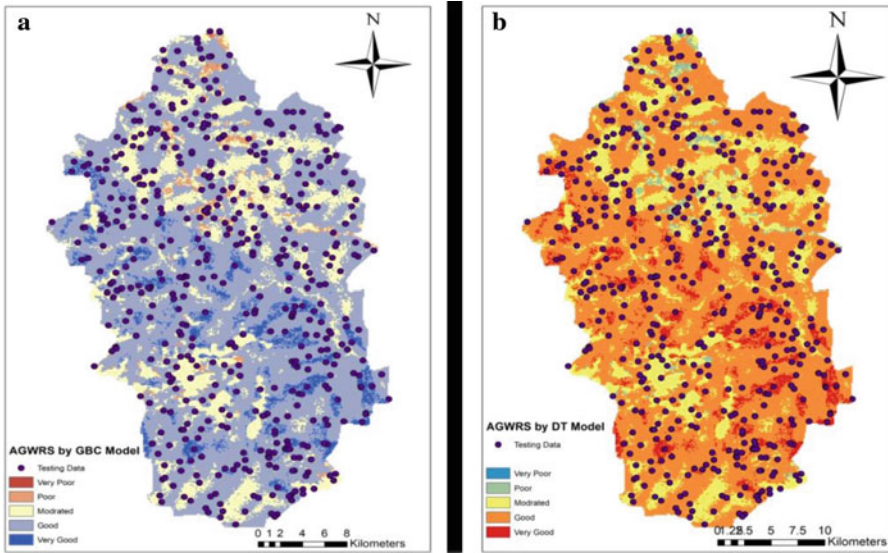


Fig. 4 a AGWRS by GBC model, b AGWRS by DT model

The performance of machine learning methods for artificial groundwater recharge sites prediction model depends on the data set and machine learning algorithms proper utilization. Picking the correct classification ML technique for the defined crisis is the basic necessity to get the most excellent possible result. Though, the proper selection of algorithms won't guarantee the best possible results. The input dataset served to build the ML model is also a vital issue and for getting the best possible results to feature engineering, the process of modifying the data for machine learning is also an important thing to be considered. Comparing the results of ML techniques applied to the testing data Table 9 shows the prediction results obtained by applying the ML techniques LR, RF, SVM, Gradient Boosting Classifier, XG Boost, KNN, and Decision Tree on the test dataset extracted from geo-data layers. The results are shown in Fig. 5 and Table 1.

3 Conclusions

The aim of the research work was to develop an artificial groundwater recharge site maps (AGWRS), which involves the model comparison of the raw dataset and engineered dataset, in order to improve the result prediction capability of the classification model. The datasets were obtained from the study area. The geo-data layers imported in Python programming language were processed using raster package and information was extracted. The selected six ML algorithms were implemented and their confusion matrix were created which served as input to the prediction model to get the best algorithm for classification. Their outcomes were compared using bench mark evaluation measures. It was observed that the best prediction results from SVM, with a prediction accuracy of 0.83% followed by Logistic Regression, Random Forest, Gradient Boosting

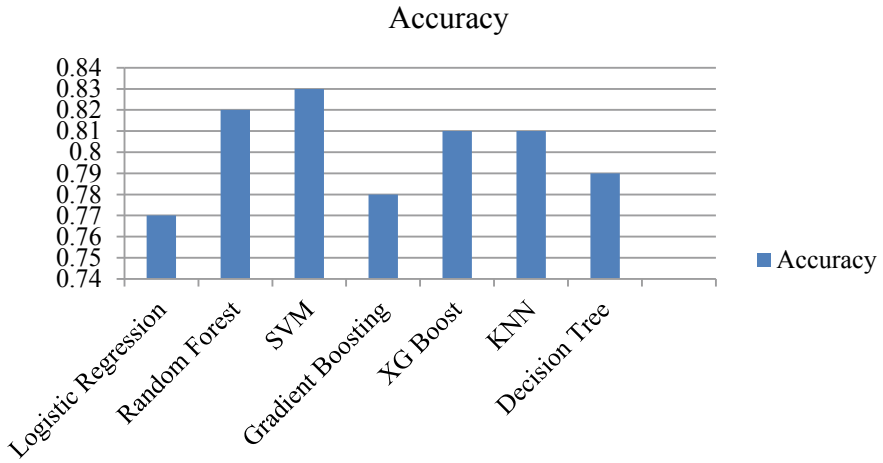


Fig. 5 Comparative results of all models

Table 1 Tabulation of efficiencies of all models

Algorithms	Results
Logistic regression	0.77
Random forest	0.82
SVM	0.83
Gradient boosting classifier	0.78
XG Boost	0.81
KNN	0.81
Decision tree	0.79

Classifier, XG Boost, KNN, and Decision Tree prediction accuracy results are 0.77, 0.82, 0.78, 0.81, 0.81, 0.79, respectively.

References

1. Jha, M.K., Chowdhury, A., Chowdary, V.M., Peiffer, S.: Groundwater management and development by integrated remote sensing and geographic information systems: prospects and constraints. *Water Resour. Manage* **21**, 427–467 (2007)
2. Lee, S., Song, K.Y., Kim, Y., Park, I.: Regional groundwater productivity potential mapping using a geographic information system (GIS) based artificial neural network model. *Hydrogeol. J.* **20**(8), 1511–1527 (2012)
3. Ozdemir, A.: Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *J. Hydrol.* **405**, 123–136 (2011)

4. Jaiswal, R.K., Mukherjee, S., Krishnamurthy, J., Saxena, R.: Role of remote sensing and GIS techniques for generation of groundwater prospect zones towards rural development—an approach. *Int. J. Remote Sens.* **24**(5), 993–1008 (2003)
5. Srivastava, P.K., Bhattacharya, A.K.: Groundwater assessment through an integrated approach using remote sensing, GIS and resistivity techniques: a case study from a hard rock terrain. *Int. J. Remote Sens.* **27**, 4599–4620 (2006)
6. Davoodi Moghaddam, D., Rezaei, M., Pourghasemi, H.R., Pourtaghie, Z.S., Pradhan, B.: Groundwaterspring potential mapping using bivariate statistical model and GIS in the Taleghan watershed. Iran. *Arab. J. Geosci.* **8**(2), 913–929 (2015)
7. Stumpf, A., Kerle, N.: Object-oriented mapping of landslides using random forests. *Remote Sens. Environ.* **115**(10), 2564–2577 (2011)
8. Olden, J.D., Lawler, J.J., Poff, N.L.: Machinelearning without tears: a primer for ecologists. *Q. Rev. Biol.* **83**(2), 171–193 (2008)
9. Web resource at <http://www.bhuvan.nrsc.gov.in>
10. Saaty, T.L.: Decision making with the analytic hierarchy process. *Int. J. Serv. Sci.* **1**(1) (2008)
11. Husen, S. et al.: Integrated use of GIS AHP and GIS techniques for selection of artificial ground water recharge site, information and communication technology for sustainable development. *Adv. Intell. Syst. Comput.* **933** (2020)