GIS-BASED GROUNDWATER POTENTIAL MAPPING USING MACHINE LEARNING MODELS, A CASE STUDY: QOM PROVINCE, IRAN

E. Masoudian [1], P. Pahlavani [1] *

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran – (ehsan.masoudian, Pahlavani)@ut.ac.ir

Commission IV, WG IV/3

KEY WORDS: Groundwater Potential Mapping, Geographic Information Systems, GIS, Random Forest, SVR, Iran.

ABSTRACT:

By considering the increasing trend in water consumption and significant reduction of water resources in most countries of the world, groundwater resources have become very important. The Target of this study is to implement machine learning models to produce a groundwater potential map (GWPM), identify areas with higher water potential, and also identify influencing factors. Therefore, two algorithms including the random forest (RF) and support vector regression (SVR), were performed that according to the literature have a good compatibility with this type of problems, compared to the other models. Of the 351 well points available throughout the study area, 70% (245 well points) were selected as the target for training the models and the rest 30% (106 well points) were used for evaluating the models. In addition, 20 effective information layers were used for modeling. In this study, an effort was made to focus more on data preparation that is one of the most important parts of model development. The variance inflation factor (VIF) and correlation coefficient were applied to identify the dependent variables. Also, feature selection was done to identify the most influential factors. Finally, two groundwater potential map(GWPM)s were created based on these two models. By calculating the area under the curve (AUC) from the receiver operating characteristic (ROC), the prediction accuracy of the two models was calculated. The values for AUC of the two maps produced by the RF and SVR algorithms were 93.4% and 89.7%, respectively. This study improves the knowledge of groundwater potential in the study area which is one of the cities with water scarcity in the country.

1. INTRODUCTION

Water resources are vital for human beings' survival. these resources include surface and underground water and are recovered from evaporation, precipitation, and surface runoff. last climate change estimates indicate increasing heterogeneity in the water cycle, which would result in water demand outstripping supply (Change, 2013). Groundwater is in a saturated zone that fills the pore spaces between mineral grains or cracks and fractures in a rock mass(Nampak, Pradhan, & Abd Manap, 2014). Currently, groundwater provides nearly 20% of the water needed by humans, and This ratio is expected to increase over the next few decades(Biswas, Mukhopadhyay, & Bera, 2020; Cho et al., 2018). GWPMs, as indicated in the literature, are created by a variety of models. This extremely broad spectrum of methods which contains Evidential Belief Function (EBF, (Nampak et al., 2014)), Frequency Ratio (FR, (Manap et al., 2014)), Weights of Evidence (WoE, (Saro Lee, Kim, & Oh, 2012)), Classification and Regression Tree (CART, (Naghibi & Pourghasemi, 2015)), Analytic Hierarchy Process (AHP, (Rahmati, Nazari Samani, Mahdavi, Pourghasemi, & Zeinivand, 2015; Singh, Jha, & Chowdary, 2018)), Multivariate Adaptive Regression Splines (MARS, (Golkarian, Naghibi, Kalantar, & Pradhan, 2018; Naghibi & Moradi Dashtpagerdi, 2017)), Fuzzy Logic (FL, (Shahid, Nath, & Maksud Kamal, 2002)), K-Nearest Neighbor (KNN, (Naghibi, Pourghasemi, & Abbaspour, 2018)), Support Vector Machine (SVM, (J. H. Lee, Zhao, & Kerr, 2017; Sunmin Lee, Kim, Jung, Lee, & Lee, 2017; Naghibi, Ahmadi, & Daneshi, 2017)), and random forest (RF, (Naghibi et al., 2017; Zabihi, Pourghasemi, Pourtaghi, & Behzadfar, 2016)), can classify to four groups as i) Bivariate Statistics; ii) Multivariate Statistics; iii) Machine Learning/ Data-Mining; and, iv)Multi-Criteria Decision Making(MCDM) (Arabameri, Rezaei, Cerda, Lombardo, & Rodrigo-Comino, 2019). The proposed method of this research (Figure 1) focused on two machine learning algorithms: RF and SVR which had high accuracy in previous researches. It should also be mentioned that both can solve either a classification or a regression problem. Since this study is performed to make a GWPM, thus methods were developed in the form of regression. The main goals of this research are: i) Create GWPM for study area, ii) Compare two methods mentioned above, iii) Identify most relevant factors.

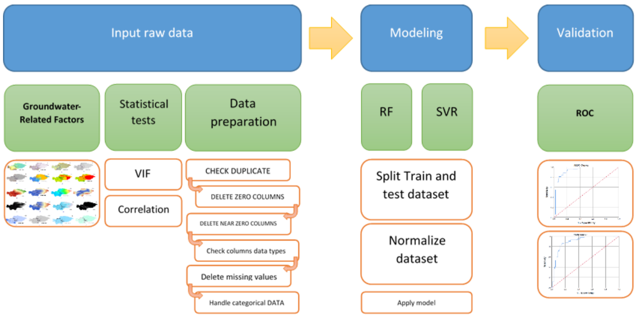

Figure 1. Proposed method

* Corresponding author

https://doi.org/10.5194/isprs-annals-X-4-W1-2022-489-2023 | © Author(s) 2023. CC BY 4.0 License.

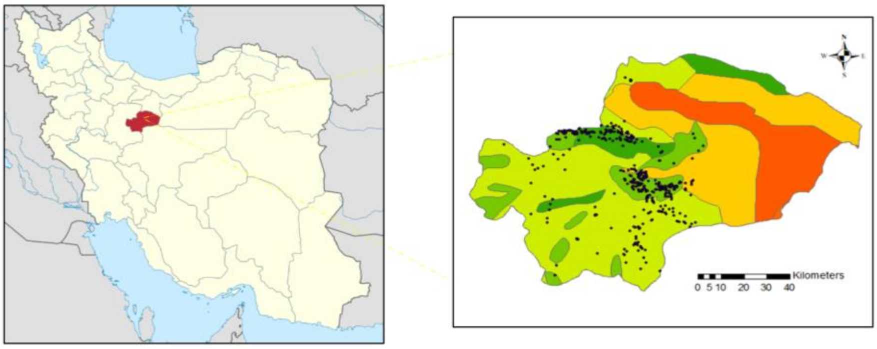

Figure 2. Study area

1.1 Study area

The study area is Qom province located in the central part of Iran. The considered area is between 34°9'N and 35°11'N latitude and 50°6'E and 51°58'E longitude (Figure 2). Qom covers a total area of 11,237 km² and has a population of over 1,151,672 inhabitants. The use of groundwater resources in this area includes wells, qanats, and springs. The climate of Qom province varies between semi-desert and desert and includes mountainous regions, foothills, and plains. According to its location close to an arid region and far inland, the climate is dry with low humidity and sparse rainfall. The elevation of the study area varies between 800 m and 3200 m, and the average height of area is 930 m. The average annual rainfall in Qom is 618.8 mm, which also varies due to the different altitudes in different areas. Due to its natural conditions, the province faces a shortage of surface and underground water resources. The Qamroud and Qarachay rivers form permanent and surface streams.

2. DATA

The groundwater potential mapping is done by modelling well or spring locations as target layer. Therefore, in this study, to create the groundwater potential map, 351 well sites throughout the study area were selected then divided into a training dataset (70% = 245 wells) and a test dataset (30% = 106 wells). However, these raw data are not yet suitable for modelling, so an upstream step is required to prepare the collected data. This step is called data preparation. This part will explain in section 2.2.

2.1 Groundwater-Related Factors:

For modelling groundwater potential, selecting more effective and relevant factors is so important. Accordingly, 20 factors that are related to GWPM were selected, which can classify as topographical, hydrological, and geological factors. These factors are DEM, slope, aspect, TWI, SPI, land cover, land use, climate type, village density, fault density, qanat density, spring density, river density, Euclidean distance(ED) of villages, ED of roads, ED of rivers, ED of creeks, ED of qanat, ED of springs, ED of faults(Figure 3). TWI which is a secondary topographic factor is calculated based on Equation (1)(Moore, Grayson, & Ladson, 1991):

$$\mathrm{TWI} = \ln(\beta/\tan\alpha)\,, \qquad (1)$$

where β = represents the Catchment Area (m2/m)
α = is the slope at the point

Also SPI can be defined as Equation (2)((Moore et al., 1991)):

$$SPI = \lambda \times \tan\eta\,, \qquad (2)$$

where λ = is the specific catchment's area
η = is the local slope angle gradient

2.2 Data Preparation:

An essential and critical step in the machine learning process is data preparation. Hence, machine learning algorithms are routines, and efforts are often to prepare data (Brownlee, 2020). Then, before train models and creating GWPM, it is necessary to clean and validate data. As shown in a Figure 1, a regular data engineering planned to check data.

2.2.1 Check Duplicate: In the raw data collected, there is a possibility that some rows are duplicates. An identical row is one that has the same value for all its columns as another row. Duplicate Rows are not only useless for the training step, but also can be misleading during model evaluation (Brownlee, 2020). These redundant rows should be identified and deleted.

2.2.2 Delete Zero and Near Zero columns: Columns that have only a single value or low variance of observation are probably useless for modelling and should be considered. These single-valued predictors are known as zero-variance predictors and should be deleted. However, columns with very few numerical values may or may not be useless for modelling, and depending on the situation, a decision should be made whether or not to remove them.

2.2.3 Check Samples Columns Data Types and Handle Categorical Data: Machine learning models only take numbers and output numbers. Therefore, it is important to consider input data types. Especially if some of the columns have categorical types. In this study, dummy encoding techniques were used for variables that are naturally non-numeric (categorical data).

2.2.4 Delete Missing Values: This step involves identifying rows that have one or more columns with no values and deleting them. In modelling, all observations should have the same size and have a value for all variables.

2.2.5 Split train and test: After data cleaning, the data were split into a training dataset (70% of observations) for modelling and a test dataset (30%) for evaluation of the model.

2.2.6 Normalize data:

The last and most important part of data preparation is data transformation. Not only the scaling of the data, but also all other procedures that need to be applied to the data should first fit on the training data set then be applied to the training and test data sets. Otherwise, the data transformation on the entire data set will lead to an issue known as data leakage, which means some information from the test data set permeates into the data set used to train the model. For this reason, the data set is first split and then normalized(Brownlee, 2020).

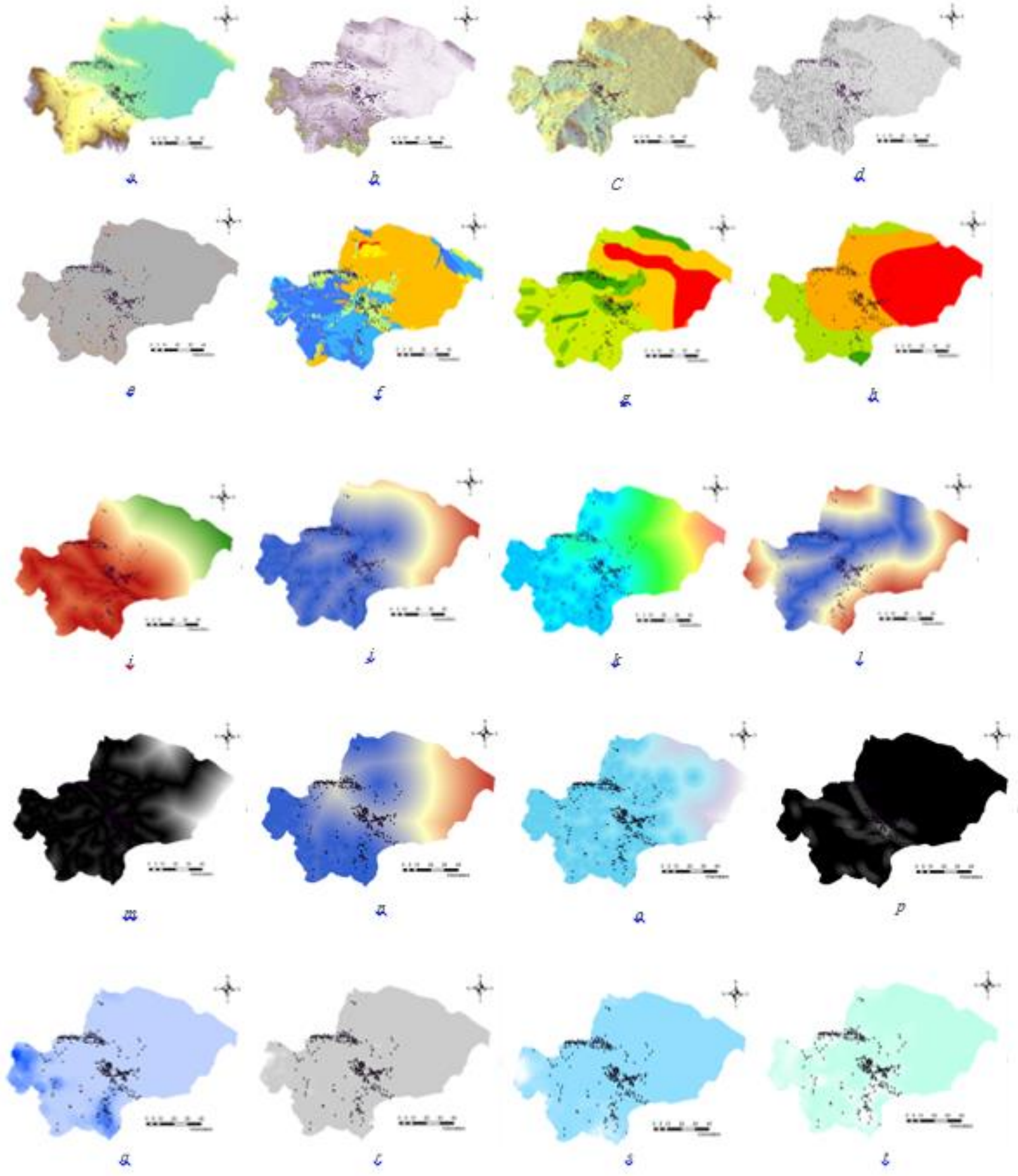

Figure 3. Groundwater-Reflated Factors
a)DEM, b)slope, c)aspect, d)TWI, e)SPI, f)landuse, g)landcoder, h)climate, i)ED-fault, j)ED-creek, k)ED-qanat, l)ED-river, m)ED-road, n)ED-spring, o)ED-village, p)fault-DEN, q)qanat-DEN, r)Creek-DEN, s)spring-DEN, t)village-DEN

3. METHODOLOGY

3.1 Random Forest:

Random forest is one of the most famous machine learning models trademarked by Leo Breiman and Adele Cutler, which incorporates the output of several decision trees to get a proper result. It's capable to solve both classification and regression problems. Due to its ease of use and flexibility, it is widely used Classification and Regression Trees represent a class of Decision Trees presented by (Breiman, Friedman, Olshen, & Stone, 1984). The main core of all Random forest algorithms has three hyperparameters, which should be tuned before training. These three hyperparameters are the number of features sampled, node size, and the number of trees. Hence, the random forest classifier is able to solve regression or classification problems (Figure 4).

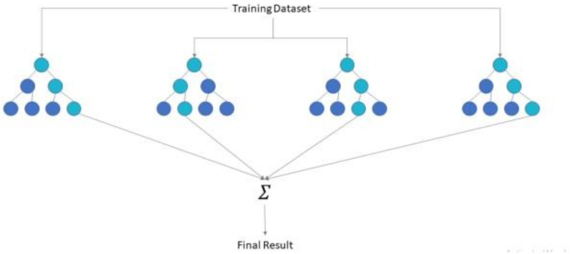

Figure 4. The Random Forest framework

3.2 Supported Vector Machine

Support vector machines (SVMs) are a set of supervised learning methods used for outliers detection, classification and regression problems. The method of SVM can be extended to solve regression problems. This method is defined as SVR. Since GWPM in this study, is a regression problem, SVR package of SVM in Scikit-learn was implemented.

4. RESULTS

In this section, the results of the models, GWPMs of two models and their evaluation are mentioned. In order to compare the models, the data preparation section is the same for both models, and after dividing the dataset into a training dataset and a test dataset, the models are trained separately and the results has been reported. The AUC was used to evaluate the models, which is a very efficient indicator of prediction accuracy.

4.1 Random Forest

Random forest algorithm implemented in Python, using the sklearn.ensemble package and GridSearchCV from the sklearn.model_selection package to find the best hyper parameters. The results of the model RF are shown in Table 1. AUC results are also shown in Figure 5.

R^2 train	R^2 test	RMSE	MAE
0.939	0.618	0.308	0.196

Table 1. Results of Random forest

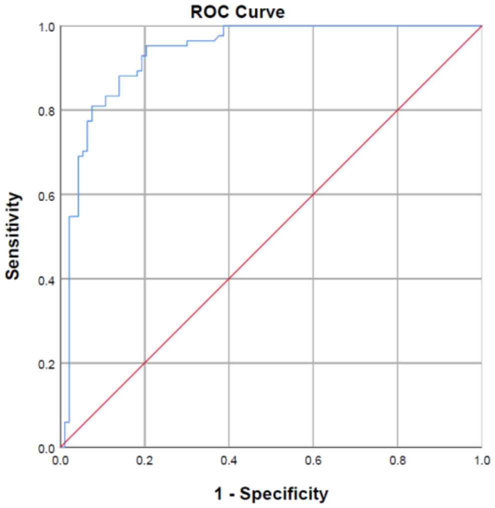

Figure 5. AUC of Random Forest (93.4%)

4.2 SVR

SVr algorithm also implemented in Python, using the sklearn.svm package and GridSearchCV from the sklearn.model_selection package to find the best hyper parameters. The results of the SVR are shown in Table 2. AUC results are also shown in Figure 6.

R^2 train	R^2 test	RMSE	MAE
0.889	0.404	0.385	0.257

Table 2. Results of SVR

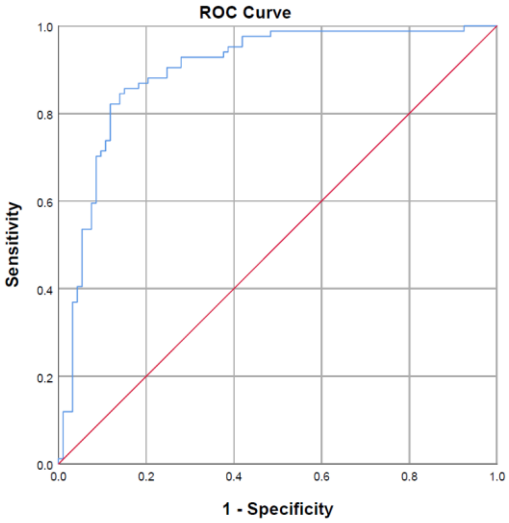

Figure 6. Results of SVR (89.7%)

4.3 Feature Importance:

The importance of the influencing factors can be obtained while training models, as shown in Figure 7 and 8. These measures help to understand the importance of features and the most important variables for map making. Thus, they can be used in dimensionality reduction, whose goal is to obtain maps with high accuracy at lower data dimensionality.

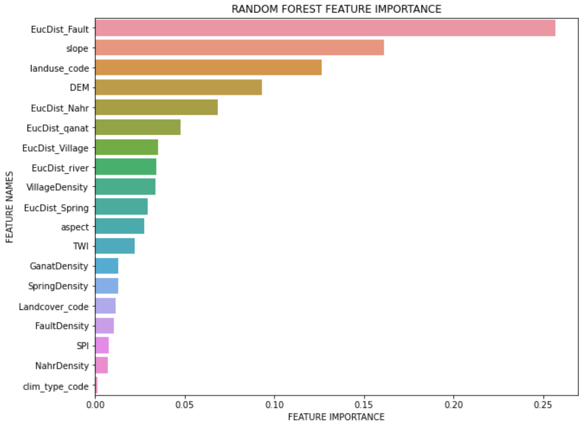

Figure 7. Random Forest Feature Importance

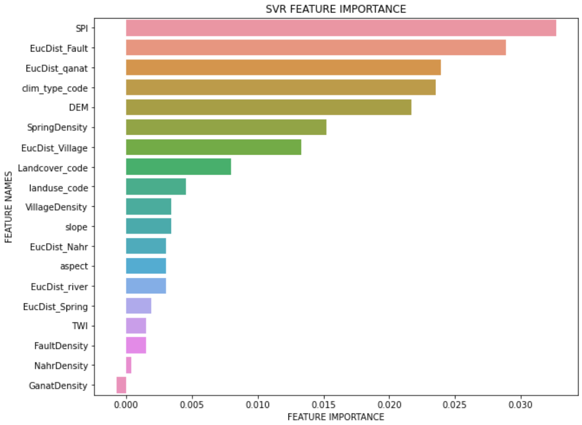

Figure 8. SVR Feature Importance

4.4 Groundwater Potential Mapping:

Finally, by use of these two models, GWPMs are created as shown in Figure 6 and 7.

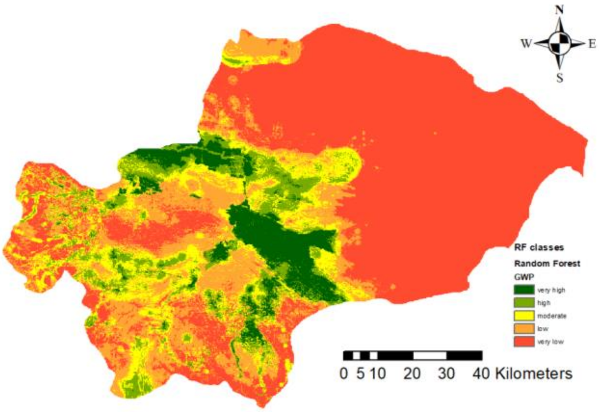

Figure 6. GWPM of Random Forest

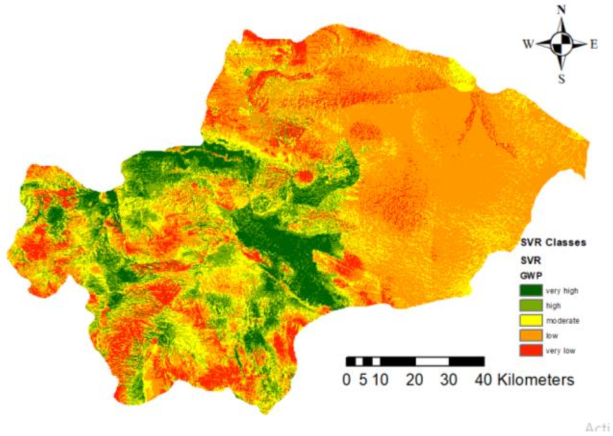

Figure 7. GWPM of SVR

5. CONCLUSIONS

The main target of this study was to create a GWPM based on two machine learning models for the study area, Qom province in Iran, which is one of the cities facing water scarcity. An attempt was also made to identify the most important factors affecting groundwater potential in both methods. The results show that RF method performs better than SVR, although both have reasonable accuracy calculated by AUC. By the use of data cleaning and dimensionality reduction (using feature importance to select more influential factors to train models) implemented in this research, prediction accuracy for RF and SVR reached from 91.3% and 86% to 93.4% and 89.7% respectively. In prior studies, these techniques were neglected or not mentioned to improve the result and quality of the maps. This study improves the knowledge of groundwater potential in the study area and shows that machine learning methods are operational and can be used instead of the old expensive methods.

REFERENCES

Arabameri, A., Rezaei, K., Cerda, A., Lombardo, L., & Rodrigo-Comino, J. (2019). GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *Science of the total environment, 658*, 160-177.

Biswas, S., Mukhopadhyay, B. P., & Bera, A. (2020). Delineating groundwater potential zones of agriculture dominated landscapes using GIS based AHP techniques: a case study from Uttar Dinajpur district, West Bengal. *Environmental earth sciences*, *79*(12), 1-25.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees.

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*: Machine Learning Mastery.

Change, I. C. (2013). The physical science basis. *Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change, 1535*, 2013.

Cho, H.-M., Kim, G., Kwon, E. Y., Moosdorf, N., Garcia-Orellana, J., & Santos, I. R. (2018). Radium tracing nutrient inputs through submarine groundwater discharge in the global ocean. *Scientific reports, 8*(1), 1-7.

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), 1-16.

Lee, J. H., Zhao, C., & Kerr, Y. (2017). Stochastic bias correction and uncertainty estimation of satellite-retrieved soil moisture products. *Remote Sensing, 9*(8), 847.

Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk, 8*(2), 1185-1203.

Lee, S., Kim, Y.-S., & Oh, H.-J. (2012). Application of a weights-of-evidence method and GIS to regional groundwater productivity potential mapping. *Journal of Environmental Management, 96*(1), 91-105.

Manap, M. A., Nampak, H., Pradhan, B., Lee, S., Sulaiman, W. N. A., & Ramli, M. F. (2014). Application of probabilistic-based frequency ratio model in groundwater potential mapping using remote sensing data and GIS. *Arabian Journal of Geosciences, 7*(2), 711-724.

Moore, I. D., Grayson, R., & Ladson, A. (1991). Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological processes, 5*(1), 3-30.

Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water resources management, 31*(9), 2761-2775.

Naghibi, S. A., & Moradi Dashtpagerdi, M. (2017). Evaluation of four supervised learning methods for groundwater spring potential mapping in Khalkhal region (Iran) using GIS-based features. *Hydrogeology Journal, 25*(1), 169-189.

Naghibi, S. A., & Pourghasemi, H. R. (2015). A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. *Water resources management, 29*(14), 5217-5236.

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.

Nampak, H., Pradhan, B., & Abd Manap, M. (2014). Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. *Journal of Hydrology, 513*, 283-300.

Rahmati, O., Nazari Samani, A., Mahdavi, M., Pourghasemi, H. R., & Zeinivand, H. (2015). Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. *Arabian Journal of Geosciences, 8*(9), 7059-7071.

Shahid, S., Nath, S. K., & Maksud Kamal, A. (2002). GIS integration of remote sensing and topographic data using fuzzy logic for ground water assessment in Midnapur District, India. *Geocarto International*, *17*(3), 69-74.

Singh, L. K., Jha, M. K., & Chowdary, V. (2018). Assessing the accuracy of GIS-based multi-criteria decision analysis approaches for mapping groundwater potential. *Ecological Indicators*, *91*, 24-37.

Zabihi, M., Pourghasemi, H. R., Pourtaghi, Z. S., & Behzadfar, M. (2016). GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. *Environmental earth sciences, 75*(8), 1-19.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W1-2022-489-2023 | © Author(s) 2023. CC BY 4.0 License.