Moukhliss, M., Taleb, A., Souabi, S., Ouallali, A., Spalevic, V. (2022). Groundwater quality forecasting using machine learning algorithms: Case Study Berrechid aquifer, central Morocco. Agriculture and Forestry, 68 (3): 35-56. doi:10.17707/AgricultForest.68.3.03

DOI: 10.17707/AgricultForest.68.3.03

Mohammed MOUKHLISS[1], Abdeslam TALEB, Sonia SOUABI, Abdessalam OUALLALI, Velibor SPALEVIC[2]

GROUNDWATER QUALITY FORECASTING USING MACHINE LEARNING ALGORITHMS: CASE STUDY BERRECHID AQUIFER, CENTRAL MOROCCO

ABSTRACT

In order to provide a recommendation on the quality of groundwater in the region of Berrechid, Morocco, we analysed the concentration of conductivity as one of the main measures to identify the salinity of the water. We applied artificial intelligence models for predicting the conductivity of water while analysing several physical parameters as input parameters of the models. To achieve this purpose, we exploited and evaluated the Random Forest (RF), Support Vector Regression (SVR), and k-nearest neighbour models using 400 data samples related to ten groundwater quality parameters in the Berrechid aquifer, Morocco. The results revealed that the overall prediction performance of the RF models is higher than the SVR and KNN models. Overall, the developed models are able to predict conductivity with high accuracy. The approaches developed in this study are promising for real-time and low-cost prediction of groundwater quality by using physical parameters as input variables.

Keywords: Groundwater quality, Artificial Intelligence, Random Forest, Support Vector Regression, k-Nearest Neighbour's.

INTRODUCTION

Groundwater is one of the most important water sources for agriculture, but also for other industries. Groundwater supplies nearly two-thirds of the world's population with drinking water and other domestic uses (Adimalla *et al*., 2018). More than 1.5 billion people worldwide rely on groundwater for basic needs such as drinking and irrigation (Adimalla and Li 2019). In addition, 946 million people lack access to adequate drinking water and have sanitation practices that are not clean (UNICEF and WHO 2015). Water availability is one of the most

[1] Mohammed Moukhliss, (corresponding author: moukhlissmohamed016@gmail.com), Abdeslam Taleb, Sonia Souabi, Abdessalam Ouallali; Faculty of Sciences and Techniques of Mohammedia, Hassan II University of Casablanca, MOROCCO;
[2] Velibor Spalevic, Biotechnical faculty, University of Montenegro, MONTENEGRO.
Received: 01/07/2022 *Accepted: 24/09/2022*

challenging tasks of the 21st century (Adimalla and Qian 2019). Water demand has been increasing due to population growth, intensive agriculture, urbanization, and industrial expansion in recent years. However, climate change, anthropogenic and natural activities have led to a decrease in groundwater quantity and degradation of its quality, making it unsuitable for domestic consumption, irrigation, and industry (Du *et al*. 2016; He *et al*. 2016; Li *et al*. 2018; Khanoranga and Khalid 2019; Su *et al*. 2019). Recent studies have estimated that 3.5-4.4 billion people are expected to live with the problem of water supply in 2050 due to climate change and increasing human water demand (Hanasaki *et al*. 2013; Wada *et al*. 2014). Therefore, assessment and prediction of groundwater quality is necessary for effective management and to keep pollution levels within allowable limits (Meguid 2019; Najah Ahmed *et al*. 2019). Conventional process-based modelling methods provide relatively accurate predictions for water quality parameters. Because of this, groundwater experts have focused their efforts on determining groundwater quality and suitability for drinking, household, irrigation, and industrial uses (Adimalla and Venkatayogi, 2018; Roy *et al*. 2018; Ehya and Saeedi, 2019; Mukate *et al*., 2019; Das *et al*., 2020).

Water quality is based on many indices and parameters adopted to meet water needs. Several studies have focused on index and statistical-based approaches. Recently, Adimalla *et al*. (2018) used Water Quality Index (WQI) to determine the adequacy of water quality for irrigation. Rao et al. (2018) used a groundwater pollution index (GPI) to assess groundwater quality for drinking purposes. Dutta et al. (2018) effectively demonstrates the use of WQI and multivariate statistical techniques to obtain simpler and more meaningful information about water quality and identify pollution sources. These methods have been applied to support groundwater quality research and promote various groundwater quality assessment and management strategies, all with the innovative and intelligent methods of reasonable cost to assess the state of groundwater quality. To this end, prediction-based approaches can be useful tools to overcome challenges in groundwater planning and management. Nowadays, application of artificial intelligence (AI) techniques has increased in many fields. In the environmental sciences, researchers proposed intelligent and adaptive dynamic water resources planning to preserve the water environment (all surface water, groundwater and wetlands) in urban areas (Xiang *et al*. 2021). Other studies conducted a systematic review of the literature on applying different types of artificial intelligence models to enable better monitoring of surface water quality (Ighalo *et al*. 2021). In the area of groundwater quality assessment, researchers investigated the performance of a few artificial intelligence techniques, including particle swarm optimization (PSO), a naive Bayes classifier (NBC) as a simple "probabilistic classifiers", and a support vector machine (SVM) with the purpose of predicting the water quality index (Agrawal *et al*. 2021). Other research scrutinize artificial intelligence technologies for assessing source water quality, disinfection, and membrane filtration, including monitoring

and identifying source water contaminants (Li *et al*. 2021). Researchers highlight a new approach based on integrating deep learning and feature extraction techniques to improve water quality classification (Dilmi and Ladjal, 2021). However, several factors are considered in a prediction model, including the nature and number of predictors used. Thus, a proper selection of input variables is required to increase the efficiency of Machine Learning (ML) models.

Morocco is a semi-arid Mediterranean country where surface and groundwater resources are essential for socio-economic sustainability. Consequently, aquifers are heavily exploited to meet increasing agricultural, industrial, and demands of local population. In addition, accelerated industrial operations, rapid population expansion, and agricultural intensification have led to significant groundwater depletion and degradation (Malki *et al*., 2017). Groundwater depletion is a significant problem in Morocco and is of particular concern to water managers. Rapid declines in groundwater levels (0.5 to 2 m per year on average) is generally caused by low groundwater recharge, marine intrusion, and excessive expansion of agricultural activities (Fadili *et al*., 2015; Najib *et al*., 2016; Ait Kadi and Ziyad, 2018; Alabjah *et al*., 2018; Mountadar *et al*., 2018; Bilali *et al*., 2021; Moukhliss *et al*., 2021; Zeynolabedin *et al*., 2021). Studied areas represent typical cases and are strongly impacted by climate variability, anthropogenic activity, and marine intrusion. In this case, the most important problems affecting groundwater recharge are intermittent river flow, decreasing reservoir capacity, the release of water by dams, removal of groundwater by pumping, soil salinity in irrigated areas. Despite these problems, the area remains an excellent example of effective quantitative and qualitative management of groundwater resources in integrated water resources management.

The main objectives of this study are: (1) predicting water conductivity using several physical parameters as model input parameters; with (2) comparing the performance of four models, including Random Forest, Support Vector Regression (SVR) and Machine learning (ML).

MATERIAL AND METHODS

Study area

The Berrechid aquifer is located in the Atlantic coastal basin between Rabat and Azemmour. It is around 10,470 km^2 in size (Royaume du Maroc, 2003). This aquifer is situated in the south of Casablanca. It differs from other aquifers in the region due to its enormous surface area of around 1,500 km^2. It is part of the quadrilateral formed by Settat, El Gara, Mediouana, and the center of Bouskoura (Figure 1).

This is a semi-arid region with annual rainfall ranging between 280 and 320 mm, with more than 90% falling between October and April. The temperature varies from 6.5 °C in January to 38 °C in August (Lyazidi *et al*., 2003; El Gasmi *et al*., 2014).

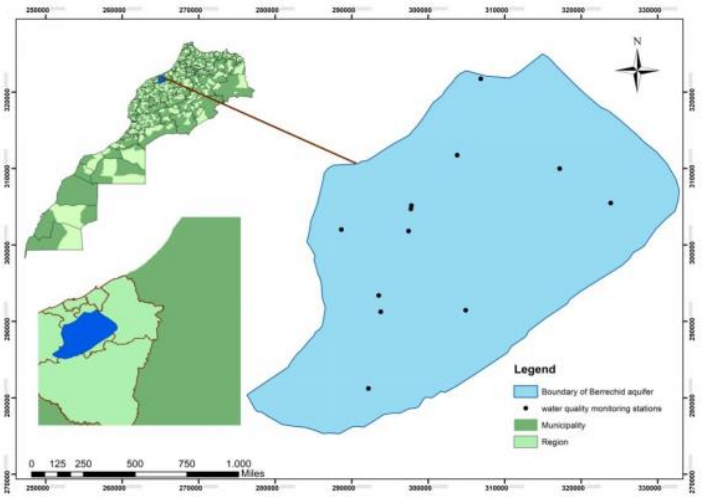


Figure 1. Study area and monitoring station

The basin is endorheic, having no outflow to an external body of water such as a river or ocean, and only losing water through evaporation or seepage into the ground; it is naturally fed by rainfall and streams that enter through the south and disappear below the plain (Moullard, 1960). The basin is flat in topography, with altitudes and slopes varying from 140 m and 0.2 % in the North to 350 m and 0.8% in the South (Naïma *et al*., 2015). In semi-arid climatic contexts, numerous geological formations were formed, combined with ancient subsidences and sedimentations spanning from primary to quaternary deposits. Nevertheless, the lithostratigraphic succession of these formations is as follows:

1) Primary, the bedrock consists of shale interbedded with layers of quartzite and sandstone, with 150 m thick outcrops Siluro-Devonian and Acadian Green to the SE and NW (El Mansouri, *et al*, 1992);

2) Triassic: Saline and siliciclastic red clays are interbedded with evaporite and basalt in these deposits. They are distributed in the eastern part of the aquifer (Bensalah *et al*., 2011);

3) The Infra-Cenomanian: with a total depth of about 40 m, including detrital red clays and gypsum abundant in the deposits. Layers of limestone and white succeeded some layers of the yellow marl conglomerate;

4) The Cenomanian, predominantly made up of dolomitic limestone and yellow marl, with green marl intercalated across a thickness of roughly 120 m (Ruhard 1975);

5) Pliocene that set the Berrechid aquifer system made up of sandstone, sand, sandy limestone, and small conglomerates with a total thickness of 5 to 40 m (Droubi *et al*. 2008);

6) Quaternary: The dominating facies is characterized by silts and conglomerates, which are followed by red silty clays, pebbles, and gravels with thicknesses varying from 0 to 50 m. As far as water quality is concerned, this

water table is classified into three hydro-geochemical facies (Na-Cl; Na-Mg-Ca-Cl; Ca-Mg-HCO_3-Cl). This region is marked by fertile soils and productive water tables, which causes the depletion of the water tables and the degradation of water quality, the latter becoming non-compliant with the consumption standard (El Ghali *et al*., 2020). Farmers mainly consume water for market gardening and livestock (Ouassissou *et al*., 2019). Water quality is monitored by the Bouregreg and Chaouia Basin Agency through fourteen monitoring stations.

Machine learning models

The models used are the Random Forest, Support Vector Regression (SVR), and KNN approaches.

Random Forest. It is a supervised machine learning algorithm that is hugely used in many classification and regression problems. It consists in building decision trees on different samples and takes their majority vote for classification and the mean in the case of regression. The concept of the sample with training data replacement is considered in each decision tree, also based on the bagging (Random forests: from early developments to recent advancements). The Random Forest offers several advantages: (a) Generally, the accuracy of Random Forest is considerably better than other types of decision trees. (b) The speed of Random Forest is high compared to bagging and boosting. (c) The estimates are significant in terms of error and correlation.

A large number of decision trees are the subject of the Random Forest. In both the construction and selection processes of the sample subsets, the random process is integrated to guarantee certain independence of each decision tree, thus improving the accuracy and performance of the model (A Review on Random Forest: An Ensemble Classifier). The following diagram (Figure 2) gives a brief description of the Random Forest model (K. Liu *et al*., 2021):

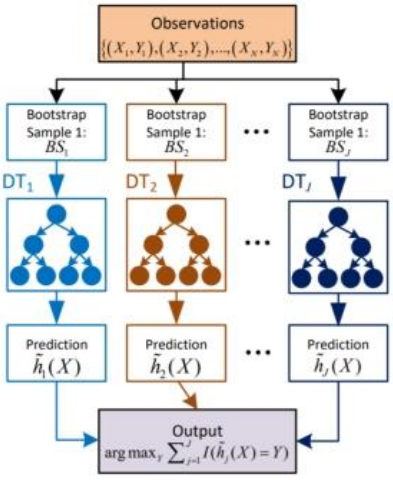


Figure 2. Random Forest process

K-nearest neighbors – KNN. The algorithm is a Machine Learning algorithm (Peterson 2009) belonging to the class of simple supervised learning algorithms, easy to implement, and can solve classification and regression problems. The KNN algorithm consists of two primary steps:
Step 1: Select the number K of neighbors.
Step 2: Calculate the distance, the Manhattan distance or the Euclidean distance.
Step 3: Take the K nearest neighbors according to the calculated distance.
Step 4: Among K neighbors, count number of points belonging to each category.
Step 5: Assign the new point to the most present category among K neighbors.

Several methods can be effective in determining the value of K. Generally, several values of K are tested to obtain the most effective value. Usually, the K number is increased. This defines the nearest neighbor region and brings more clarity and transparency to the results. However, when the data dispersion is a problem, the value is difficult to specify. The diagram of the KNN algorithm summarizes all the steps (Figure 3).

Figure 3. KNN process

Support vector regression SVR. It is an evolution of support vector classification to perform regression tasks. To perform this task, the ε-insensitive loss function is implemented to analyze the data.

Data collection and analysis. As part of this research, four hundred quality samples were collected at the following monitoring sites in the Berrechid aquifer. The sample methodology was carried out twice in the spring and summer during irrigation pumping. The wells varied in depth from 40 to 140 meters. All of the wells are utilized on a regular basis for residential usage, irrigation, and industrial purposes.

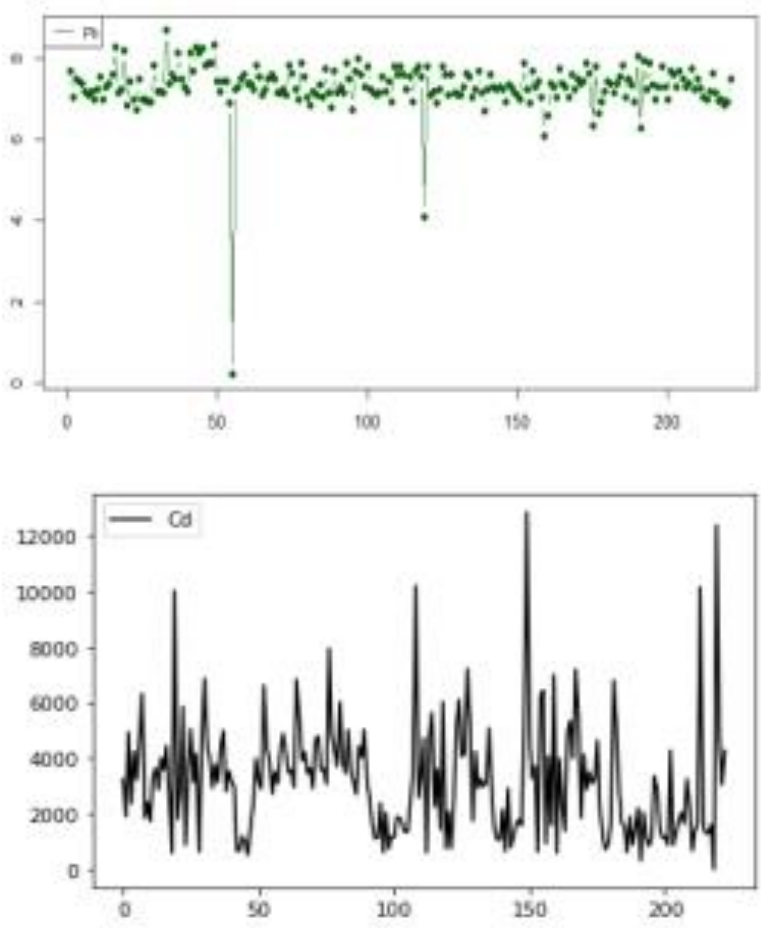
For sample collection, we utilized one-liter polyethylene bottles that had been prewashed and labeled. The following parameters were measured in these samples: Electrical Conductivity (EC), Chlorides (Cl^-), Calcium (Ca^{2+}), Magnesium (Mg^{2+}), pH, Sulfate (SO_4^{2-}), Sodium (Na^+), Potassium (K^+), Carbonates (CO_3^{2-}), and Bicarbonates (HCO_3^-). The collecting points' location coordinates (X, Y coordinates) were determined using a portable global positioning system (GPS). The samples are sent immediately to the laboratory for analysis.

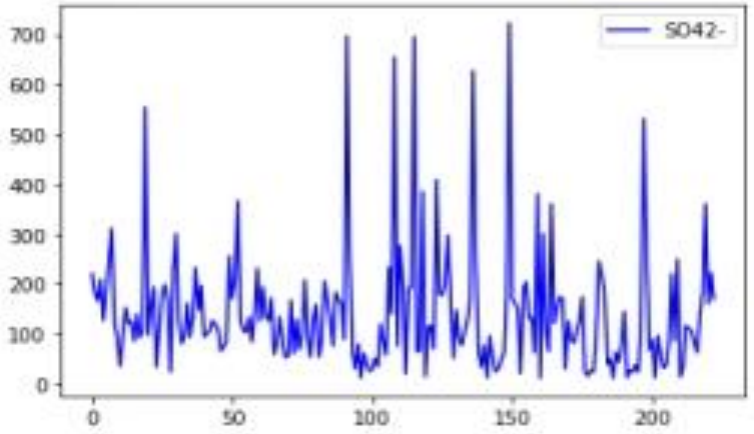
Conductivity was measured with an intelligent conductivity meter YK-2001PH. The determination of chloride was achieved by the MOHR method, according to AFNOR 90-014. The determination of bicarbonate and carbonate was carried out by the titration method in the presence of H_2SO_4 (0.02 N), a solution of NaOH (0.1 N), phenolphthalein and methyl orange indicator. The sulfate determination was performed using the spectrophotometric method (Shimadzu UV 1800 model). The determination of potassium was conducted by a flame photometer (Elico flame photometer model CL 22 B) and KCl reagents. Sodium was determined by a flame photometer (Elico flame photometer model CL 22 B) and by NaCl reagent. Calcium was determined by EDTA titrimetric analysis. The magnesium determination was performed by TH and Ca concentration. The solutions' hydrogen potential (pH) was determined by a "Accumet Basic AB15" pH meter.

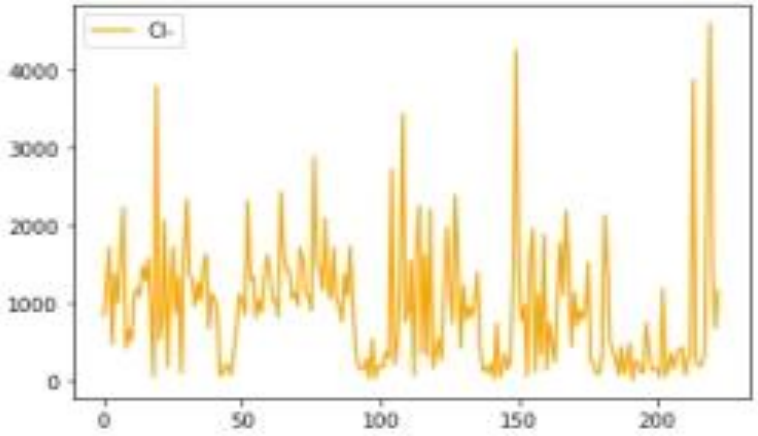
Data processing and model performance evaluation methods. From an analysis of all the parameters, there are aberrant values on almost all the parameters. Therefore, it is important to analyze the data before entering it. Indeed, data exploration and cleaning are among the main steps to obtaining efficient models. Moreover, in ML model development, data mining and cleaning are the significant steps to obtaining accurate and reliable models. In this study, data pre-processing was performed in several steps:

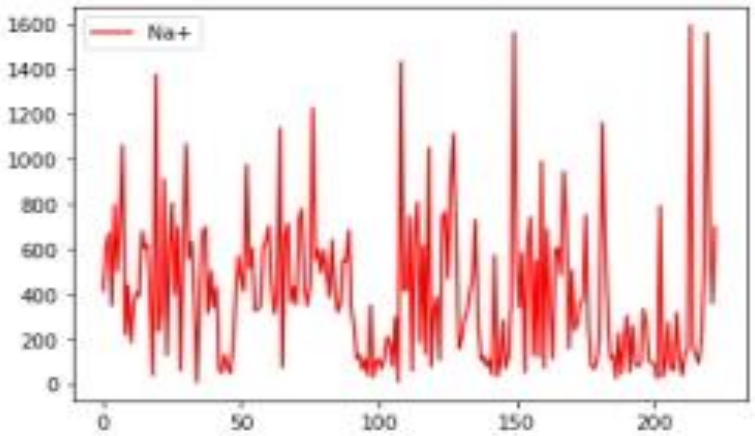
Step 1: Checking for outliers and missing values. In our database, there are two scenarios. The first scenario concerns outliers. An outlier is a value or observation that is "distant" from other observations made on the same phenomenon, i.e. it contrasts significantly with the "normally" measured values. Therefore, any value that appears to be an outlier in the database should be removed. They can either be replaced for missing values the row related to the missing data can be deleted. Since there are only two missing values in our case, we chose to delete the other data related to these missing values.

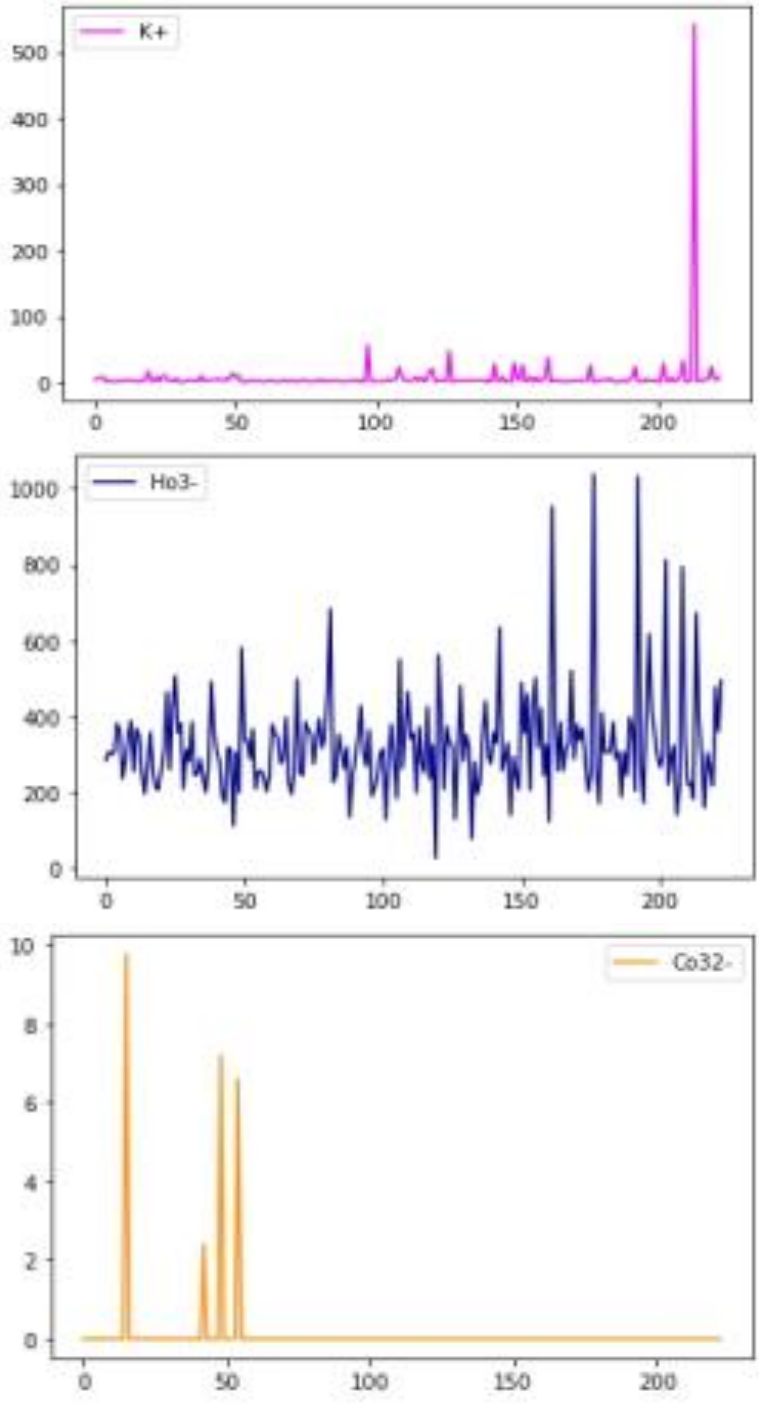
Step 2: A complete description of the different parameters involved. This step consists of making a complete description of all the parameters we consider in creating the machine learning models. This first description includes the evolution of each parameter and the min, max, average and median values, etc. A general view, but also diagrams showing the range for each parameter. This is achieved firstly through graphs visualizing the data for each variable. Figure 4 illustrates the following variables: CE, pH, Cl^-, SO_4^{2-}, Na^+, K^+, CO_3^{2-}, HCO_3^-, Ca^{2+}, Mg^{2+}. In the conductivity, for example (CE), there are places where the conductivity is significantly high and where the conductivity remains very limited. This creates ups and downs throughout the graph. This applies to all variables. However, there are variables where the values remain more or less like K^+ and CO_3^{2-}.




SO42-
700
600
500
400
300
200
100
0
0
50
100
150
200


Cl-
4000
3000
2000
1000
0
0
50
100
150
200


Na+
1600
1400
1200
1000
800
600
400
200
0
0
50
100
150
200


K+
500
400
300
200
100
0
0
50
100
150
200
Ho3-
1000
800
600
400
200
0
0
50
100
150
200
Co32-
10
8
6
4
2
0
0
50
100
150
200

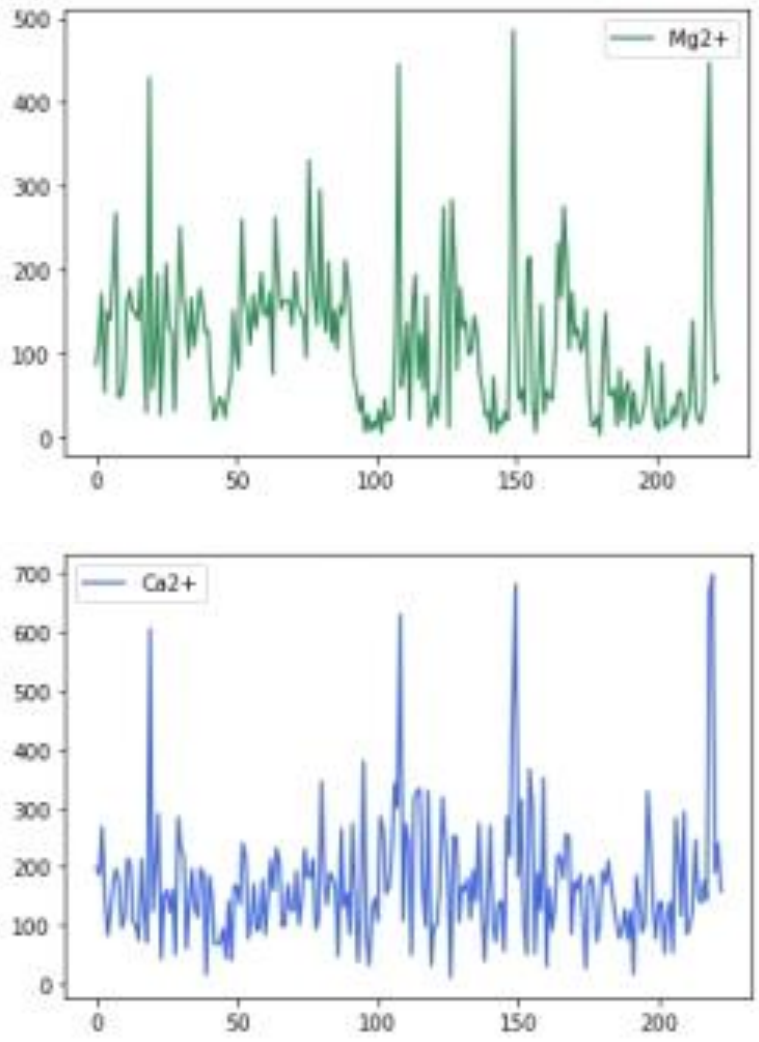


Figure 4. Plots of the raw data related to ten parameters
(X line represents the identification of samples while Y line represents the parameters)

The boxplot of the variables (Fig. 5) presents the evolution of the different variables, focusing mainly on the first quartile, the third quartile, etc. According to the obtained boxplot, 25% of the conductivities are between 0 and below 2000. However, the other values exceed 2000. pH and CO_3^{2-}, on the other hand, take the lowest values among all the variables.

In a third step, we made a general description of the variables using the following characteristics (Table 1): minimum, maximum, first quartile, median, average, third quartile, and maximum. The variables with the lowest minimum values are CO_3^{2-} and pH, and the highest maximum values are Ca^{2+} and Cl^-.

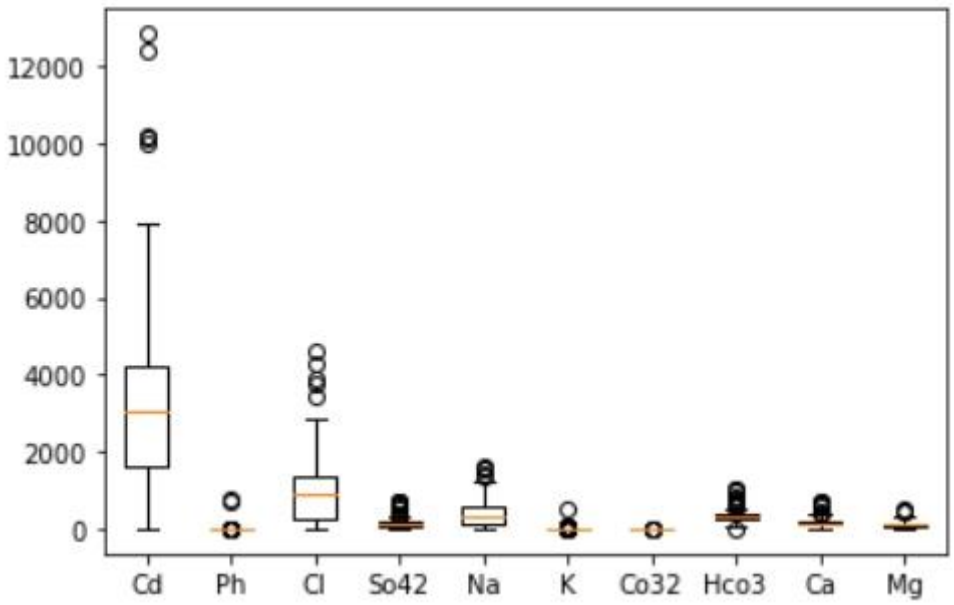


Figure 5. Boxplots of the used variables in the database.

Table 1. Descriptive statistical characteristics of the used variables.

	CE	pH	Cl^-	SO_4^{2+}	Na^+	K^+	CO_3^{2}-	HCO_3^-	Ca^+	Mg^{2+}
Min	7.54	0.2	12	11	5	0.29	0	26	11.2	3.68
1st Quartile	1595	7.138	276.2	64.67	124	1.157	0	244	102.8	37.83
Median	3014	7.3	867	117.5	349	2.640	0	305	154	101
Mean	3193	7.32	947.3	140.11	406.2	7.191	0.1159	323.3	171.6	109.07
3rd Quartile	4240	7.6	1352.8	174.75	604	4.030	0	363.8	212.0	155.5
Max	12835	8.7	4603	723	1595	542	9.76	1037	697	486

Step 3: Training and model validation process. In Machine Learning, and to prove the performance and reliability of an ML model, the database is divided into two main parts: The first part is dedicated to creating the ML model and the second part is dedicated to validating the model. Generally, the percentage of data dedicated to validation is between 20% and 30%.

Metric evaluation models. Model performance is evaluated using three statistical criteria: correlation coefficient (Asuero *et al*., 2006), root mean square error (RMSE) (Wang and Lu, 2018) and accuracy (Yin *et al*., 2019).

Correlation coefficient formulas are used to determine the strength of a relationship between data. The formulas return a value between -1 and 1, where: 1 indicates a strong positive relationship, -1 indicates a strong negative relationship. A result of zero indicates that there is absolutely no relationship whatsoever (Eq. (2)):

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 (Y_i - \overline{Y})^2}} \qquad \text{Eq. (2)}$$

RMSE is the square root of the variance of the residual errors (Eq. (3)). The lower value of the RMSE compared to the output ranges indicates a better fit of the model.

$$RSME = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \qquad \text{Eq. (3)}$$

Machine learning model accuracy is used to determine which model best identifies relationships and patterns between variables in a dataset based on the input or training data. Accuracy is defined as follows (Eq. (4)):

$$Accuracy = \frac{Correct\ predictions}{All\ predictions} \qquad \text{Eq. (4)}$$

RESULTS AND DISCUSSION

Exploratory data analysis

A correlation matrix and evaluation of the input variables' significance were performed for further exploration of the variable. Table 2 shows the matrix correlation. Figure 8 highlights the correlation plot. Our purpose is to predict the conductivity as a function of the other parameters: a complete description of the different parameters involved in the study.

The results revealed that the conductivity (CE) has a strong correlation with the parameters Cl^- (0.92), Na^+ (0.91) and Mg^{2+} (0.87) and a moderate correlation with the parameters SO_4^{2-} (0.58) and Ca^{2+} (0.6). However, the conductivity correlates poorly with the parameters pH (0.06), K^+ (0.26), CO_3^{2-} (-0.008) and HCO_3^- (0.17). These results indicate that the variables Cl^-, Na^+, Mg^{2+}, SO_4^{2-} and Ca^{2+} better predict the conductivity. Although the other parameters are less correlated with the conductivity, they are needed to improve the prediction accuracy of the ML models.

Table 2. Correlation matrix of the used variables

	CE	pH	Cl^-	SO_4^{2+}	Na^+	K^+	CO_3^{2-}	HCO_3^-	Ca^+	Mg^{2+}
CE	1	0.0606	0.925	0.583	0.919	0.26	-0.008	0.17	0.607	0.87
pH	0.06	1	0.062	-0.038	0.111	-0.017	-0.013	0.039	-0.071	0.05
Cl^-	0.92	0.062	1	0.507	0.908	0.27	0.013	0.084	0.616	0.89
SO_4^{2+}	0.583	-0.03	0.507	1	0.541	0.024	-0.049	0.076	0.616	0.55
Na^+	0.9190.111	0.111	0.908	0.541	1	0.293	0.035	0.185	0.475	0.797
K^+	0.261	-0.0179	0.27	0.024	0.293	1	-0.015	0.232	0.071	0.034
CO_3^{2-}	-0.008	-0.0137	0.0135	-0.049	0.035	-0.0156	1	-0.0899	-0.122	0.004
HCO_3^-	0.17	0.00396	0.0842	0.076	0.185	0.232	-0.089	1	0.094	0.052
Ca^+	0.607	-0.071	0.616	0.6168	0.475	0.071	-0.122	0.094	1	0.608
Mg^{2+}	0.873	0.0523	0.89	0.554	0.797	0.0344	0.004	0.052	0.608	1

Simulated and real values. This part is dedicated to schematizing the simulated values created by the different models and comparing them to the real conductivity values in part dedicated to the validation. For example, starting with the Random Forest (Fig. 6), it turns out that a large percentage of the simulated values coincide with the actual values, except for a few points that deviate sharply and remarkably.

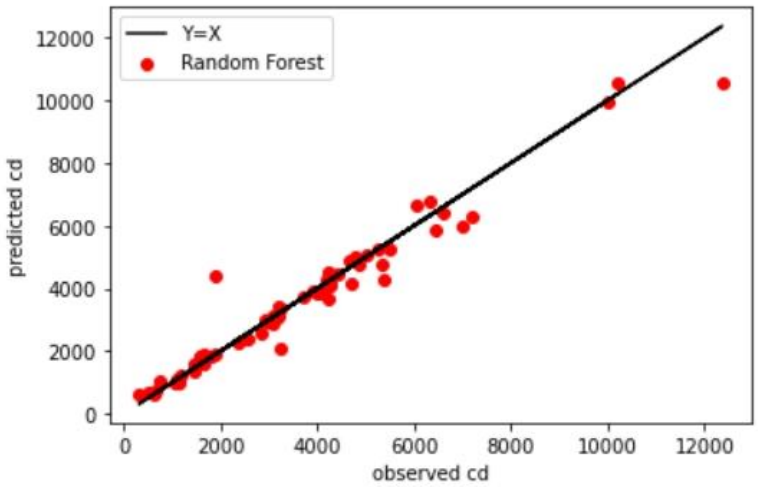


Figure 6. Scatter plots of the observed and simulated values for predicting the conductivity using the Random Forest Model

For the SVR model (Fig. 7), it is likewise found that several predicted values coincide approximately with real values, which shows a remarkable ability of the model to predict the conductivity in the groundwater. However, there are some locations where there is a considerable discrepancy between the predicted and actual values, especially in the last part of the plot.

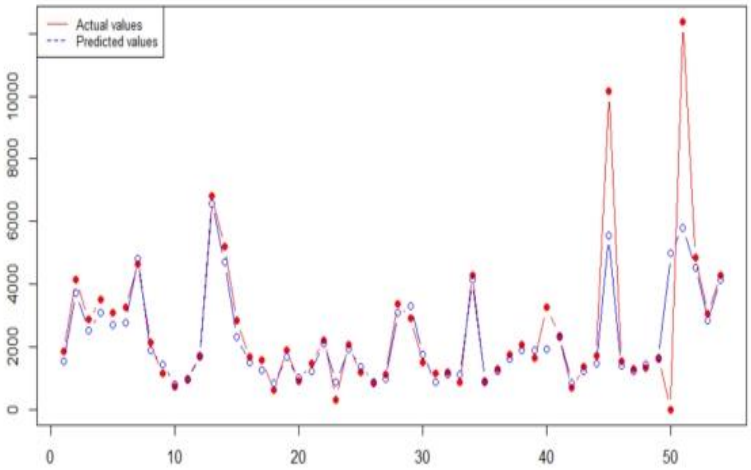


Figure 7. Scatter plots of the observed and simulated values for predicting the conductivity using the SVR Model

For the KNN model (Fig. 8), we notice that the points are more scattered compared to the model considering the RF. In addition, several points are far away from the actual values, namely the last points.

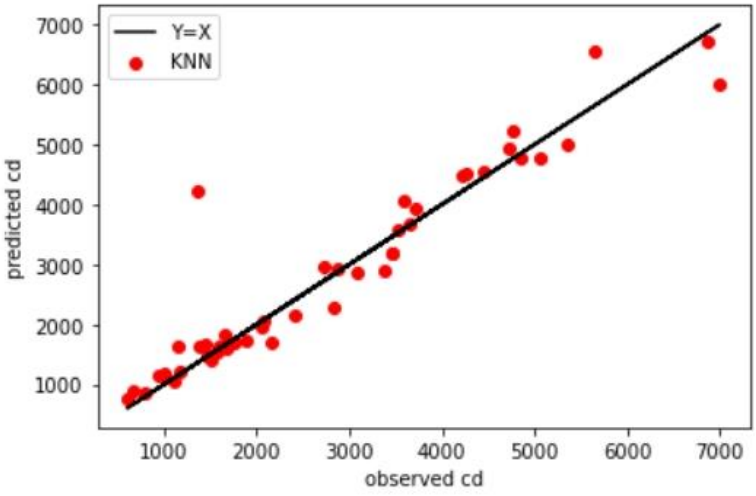


Figure 8. Scatter plots of the observed and simulated values for predicting the conductivity using the KNN Model

If we compare the three figures, we can see that the RF model is the model with the most overlap between the predicted and actual values. Most of the points rotate around the X=Y line, proving that this model can predict the conductivity with a high accuracy depending on the selected parameters.

General results

Before the training phase, the selected data were organized in a CSV file for processing. Thus, they were divided into two datasets: 372 samples were used for the training process and 93 samples were run in the validation phase. Then, the ML models were built in Spyder using the anaconda platform. In this sense, the evaluation metrics are measured by comparing the outputs of the validation part generated by the created model and the actual outputs. By comparing the two, we will be able to specify which model or algorithm performs better in terms of correlation, squared error and accuracy (Fig. 9).

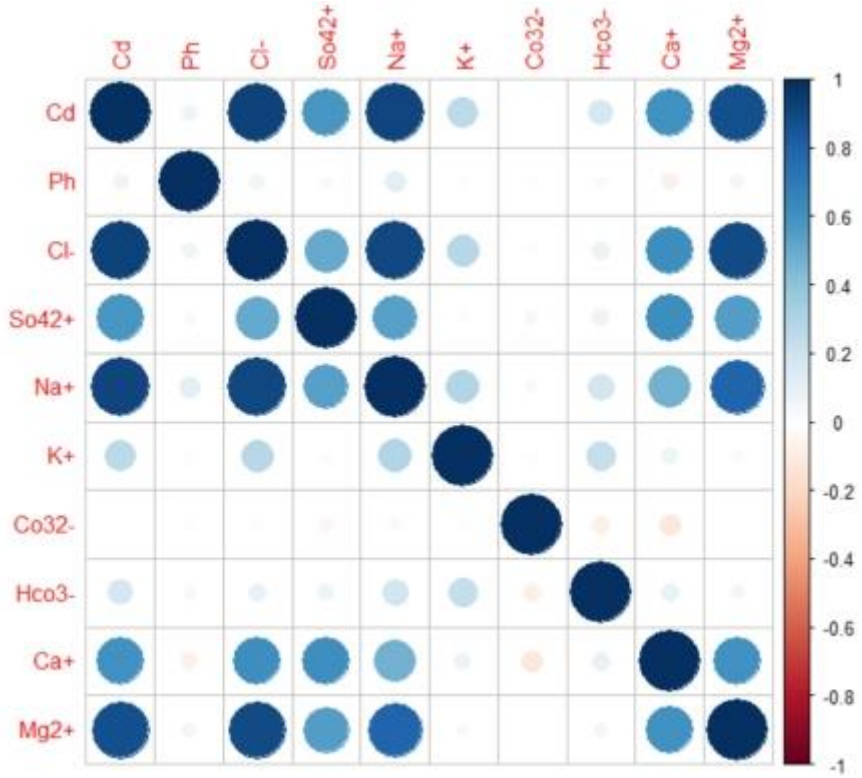


Fig. 9. The correlation plot of the used variables

Table 3 represents the RMSE, R-values and accuracy of the models during the training process. These results reveal that the RF model yields the best result in correlation r with a value of 0.966 compared to a value of 0.949 for the KNN algorithm and a value of 0.825 for the SVR model. For the value of RSME, we

calculated the normalized RMSE (Eq. (5)) using the following formula with y is the output parameter (Conductivity).

$$RMSEnormalized = \frac{RMSE}{y_{max} - y_{min}} \qquad Eq.(5)$$

The smaller the RMSE value, the lower the error rate and the ML model performs better. Through this indication, we can see that in terms of RMSE, the RF model performs better with an RMSE of 0.042 against an RMSE value of 0.082 for the KNN and a value of 0.102 for the SVR. Regarding accuracy, the RF model is the most accurate among the three models. The KNN algorithm provides an accuracy of 85.56% against accuracy of 83.86% for the SVR model.

Table 3. Evaluation metrics for RF, SVR and KNN models

Statistical indices	Random Forest (RF)	Support Vector Regression (SVR)	K-Nearest Neighbor (KNN)
Correlation r	0.966	0.825	0.949
RSME	0.042	0.102	0.082
Accuracy	89.19 %	83.86 %	85.56 %

Field practical implication in water management

Our study built and validated ML models to predict the conductivity using several parameters as predictors: pH, Cl^-, K^+, Na^+, etc. Following the results of this study, the ML models have proven to be accurate and robust in predicting conductivity (Table 3). In addition, we opted for several evaluation measures. These analyses reveal that the capabilities of the RF and KNN models are significantly higher than the SVR model in the present study. In terms of correlation and squared error and accuracy, the RF performs better. For the SVR, we notice a high mean square error of 0.102, which presents a difference of more than 50% between the RF and the SVR in terms of square error. The objective of this paper was to predict the conductivity level to assess the groundwater quality with the minimum costs.

To measure the conductivity of Water, it requires a whole measurement process and equipment such as a probe or a conductivity meter. The basic principle of conductivity measurement consists of several steps: the measuring instrument applies an electrical voltage to the solution to be measured. An electric current flows according to the conductivity. Depending on the method or application, either the meter applies a constant voltage. It records the change in electric current, or the meter applies a constant current and evaluates the voltage change. In addition to the measuring equipment costs, there is also the necessity to maintain the equipment. Indeed, wear and tear and dirt reduce its reliability and introduce measurement errors. A regular calibration allows detecting of unreliable components. It restores them to their normal state, for example, by cleaning them. Non-maintenance can therefore cause divergent measurements

and thus a dispersion of the measurements. Cleaning is also a possible cost of measuring the conductivity of Water. Cleaning can be achieved with hot Water or vinegar. All this leads to a set of costs and expenses that is summarized in the following equation (Eq(6)):

$$General\ cost = Equipement\ cost + Maintaining\ cost + Cleaning\ cost + Other\ possible\ costs \qquad Eq.\ (6)$$

Equation 5 highlights the high cost of measuring the conductivity degree. Our study, leveraging Machine Learning models, allows us to reduce these costs distinctively, with high accuracy and performance.

CONCLUSION

In this paper, our objective was to predict the conductivity contributing in the evaluation of grounded water quality as a function of several parameters such as: Mg2+, Na+, Cl-... In this perspective, we used several ML models using several input parameters. Also, we performed a statistical description that calculates the correlation matrix and the different measurements of each variable. Therefore, Random Forest, SVR and KNN models were developed and evaluated to predict the conductivity using several input parameters. This study leads to several results:

ML models based on several input parameters are All three models (RF, SVR and KNN) produced a correlation more significant than 80%.

All three models (RF, SVR and KNN) give more than 80% accuracy.

The RF model generates more significant results compared to the other models.

The ML models allowed us to predict the conductivity with high accuracy and low error as a function of several other parameters.

The models created to minimize the costs of measuring conductivity with special equipment are highly effective tools. Therefore, they should be recommended to predict water quality parameters. These models will improve groundwater quality monitoring for irrigation purposes in real-time and at a low cost.

REFERENCES

Adimalla, N., Li, P. (2019). Occurrence, health risks, and geochemical mechanisms of fluoride and nitrate in groundwater of the rock-dominant semi-arid region, Telangana State, India. Human and Ecological Risk Assessment: An International Journal. 25(1–2): 81–103. doi:10.1080/10807039.2018.1480353. [accessed 2022 Jan 30]. https://www.tandfonline.com/doi/full/10.1080/10807039.2018.1480353.

Adimalla, N., Li, P., Venkatayogi, S. (2018). Hydrogeochemical Evaluation of Groundwater Quality for Drinking and Irrigation Purposes and Integrated Interpretation with Water Quality Index Studies. Environ Process. 5(2):363–383. [accessed 2022 Jan 30]. http://link.springer.com/10.1007/s40710-018-0297-4.

Adimalla, N., Qian, H. (2019). Hydrogeochemistry and fluoride contamination in the hard rock terrain of central Telangana, India: analyses of its spatial distribution and health risk. SN Appl Sci. 1(3): 202. doi:10.1007/s42452-019-0219-8. [accessed 2022 Feb 2]. http://link.springer.com/10.1007/s42452-019-0219-8.

Adimalla, N., Venkatayogi, S. (2018). Geochemical characterization and evaluation of groundwater suitability for domestic and agricultural utility in semi-arid region of Basara, Telangana State, South India. Appl Water Sci. 8(1):44. [accessed 2022 Jan 5]. https://doi.org/10.1007/s13201-018-0682-1.

Agrawal, P., Sinha, A., Kumar, S., Agarwal, A., Banerjee, A., Villuri, V.G., Annavarapu, C.S., Dwivedi, R., Dera, V.V., Sinha, J. (2021). Exploring Artificial Intelligence Techniques for Groundwater Quality Assessment. Water 2021, 13, 1172. https://doi.org/10.3390/w13091172.

Ait Kadi, M., Ziyad, A. (2018). Integrated Water Resources Management in Morocco. In: World Water Council, editor. Global Water Security. Singapore: Springer Singapore. Water Resources Development and Management, p. 143–163. [accessed 2022 Feb 3]. http://link.springer.com/10.1007/978-981-10-7913-9_6.

Alabjah, B., Amraoui, F., Chibout, M., Slimani. M. (2018). Assessment of saltwater contamination extent in the coastal aquifers of Chaouia (Morocco) using the electric recognition. Journal of Hydrology. 566: 363–376. [Accessed 2022 Jan 5]. https://www.sciencedirect.com/science/article/pii/S0022169418306899.

Asuero, A.G., Sayago, A., González, A.G. (2006). The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1):41–59. doi:10.1080/10408340500526766.

Bensalah, M.K., Youbi, N., Mahmoudi, A., Bertrand, H., Mata, J., Hachimi, H.E., Martins, L., Marzoli. A., Bellon, H., Medina, F. (2011). The Central Atlantic Magmatic Province (CAMP) volcanic sequences of Berrechid and Doukkala basins (Western Meseta, Morocco). Volcanology and geochemistry. Comunicações Geológicas 98(1): 15-27.

Bilali, A.E. Moukhliss, M., Taleb, A., Nafii, A. Alabjah, B. Brouzyine, Y., Mazigh, N., Teznine, K., Mhamed, M. (2021). Predicting Daily Pore Water Pressure in Embankment Dam Using Machine Learning Models and Hydrostatic Seasonal Time Approach. In Review. https://www.researchsquare.com/article/rs-947164/v1.

Das, A., Das, S.S., Chowdhury, N.R., Joardar, M., Ghosh, B., Roychowdhury, T. (2020). Quality and health risk evaluation for groundwater in Nadia district, West Bengal: An approach on its suitability for drinking and domestic purpose. Groundwater for Sustainable Development. Volume 10, April 2020, 100351.

Dilmi, S., Ladjal, M. (2021). A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques. Chemometrics and Intelligent Laboratory Systems. Volume 214, 15 July 2021, 104329. doi:10.1016/j.chemolab.2021.104329.

Droubi, A., Al-Sibai, M., Abdallah, A., Zahra, S., Obeissi, M., Wolfer, J., Huber, M., Hennings, V., Schelkes, K. (2008). A Decision Support System (DSS) for Water Resources Management, – Design and Results from a Pilot Study in Syria. In: Zereini F, Hötzl H, editors. Climatic Changes and Water Resources in the Middle East and North Africa. Berlin, Heidelberg: Springer. (sur les sciences et l'ingénierie, environnement). p. 199–225. [accessed 2022 Jan 5]. https://doi.org/10.1007/978-3-540-85047-2_16.

Du, S., Zheng, L., Zhang, W. (2016). Assessment of shallow aquifer remediation capacity under different groundwater management conditions in CGS field. Arab J Geosci. 9(6):448. doi:10.1007/s12517-016-2479-6. [accessed 2022 Feb 1]. http://link.springer.com/10.1007/s12517-016-2479-6.

Dutta, S., Dwivedi, A., Suresh Kumar, M. (2018). Use of water quality index and multivariate statistical techniques for the assessment of spatial variations in water quality of a small river. Environ Monit Assess. 190(12):718. doi:10.1007/s10661-018-7100-x. [accessed 2022 Feb 2]. https://doi.org/10.1007/s10661-018-7100-x.

Ehya, F., Saeedi, F. (2019). Assessment of groundwater quality in the Garmez area (Southeastern Khuzestan province, SW Iran) for drinking and irrigation uses. Carbonates Evaporites. 34(4): 1443–1454. doi:10.1007/s13146-018-0481-7. [accessed 2022 Feb 2]. https://doi.org/10.1007/s13146-018-0481-7.

El Gasmi, E.H., Mridekh, A., El Mansouri, B., Tammal, M., El Bouhaddioui, M. (2014). Contribution of geophysical and geological data for the identification of new structural elements related to the Settat flexure (central Morocco). 36: 109–121.

El Ghali, T., Marah, H., Qurtobi, M. Raibi, F., Bellarbi, M., Amenzou, N., El Mansouri, B. (2020). Geochemical and isotopic characterization of groundwater and identification of hydrogeochemical processes in the Berrechid aquifer of central Morocco. Carbonates Evaporites. 35(2): 37. doi:10.1007/s13146-020-00571-y. [accessed 2022 Jan 5]. https://doi.org/10.1007/s13146-020-00571-y.

El Mansouri, B., Dzikowski, M., Delay, F., Carlier, E., & Crampon, N. (1992). Calage d'un modèle mathématique en régime permanent appliqué à la nappe de Berrechid (Maroc). Ann. Soc. Géol. Du Nord, 189–193.

Fadili, A., Khalid, M., Joëlle, R., Najib, S., Makan, A., Boutayab, K. (2015). Evaluation of groundwater mineralization processes and seawater intrusion extension in the coastal aquifer of Oualidia, Morocco: Hydrochemical and geophysical approach. Arabian Journal of Geosciences. 8. doi:10.1007/s12517-015-1808-5.

Hanasaki, N., Fujimori, S., Yamamoto, T., Yoshikawa, S., Masaki, Y., Hijioka, Y., Kainuma, M., Kanamori, Y., Masui, T., Takahashi, K. (2013). A global water scarcity assessment under Shared Socio-economic Pathways – Part 1: Water use. Hydrol Earth Syst Sci. 17(7):2375–2391. doi:10.5194/hess-17-2375-2013. [accessed 2022 Feb 1]. https://hess.copernicus.org/articles/17/2375/2013/.

He, X., Liu, Z., Qian, J., Zhao, W., Liu, Y. (2016). Distribution of Nitrate in Different Aquifers in the Urban District of Zhanjiang, China. Bull Environ Contam Toxicol. 97(2): 279–285. doi:10.1007/s00128-016-1822-7. [accessed 2022 Feb 1]. https://doi.org/10.1007/s00128-016-1822-7.

Ighalo, J.O., Adeniyi, A.G., Marques, G. (2021). Artificial intelligence for surface water quality monitoring and assessment: a systematic literature analysis. Modeling Earth Systems and Environment. 7(2): 669–681. doi:10.1007/s40808-020-01041-z.

Liu, K., Hu, X., Zhou, H., Tong, L., Widanage, W. D., Marco, J. (2021). Feature Analyses and Modeling of Lithium-Ion Battery Manufacturing Based on Random Forest Classification. IEEE/ASME Transactions on Mechatronics. 26(6):2944–2955. doi:10.1109/TMECH.2020.3049046.

Khanoranga Khalid, S. (2019). An assessment of groundwater quality for irrigation and drinking purposes around brick kilns in three districts of Balochistan province, Pakistan, through water quality index and multivariate statistical approaches. Journal of Geochemical Exploration. 197:14–26.

Li, L., Rong, S., Wang, R., Yu, S. (2021). Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. Chemical Engineering Journal. 405:126673. doi:10.1016/j.cej.2020.126673.

Li, X., Wu, H., Qian, H., Gao, Y. (2018). Groundwater Chemistry Regulated by Hydrochemical Processes and Geological Structures: A Case Study in Tongchuan, China. Water. 10(3):338. doi:10.3390/w10030338. [accessed 2022 Jan 30]. http://www.mdpi.com/2073-4441/10/3/338.

Lyazidi, A., El Wartiti, M., Fadli, D. (2003). Évolution géodynamique du bassin triasique de Berrechid-ElGara-BenSlimane : Dynamique sédimentaire et géométrie des dépôts (Méseta nord occidentale, Maroc) Geodynamic evolution of the Berrechid-ElGara-BenSlimane triassic basin (north-western Meseta, Morocco): Depositional dynamics and geometry of sedimentary bodies. Pangea infos. 39/40:23–36. [accessed 2022 Jan 5]. https://hal-insu.archives-ouvertes.fr/insu-00948279.

Malki, M., Choukr-Allah, R., Bouchaou, L., Hirich, A., Ait Brahim, Y., Krimissa, S., Hssaisoune, M., Nghira, A., Barceló, D. (2017). Assessment of Groundwater Quality: Impact of Natural and Anthropogenic Contamination in Souss-Massa River Basin. In: Choukr-Allah R, Ragab R, Bouchaou Lhoussaine, Barceló Damià, editors. The Souss-Massa River Basin, Morocco. Cham: Springer International Publishing. (The Handbook of Environmental Chemistry). p. 143–162. [accessed 2022 Feb 3]. https://doi.org/10.1007/698_2016_72.

Meguid, M.A. (2019). Key Features of the Egypt's Water and Agricultural Resources. In: Negm AM, editor. Conventional Water Resources and Agriculture in Egypt. Cham: Springer International Publishing. (The Handbook of Environmental Chemistry). p. 39–99. https://doi.org/10.1007/698_2017_41.

Moukhliss, M., Taleb, A., Khattabi Rifi, S., El fathi, B., El bilali, A., Souabi, S. (2021). Assessment of groundwater quality in Berrechid Aquifer, Morocco. Moroccan Journal of Chemistry. 9(4): 9–812. doi:10.48317/IMIST.PRSM/morjchem-v9i3.26974. https://revues.imist.ma/index.php/morjchem/article/view/26974.

Moullard, L. 1960. Nappe phréatique de la plaine de Berrechid. Assoc. Inter. Hydrol. Sc. Helsinki, 52, 105–142.

Mountadar, S., Younsi, A., Hayani, A., Siniti, M., Tahiri, S. (2018). Groundwater salinization process in the coastal aquifer Sidi Abed-Ouled Ghanem (Province of El Jadida, Morocco). Journal of African Earth Sciences. 147:169–177. doi:10.1016/j.jafrearsci.2018.06.025.

Mukate, S., Wagh, V., Panaskar, D., Jacobs, J.A., Sawant, A. (2019). Development of new integrated water quality index (IWQI) model to evaluate the drinking suitability of water. Ecological Indicators. 101: 348–354. doi:10.1016/j.ecolind.2019.01.034.

Naïma, E.A., Amraoui, F., El Mansouri, B. (2015). Modélisation numerique de l'effet des changements climatiques sur la nappe de berrechid (Maroc). European Scientific Journal. 11: 1857–7881.

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology. 578: 124084. doi:10.1016/j.jhydrol.2019.124084.

Najib, S., Fadili, A., Mehdi, K., Riss, J., Makan, A., Guessir, H. (2016). Salinization process and coastal groundwater quality in Chaouia, Morocco. Journal of African Earth Sciences. 115:17–31. doi:10.1016/j.jafrearsci.2015.12.010.

Ouassissou, R., Kuper, M., Hammani, A., El Amrani, M. (2019). Le contrat de gestion participative pourrait-il résoudre la crise de gouvernance des eaux souterraines ? Cas de la nappe de Berrechid au Maroc. Alternatives Rurales. [accessed 2022 Jan 5]. https://agritrop.cirad.fr/594402/.

Peterson, L. (2009). K-nearest neighbour. Scholarpedia. 4(2): 1883. doi:10.4249/scholarpedia.1883.

Rao, N.S., Sunitha, B., Rambabu, R., Rao, P.V.N., Rao, P.S., Spandana, B.D., Sravanthi, M., Marghade, D. (2018). Quality and degree of pollution of groundwater, using PIG from a rural part of Telangana State, India. Appl Water Sci. 8(8): 227.

Roy, A., Keesari, T., Mohokar, H., Sinha, U.K., Bitra, S. (2018). Assessment of groundwater quality in hard rock aquifer of central Telangana state for drinking and agriculture purposes. Appl Water Sci. 8(5): 124. doi:10.1007/s13201-018-0761-3. [accessed 2022 Feb 2]. https://doi.org/10.1007/s13201-018-0761-3.

Royaume du Maroc. (2003). Ministère de l'Aménagement du Territoire, de l'Eau et de l'Environnement. Département de l'Aménagement Du Territoire, Le Schéma National d'Aménagement Du Territoire, Groupe Huit-Sud.

Ruhard (1975). Chaouia et la plaine de Berrechid (Tome 1). Ressources en Eau du Maroc. Notes et Mémoire Du Service Géologique, 231.

Su, F., Wu, J., He, S. (2019). Set pair analysis-Markov chain model for groundwater quality assessment and prediction: A case study of Xi'an city, China. Human and Ecological Risk Assessment: An International Journal. 25(1–2):158–175. doi:10.1080/10807039.2019.1568860. [accessed 2022 Jan 31]. https://www.tandfonline.com/doi/full/10.1080/10807039.2019.1568860.

UNICEF and WHO. (2015). Progress on Sanitation and Drinking Water : 2015 Update and MDG Assessment. World Health Organization.

Wada, Y., Gleeson, T., Esnault, L. (2014). Wedge approach to water stress. Nature Geosci. 7(9):615–617. doi:10.1038/ngeo2241. [accessed 2022 Feb 1]. http://www.nature.com/articles/ngeo2241.

Wang, W., Lu, Y. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. IOP Conference Series: Materials Science and Engineering. 324:012049. doi:10.1088/1757-899x/324/1/012049.

Xiang, X., Li, Q., Khan, S., Khalaf, O.I. (2021). Urban water resource management for sustainable environment planning using artificial intelligence techniques. Environmental Impact Assessment Review. 86: 106515.

Yin M, Wortman Vaughan J, Wallach H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow Scotland Uk: ACM. p. 1–12. [accessed 2022 Jan 23]. https://dl.acm.org/doi/10.1145/3290605.3300509.

Zeynolabedin, A., Ghiassi, R., Norooz, R., Najib, S., Fadili, A. (2021). Evaluation of geoelectrical models efficiency for coastal seawater intrusion by applying uncertainty analysis. Journal of Hydrology. 603: 127086. doi:10.1016/j.jhydrol.2021.127086.