WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparative Analysis of Machine Learning Techniques for Assessment of Water Quality

Vishnu Kumar Singh

Geological Survey of India, Hyderabad,500068

ABSTRACT

Several machine learning technique were used for determination of water quality is last past decade such as fuzzy logic, SVM, KNN, neural network, random forest, decision tree etc but very few papers had discussed about effective machine learning technique in assessment of water quality.PH, Total hardness, (total dissolved solid), Major cation and Anion were used as input by machine learning technique to draw inferences. In this paper SVM and random forest techniques were stand out as efficient Techniques in assessment of water quality

Keywords: Machine learning (ML), SVM (Support Vector Machine). KNN (K nearest neighbour), Random forest

INTRODUCTION

Although water constitutes 71% of the earth's surface, yet only 0.3% of it is available as fresh water for human use. Moreover, it is the quality of such freshwater in ground and surface system that is of utmost concern to us as potable water needs to be appropriate in terms of calories and organic nutrient content. Because of this, the water is becoming scarce day by day as the population increases across the world. Changes in water quality due to natural processes are because of weathering of rocks; evapotranspiration; depositions due to wind; leaching from soil' runoff due to hydrological factors; and biological processes happening in the aquatic environment. The changes that occur due to natural processes bring about change in PH and alkalinity in water while at places also causing phosphorus loading, increase in fluoride content and high concentrations of sulphates. The anthropogenic factors affecting the water quality includes impacts due to agriculture; use of fertilizers, manures and pesticides; animal husbandry activities; inefficient irrigation practices; deforestation of woods ; aquaculture; pollution due to industrial effluents and domestic sewage; mining and recreational activities. The anthropogenic influence causes elevated concentrations of heavy metals, Mercury, coliforms and nutrient loads.to solve the this issue efficient techniques are required and machine learning techniques has revolutionized the every aspect of human life . in this paper we will throw light upon the some of the selected machine learning technique and their suitability in Assessment of water quality

Sampling and analytical Procedure

7 samples were collected as a part geological survey of India's annual program. Sample collection, transportation, and analysis were carried out according to standard methods and procedures (APHA, 2006). To avoid the effect of floating debris, the samples were collected at depths greater than 20 cm below the water surface (Goldscheider and Drew,2007). Prior to sample collection, the containers were washed with concentrated HNO3 and completely rinsed with distilled water. The samples were collected and packed in these plastic water bottles for further analysis at GSI Geochemical Laboratory, SR, Hyderabad

Geology of the Study Area

1. Geology of Toposheet 57O/1

The area diagonally separates into two halves' and exposes two divergent group of litho units viz. Peninsular Gneissic Complex-II (PGC-II) rocks of Archaean to Palaeo- proterozoic age and Cuddapah Supergroup of rocks of Meso-proterozoic age. The Cuddapah supergroup rocks occupy the northern half and in the southern half PGC-II rocks are exposed.The PGC-II unit is composed of hybrid granite gneiss, grey biotite granite gneiss, porphyritic grey granite, quartz veins and dolerite dykes. Cuddapah Supergroup comprises the Nallamalai group of rocks consisting of Bairenkonda/Nagari and Cumbum formations. The Bairenkonda Formation is composed of quartzite and shale/phyllites and the Cumbum Formation consists of shale and quartzite units. hybrid granite gneiss is composed of small enclaves of amphibolite, quartz mica schist and banded ferruginous quartzite. This unit of rocks are exposed all along the contact with the Cuddapah supergroup of rocks between Reddivaripalle in the northwest to Turkapalle in southeast. Generally, the rock is greenish grey in colour, coarse grained and composed of biotite, plagioclase, hornblende and quartz. Under thin section the rock consists major minerals like acid plagioclase, quartz, antiperthite, greenish biotite and sphene whereas apatite, pyrite and magnetite occur as accessory minerals. Grey biotite

granite gneiss which is the dominant member in the PGC-II unit is exposed in the rest of southern half of the area. The rock is medium to coarse grained, grey to whitish grey in colour and consists of quartz, k-feldspar, biotite, hornblende and plagioclase. Petrographical studies indicate that this rock consists of microcline, intermediate plagioclase and quartz as major minerals with minor amounts of hornblende, biotite, epidote, sphene, zircon, tourmaline and chlorite. Porphyritic grey granite is exposed as an intrusive in both the hybrid granite gneiss and grey biotite granite gneiss at one/two places in the area. The rock is exposed between Kottagollapalle and Dinnelapalle in the central part and around Amma Bhavi in the northern part of the area. The rock is in grey colour porphyritic and composed of hornblende, biotite, quartz and plagioclase. Quartz veins occur in the PGC-II unit of rocks with varying lengths. Generally, the veins color varies from white to light rose which is trending E-W and NE-SW directions. basic intrusive which occur in the PGC unit are represented by dolerite. These dykes are mainly medium to fine grained in nature and traverse broadly along E-W, NW-SE, N-S and NE-SW directions. The length of the dykes varies from few meters to more than 2 to 3 kilometers. The northern half of the area is occupied by the Bairenkonda/Nagari and Cumbum Formations. The Bairenkonda/ Nagari Formations consists of quartzite at the bottom and shale/phyllite at the top. At places shale is making a direct contact with the PGC unit of rocks. The Bairenkonda Formation occurs as an alternate sequence of quartzite and shale/phyllite. The Cumbum Formation conformably overlies the Bairenkonda/ Nagari Formation and mainly consists shale horizon which is succeeded by quartzite unit.Primary bedding is noticed in the Cuddapah Supergroup of rocks. The general strike of the bedding is NW-SE with dips 10° - 35° towards northeast.

Crudely developed foliation is noticed at some places in the PGC rocks. The general trend of the foliation is N-S and dips 50°-60° towards east. Joints have developed predominantly in the area along N-S, E-W, NNW-SSE and NNE-SSW directions with vertical to sub-vertical dips on either direction. However, E-W trending joints are more profuse in nature. NE-SW trending faults are recorded in the PGC unit of rocks of the area. Few faults are extended from few meters to few kilometers. A NE-SW trending fault is noticed in PGC unit between Dasaragudem to Venkataramana colony. Few NE-SW and NW-SE trending faults are noticed which are cutting both the PGC and Cuddapah Supergroup of rocks.

2. Geology of Toposheet 57O/5

In Toposheet No. 57O/5 The Cuddapah Supergroup comprises the Nallamalai group of rocks consisting of Bairenkonda/Nagari and Cumbum formations. The Bairenkonda Formation is predominantly an arenaceous unit consisting of fine to medium-grained, thick-bedded quartzite (orthoquartzite), dazzling white in color with a vitreous (Pearly) appearance. Shale occurs as persistent bands and shows facies variation to quartzite and vice-versa. The Cumbum Formation is essentially an argillaceous sequence comprising shale, slate, and phyllite with intercalations of quartzite, dolomite, and limestone

3. Geology of Toposheet 57J/8

the area exposes Veligallu schist belt rocks in the eastern part as two separate lenticular bands, and the rest of the area is occupied by the PGC-II rocks of the Archaean age. Both the schistose as well as PGC-II rocks are intruded by quartz veins and dolerite dykes of the Paleo-Proterozoic age. Evidence of gold mineralization is found in the area in acid volcanic rocks, meta basic rocks, and BIF.The regional trend of different litho-units of the Veligallu Schist belt varies between NNW-SSE and NNE-SSW with steep clips on either side. The metasedimentary and the meta-volcanic members show well-developed foliation and banding. Three generations of folding are reported by earlier workers (Srinivasan, B.V. et.al 1985) of which two generations are tight oppressed major folds trending NNW-SSEE to N-S. In the Tsadukonda (.3691 hill) the schist belt units are folded into tight SSE plunging synform. These are further complicated by NE-SW cross folds as noticed in Inumukonda (.3051 hill). In the northern part of the area i.e. north of the Nambulapulakunta-Galiveedu road litho-units dominantly exhibit NNW-SSE to N-S trends with fold axes generally parallel to foliation. The third generation of broad open folds is developed in the E-W direction. The E-W trending open warps/folds are clearly seen on the Nambulapulakunta- Veligallu main road which runs parallel to the fold axis. A number of shears, fractures, faults, and regional joints have developed parallel to the different fold axes mentioned above

Table 1: Descriptive statistics of Hydrogeochemical properties

	PH	*EC*	*TDS*	*TH*	*Ca*	*Mg*	*Na*	*K*	*HCO 3*	*Cl*	*SO4*	*NO3*	*PO 4*	*F*	*alkalinity*
Mean	7.40	384.57	240.69	154.29	27.48	20.85	43.94	2.50	197.01	42.54	8.14	6.57	1.14	0.42	236.41
Standard Error	0.13	111.28	65.10	25.90	5.56	3.31	14.03	0.00	43.55	8.33	2.25	1.29	0.14	0.25	52.26
Median	7.50	408.00	265.20	150.00	24.05	17.02	39.60	2.50	207.47	42.54	8.00	8.00	1.00	0.12	248.96
Standard Deviation	0.34	294.41	172.25	68.52	14.71	8.75	37.13	0.00	115.22	22.04	5.96	3.41	0.38	0.67	138.27
Range	0.90	872.20	501.93	190.00	40.08	21.89	105.20	0.00	341.71	56.72	16.00	7.00	1.00	1.91	410.05
Minimum	6.80	48.80	31.72	70.00	8.02	12.16	8.40	2.50	73.22	14.18	3.00	3.00	1.00	0.01	87.87
Maximum	7.70	921.00	533.65	260.00	48.10	34.05	113.60	2.50	414.94	70.90	19.00	10.00	2.00	1.91	497.92

SVM

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

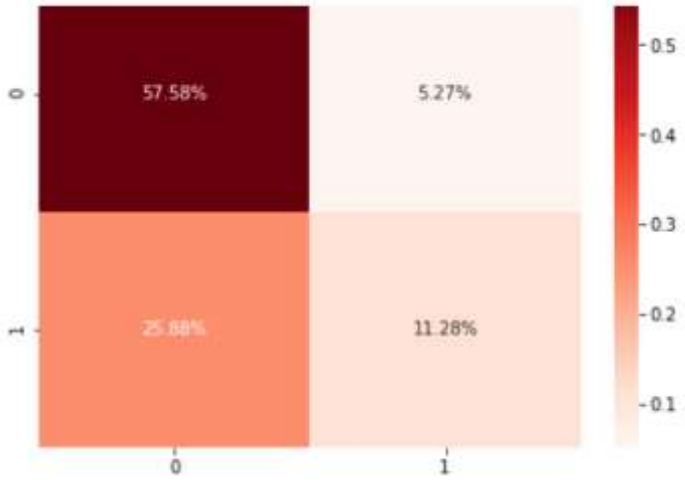
Result obtained by SVM (by use of python code)

Table 2: classification report by SVM techique

```
print(classification_report(y_test,pred_ada))
```

	precision	recall	f1-score
0	0.63	0.99	0.77
1	0.62	0.04	0.07
accuracy			0.63
macro avg	0.62	0.51	0.42
weighted avg	0.63	0.63	0.51

Fig 1: Confusion matrix generated by SVM technique



Random forest

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

the random forest algorithm in machine learning, we must look into the ensemble learning technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

Bagging

Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forest.Bagging chooses a random sample/random subset from the entire data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently, which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting, is known as aggregation.

Boosting

Boosting is one of the techniques that use the concept of ensemble learning. A boosting algorithm combines multiple simple models (also known as weak learners or base estimators) to generate the final output. It is done by building a model by using weak models in series.

Steps Involved in Random Forest Algorithm

Step 1: In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

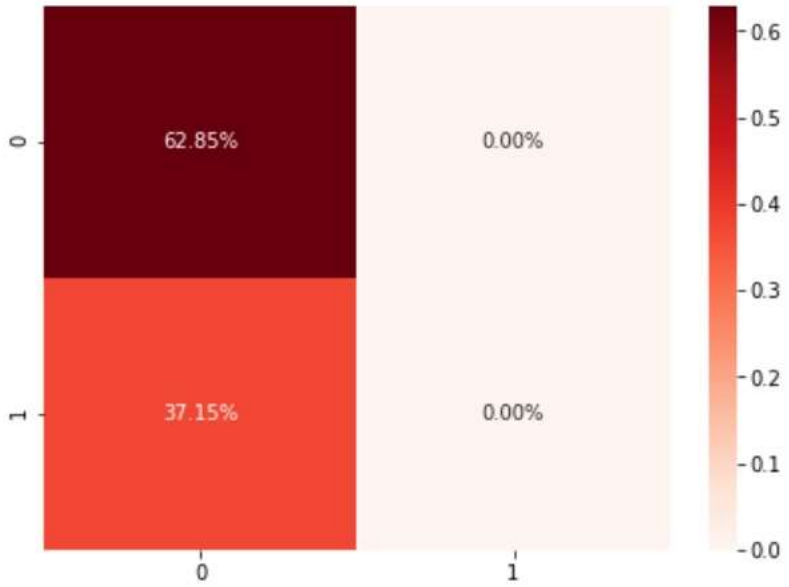
Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

Table 3: classification report by Random forest technique

```
print(classification_report(y_test,pred_rf))
```

	precision	recall	f1-score	support
0	0.63	1.00	0.77	
1	0.00	0.00	0.00	
accuracy			0.63	
macro avg	0.31	0.50	0.39	
weighted avg	0.39	0.63	0.49	

Fig 2: Confusion Matrix by Random Forest Technique



KNN (K-nearest neighbours (KNN))

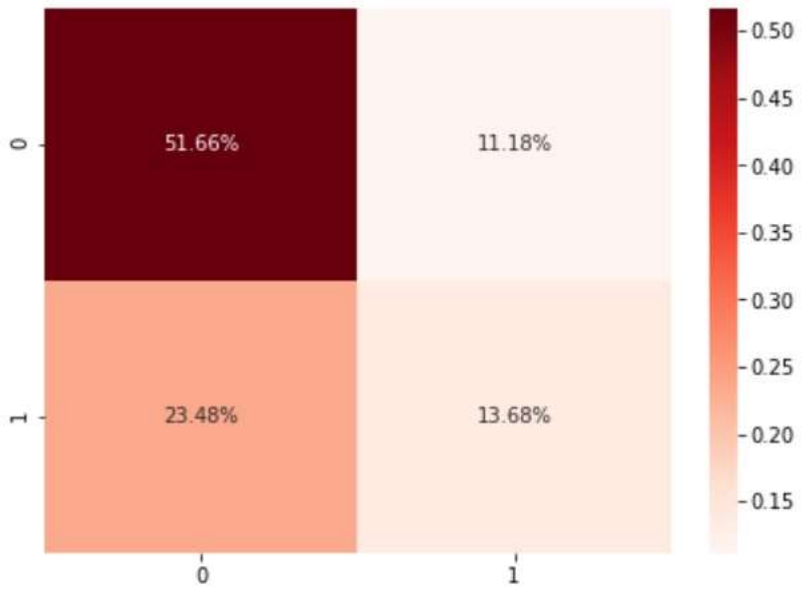
KNN is a simple, supervised machine learning (ML) algorithm that can be used for classification or regression tasks - and is also frequently used in missing value imputation. It is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points. By choosing K, the user can select the number of nearby observations to use in the algorithm.

Table 4: classification report by KNN technique

```
print(classification_report(y_test,pred_kn))
```

	precision	recall	f1-score
0	0.69	0.82	0.75
1	0.55	0.37	0.44
accuracy			0.65
macro avg	0.62	0.60	0.59
weighted avg	0.64	0.65	0.63

Fig 3: Confusion Matrix by KNN Technique



Conclusion:

From the confusion matrix and classification report it is quite clear that SVM and Random forest shows actual positive and predictive positive nearly 60 % which means both the machine learning technique are effective in quality assessment of water in comparison with KNN

References

1. Biau , Gerard (2012), “Analysis of a Random Forests Mode, The Journal of Machine Learning Research , No.1 ,pp 1063-1095.

2. Boateng TK, OpokuF, Acquaah So,Akoto O (2016), “Ground water quality assessment using statistical approach and water quality index in Ejsujuaben Municipality, Ghana”, Environmental Earth Sciences, pp 75-489.

3. Breiman, Leo (2001) , “Random Forests” Machine Learning, No. 1 , 5–32. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 6S4, April 2019 1466Published By: Blue Eyes Intelligence Engineering

& Sciences Publication Retrieval Number: F12980486S419/19©BEIESP DOI: 10.35940/ijitee. F1298.0486S419

4. Farhad Howladar, Md Abdullah, AI Numanbakth, Mohammad Omar Faruque (2017), "An application of Water quality index (WQI) and multivariate statistics to evaluate the water quality around Maddhapara Granite Mining Industrial area,Dinajpur,Bangladesh" ,Environmental System Research, Springer Open.

5. Hamilton, Howard (2012) , "Confusion Matrix", Knowledge Discovery in Databases.

6. Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer.

7. J. Lu, T. Huang (2009), "Data Mining on Forecast Raw Water Quality from Online Monitoring Station Based on Decision making Tree", Fifth International Joint Conference on INC, IMS and IDC.

8. Jorge Camejo, Osvaldo Pacheco, Miguel Guevara (2013), "Classifier for Drinking water Quality in real time", Foundation for Science and Technology, IEEE.

9. M.J. Diamantopoulou, V.Z. Antonopoulos and D.M. Papamichail (2005), "The use of Neural Network technique for the prediction of water quality parameters of Axios River in Northern Greece", EuropeanWater,11/12