Forecasting potabilityof Groundwater using Univarient and Multivarient model and its Spatial Representation

Sai Praneeth G, Dr. Shobha G, V Anantharama

Abstract— Ground water quality index is forecasted using machine learning tecnique. Water sample data for forecasting is collected from the wells of chikballapur. Two models are developed for the prediction and forecasting ground water. The first model, which is an univariate analysis model, is constructed using the technique of Artificial Neural Networks integrated with Decision tree algorithm to predict the salt content of the water sample and also to classify it as potable or not. The second model, which is multivariate analysis is constructed using Linear Regression integrated along with Pearson Coefficient, predicts the values of the parameters. Classification of that water sample, whether it is potable or not is checked using the Naïve Baye's Classifier method. In this study, the accuracy of prediction of both the models is calculated and a comparison between them is drawn, thus determining the better model. Finally, for the multivariate analysis model, the forecasted values are spatially represented in GIS maps using ARCGIS software. The model developed was tested with the water sample data collected for 20 years from centre board of ground water and it has been found that Multivariate model had an overall accuracy of 87% for forecasting and 92% for classification compared to Univariant model which had an accuracy of 85%.

Index Terms— Artificial Neaural Networks(ANN), Central Ground Water Board(CGWB),Decision Tree Algorithm, Geographical Information System(GIS), groundwater quality, Instance Distance Weighting (IDW), Linear Regression, Machine learning, Pearson Co-efficient, water level.

1 INTRODUCTION

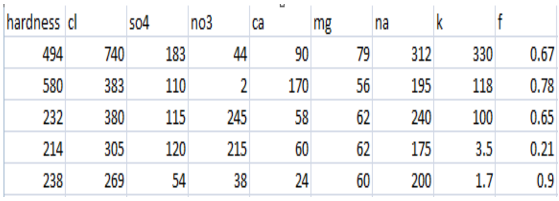
Machine learning relates on study of systems which learns from previous available data by training this data into the system. Machine learning concentrates on prediction which establishes on experienced properties which learns from the training data in the system[5].

To determine potability of ground water samples and its prediction has vast scope in today's market. It can be used by the government to keep track of various viable ground water sources in the state. It can be used by the pollution control board to keep track of the levels of pollution in bore wells that contains water that is not fit for consumption. Hence the scope of usage of this project sums to be limitless. Ground water is ever changing and its quality can never be accurately predicted. Most environment phenomenon is cyclic in nature. They follow a particular cycle and their compositions may change according to that cycle. These cycles usually have a long time period and are hard to recognize. Various conventional methods for prediction and forecasting of water quality, such as gray system theory, neural network etc., depend on chaos theory. Most environment phenomenon is cyclic in nature. They follow a particular cycle and their compositions may change according to that cycle. These cycles usually have a long time period and are hard to recognize.

To determine the potability of ground water is very difficult task. It is significant issue to predict water quality which increases economic efficiency as a result. Nevertheless, prediction of water quality has been complicated issue due to complexity and diversity. The data affects the quality of water and accuracy to forecast. Meantime, the models that are constructed on predicted accuracy in monitoring the data is very hard[4].

2 METHODOLOGY FOR THE UNIVARIATE MODEL

The developed model is an univariate model [1],[2], which is constructed using ANN methods integrated with Decision Tree Algorithm for the prediction of the parameter values and to classify if that water sample is potable or not [6],[7]. In this model, nine parameters are considered for the classification of potability. The parameters considered are Alkalinity, Total hardness, Calcium, Chloride, Magnesium, Sulphate, Fluorine ,Sodium, Potassium. These parameters' values have been collected from CGWB and analysed[3]. This analysed data is the input to this model. The system is trained using the values from the training data set.Water Quality index is calculated. The result that is obtained is compared with the WHO standards and then the water sample is classified as potable or not. Then the accuracy of the system is determined using the test data set. The training data set is shown in fig 1.



hardness	cl	so4	no3	ca	mg	na	k	f	
494	740	183	44	90	79	312	330	0.67	
580	383	110	2	170	56	195	118	0.78	
232	380	115	245	58	62	240	100	0.65	
214	305	120	215	60	62	175	3.5	0.21	
238	269	54	38	24	60	200	1.7	0.9	

Fig. 1. Training data set

2.1 ANN model

In this model, two years'data is used for training and the parameters are fed as input in the input layer, then hidden layer equation is constructed, thus determining the water quality index[4] and hence predicting the water quality index for subsequent year[9].

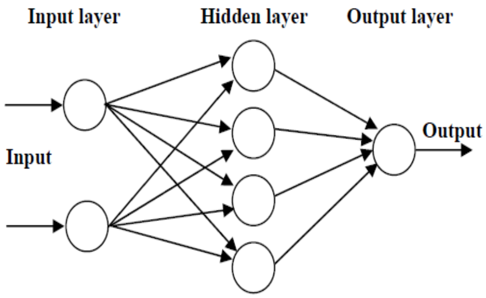


Fig. 2. ANN model

2.2 Classification

The Decision tree algorithm is used for classification[9]. The index values that are obtained using the ANN in fig 2, are inputted into decision tree. The output of this tree will determine if the water sample is potable or not[10].

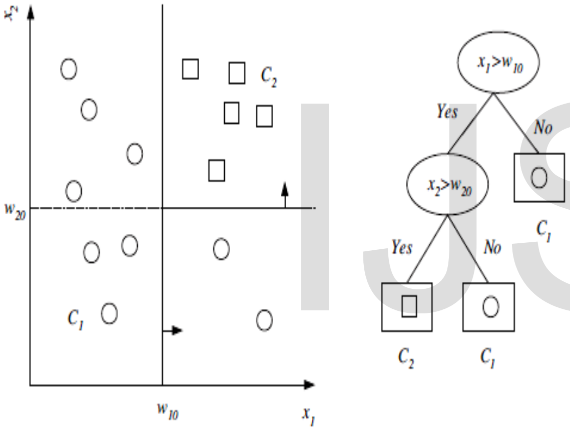


Fig .3. Decision tree

The index values for different classifications of the sample is as follows[2]:

Table 1. Ground water Quality index

GWQI	Class
0-25	Excellent
26-50	Good
51-75	Moderate
76-100	Poor
101-125	Very poor
125 and above	Unfit

The main task in decision tree lies in selecting the root nodes. The root node is decided by the parameter called information gain, which inturn depends on the entropy. To find these parameters we should know the propertion of positive and negative samples in the training data.

For a collection S, with the positive proportion p1, and negatie proportion p2, its entropy is calculated as,

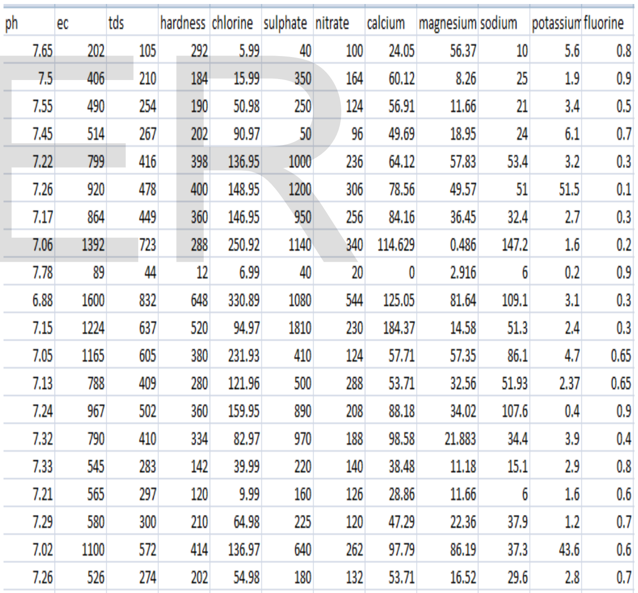
$$entropy(S) = -p1log_2p2 - p2log_2p1 \quad (1)$$

The information gain of each attribute A is calculated as,

$$Gain(S, A) = Ent(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Ent(S_v) \quad (2)$$

3 Methodology for Mutlivariate Model

Our model is an multivariate analysis model, which is constructed using Linear Regression technique improved with Pearson's co-efficient model for the prediction of the parameter values. Classification of that water sample, whether it is potable or not is checked using the Naïve Baye's Classifier method[8],[11]. In this model, twelve parameters are considered for the classification of potability. The parameters considered are Nitrate, Total hardness, Calcium, Chloride, Magnesium, Sulphate, Fluorine,Sodium, Potassium.These parameters' values have been collected from Central Ground Water Board (CGWB) and analysed[3]. This analysed data is the input to this model. The system is trained using the values from the training data set. The result that is obtained is compared with the WHO standards and then the water sample is classified as potable or not. Then the accuracy and the efficiency of the system is determined using the test data set.The training data set is shown in fig 4.



ph	ec	tds	hardness	chlorine	sulphate	nitrate	calcium	magnesium	sodium	potassium	fluorine
7.65	202	105	292	5.99	40	100	24.05	56.37	10	5.6	0.8
7.5	406	210	184	15.99	350	164	60.12	8.26	25	1.9	0.9
7.55	490	254	190	50.98	250	124	56.91	11.66	21	3.4	0.5
7.45	514	267	202	90.97	50	96	49.69	18.95	24	6.1	0.7
7.22	799	416	398	136.95	1000	236	64.12	57.83	53.4	3.2	0.3
7.26	920	478	400	148.95	1200	306	78.56	49.57	51	51.5	0.1
7.17	864	449	360	146.95	950	256	84.16	36.45	32.4	2.7	0.3
7.06	1392	723	288	250.92	1140	340	114.629	0.486	147.2	1.6	0.2
7.78	89	44	12	6.99	40	20	0	2.916	6	0.2	0.9
6.88	1600	832	648	330.89	1080	544	125.05	81.64	109.1	3.1	0.3
7.15	1224	637	520	94.97	1810	230	184.37	14.58	51.3	2.4	0.3
7.05	1165	605	380	231.93	410	124	57.71	57.35	86.1	4.7	0.65
7.13	788	409	280	121.96	500	288	53.71	32.56	51.93	2.37	0.65
7.24	967	502	360	159.95	890	208	88.18	34.02	107.6	0.4	0.9
7.32	790	410	334	82.97	970	188	98.58	21.883	34.4	3.9	0.4
7.33	545	283	142	39.99	220	140	38.48	11.18	15.1	2.9	0.8
7.21	565	297	120	9.99	160	126	28.86	11.66	6	1.6	0.6
7.29	580	300	210	64.98	225	120	47.29	22.36	37.9	1.2	0.7
7.02	1100	572	414	136.97	640	262	97.79	86.19	37.3	43.6	0.6
7.26	526	274	202	54.98	180	132	53.71	16.52	29.6	2.8	0.7

Fig.4. Training data set

3.1 Analysis of the data set

The training data set is as shown in fig 5. Here, 20 years of data for a particular well is considered from 1995-2015. The data set is analyzed and it is found that the variation of the parameters over the years considered are inconsistent and hence the method of linear regression is considered. The correlation among the parameters are derived using the Pearson Co-efficient formula.

3.2 Pearson Co-efficient

The Pearson Co-efficient is calculated using the formula :

$$r = \frac{\sum(X - X')(Y - Y')}{\sqrt{\sum(X - X')^2 \sqrt{\sum(Y - Y')^2}}} \quad (3)$$

Where, r is the PearsonCo-efficient. X and Y are the parameters.

The Pearson Co-efficient value 'r' can range from -1 to +1, where, -1 indicates no correlation and +1 indicates maximum correlation.

In this model, the Pearson Coefficient is used to derive the correlation between the parameters. For each parameter, the other parameter[8], which has the highest correlation with the parameter in consideration, is determined.

Example: For the parameter Ph, fluorine has the highest Correlation with Pearson Co-efficient being equal to 0.82. Here, X is fluorine and Y is pH.

3.3. Linear Regression

The formula used in developing the model is:

$$h_\theta(x) = \theta^T.x \qquad (4)$$

$$J(\theta) = \frac{1}{m}\sum_{i=0}^{m}(h_\theta(x) - y)^2 \qquad (5)$$

$$\theta_i = \theta_j - \alpha.\frac{1}{m}\frac{\delta}{\delta_j}J(\theta) \qquad (6)$$

Where 'h' is the hypothesis function.

'J' is the cost function

'y' is the actual output values.

The final formula is the gradient descent formula.

This method used as it is one the most efficient method due to the inconsistency in the data[12]. This is because the error factor is considered, and in this technique the difference between the predicted values(hypothesis function) and actual values is considered for every year's data and this error decreases for more number of data samples.

For the training of data, the process of linear regression is applied on the parameters that are obtained using the Pearson Co-efficient.

For the forecasting of the values of the parameters for the (N+1)th year, the previous N years(20) data are trained, thus using the linear regression technique as a machine learning concept. In this model 20 training data sets are considered. The predicted values are compared with the WHO standards and then classified as potable or not.

The WHO standards for the parameters are:[14]

Table 2. WHO standards for the parameters.

Water Quality Parameters	WHO Standards (Si)
pH	8.5
Ec	2000
TDS	1000
Total Hardness	300
Cl	250
So4	250
No3	50
Ca	75
Mg	30
Na	200
K	12
F	1.5
Total	**4177**

3.4 Classification

Naves bayes classifier is the classification algorithm based on the baye's theorem of probability. Given the set of attributes, we calculate the posterior probability of the event using the baye's theorem[11]. Posterior probability that water is p1otable is,

$$Posterior(potable) = P(potable)P(potable|ph)P(potable|ec)\ .../evidence, \qquad (7)$$

Evidence is the sum of numerators in Posterior(potable) and Posterior(non-potable).

Given the Standard deviation σ, and mean μ, of the parameter p, P(potable | p) is calculated as,

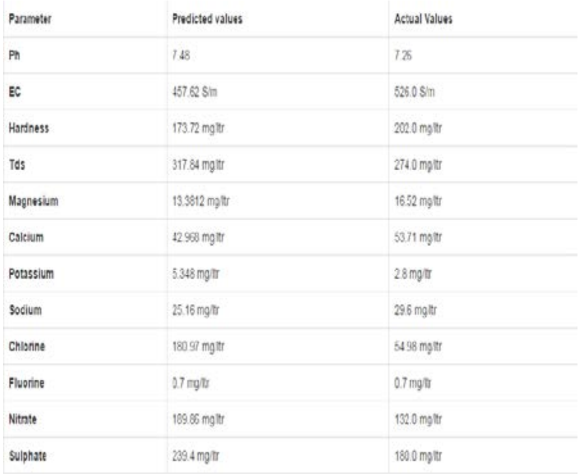
$$p(potable \mid p) = \frac{1}{\sqrt{2\pi\sigma\char`^2}}\char`^(\frac{-(val-\mu)\char`^2}{2\sigma\char`^2}) \qquad (8)$$

Similarly the posterior probability of not potable is also calculated, i.e. posterior(notpotable).

Finally the water is classified potable or not based on the posterior probability. If the posterior probability of potable is greater, then water is potable otherwise it is not.

3.5 Experimental Analysis

During the experimental analysis, the result of the year 2015 is considered. The actual values are collected from the CGWB. The predicted values are compared with the actual values and an accuracy of forecasting the values is 87%. Naïve Baye's classifier is used for classification and an accuracy of 92% is obtained. The test data set is as shown in Fig 5.



Parameter	Predicted values	Actual Values
Ph	7.48	7.26
EC	457.62 S/m	526.0 S/m
Hardness	173.72 mg/ltr	202.0 mg/ltr
Tds	317.84 mg/ltr	274.0 mg/ltr
Magnesium	13.3812 mg/ltr	16.52 mg/ltr
Calcium	42.968 mg/ltr	53.71 mg/ltr
Potassium	5.348 mg/ltr	2.8 mg/ltr
Sodium	25.16 mg/ltr	29.6 mg/ltr
Chlorine	180.97 mg/ltr	54.98 mg/ltr
Fluorine	0.7 mg/ltr	0.7 mg/ltr
Nitrate	189.86 mg/ltr	132.0 mg/ltr
Sulphate	239.4 mg/ltr	180.0 mg/ltr

Fig. 5. Testing data set

The accuracy obtained in univariate model that was developed using ANN and decision tree was found to be 88.87% for the classification.

3.6 Spatial Representation

For representing the data in spatial form (ie. GIS maps), the values of three wells are considered. The data is obtained from CGWB. The spatial representation of data on the GIS maps are obtained using the technique of IDW[13]. The maps are generated using the ARCGIS 10.0 software. The maps are as shown in Fig 6 & Fig 7.

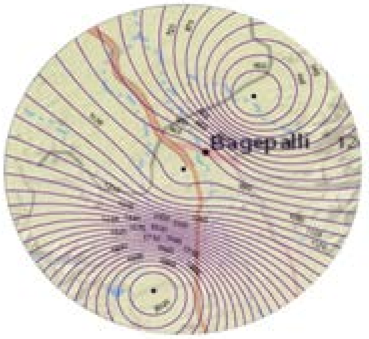


Fig. 6. GIS map for EC, for 3 wells in Bagepalli taluk

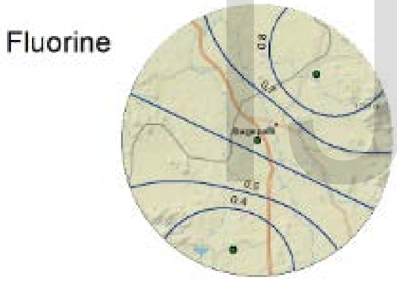


Fig. 7. GIS map for fluorine, for 3 wells in Bagepalli taluk

4 CONCLUSION

The conclusion hence derived is that, the univariate model is developed using the methods of ANN for forecasting the parameters' values and decision tree algorithm for classifying as potable or not. In this model, the parameters are individually considered and forecasted. The effect of one parameter on the other is not taken into account. On the other hand, the multivariate analysis model is developed using Linear Regression method improved with Pearson Co-efficient for forecasting the parameters' values. Then the classification of the water sample is determined using Naïve Baye's classifier method. In this multivariate model, one parameter that affects the parameter in consideration is taken into account. Also the quality of ground water is determined at different levels.

In the univariate model that uses artificial neural network and decision tree algorithm, the accuracy of the model for prediction and for classification was found to be 88.87%. In the multivariate model that uses Linear Regression improved with Pearson Co-efficient. Naïve Baye's Classifier method is used for classification. The accuracy is found to be 87% for forecasting and 92% for classification.

REFERENCES

[1] Shoba G,Shobha G, "Rainfall Prediction Using Data Mining Techniques", *International Journal of Engg and Computer Science*, ISSN: 2319-7242, Volume 3,Issue 5 May, 2014, PP: 6206-6211.

[2] Shoba G,Shobha G, " Water Quality Prediction Using Data Mining techniques: A Survey", *International Journal Of Engineering And Computer Science* , ISSN: 2319-7242 Volume 3 Issue 6 June, 2014 Page No. 6299-6306.

[3] Ground Water Information Booklet Chickballapur District, Central Ground Water Board, Karnataka, Bangalore, August 2014.

[4] Ali and Qamar, "Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan", *In proceedings of Digital Information management (ICDIM),* Islamabad, January 2014, pp. 108-113.

[5] Vahid Nourani, Tohid Rezapour Khanghah and Milad Sayyadi, "Application of the Artificial Neural Network to monitor the quality of treated water", *International Journal of Management and Information Technology,* Vol. 3, no 1, January 2013, pp. 94-96.

[6] Rahimi, Mokarram, "Assessing the groundwater quality by applying fuzzy logic in GIS environment- A case study in Southwest Iran", *International journal of Environmental Sciences,* February 2012, pp. 153-158.

[7] V. Kumar, N. S. Mathew and G. Swaminathan, "Fuzzy Logic and GIS based Information Processing for as Assessment of Groundwater Quality," *Journal of Geographic Information System*, July 2010, pp. 152-162.

[8] Soo-Yeon Ji, Sharad Sharma , Byunggu Yu and Dong Hyun Jeong, "Designing a Rule-Based Water Quality Model", *IEEE IRI*, August 2012, pp. 8-19.

[9] Hao Liao, Wen Sun. Forecasting and Evaluating Water Quality of Chao Lake based on Improved Decision Tree Method, 2, October 2010, pp. 970-979.

[10] Jinsuo Lu and Tinglin Huang, "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*, June 2009, pp. 233-242.

[11] James N.K.Liu, Bavy N.L.Li, and Tharam S.Dillon, "An Improved Naïve Bayesian Classifier Technique Coupled with a Novel Input Solution Method", *IEEE Transactions on systems, man, and Cybernetics – Part C: Applications and Reviews,* Vol.31, No.2, May 2001, pp.167-176.

[12] G. Shobha , Jayavardhana Gubbi, Krishna S Raghavan, Lakshmikanth K Kaushik, M. Palaniswami, "A novel fuzzy rule based system for assessment of ground water potability: A case study in South India",*IOSR Journal of Computer Engineering (IOSR-JCE)* e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 2 Nov. 2013, pp. 35-41.

[13] Muzafar N. Teli, Nisar A. Kuchhay, Manzoor A. Rather , " Spatial Interpolation Technique For Groundwater Quality Assessment Of District Anantnag J&K", *International Journal of Engineering Research and Development* e-ISSN: 2278-067X, p-ISSN: 2278-800X, Volume 10, Issue 3, March 2014, pp.55-66.

[14] Dehghani, Mohammad Hadi, et al. "Microbiological Quality of Drinking Water in Shadegan Township, Iran." *Iran J Energy Envi-*

ron 2.3, 2011, pp. 286-290.

Authors Profile:

Sai Praneeth G – Currently pursuing B.E course in Computer Science and Engineering (8th semester), R V College of Engineering, Bangalore; placed in SPRINKLR Systems through campus selection; won the best project award for Database Management System (DBMS) projecttitled Agricultural Management System in 6th semester.

Dr Shobha G – Ph.D from Mangalore University; currently HOD, Department of Computer Science and Engineering, R V College of Engineering, Bangalore; awards – career award for best teacher by AICTE, excellence award in education from RSST; Won best paper award for paper titled "Egocentric Network Analysis to score the Telecom Customers", IEEE International Conference on Computational Intelligence and Computing Research(ICCIC), Madurai, India; Won best paper award for paper titled " Centrality Measures for Predicting churn in Telecom Social Networks", International Conference on Advanced Computer Science and Information Technologies (ICACSIT), December 2nd 2013, Pune, India.

Prof V Anantharama – M Tech in Con Engineering and Management, Ph.D from Kuvempu University; currently, Associate professor, Civil Engineering, R V College of Engineering, achievements - ISRO/DOS sanctioned Rs 26 Lakhs towards setting up of Geoinformatics lab under the co-ordination of Sri. Anantha Rama.V (PI) & Dr Prakash. P, (Co-PI) Asst.Professors.Dept.of Civil Engg, RVCE.