Ground Water Quality Prediction using Machine Learning Algorithms in R

[1] S.Vijay & [2]Dr.K.Kamaraj
[1]Assistant Professor, [2]Assistant Professor
[1] Department of Computer Science, [2]Department of Computer Science
[1]Vivekanandha College of Arts and Sciences for Women, Tiruchengode, Namakkal Dt
[2]SSM College of Arts and Science, Komarapalayam, Namakkal.

Received: January 25, 2019 Accepted: March 01, 2019

ABSTRACT: *Water plays a dominant role in the growth of the country's economy and essential for all the activities. The present study deals with the physico-chemical characteristics of ground water quality in Ranipet, Arcot, Walljah pet, towns in vellore district. Such a water samples were collected from different identified bore wells for the purpose of studying the quality of groundwater . The bore wells from which the samples were collected are extensively used for drinking purpose. The water quality parameters such as PH, TDS, EC, Chloride, Sulphate, Nitrate, Carbonate, Bicarbonate, metal ions, trace elements have been estimated. There are two major classifications like High , Low level of water contamination observed in Vellore district. This paper focus on predicting water quality by using Machine Learning classifier algorithm C5.0, Naïve Bayes and Random forest as leaner for water quality prediction with high accuracy and efficiency.*

Key Words: *: Groundwater quality, C5.0, Naïve Bayes, Machine Learning*

I. Introduction

Data Mining is an emerging technique for extracting important and useful information from large sets of data. The ultimate goal of knowledge discovery and data mining is to find the patterns that are hidden among the huge sets of data and interpret them to useful knowledge and information. This information is used to improve the efficiency of the system. Data Mining contributes a lot of benefits to business, scientific and medical research . In order to identify particular pattern from the large data sets, an application is developed by using specific computerized algorithm in the domain of Data Mining. Given large data sets, prediction of new sets of data are developed using learning concept by this model through training and testing. The classification in data mining process is predicting the value of a target variable by generating a model based on some attributes categorical variable. By this process, classification of a given data is based on class labels and training. Data mining is a process of discovering previously unknown patterns that are used for strategic decision making. There are different stages:

- Data Collection
- Data Cleaning
- Data Transformation
- Application of Data mining algorithms
- Model construction and pattern evaluation

Knowledge gain used for decision making Water Quality prediction is an important environmental problem. This paper proposes an idea to develop efficient model to improve the efficiency and accuracy of water quality prediction using popular Machine Learning algorithms such as C5.0, Naïve Bayes and Random forest.

This paper proposes an idea to develop efficient model to improve the efficiency and accuracy of water quality prediction using popular Machine Learning algorithms such as C5.0, Naïve Bayes and Random forest .This paper is organized as follows. Section 2 represents outline of the work. Section 3 Literature Review. Section 4 Description of the study area. Section 5 Materials and Methods Section 6 Section Result and Discussion 7 Conclusion and future work.

II. OUTLINE OF THE WORK

The outline of this work describes the overview of the proposed work that includes data collection. At the beginning stage, the collected data set is preprocessed and models are generated by using Machine

Learning Classifier algorithms. At the final stage, validating the model is done by comparing the result of proposed algorithm with existing. The overview of the proposed work is illustrated in Figure 1

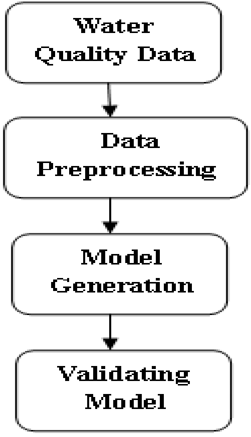


Figure 1: Overview of the proposed work

III. LITERATURE REVIEW

Sundarambal Palani et.al [1] proposed ANN models to predict water quality parameters whereas salinity, temperature, dissolved oxygen and Chl-a concentrations using continuous weekly measurements at different locations. It has been observed that the GRNN and Ward Net architecture shows the best performance based on different water quality variables. Depending on their performance, Ward Net is the superior architecture for the temperature and salinity models, but the GRNN is superior for DO and Chl-a models. Wen-Heun Chine et.al [2] proposed ANN model with back propogation algorithm which represents a non-linear relationship toconclude and predict the total nutrient concentration in reservoir in Taiwan. The BPNN accesses the concluded results via a complex structure, but does not able to express the relationships by well-defined precise and explicit functions.

Changjun Zhu et.al [3] proposed fuzzy neural network(FNN) model to evaluate and classify outer water quality in suzhou. The FNN model is reliable and effective and can deal with the problem of solitary elements which reflects the water quality at current stage. Therefore, this methodology is not convenient for the assessment of river water quality.

Yafra Khan et.al [9] has developed a water quality forecast model using the support of water quality components applying Artificial Neural Network (ANN) and time-series analysis with ANN-NAR. The performance measures such as Regression, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) indicated the best prediction accuracy results with ANN-NAR time series algorithm. Chadaphim Photphanloet et.al [10] proposed an α − trimmed ARIMA model which is often practiced to calculate the BOD value of the up-coming year making use of assortment of BOD data from the past. The accuracy of BOD prediction attained from the proposed α -trimmed ARIMA model is greater than 70% and the results are better than the smoothing method.

S.Wechmongkhonkon et.al [13] has developed a MLP neural network using the Levenberg-Marquardt algorithm is employed to analyze and distribute the water quality of Dusit district canals of Bangkok,Thailand. MLP results with a very high accuracy with the help of which cost and time is minimized.

IV. DESCRIPTION OF THE STUDY

The study area lies between Latitude N 12°52'30'' – 12°57'30" and Longitude E 79°15'00''– 79°25'00'' is located in North of TamilNadu in India, covering about 154.52 Sq. Km area The area includes Ranipet, Walajapet, and Arcot. The drainage of the study area is mainly Palar River and Ponnai River. The Ranipet area is a chronic polluted area and one of the biggest exporting centers of tanned leather. Many small-scale tanneries are processing leather in the study area and discharging their effluents on the open land and surrounding water bodies.

V. MATERIALS AND METHODS

Groundwater samples were collected from 30 Locations within study area during month of May 2014, Sampling is done at each station in polythene bottles of two-litre capacity. The samples were analyzed various water quality parameters such as pH, electrical conductivity (EC), Total Dissolved Solids (TDS), Alkalinity, Total Hardness (TH), Chloride, Biological Oxygen Demand, Chemical Oxygen Demand, sulphate, Nitrate, iron, calcium and magnesium were determined using standard method.[8] The method used for estimation of various Physico-chemical parameters are shown in Table-1. Reagent used for the present investigation was A.R. Grade and double distilled water was used for preparing various solutions. Methods used for estimation of various Physico-chemical parameters are shown in Table 1.

Table-1 Physico - Chemical Parameters

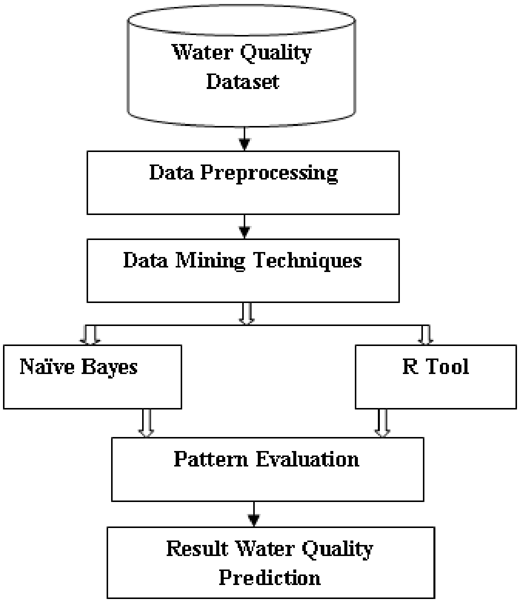
S.No	Parameter	Methods
1	pH	pH Meter
2	Electrical conductivity	Conductivity meter
3	Total Hardness	EDTA Titration
4	TDS	Filtration method
5	Alkalinity	Indicator method
6	Chloride	Argentometric method
7	Nitrate	Phenol disulphonic acid method
8	Sulphate	Nephelometry Method
9	Fluoride	SPADN spectrophotometric method
10	Calcium	EDTA titration
11	Magnesium	EDTA Titration
12	Iron	PHENANTHROLINE Spectrometry
13	COD	Open reflux method
14	BOD	Winkler"s method

5.1 Machine Learning Algorithm:

5.2 Tool Used

RStudio was founded by JJ Allaire, creator of the programming language cold fusion. R is the leading tool for statistics, data analysis and machine learning. It makes statistical computing easy and the programming effort is reduced. The graphs are easy to plot and depict. It is more than a statistical package: it is a programming language so it is possible to create own objects, functions and packages. It is a platform independent so it can be used on any operating system and open source. R programs explicitly document the steps of analysis and make it easy to reproduce and /or update analysis which means it can quickly try many ideas to correct issues. R is used in data preprocessing, data visualization, predictive analysis, statistical modeling and deployment.

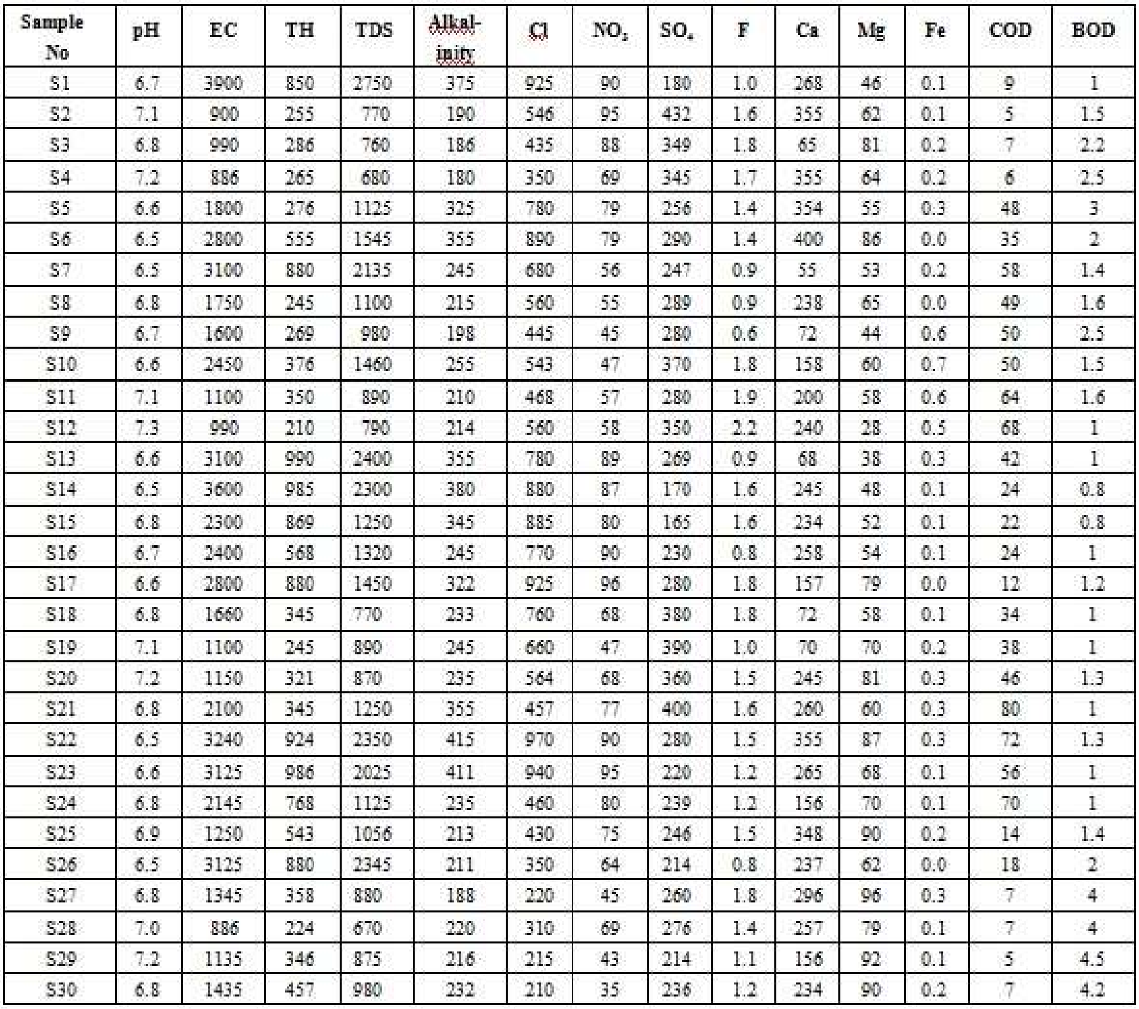
5.3 Flowchart of Proposed Work



VI. RESULT AND DISCUSSION

Results and analysis are presented in Table-2, and Table-3 compared with the permissible drinking water standards specified by WHO Standard Specification as per 2011, and the number of samples exceeding the limits parameter wise and their values are given.

Table -2 Physico-chemical parameter of ground water during month of May 2014.



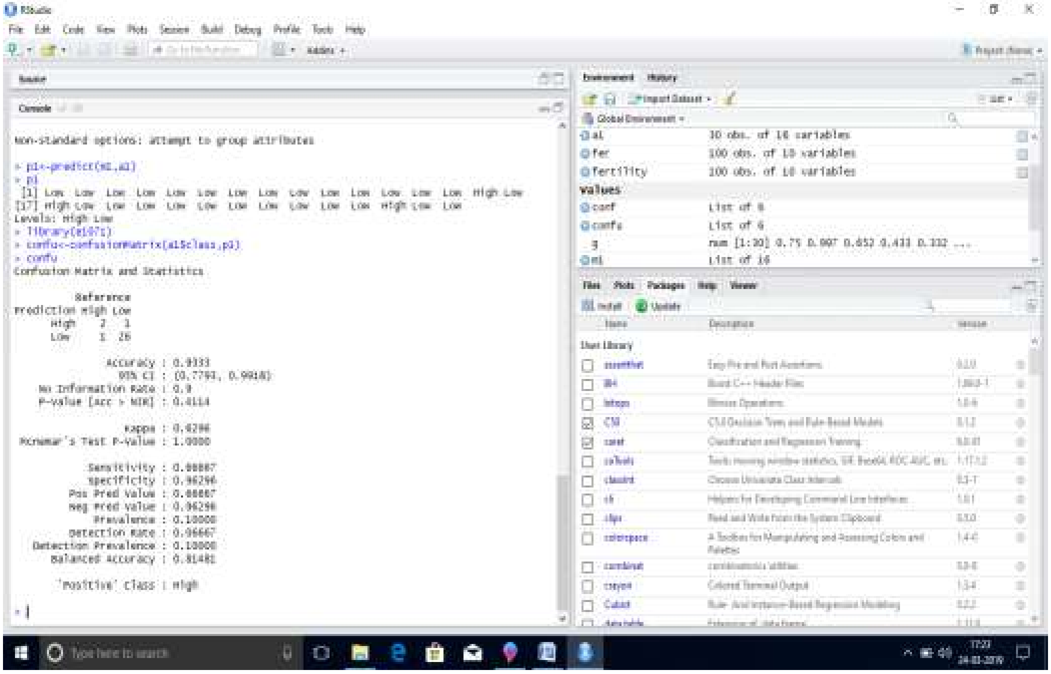
Sample No	pH	EC	TH	TDS	Alkal-inity	Cl	NO_3	SO_4	F	Ca	Mg	Fe	COD	BOD
S1	6.7	3900	850	2750	375	925	90	180	1.0	268	46	0.1	9	1
S2	7.1	900	255	770	190	546	95	432	1.6	355	62	0.1	5	1.5
S3	6.8	990	286	760	186	435	88	349	1.8	65	81	0.2	7	2.2
S4	7.2	886	265	680	180	350	69	345	1.7	355	64	0.2	6	2.5
S5	6.6	1800	276	1125	325	780	79	256	1.4	354	55	0.3	48	3
S6	6.5	2800	555	1545	355	890	79	290	1.4	400	86	0.0	35	2
S7	6.5	3100	880	2135	245	680	56	247	0.9	55	53	0.2	58	1.4
S8	6.8	1750	245	1100	215	560	55	289	0.9	238	65	0.0	49	1.6
S9	6.7	1600	269	980	198	445	45	280	0.6	72	44	0.6	50	2.5
S10	6.6	2450	376	1460	255	543	47	370	1.8	158	60	0.7	50	1.5
S11	7.1	1100	350	890	210	468	57	280	1.9	200	58	0.6	64	1.6
S12	7.3	990	210	790	214	560	58	350	2.2	240	28	0.5	68	1
S13	6.6	3100	990	2400	355	780	89	269	0.9	68	38	0.3	42	1
S14	6.5	3600	985	2300	380	880	87	170	1.6	245	48	0.1	24	0.8
S15	6.8	2300	869	1250	345	885	80	165	1.6	234	52	0.1	22	0.8
S16	6.7	2400	568	1320	245	770	90	230	0.8	258	54	0.1	24	1
S17	6.6	2800	880	1450	322	925	96	280	1.8	157	79	0.0	12	1.2
S18	6.8	1660	345	770	233	760	68	380	1.8	72	58	0.1	34	1
S19	7.1	1100	245	890	245	660	47	390	1.0	70	70	0.2	38	1
S20	7.2	1150	321	870	235	564	68	360	1.5	245	81	0.3	46	1.3
S21	6.8	2100	345	1250	355	457	77	400	1.6	260	60	0.3	80	1
S22	6.5	3240	924	2350	415	970	90	280	1.5	355	87	0.3	72	1.3
S23	6.6	3125	986	2025	411	940	95	220	1.2	265	68	0.1	56	1
S24	6.8	2145	768	1125	235	460	80	239	1.2	156	70	0.1	70	1
S25	6.9	1250	543	1056	213	430	75	246	1.5	348	90	0.2	14	1.4
S26	6.5	3125	880	2345	211	350	64	214	0.8	237	62	0.0	18	2
S27	6.8	1345	358	880	188	220	45	260	1.8	296	96	0.3	7	4
S28	7.0	886	224	670	220	310	69	276	1.4	257	79	0.1	7	4
S29	7.2	1135	346	875	216	215	43	214	1.1	156	92	0.1	5	4.5
S30	6.8	1435	457	980	232	210	35	236	1.2	234	90	0.2	7	4.2

Table-3 Results of water analyzed in comparison with WHO standards

	Permissible limit as per,	**Concentration Observed**		**No of samples exceeding**	**Percentage**
Parameters	**WHO 2011**	**Minimum**	**Maximum**	**permissible limit**	**%**
pH	7.0-8.5	6.5	7.3	22	73.3
EC	1000	886	3900	25	83.3
Total Hardness	300	210	990	21	70
TDS	1000	670	2750	17	57
Alkalinity	200	180	415	25	83.3
Chloride	250	210	970	27	90
Nitrate	45	35	96	26	87
Suphate	200	165	432	27	90
Fluoride	1.5	0.6	2.2	12	40
Calcium	75	55	400	24	80
Magnesium	50	28	96	25	83.3

Iron	0.3	0	0.7	4	13
COD	10	5	80	22	73.3
BOD	5	0.8	4.5	Nil	-

C5.0, Naïve Bayes and Random Forest are the classification methods used for water quality data analysis in this paper. Two groups are separated from the data set for training and testing the algorithms of classification. R Tool is used to implement the classification algorithms.

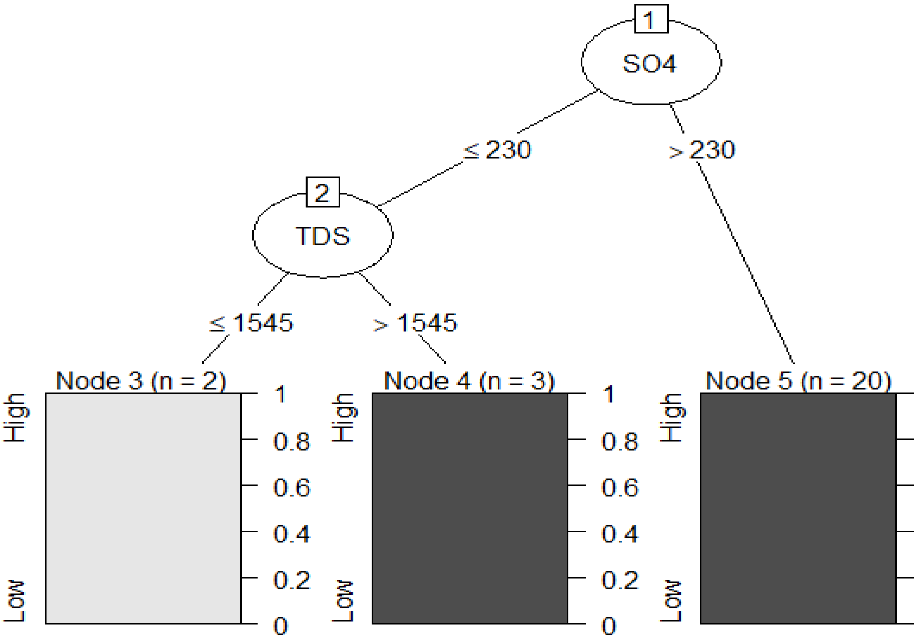


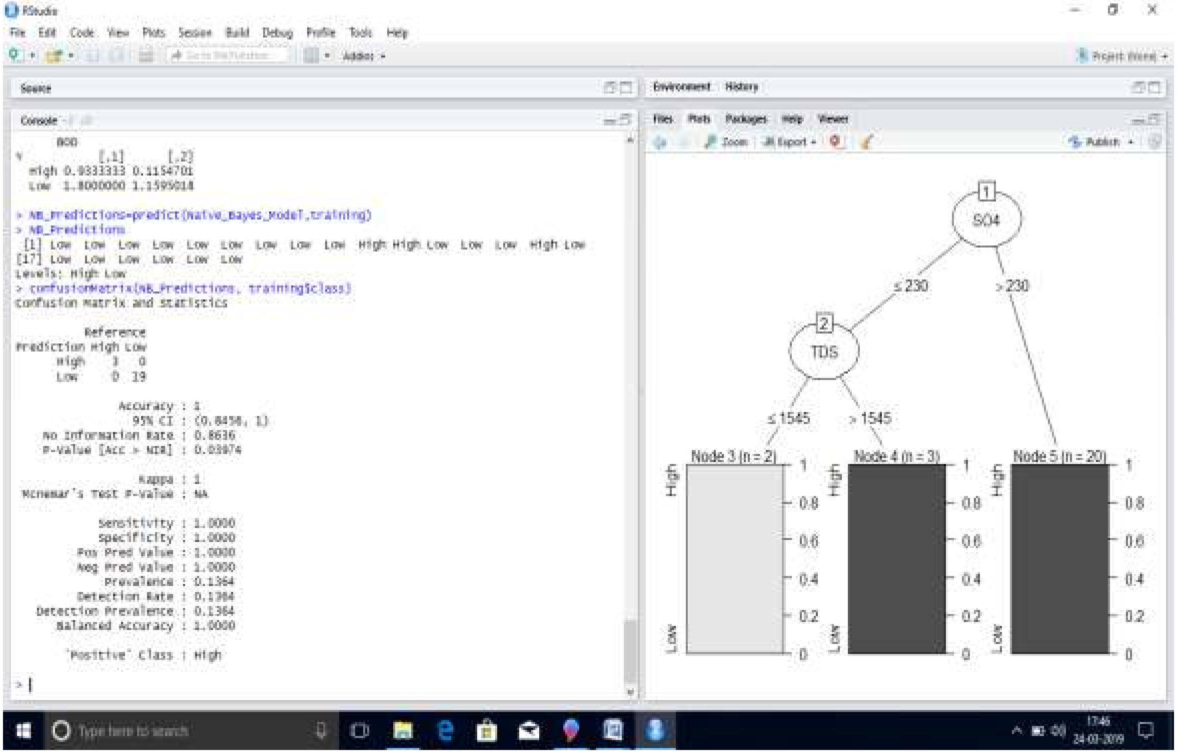
```
Non-standard options: attempt to group attributes

> p1<-predict(m1,a1)
> p1
 [1] Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  High Low
[17] High Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  High Low  Low
Levels: High Low
> library(e1071)
> confu<-confusionMatrix(a1$class,p1)
> confu
Confusion Matrix and Statistics

          Reference
Prediction High Low
      High    2   1
      Low     1  26

               Accuracy : 0.9333
                 95% CI : (0.7793, 0.9918)
    No Information Rate : 0.9
    P-Value [Acc > NIR] : 0.4114

                  Kappa : 0.6296
 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.66667
            Specificity : 0.96296
         Pos Pred Value : 0.66667
         Neg Pred Value : 0.96296
             Prevalence : 0.10000
         Detection Rate : 0.06667
   Detection Prevalence : 0.10000
      Balanced Accuracy : 0.81481

       'Positive' Class : High

> |
```

```
a1           30 obs. of 16 variables
fer          100 obs. of 10 variables
fertility    100 obs. of 10 variables
Values
conf         List of 6
confu        List of 6
g            num [1:30] 0.75 0.997 0.652 0.411 0.332 ...
m1           List of 16
```

C5.0 Classifier Result

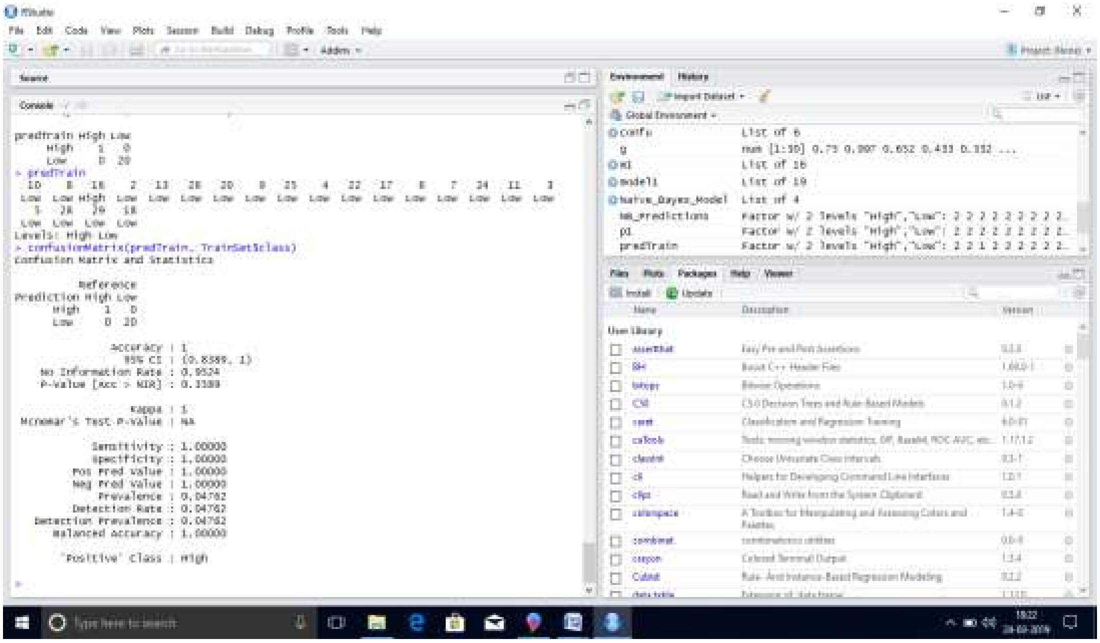
Visualization tree of C5.0





```
     BOD
Y         [,1]      [,2]
  High 0.9333333 0.1154701
  Low  1.8000000 1.1595018

> NB_Predictions=predict(Naive_Bayes_Model,training)
> NB_Predictions
 [1] Low  Low  Low  Low  Low  Low  Low  Low  Low  High High Low  Low  Low  High Low 
[17] Low  Low  Low  Low  Low  Low 
Levels: High Low
> confusionMatrix(NB_Predictions, training$class)
Confusion Matrix and Statistics

          Reference
Prediction High Low
      High    3   0
      Low     0  19
                                     
               Accuracy : 1          
                 95% CI : (0.8456, 1)
    No Information Rate : 0.8636     
    P-Value [Acc > NIR] : 0.03974    
                                     
                  Kappa : 1          
 Mcnemar's Test P-Value : NA         
                                     
            Sensitivity : 1.0000     
            Specificity : 1.0000     
         Pos Pred Value : 1.0000     
         Neg Pred Value : 1.0000     
             Prevalence : 0.1364     
         Detection Rate : 0.1364     
   Detection Prevalence : 0.1364     
      Balanced Accuracy : 1.0000     
                                     
       'Positive' Class : High       
                                     
> |
```

Naïve Bayes Classifier Result



```
predTrain High Low
     High    1   0
     Low     0  20
> predTrain
  10    8   16    2   13   26   20    9   25    4   22   17    6    7   24   11    3 
 Low  Low High  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low  Low 
   5   28   29   18 
 Low  Low  Low  Low 
Levels: High Low
> confusionMatrix(predTrain, TrainSet$class)
Confusion Matrix and Statistics

          Reference
Prediction High Low
      High    1   0
      Low     0  20
                                     
               Accuracy : 1          
                 95% CI : (0.8389, 1)
    No Information Rate : 0.9524     
    P-Value [Acc > NIR] : 0.3589     
                                     
                  Kappa : 1          
 Mcnemar's Test P-Value : NA         
                                     
            Sensitivity : 1.00000    
            Specificity : 1.00000    
         Pos Pred Value : 1.00000    
         Neg Pred Value : 1.00000    
             Prevalence : 0.04762    
         Detection Rate : 0.04762    
   Detection Prevalence : 0.04762    
      Balanced Accuracy : 1.00000    
                                     
       'Positive' Class : High       
                                     
> 
```

```
confu                List of 6
g                    num [1:30] 0.75 0.997 0.652 0.433 0.332 ...
m1                   List of 16
model1               List of 19
Naive_Bayes_Model    List of 4
NB_Predictions       Factor w/ 2 levels "High","Low": 2 2 2 2 2 2 2 2 2 ...
p1                   Factor w/ 2 levels "High","Low": 2 2 2 2 2 2 2 2 2 ...
predTrain            Factor w/ 2 levels "High","Low": 2 2 1 2 2 2 2 2 2 ...
```

	Name	Description	Version
User Library			
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	BH	Boost C++ Header Files	1.69.0-1
□	bitops	Bitwise Operations	1.0-6
□	C50	C5.0 Decision Trees and Rule-Based Models	0.1.2
□	caret	Classification and Regression Training	6.0-81
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1.2
□	classInt	Choose Univariate Class Intervals	0.3-1
□	cli	Helpers for Developing Command Line Interfaces	1.0.1
□	clipr	Read and Write from the System Clipboard	0.5.0
□	colorspace	A Toolbox for Manipulating and Assessing Colors and Palettes	1.4-0
□	combinat	combinatorics utilities	0.0-8
□	crayon	Colored Terminal Output	1.3.4
□	Cubist	Rule- And Instance-Based Regression Modeling	0.2.2
□	data.table	Extension of `data.frame`	1.12.0

Random forest Classifier Result

Table-4 Results Produced by Three Data Mining Algorithms on Groundwater Dataset

Classifiers	Accuracy	Kappa	Sensitivity	Specificity
Naïve Bayes	100%	1	1.0000	1.0000
C5.0	93%	0.6296	0.66667	0.96296
Random forest	100%	1	1.0000	1.0000

VIII. CONCLUSION AND FUTURE WORK

The ground water plays a prime role in a country like India. In this paper, we proposed three classification algorithms like C5.0, Naive Bayes and Random forest with data analytics tool R to generate effective predictive model which predicts whether water is “High ” of “Low ” for drinking purpose based on water quality parameters. Naïve Bayes and Random forest produced better result with accuracy and classification error. In future we intend to use more classification algorithm with extended dataset to analyze the ground water quality Hence proper water treatment is required in terms of community health.

References

1. Shoba Nath., Raju M.B., Rajagopalan K. and Nandakumar P., Central Groundwater Board, World Bank aided Hydrology Project, Water Resources Organisation, PWD, Chennai, Workshop on Water Quality issues in Tamil Nadu, India, 25 (1988).
2. Kulasekaran A., ChettiaGounder K. and Chellapandian K.,Water Quality Problems in Tamil Nadu, Workshop on Water Quality issues in Tamilnadu, World Bank aided Hydrology Project, organised by Water Resources Organisation, PWD, 14 (1998)
3. APHA., Standard methods for the examination of water and waste water, Edition. American Public Health Association, Washington D.C. 21st. Edition (2005)
4. BIS, Indian Standards Specifications for drinking water,Bureau of Indian Standards, New Delhi (2012).
5. Sinha D.K., Shilpi S. and Ritesh S., Water Quality Index for Ram Ganga River at Moradabad, Pollution Research, 23(3), 527-531 (2004)
6. Boakye.E, S. N. Odai, K. A. Adjei, F. O. Annorse (2008) Landsat Images for Assessment of the Impact of Land Use and Land Cover Changes on the Bareke, Catchment in Ghana, European Journal of Scientific Research.
7. Pandey Sandeep K, Tiwari S, Physico-chemical analysis of ground water of selected area of Ghazipur city-A case study. Nature and Science., 2009; 7(1).
8. Gupta. Acute toxicity to as estuarine toleost of mixtures of Cd, Cu, and Zn salts.1984
9. Anurag tewari, Ashutosh dubey and Aviral trivedi, A study on physico-chemical characteristics of ground water quality Journal of Chem. Pharm. Res., (2010) 2(2): 510-518.
10. Gursimran Singh, Dapinder Deep Singh, Prof. S.K.Sharma,“Effect Of Polluted Surface Water On Groundwater: A CaStudy Of Budha Nullah”. IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)e-ISSN: 2278-1684Volume 5, Issue 5 (Mar. - Apr. 2013), PP 01-08.
11. K. Prakash, V. Hanuman Reddy, R. Chenna Krishna Reddy, P. M. N Prasad, V. Krishanaiah and Y. V. Rami Reddy, Journal of Chemical a and Pharmaceutical Research, (2012) 4(2):1239-1245.
12. Johansson, J., and Rasmussen, L., Retrospective study (1944–1976) of heavy metals in the epiphyte Pterogonium racile collected from one phorophyte., Bryologist,(1977) 80(3),625-629.
13. R. Jayalakshmi , M. Savitha Devi, Soil Fertility Prediction for Yield Productivity and Identifying the Hidden Factors through Machine Learning Algorithms, , International Journal of Computer Sciences and Engineering Vol.-7, Issue-1, Jan 2019,pp 596-600.