A Novel Approach on Various Machine Learning Algorithms for Predicting Ground Water Quality

S. Vijay[1], Dr. K. Kamaraj[2]

[1]*Department of Computer Science, Vivekanandha College of Arts and Sciences for Women, Tiruchengode.*

[2]*Department of Computer Science, SSM College of Arts and Science, Komarapalayam, Namakkal(Dt).*

ABSTRACT

Data mining is the process of extracting useful or hidden information from a large database. Extracted information can be used to discover relationships among features, where data objects are grouped according to logical relationships; or to predict unseen objects to one of the predefined groups. In this paper, we aim to investigate four well-known data mining algorithms in order to predict groundwater quality. These algorithms are C5.0 Algorithm, Random Forest, K-Nearest Neighbour (KNN), The experimental results indicate that the C5.0 algorithm outperformed other algorithms in terms of classification accuracy. This work is to build a quality water for people usage and as well as for drinking purposes using data mining techniques such classification and clustering to find suitable data models with high accuracy.

Keywords: Water Quality, C5.0, K-Nearest Neighbour and Random Forest.

I. INTRODUCTION

Ground water has made significant contributions to the growth of India's Economy and has been an important catalyst for its socio economic development. Contamination of groundwater, both in terms of quality and quantity, are increasing rapidly to growing demands, significant changes in land use pattern, industrial effluents, domestic effluent etc. The need to assess the groundwater quality is becoming increasingly important as groundwater sources become more and more contaminated by industrial effluents and unsustainable agricultural practices.

Water is the most vital element among the natural resources, and is critical for the survival of all living organisms including human, food production, and economic development. Today there are many cities worldwide facing an acute shortage of water and nearly 40 percent of the world's food supply is grown under irrigation and a wide variety of industrial processes depends on water. The environment, economic growth, and developments are all highly influenced by water-its regional and seasonal availability, and the quality of surface and groundwater. The quality of water is affected by human activities and is declining due to the rise of urbanization, population growth, industrial production, climate change and other factors. The resulting water pollution is a serious threat to the well-being of both the Earth and its population.

The specific contaminants leading to pollution in water include a wide spectrum of chemicals, pathogens, and physical changes such as elevated temperature and discoloration. While many of the chemicals and substances that are regulated may be naturally occurring (calcium, sodium, iron, manganese, etc.) the concentration usually determines what is a

natural component of water and what is a contaminant. High concentrations of naturally occurring substances can have negative impacts on aquatic flora and fauna. Agricultural wastewater treatment for farms, and erosion control from construction sites can also help prevent water pollution. Nature-based solutions are another approach to prevent water pollution.

The research aims to investigate four well-known data mining techniques: C5.0, Naïve Bayes (NB) and Random Forest, to determine which better data mining techniques that can predict groundwater.

II. LITERATURE REVIEW

This section presents related works of machine learning and data mining in groundwater applications.

Ruchi Gupta, Anil Kumar Misra developed a method to detect the quality of ground water is the availability of salt rich geological formation in subsurface in Vellore district. The groundwater quality is totally unsuitable for domestic purposes as Water Quality Index is more than 100 for the region. Long term intake of fluoride above the permissible limit in drinking water is causing dental fluorosis diseases in the study areas.

Gorai, Hasni, JasedIqbal proposed Fuzzy rule-based approach to predict ground water quality index to assess suitability for drinking purpose. The study suggests a robust decision-making tool for drinking water quality management in the form of the fuzzy water quality index (FWQI). The developed methodology demonstrates to determine a single index value to make assessment of drinking water quality more understandable especially in public consideration. The fuzzy model developed is applicable only for specific number of water quality parameters in specified range selected.

Mallika et.al proposed linear regression to predict ground water quality model for irrigation using data mining techniques. An effective data mining technique was used to predict water quality and thereby forecast the crop yield. It is decisively concluded that this may help the decision maker to predict the crop yield with respect to water quality before harvesting the crop.

III.DATA MINING TECHNIQUES

Data Mining is the process of turning raw data into appropriate and meaning information. Various researchers have studied and work on data mining techniques to evaluate and classify the water quality.

A. Decision Tree (C5.0)

Decision tree is one of the predictive modeling techniques used in data mining. It aids to divide the larger dataset into smaller dataset indicating a parent-child relationship. Each internal node defined as inner node is labeled with an input feature. The inner nodes which exhibit many types of attribute test, bifurcations exhibit the test outcomes and leaf nodes particularly exhibit the category of a specific type[4]. Decision tree can handle both numerical and categorical data. It is well suited with large datasets. Higher accuracy in decision tree classification technique depicts that the technique can simulate. It is able to optimize variety of input data such as nominal, numeric and textual. It is a successful supervised learning approach which has the capability of extracting the information from vast amount of data based on decision rules.

B. K- Nearest Neighbour (KNN)

The KNN algorithm is simple. Based on training and test data, the KNN finds the kNNs of the training data, and uses classes of the k-neighbors to assign the class of the test instance. The scores of similarity of each neighbor instance to the test instance are used as a weight of the classes of the neighbor instance. When several

kNNs share a class, then the pre-neighbor weights of that class should be added together, and the result of the added weights should be used as the likelihood score of that class with regard to the test instance.

C. Random Forest

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

IV.MATERIAL AND METHODS

Study area

Vellore is the second most populous district in TamilNadu and had population of 906,745 as per 2011 census. In terms of urbanization level, Vellore district ranks 8thplace among the other districts in Tamil Nadu. Vellore is a major transit point for travellers, a hub for medical tourism and is emerging as a tourism hot spot. This place is known for its extreme climatic conditions. Vellore has an arid and dry climate, reaching high temperatures during summer. The city experiences wet winters and dry summers and has an elevation of about 224 meters with the north-east monsoon the highest contributor to rainfall. The mean maximum and minimum temperatures during summer and winter varies between 38.3°C and 18.95°C. The District lies between 12º15’23’’to 13º12'32’’ N Latitude, 78º24’16’’ to 79º54’56’’E Longitude and has an aerial extent of 6077 sq.km.

In the present research work, the study area in Vellore district have large industrials profiles such as textiles, leather tanneries and small-scale dying industries. The effluents of the leather industries, usage of the chemical fertilizers for

Table 1. The analytical results showing quality of ground water in the study area

Sample Station	Type	TDS	Hardness	Magnesium	Nitrate	Sulphate
ANAICUT	Deep Tubewell	3150	880	154	53	420
ARAKKONAM	Deep Tubewell	2349	880	103	56	461
ARCOT	Deep Tubewell	2067	720	106	49	604
GUDIYATHAM	Deep Tubewell	2520	950	118	49	410
KANIYAMBADI	Deep Tubewell	3080	960	106	52	405
KAVERIPAKKAM	Deep Tubewell	2426	695	113	50	512
PERANAMBATTU	Deep Tubewell	2961	840	120	53	436
THIMIRI	Deep Tubewell	2188	720	110	48	408
TIRUPATHUR	Deep Tubewell	2092	900	108	54	443
VELLORE	Deep Tubewell	2443	840	101	58	517
WALAJAPET	Deep Tubewell	2468	630	106	55	478

VI. EXPERIMENTAL RESULTS AND ANALYSIS

In this work, different classifier techniques of Machine Learning such as C5.0, Random forest and K-Nearest Neighbour were compared and evaluated on the basis of accuracy to build a model.

Two groups are separated from the data set for training and testing the algorithms of classification. R Tool is used to implement the classification algorithm.

TABLE II

Results Produced by Three Data Mining Algorithms on Groundwater Datasets

Classifier	Accuracy Rate
C5.0	96%
Random forest	94.2%
K-Nearest Neighbour	88.8%

VII. CONCLUSION

We implement three classification algorithms like C5.0, Random forest and K-Nearest Neighbour with data analytics tool R to generate effective predictive model which predicts whether ground water "Yes" or "No" for drinking purpose based on quality parameter C5.0 Produced better result with accuracy 96%.

In future we intend to use more classification algorithms with extended dataset to analyze the ground water quality.

REFERENCES

[1] Jordan Ministry of Water and Irrigation – Reports 2013-2016.
[2] Nortcliff A, Carr G, Potter RB, Darmame K. (2008) Jordan's water Resources: Challenges for the Future. Geographical Paper No. 185, The University of Reading.
[3] Karthik, D., & Vijayarekha, K. (2014). Multivariate Data Mining Techniques for Assessing Water Potability. Rasayan Journal of Chemistry, 7 (3):256-259.
[4] Maatta, S. (2011). Predicting groundwater levels using linear regression and neural networks, CS229 final project, December 15, 2011.
[5] Al Kuisi, M., El-Naqa, A., & Hammouri, N. (2006). Vulnerability mapping of shallow groundwater aquifer using SINTACS model in the Jordan Valley area, Jordan. Environmental Geology, 50(5), 651-667.
[6] Salah, H., (2009). Geostatistical analysis of groundwater levels in the south Al Jabal Al Akhdar area using GIS. GIS Ostrava.
[7] Kumar, S., Dirmeyer, P. A., Merwade, V., DelSole, T., Adams, J. M., & Niyogi, D. (2013). Land use/cover change impacts in CMIP5 climate simulations: A new methodology and 21st century challenges. Journal of Geophysical Research: Atmospheres, 118(12), 6337-6353.
[8] Cook, J.B., Roehl, E.A. and Daamen, R.C., 2013. Predicting the Impact of Climate Change on Salinity Intrusions in Coastal South Carolina and Georgia. *Proceedings of the 2013 Georgia Water Resources Conference*, held April 10–11, 2013, at the University of Georgia.