International Journal of Early Childhood Special Education (INT-JECSE)
DOI:10.9756/INTJECSE/V14I5.399 ISSN: 1308-5581 Vol 14, Issue 05 2022

COMPARATIVE ANALYSIS OF DECISION TREE AND RANDOM FOREST TECHNIQUE FOR ANALYSIS OF WATER IN MAHARASHTRA

Ms. Swapnali D. Mahadik[1] and Dr.Anup Girdhar[2]
[1]PH.D Research Scholar, TMV,Pune, Assistant Professor, MCA, DES's NMITD, Mumbai, India (Affiliated to University to Mumbai)
[2]CEO-Founder, Sedulity solutions & Technologies, Ph.D Guide, TMV, Pune, India
[1]Swapn.mahadik30@gmail.com, [2]anupgirdhar@gmail.com

ABSTRACT
This study presents a data model of Classification based onDecision Tree and Random Forest for considering Maharashtra water quality data from several different places. Data model analysis is done in WEKA software.It deals with ten different samples of water with various parameters like pH, Taste, Turbidity, TDS, Aluminum, Calcium and Iron. In this the comparative analysis is done by applying Decision Tree and Random Forest Technique on the sample. By implementing this evaluation technique of Decision Tree and Random Forest,it can be analyzed that few parameters cannot fit into the model. But the variations of the property values affect the quality or productivity of the water. The parametric value must fall within the permitted range of Indian Guidelines and World Health Organization (WHO) drinking water standards.

Keywords: Data Model, Classification,Decision Tree, Random Forest,

I. INTRODUCTION

The environment is changing significantly as a result of increased development. Generally, for safety purposes, most people choose packaged water when traveling. [1] As a result, given the numerous types of health difficulties, water quality should be examined on a regular basis for safety.Getting access to clean water and hygiene practice is the necessary for a healthy population.[4] At the point of supply to users and any other usage for manufacturing or manufacture, the water must meet the required chemical, biological, and physical quality criteria.

Water quality in metropolitan regions, particularly in Mumbai, Pune, and Nashik, has deteriorated due to the discharge of domestic and industrial wastewaters, as well as urbanization and other causes. So the trained data set is used to classify the parameters based on their values which can be easily map and predict the water quality.

The complexity of water quality is reflects in many types of measurement techniques of water quality indicator. So the water quality inclines to be focused on water that is treated for human consumption.

II. LITERATURE REVIEW

Different countries regulate drinking water differently based on the quality of their water source, according to MegersaOlumana Dinka.[1] [5] Data mining is a technique for extracting information from large amounts of data included in a dataset. [2] Data mining is used to extract and interpret information contained in datasets, as well as discover required information and relationships between attributes, thus according current trends and applications.

According to Shailesh Jaloreeet.alTo classify/predict the pollution class of water, decision tree classification was used.[2]Alexander finds a significant link between water source sanitation and microbiological (TC and E. coli) water quality, showing that the protected water source was safer than the unprotected water source.[4]

Although the report revealed that storage methods affect water quality after collection, it also observed progressive contamination during storage, with nearly three-quarters of stored samples collected contaminated with enteric bacteria.

As a result, disturbances in this ecological and biological system may have an impact on the health of birds, animals, and aquatic life. [3]

III. Sample Collection, Material and Methodology:

Area of the Study: The study area is of Maharashtra's urban and rural area. This location was chosen because of its high population and diverse crowd.

A total 10 samples were collected from various locations. Here in table 1.1 10 samples are shown for reference. Which includes 7 different parameters which is shown in Table 1.1. Between the moment of collection of samples and the time of analysis, many physical and chemical events occur, altering the status of water sample.

Therefore to analyze that water samples are preserved and then only it is being tested to check variations for parametric values. Analysis is done on 7 different parameters as shown in table 1.1 parameters categories are the combination of

essential and General Parameter. In the absence of an alternate source, the specification defines the desirable and allowed boundaries for parametric values. In India it is recommended that the Drinking water quality standards has to be followed as per the Indian Standard Bureau and World Health Organization (WHO). However, if the value exceeds the permitted limitations in the absence of an alternative resource, the sources should be disregarded. [10] [11] [12]

METHODOLOGY:

"Decision tree is a machine learning technique that allows to estimate a quantitative target or classify observation into one category of a categorical target variable by repeatedly dividing observations into mutually exclusive groups".A decision tree is a decision-making tool that uses a tree-like graph and its likely outcomes, such as chance event outcomes, cost objects, and utility, to make decisions."Random forest is a Supervised Machine Learning Algorithm which is used in Classification and Regression". It creates decision trees from various samples and classifies them based on the majority vote.

Here Data were collected, analyzed and processed in Weka tool for implementing Decision Tree and Random Forest Technique. Using the Standard Methods for the Examination, the water samples were examined for the presence and correlation of seven different parameters.

Taste, pH, Turbidity, TDS, Calcium, Aluminum, and Iron are just a few examples.

In preprocessing of data, Data was not directly suited for knowledge acquisition since that's including additional components such as missing values and an inconsistent data set. As a result, we must considerably reduce noise and, as a result, verify and validate the training sets that will be used to build a model and, finally, process the continuous phases.

Parameters Type	**Sr No**	**Parameters**	**Sample**									
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Parameters	**1**	**Taste**	Agreeable	Agreeable	Agreeable	Agreeable	Agreeable	Agreeable	Agreeable	Agreeable	Agreeable	
	2	**PH**	7.5	7.6	7.4	7.6	7.7	7.5	7.5	7.6	7.4	7.6
	3	**Turbidity**	1	1	1	1	1	1	1	1	1	1
	4	**TDS**	49	52	47	51	52	51	51	48	47	52
	5	**Aluminium**	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	6	**Calcium**	92	80	70	83	80	86	89	91	85	98
	7	**Iron**	0.2	0.3	0.2	0.3	0.3	0.2	0.21	0.3	0.2	0.3

Table 1.1 Water Samples and their Actual Parametric Values

Taste	PH	Turbidity	TDS	Aluminium	Calcium	Iron
Agreeable	7.5	0.98	49	0.029	92	0.2
Agreeable	7.6	1	52	0.03	80	0.3
Agreeable	7.4	1.02	47	0.031	70	0.2
Agreeable	7.6	1	51	0.028	83	0.3
Agreeable	7.7	0.99	52	0.03	80	0.3
Agreeable	7.5	0.98	51	0.032	86	0.2
Agreeable	7.5	1.01	51	0.031	89	0.21
Agreeable	7.6	1	48	0.029	91	0.3
Agreeable	7.4	0.99	47	0.03	85	0.2
Not	7.6	1.01	52	0.031	98	0.3

Table 1.2: Dataset

IV. DISCUSSION

The main motive of decision tree is data can split continuously according to certain parametrs. The tree can be explained here by showing sample of Iron. Many parameters influence the quality of water. Seven parameters are selected for invesigation. The main motive of decision tree is data can split continuously according to the certain parametrs. The parameters and settings of the tree can be explained here.

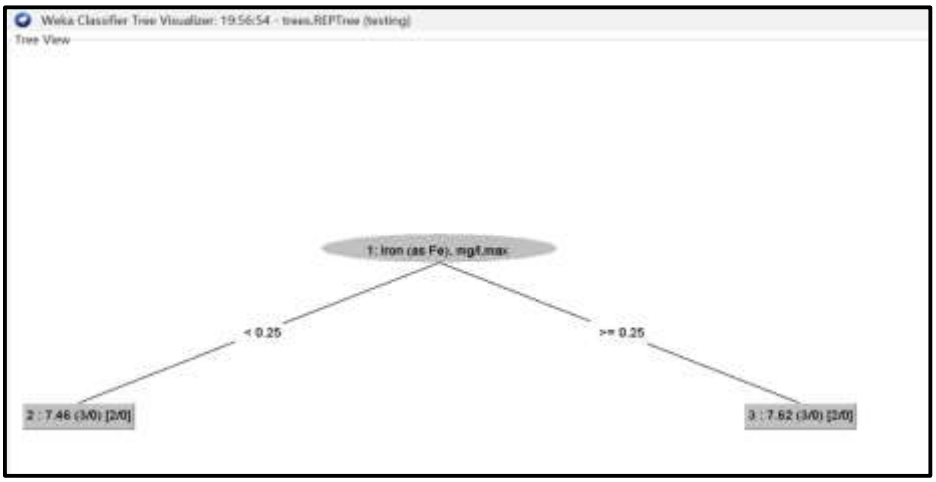

Figure 1.1 Decision Tree

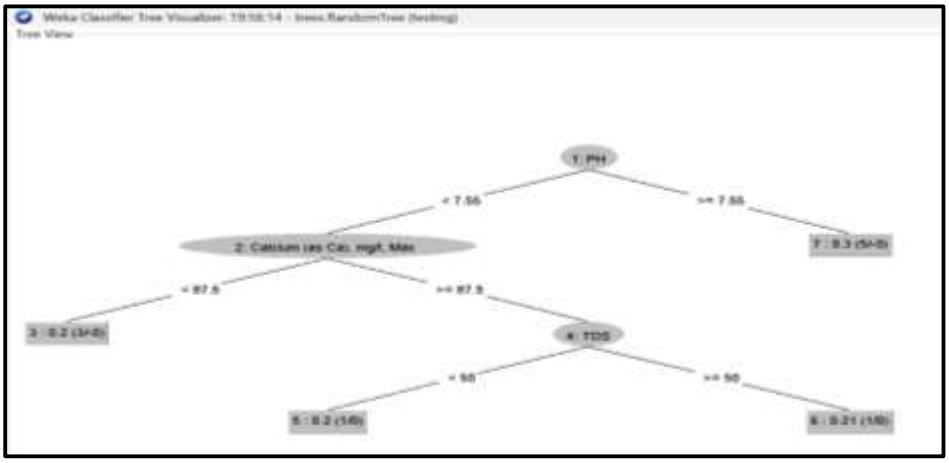

Figure1.2. Random Forest Technique

The entire training set is taken into account. Feature values are usually categorical, and records are dispersed according to their properties. And the ordering is done using a statistical method. This method use numerous algorithms to determine if a node should be split into two or more sub-nodes.

It's possible that the node's purity improves as the target variable increases. Random Forest is a supervised machine learning technique that can be used to classify and predict data. The hyper parameters are nearly identical to those of a decision tree. Its ensemble approach of decision trees is based on data that has been randomly partitioned. Each tree in this category has a separate independent random sample.Many trees can make the random forest algorithm excessively sluggish and inefficient for real-time prediction in the case of the random forest algorithm.

A decision tree integrates certain decisions in a comparative analysis, whereas a random forest mixes numerous decision trees. The fundamental advantage of a decision tree is that it is simple to understand. We know which variable and which value the variable uses to split the data in the decision tree. The ability to predict the outcome is really quick. Random forest algorithm models, on the other hand, are more difficult because they integrate decision trees.

The trees to create and the set of variables required for each node must be given when creating a random forest algorithm model. As a result, taking into account more trees increases performance and makes predictions more stable, but it also slows down computation speed.

V. CONCLUSION

This comparative analysis of Decision Tree and Random Forest shows that random forest contains multiple decision tree so it is diificult to interpret than the decision tree. The dataset used to obtain the outcome is smaller in this case, allowing the Random forest technique to interpret it. But if the dataset having larger value set it will become complicated to understand.So it is better to use Decision Tree to make quick prediction of water qualty.

REFERENCES

1. SwapnaliMahadik,Dr.Anup Girdhar, "Analysis of pH, and TDS in the available Packaged Druinking water by KNN and Decision Tree Technique in Mumbai,India", International Journal of Advanced Science and Technology , Vol.29, No.7s (2020), pp541-547
2. Shailesh jaloree, Anil Rajput ,DanjeevGour , "Decision Tree approach to build a model for water quality",Binary Journal of Data Mining and Networking 4(2014)25-28, ISSN 2229-7170
3. Devangeeshukla*, Kinjal Bhadresha , Dr.N. K. Jain , Dr.H. A. Modi, "Physicochemical Analysis of Water from Various Sources and Their Comparative Studies", IOSR Journal Of Environmental Science, Toxicology And Food Technology (IOSR-JESTFT)e-ISSN: 2319-2402,p-ISSN: 2319-2399.Volume 5, Issue 3 (Jul. -Aug. 2013), PP 89-92
4. Alexander Agensi, Julius Tibyangye , Andrew Tamale, EzeraAgwu, Christine Amongi , "Contamination Potentials of Household Water Handling and Storage Practices in Kirundo Subcounty, Kisoro District, Uganda", HindawiJournal of Environmental and Public Health, Volume 2019 |Article ID 7932193,March 2019
5. MegersaOlumana Dinka, "Safe Drinking Water: Concepts, Benefits, Principles and Standards", March 2018
6. https://www.copperutensilonline.com/blog/the-best-option-for-drinking-water-copper-stainless-steel-glass-or-plastic/
7. https://doctor.ndtv.com/living-healthy/this-is-why-you-should-use-a-clay-pot-to-store-drinking-water-1712265
8. https://www.copperh2o.com/blogs/blog/the-ultimate-guide-to-copper-vessels
9. Hydrology and Water Resources Information System for India, http://117.252.14.242/rbis/india_information/water%20quality%20standards.htm
10. Indian Standard Drinking Water- Specification (Second Revision) , IS 10500 : 2012
11. Guidelines For Drinking water Quality , Third Edition ,Volume 1 , WHO
12. Guidelines For Drinking water Quality , Fourth Edition ,Volume 1 , WHO