Packaged Drinking Water Analysis by Classification Technique of Data Mining

Swapnali D. Mahadik, Anup Girdhar

Abstract: *In a world where sufficient and safe water is not available everywhere, either because of harmful substances are found in the layer of earth surface which enter into a water or may be because of some other microorganisms. However what about packaged drinking water is safest is the big question. To analyze this proposed a model which helps to compare few brands of packaged drinking water and checked certain water quality level through customized classification technique of data mining. Proposed customized classification model can predict safe water based on the parameters of water quality which make ease of work of the laboratory technician by predicting which packaged water will be safest to drink.*

Keywords: *Packaged Drinking Water, Microorganism, Customized Classification, Data Mining.*

I. INTRODUCTION

In perspective on the immediate consumption of water by humans, the local water supply is viewed as the most significant utilization of water and drinking use has been given need on usage of water asset in the National Water Policy.[1]But sometimes people avoid water of natural resources as well as domestic water supply and they prefer packaged Drinking Water. It is always not necessary that packaged drinking water is also always good to use. To find out such a brands in the Indian market which sold out unhealthy water, there are certain government bodies like National Accreditation Board for Testing and Calibration Laboratories NABL- accredited laboratories test the water samples and produce the results as per the consumer voice report 2017. [2] Packaged drinking water is derived from various sources like rivers, lake, ground, surface, sea and other resources subject to the treatment like filtration, demineralization, mineralization etc.

To improve the quality for the same not only the Water Quality Index but various data mining techniques can be used. The classification model is a systematic approach to build classification models from various input data sets. For analysis three Brands of water considered: Rail Neer, Mc Dowell's and Bisleri.

Parametric values i.e. pH, turbidity, Total Dissolved Solid, Color and Toxic / Heavy Metals values are stored in a dataset.

Table 1.1 Sample Water Brands List

Brand Name	Alternate Name
Rail Neer	Brand A
Mc Dowell's	Brand B
Bisleri	Brand C

Table 1.2 Dataset: Comparative analysis on Parameters of Water [2]

Brand/Parameters	PH	Turbidity	Color	TDS	Heavy Metals
Brand A	6.16	3	1	9.04	12
Brand B	6.02	3	1	7.6	12
Brand C	5.74	3	1	7.4	12

Based on this classifier built their associated labels of class. If all parametric values are fitting into a class which proves a good / safest drinking quality water that could be the one class and if not that could be the other one.

However Packaged Drinking Water plays a major role in populated country like India where packaged water is very vital.

II. LITERATURE REVIEW

As per the consumer voice report 2017, packaged drinking water is consumed by measuredly increasing population. In order to find out packaged drinking water brands which are not good for health are being sold in overall Indian society. To check this Consumer Voice tested few water provider brands on the basis of their market share in an NABL laboratory. This brands includes Rail Neer, Mc Dowell's, Bisleri, Kinley, Bailey, Aquafina.[2]

As per the World Health Association as well as Indian Standard guidelines there is a specific range for every parameter according to use of water. Specific range of ph value is nearby 6.5-8.5 to fall within Environmental Protection Agency (EPA) standards water is a good solvent but picks up impurities easily. As per World Health Organization guidelines for quality of drinking water, water with extremely low concentrations of Total Dissolved Solids may be unacceptable to consumers because of its flat taste. The European standards for turbidity also states that it must be not more than 4 NTU. The World Health Organization, establishes that the turbidity of drinking water should not be more than 5 NTU.

For trace and toxic /Heavy metals (BIS-10500-2012), Toxic/Heavy metals covers Totals arsenic as 0.01, cadmium as cd 0.003, total chromium as cr 0.05,Copper as Cu 2.0, Lead as Pb 0.01, Nickel as Ni 0.02, Zinc as Zn 3.0 total around 5.093 mg/L.[2]

As author Swapan Shaakhari and Indrajit Banerjee proposed Water quality monitoring device consisting of existing sensors to collect water quality data from three different resources and applied classification method and decision tree to finding out better results.[4]

According to author Shah Christirani Azhar , Ahmad Zaharin Aris The multivariate analysis confirmed the measurement results of the water quality status by the Department of Environment in the sense of water classification. Mainly author have utilized multivariate analysis of cluster analysis (CA), principal component analysis (PCA) and discriminant analysis (DA). [5]

Salisu Yusuf Muhammad and Mokhairi Makhtar has suggested in his research about, the characterization models, for example, Bayes, Rule, Trees, Lazy and Meta were utilized to recognize the best model among them to arrange the water quality. [6]

III. RESEARCH METHODOLOGY

Data cleaning and Data preprocessing is an important step before choosing the Data Mining algorithm. This cleaning and processing includes basic operations like strategies, methodologies and appropriate handling missing data removing noise. In water quality analysis also Data mining plays a very important role. Data mining is the process of discovering patterns from hidden datasets. [7] In this case, a classifier is required to predict class label such as safety and risky.

IV. CLASSIFICATION

Classification models predict categorical class labels. Here classification Model can be categorized as either safe or risky based on the parametric values provided by the World Health Organization. Here classification algorithm build the classifier based on parametric values, the classifier is built from the training set here all values of respective brands made up of database tuples and their associated class labels. The test data (brands with their parametric values) are used to estimate the accuracy of classification rules.

V. PROCESS MODEL

Here in this research three well-known brands on India have been considered and from each of them unique five parameters are taken. Measure criteria for selection of water parameter is determine the best attribute test condition while implementation classification technique. Each parameter is full filing nearby value as per the standard given by the government bodies. But if the parametric values of any brand is not upto the mark it will decrease the water quality. As mentioned above classifier is built based on this parameters and according to value difference that will go to the particular class with respect to classification rule.

A. Data Collection and Processing

Table 1.2 shows the collection of water quality data of total three brands with only 5 water quality parameters i.e. pH, Turbidity , color , TDS and Heavy Metals by consumer voice report 2017. So to validate proposed method for packaged drinking water classifier is used. Before using any Data set directly data cleaning is required. As shown in Figure 1, proposed Customized classification model is used for water quality monitoring system in the datasets as mentioned earlier in the table 1. And then compared the results of proposed classification model with standard classification which shows the minor change of any parametric value can change the level of water quality from Good to Poor.
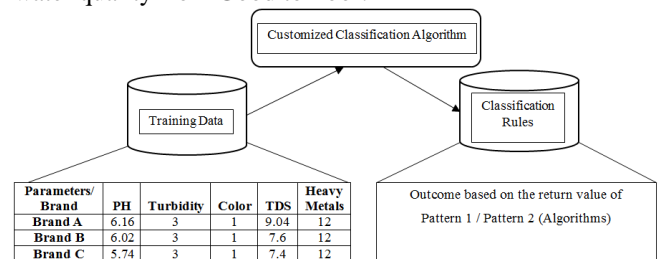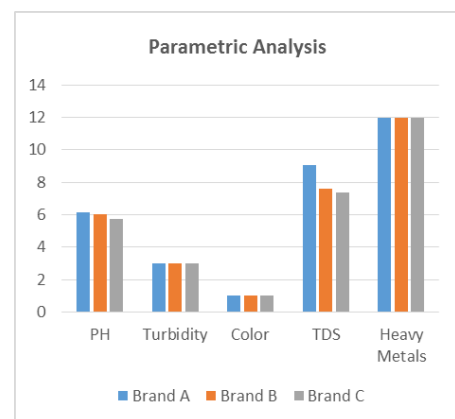

Parameters/ Brand	PH	Turbidity	Color	TDS	Heavy Metals
Brand A	6.16	3	1	9.04	12
Brand B	6.02	3	1	7.6	12
Brand C	5.74	3	1	7.4	12

Fig 1.1 Water Quality Parameter Classifier for Customized Classification

VI. RESULT AND OUTCOME

Graph 1.1 represents graphical representation of each brands parametric values. Turbidity, Color and Toxic/Heavy Metals are having same values but because of ph and TDS water quality grade differentiates heavily. Similarly as shown in Graph 1.2, because of the differentiation of ph and TDS value Brand A is at the top level, Brand B and Brand C at the second level as per the Analysis Report of consumer voice 2017.
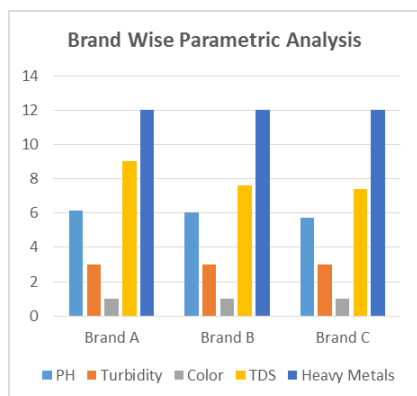

Graph 1.1 Parametric Analysis

In particular, Total Dissolved Solids in packaged drinking water shall be maximum 500 mg/liter.

As per the consumer Voice report 2017, majorly in all brands it is lower than the maximum permissible limit. But Brand A found the highest quantity of TDS i.e. 126 mg/liter. Consumer Voice suggests that there be a lower-side limit for TDS – no TDS means there are no minerals as well. Water with extremely low concentrations of TDS may also be unacceptable because of its flat taste.[2]

Similarly ph level of drinking water shows how acidic it is referring to the amount of hydrogen mixed with water. According to WHO range of ph should be in between 6.5 to 8.5.[10]

Graph 1.2 Brand wise Parametric Analysis

Pattern 1
if ((PH == 6.16) AND (Turbidity == 3) AND (Color == 1) AND (TDS == 9.04) AND (HeavyMetals == 12))
  then
  Water_Quality= Good
  end if
  Pattern 2
  if ((PH == 5.74) AND (Turbidity == 3) AND (Color == 1) AND (TDS == 9.04) AND (HeavyMetals == 12))
  then
  Water_Quality = Poor
  end if

According to customized classification model shown in fig 1.1, if all five parameters are falling in a specific range provided by WHO and Indian standard it will give result as good water quality but if only ph value is less that 6.5 it means it is more acidic in nature so it will be considered as unhealthy. Same thing will be applicable for TDS. So model will result in showing degrade of the water quality.

VII. LIMITATION

Here in Packaged drinking water the range of value is fixed for every brand with respect to every parameter. But after the expiry of a bottle or old water samples can give variations in range of specific values. So above specified proposed customized classification algorithm can be used to check the water quality. Or else As mentioned in [8] [9], to evaluate Natural Resource Krishna River, K-Nearest Neighbor model is examined using the correlation efficient, Root Mean square error and Mean Absolute error. In particular the predictive performance of parametric values with respect to its readings at different time and location, it is evaluated in terms of their root-mean-square-error (RMSE).

$$RMSE = \sqrt{\frac{1}{N}\sum_{l=1}^{M}(\mathbf{y}_l - \hat{\mathbf{y}}_l)^2}.$$

(1)

Where $\sum$ is the sum y denotes the observed values and y^ denotes the computed values.

Here Root Mean Square Error can be used for measuring the difference between values which are actually observed. This represents the Standard Deviation of the differences between predicted values and observed values. It is a measure that is used to quantify the amount of variation of a set of data values. [9]

Minor Variation in parametric value of any of the parameter which is not falling into a particular range will degrade the quality of water. So only KNN cannot be suitable in every situation with respect to packaged drinking water.

VIII. FUTURE SCOPE

Future work concerns deeper analysis of any particular mechanism, new proposals to try different methods. Here in this study classifier are made based on the parametric values of different brands of water. Classification using decision tree method can be applied to classify /predict the clean and not clean water which will be unhealthy for society.[3] Decision Tree Classifier, repetitively divides the working area into sub part repetitively because there may be two distant regions of same class divided by other.

Based on this decision tree results can be generated in better and consistent way. With the help of Decision Tree Technique the results will have clear separation of classes and less number of impurities. There can be analysis of more brands with more parameters of different importance with respect to quality. The same technique can be applicable for mineral water, Packaged Drinking water as well as other source of water. The performance of Customized Classification and Decision tree together will result into exact quality measurement. Because the main objective of Decision tree is, Distributing efficiently based on maximum information expansion is a good key to decision tree classifier.

IX. CONCLUSION

Based on the Indian standard and overall test findings, Rail Neer is purest as per the record of Consumer Voice Report 2017. Next in Ranking are Mc Dowell's and Bisleri. The model which is proposed here based on classification approach for predicting water quality with different combination of parameter made easier to find out safest water. It can be observed that through this, proposed customized classification technique of data mining supports water quality indications based on the given parameters. Further with the help of KNN Decision Tree technique it can be easier to predict the water quality of packaged drinking water.

REFERENCES

1. Jakir Hussain, Ikbal Husain, Mohammed Arif "Studies On Heavy Metal Contamination In Godavari River Basin" Volume 7, Issue 8, Pp 4539–4548 December 2017.
2. Consumer Voice Report "Packaged Drinking Water" 2017.
3. Consolata Gakii And Jeniffer Jepkoech, "Classification Model For Water Quality Analysis Using Decision Tree", European Journal Of Computer Science And Information Technology, Print ISSN: 2054-0957 (Print), Online ISSN: 2054-0965 (Online)Vol.7, No.3, Pp.1-8, June2019.
4. Swapan Shakhari, Indrajit Banerjee, "A Multi-Class Classification System Data mining is the process of discovering patterns from hidden datasets. for Continuous Water Quality Monitoring" Volume 5, Issue 5, May 2019, E01822.

5. Shah Christirani Azhar , Ahmad Zaharin aris , Mohd Kamil yusoff, Mohammad Firuz Ramli , Hafizanjuahir, "Classification Of River Water Quality Using Multivariate Analysis " Procedia Environmental Sciences Elsevier , Volume 30, 2015, Pages 79-84.

6. Salisu Yusuf Muhammad, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz and Azrul Amri Jamal, "Classification Model for Water Quality using Machine Learning Techniques" , International Journal of Software Engineering and Its Applications Vol. 9, No. 6(2015), pp. 45-52.

7. P. Tamilarasi, D. Akila, "Ground Water Data Analysis Using Data Mining: A Literature Review", International Journal of Recent Technology and Engineering (IJRTE)ISSN: 2277-3878,Volume-7, Issue-5C, February 2019.

8. Swapnali Mahadik, Dr.Anup Girdhar, "Water Quality Parameters Estimation Through Data Driven Model", Cyber Times International Journal of Technology & Management,Vol. Issue 1,October 2017-March 2018 2278-7518.

9. Swapnali Mahadik. Dr.Anup Girdhar, "Analysis of Principal Water Quality Parameters by Data Driven Technique", Research Dimension, 2249-3867, Feb 2018.

10. WHO Library Cataloguing-in-Publication DataGuidelines for drinking-water quality: fourth edition incorporating the first addendumISBN 978-92-4-154995-0, 2017.