Water Potability Analysis and Prediction

Heming Gao [1,†], Yuru Li [2,*,†], Handong Lu [3,†], Shuqi Zhu [4,†]

[1] Bengbu No.2 Middle School, Bengbu, China

[2] University of Illinois, Urbana-Champaign, Shandong, China

[3] Hangzhou Renhe Foreign Language School, Hangzhou, China

[4] University College Cork, Cork, Ireland

* Corresponding author: yurul2@illinois.edu

† These authors contributed equally.

Abstract. Water is one of the largest resources on earth. People need water to sustain life, including drinking water. It is important to know whether drinking water - human life resource - is enough for everyone now and in the future. However, water resources are not evenly distributed everywhere on the planet. While the water resource is rich in some countries and regions, it is not enough for some other regions. The analysis of different region's water resources should be done individually. In this paper, the authors analyze the potability of water by using an Indian water potability dataset from Kaggle. More specifically, this paper talks about each factor of water that influences water potability through statistical methods - binomial distribution and the k-nearest neighbor algorithm. Also, the authors build a model that allows people to predict the potability of a water resource by the data of each factor of that resource. According to the research, the features of water are not related to each other. All the features should meet a specific standard in order to get potable water.

Keywords: Big Data, water potability in India, binomial distribution, KNN, regression model.

1. Introduction

Western countries have made great use of online water quality monitoring technology. From their development trend and advanced experience, it can be seen that relevant environmental supervision departments have taken online water quality monitoring technology as a necessary means to obtain continuous water quality monitoring data. In just a few minutes, water quality information from water sources can be transmitted to computer servers in an online water quality monitoring system, where data reflecting water quality can be extracted for analysis. If it is found that the amount of some toxic and harmful substances in the water changes abruptly, the environmental supervision department can immediately conduct sampling analysis to find out the root cause of the problem, and then take corresponding countermeasures to eliminate the hidden dangers of water quality safety [1]. Water pollution is very serious in China. There are many problems in the monitoring and management of water quality [2]. The application of online monitoring technology and the development and utilization of online monitoring equipment are relatively backward. Therefore, it is urgent to control water pollution, which requires us to first do a good job of water quality monitoring, especially the online water quality monitoring technology that is more suitable for the current social needs, and gradually standardize it [3]. The core of the water quality online monitoring system is the online monitoring equipment. The system uses many technologies, including sensors, automatic measurement and control, computer applications, et al. By using these technologies, the system can sort out and analyze the detection data and output the maximum and minimum monitoring data in any time period. The average value of monitoring data in any time period can also be output by calculation [3, 4]. By analyzing the data, the system can automatically synthesize the data map reflecting the water quality, and finally store the information to the system data control center for research. The system has many functions, such as early warning whether the monitoring object exceeds the standard; The function of signal output and warning for lower monitoring stations; Automatic operation, power failure protection, automatic call recovery function [5].

Using the on-line water quality automatic monitoring technology, can achieve for all-weather monitoring of water quality and the purpose of the remote control, an understanding of major river systems of our country's main monitoring of water quality, timely grasp the changing situation of water quality and the mutation and water quality deterioration of events occurring early warning, solve for cross-regional local regulatory responsibility blaming caused by pollution, and differences, Improve the level of water quality monitoring, and finally promote water quality security [6, 7]. The all-weather monitoring and data transmission of the water quality online monitoring system for its subordinate sub-stations are realized through satellite and telephone dialing of the central station. Similarly, the hosting station can also implement the above functions for its subordinate sub-stations through telephone dialing [8].

2. Methodology

2.1. Machine Learning Models

In statistics, the binary logistic model is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). In this paper, we perform the binary logistic regression to each single variable. Specifically, we compute Hosmer-Lemeshow Goodness-of-Fit (HLGOF), and the significance level is set to be equal to or less than 0.05. For those features which do not pass the HL test, this paper considers it having no predictive power in the logistic model. Then the binary logistic regression is performed to those features which pass the HL test, and the HLGOF, the Exp(B), and the regression coefficient is computed [9].

In statistics, the k-nearest neighbor algorithm (KNN) is a non-parametric supervised learning method. The input consists of the k closest training examples in a data set, and in K-NN classification, the output is a class membership [10]. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. In this paper, as for those parameters having no predictive power in the logistic model, we perform the K-Nearest Neighbor model and use K-fold Cross Validation to find the best K.

2.2. Data

The dataset used in this study was downloaded from Kaggle. A total of 3276 samples were collected and analyzed for 9 important hydro-chemical parameters, which are pH value, hardness, total dissolved solids (TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes and turbidity. The potability of each sample, which indicates if water is safe for human consumption, is given, where 1 means Potable and 0 means Not potable. This dataset is randomly divided in such a way that 2457 (75%) samples are used for the training, whereas the remaining 819 (25%) were used for testing the models. This paper uses single imputation to deal with the missing value, i.e., all missing values that are labeled "potable" will be imputed using the mean of all non-missing "potable" samples, and the same action will be applied to "non-potable" samples with missing values.

2.3. Training process

This paper computes Pearson Correlation Coefficient between any two features to find out the degree to which these variables are correlated. Images of the distribution of all features divided by target label are shown in figure. Judging from whether the sample is potable, this paper divides all the samples into two populations. We perform a two-tailed t-test to check if there is any significant difference between the two samples' means, considering the sample size differences and unequal variance. The significance level alpha is set to be equal to or less than 0.1. This paper performs the binary logistic model and K-NN to all the features.

This paper uses the trained model above to treat the dataset used for prediction and find the false positive rate. In this process, if all the features show that the sample is potable, then output 1. Otherwise output 0.

3. Results and Discussion

3.1. The establishment of simulation model

From figure 1, the overall data clearly shows that water resources are 61% non-potable, while only 39% are potable. This contrast is a strong indication that the amount of water currently potable is under serious threat. More and more pollution which breeds bacteria occupies large space. Such bacteria pollute the groundwater and air, and then threaten safety of drinking water. At the same time, because of over-exploitation of groundwater, some areas have produced phenomena like shrinkage of lakes and disappearance of the mud flats. These all would cause the reduction of water storage capacity and water self-purification. This will deteriorate the status of water potability.

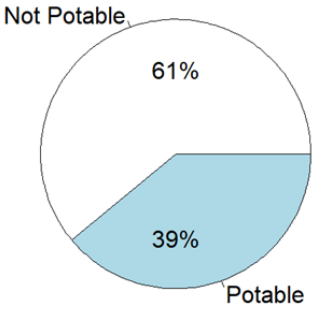

Figure 1. Pie chart of the water potability

3.2. The density distribution of variables

In this data, the nine variables which will affect water resources are considered. The data is divided into two groups by potability and their corresponding density distribution curves will be produced by analyzing nine variables. In this part, nine variables which affect water potability are split into three parts to describe. All density distributions are approximately normal distributions because their density curves are similar to normal distribution curves. The differentiation conditions are: (1) means are same but two curves are not; (2) the mean values are different; (3) two curves are almost the same.

3.2.1 Means are the same while two curves are not

PH: PH value which is a very significant parameter to evaluate the acid–base balance of water. The current maximum limit ranges using in this data of PH are 6.52-6.83 which fit the recommended range 6.5-8.5 from WHO. The number meeting this criterion in the data is 232 which is only 8.3% of the total. Hardness: Hardness is mostly caused by magnesium salts and calcium. Dissolve salts are deposited by geologic when water travels from cradle to sea. Typically, the longer the water is in contact with dis-solvable minerals, the higher the mineral content of the water will be, making it harder. Chloramines: Chlorine and chloramine are the mainly disinfectants used in public water systems. People add ammonia treat the water and chloramines might be created. The level of chlorine needs to be smaller than 4 mg/l, which is considered safe in drinking water. There is only 2.72% data that meets the requirement, and it needs to be changed desperately. Sulfate: Sulfates are naturally occurring substances that are present in minerals, soil, and rocks and they are also present in many places like groundwater, air, plants, and common food. The appropriate ranges of Sulfate are from 3 to 30 mg/L in treated water in most freshwater supplies. It is striking to find that in these tests, there is hardly any water source that meets the requirements perfectly, so the safety of drinking water is an issue that needs to be addressed. Conductivity: Pure water is a good insulator because it is less irons but water is always considered that it is not a good electric conductor. The more ions concentration is, the better the electrical conductivity is. Generally, the amount of dissolved solids are the main reason of electrical conductivity in the pure water. According to WHO standard, electrical conductivity value should not exceed 400 μS/cm. There are 1314 data that meet the requirement, and it occupies 40.11% among the data. It is better than other categories in this group.

The ones that fit this scenario are PH, Hardness, Chloramines(Chlorine and chloramine), Sulfate and Conductivity. From figure 2, figure 3, figure 4 and figure 5, all density distributions are approximately normal distributions. The symmetrical axis among four figures are the same between potable and non-potable groups that means the samples' means are the same. From figure 3, figure 4 and figure 5, the surface for non-potable is steeper than potable, especially Hardness and Sulfate. In contrast, from figure 2, the surface for potable is significantly steeper than non-potable, suggesting that the PH values under the potable category are more concentrated than those under no potable. From figure 4, the curve of potable in sulfate is slightly skewed to the left.

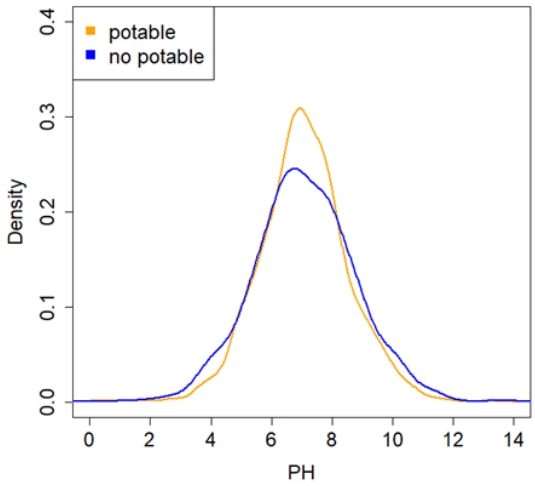

Figure 2. The density distribution curves of PH

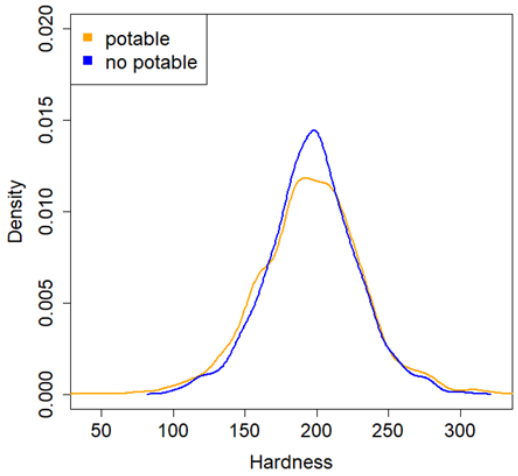

Figure 3. The density distribution curves of Hardness

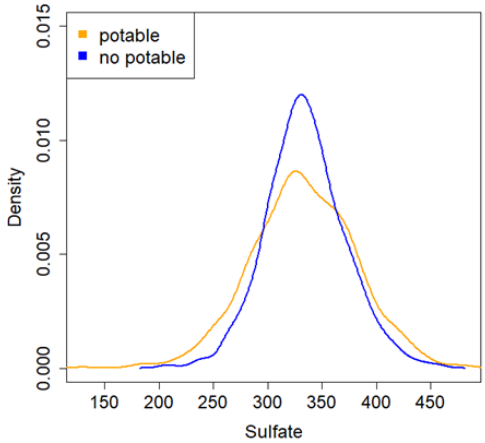

Figure 4. The density distribution curves of Sulfate

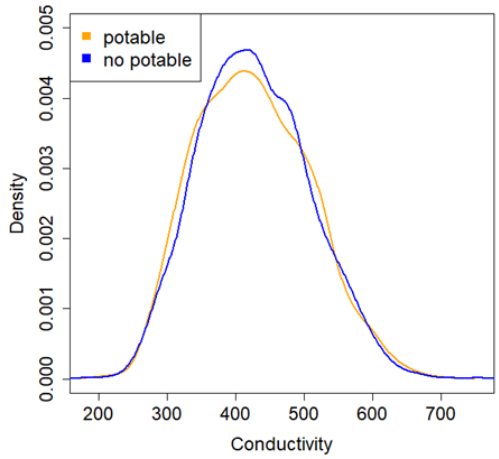

Figure 5. The density distribution curves of Conductivity

3.2.2 The mean values are different

Solids: A wide range of inorganic and some organic minerals or salts can be dissolved in the water. These minerals produce color and taste which are not wanted and influence water potability. The permitted maximum limit for solids is 1000 mg/l. There are only 2 data meeting this requirement. In the density curve in figure 6, we can clearly see that the two curves resemble a normal distribution but have a significant right skew, which is consistent with the results reflected in the outlier. The obvious difference between this data set and the others is that the symmetrical axis of the two curves is different.

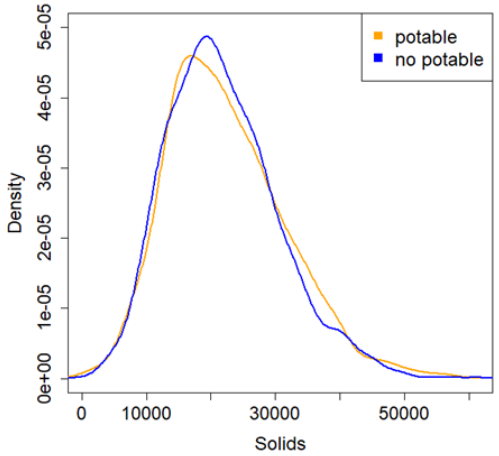

Figure 6. The density distribution curves of Solids

3.2.3 Two curves are almost the same

Organic Carbon: Natural organic matter and synthetic sources are the main component of total organic carbon which can simplify to TOC, in orginal source waters. Total amount of carbon in original untreated and treated water can be measured by TOC. According to the US regulation, total organic carbon is smaller than 4 mg/L in source water using for treatment and after treated, the organic carbon in drinking water is asked to be smaller than 2 mg/L. In the original data, there is only one set fitted to the requirement. Trihalomethanes: THMs are chemicals which might be found in water which are treated with chlorine. The level of organic material, the amount of chlorine and the temperature of the water all affect the concentration of THMs in treated water. The level of THMs is required to be smaller than 80 ppm. There are 2,512 eligible data, and 80.67% of the overall data meet the requirement. Compared to the other data, this one performed better, with most of the tested values meeting the drinking requirements. Turbidity: The water turbidity relies on the quantity of solid matter presenting in the suspended state. It is a method to measure of water light emitting properties and it can be used to present the quality of waste discharge with respect to colloidal matter. The level

of turbidity is asked to be smaller than 5.00 NTU by WHO. There are 2,962 eligible data, and 90.42% of the overall data meet the requirement. Compared to the other data, this is the best performed, with most of the tested values meeting the drinking requirements.

From figure 7, the overlap between the density curves of the two parts is very good. The figures of Trihalomethanes and Turbidity are similar to the figure 7. Their density distribution curves are almost same and better than the figure of Organic carbon. High overlap shows that the water potability has little effect on the distribution of these three data sets.

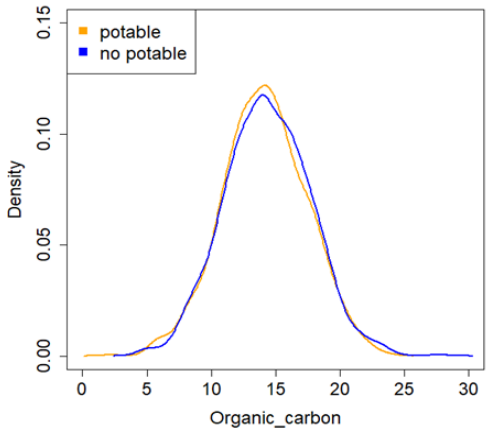

Figure 7. The density distribution curves of Organic Carbon

Overall, in these plots we can see that the distributions of the Potable and not potable images are relatively similar. We can learn that there is some difference between the distribution of the same features in different labels, but the two means are quite close.

3.3. Correlation of variables

From table 1, the correlation coefficients between the features were very low. A negative value means a negative correlation and a positive value means a positive correlation. Larger values represent stronger correlations and smaller values represent weaker correlations. Of these, only a very weak negative correlation is found between solid and sulfate. Other values are too small to be of reference value. It can be inferred from the picture that these variables are independent and hardly interact with each other.

Table 1. Comparison of coefficient in logistic regression

Correlation	pH	Hard	Solids	Chlor	Sulfate	Conduct	Organic	Trihal	Turbi
pH	1	0.109	-0.088	-0.025	0.011	0.014	0.028	0.018	-0.036
Hard	0.109	1	-0.053	-0.023	-0.109	0.012	0.013	-0.015	-0.035
Solids	-0.088	-0.053	1	-0.052	-0.163	-0.005	-0.005	-0.016	0.019
Chlor	-0.025	-0.023	-0.052	1	0.006	-0.028	-0.024	0.015	0.013
Sulfate	0.011	-0.109	-0.163	0.006	1	-0.016	0.027	-0.023	-0.01
Conduct	0.014	0.012	-0.005	-0.028	-0.016	1	0.016	0.005	0.012
Organic	0.028	0.013	-0.005	-0.024	0.027	0.016	1	-0.006	-0.015
Trihal	0.018	-0.015	-0.016	0.015	-0.023	0.005	-0.006	1	-0.02
Turbi	-0.036	-0.035	0.019	0.013	-0.01	0.012	-0.015	-0.02	1

Low p-values for the features that indeed are significantly different between the labels. When processing the data, we set the significance level to be equal to 0.05. By the results of R, we can see that only Solid qualifies significance level, suggesting that Solid has a significant impact on whether they are potable or not. The degrees of freedom are 2001 and the corresponding p-value is less than 0.05, so we can consider that it is a good model.

Multicollinearity needs to be considered because there are 9 variables in this logistic regression model. From the results of VIF test, all VIF are smaller than 10, so the model does not have multicollinearity.

Since outliers appear in the data, it needs to consider the case of over-dispersion. However, there is no significant change in the degrees of freedom, etc., after considering overdispersion in both sets of data. This means that there is no need to consider over-dispersion here and binomial is still used for the analysis.

Table 2. Results of relevant data fitted by KNN

Water prediction	Test label	Test label	Row Total
Train label	0	1	
0	391	103	494
1	209	108	317
Column Total	600	211	811

When the test label and the train label are the same, the prediction is considered successful. As can be seen from table 2, the number of successful predictions reaches 499. 61.52% of the overall prediction accuracy is achieved.

4. Conclusion

Water quality relates to everyone's life. The adequacy of water resources not only affects people's life safety, but also deeply affects the development and stability of society. Also, the government of each region should predict the sustainability of their water resources. Water used by humans should be guaranteed to be enough and safe. Different standards apply to different uses of water. When it comes to drinking water, the standards need to be particularly strict. Not every water resource meets these standards. In addition to the special substances contained in some specific type of water, there are nine main factors that affect the potability of water, including, PH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes and Turbidity.

According to the result of the research, the features of water are hardly related to each other. There are no general rules between each distribution of the factors. For example, the range of the non-potable distributions of PH value and organic are larger than the range of the potable distributions, while for the other factors except Trihalomethanes (the distribution of potable and non-potable is almost the same), the range of potable distribution is larger than the non-potable distributions. Qualification of one feature does not increase the possibility of the qualification of another feature. Only if every factor meets the potable standard, the water can be drunk.

By using statistical models, including binomial regression and the K-nearest neighbor algorithm (K-NN), the authors got a model in the end which can help to predict whether a water resource is potable. Although HLGOF of the logistic regression model is quite high, which indicates that there are not much differences between real values and predicting ones, the significance levels of most features treated by logistic regression are also very high. As a result, the false positive rate is high. The model still needs improvement.

References

[1] Roberto F., et al. Evaluation of a GFP reporter gene construct for environmental arsenic detection. Talanta, 2002, 58(1): 181-188.

[2] Erdogan O., et al. Critical evaluation of wastewater treatment and disposal strategies for Istanbul with regards to water quality monitoring study results. ELSEVISE, 2008, 226: 231-248.

[3] Lourenco N.D., et al. UV spectra analysis for water quality monitoring in a fuel park wastewater treatment plant. Chemosphere, 2006, 65: 786-791.

[4] Kim B. C. Multi-channel continuous water toxicity monitoring system: its evaluation and application to water dis-charged from a power plant. Environmental Monitoring and Assessment, 2005, 109(3): 156-164.

[5] ISO/TC 147 /SC5. ISO 11348-1 2007(E)Water quality – Determination of the inhibitory effect of water samples on the light emission of Vibrio fischeri (Luminescent bacteria test) - Part 1: Method using freshly prepared bacteria (ISO 11348-1). Geneva, Switzerland: ISO, 2007.

[6] Filho W. L., et al. Handbook of Theory and Practice of Sustainable Development in Higher Education, 2017.

[7] Sulimov A. V., et al. Long-term changes of heavy metal and Sulfur concentrations in ecosystems of the Taymyr Peninsula (Russian Federation) North of the Norilsk Industrial Complex, Environmental Monitoring & Assessment, 2011, 181(1-4): 539-553.

[8] Gao P. and Yin C. F. Study on Establishment of Water Resource Tax System: Based on the Analysis on Practice of Water Resource Fee Collection System, Journal of Central University of Finance & Economics, 2016.

[9] Fuchs Critical Social Theory and Sustainable Development: The Role of Class, Capitalism and Domination in a Dialectical Analysis of Un/Sustainability, Sustainable Development, 2017.

[10] Chevalier L. R., et al. Evaluation of in Spectra UV analyzer for measuring conventional water and wastewater parameters. Advances in Environmental Research, 2002, 6(3): 369-375.