Proceedings of Machine Learning Research 1–5

Safe Water Prediction Using Supervised Machine Learning

Dr. Sifat Momen SIFAT.MOMEN@NORTHSOUTH.EDU
Sabbir Ahmed Sozol SABBIR.SOZOL@NORTHSOUTH.EDU
Hasibur Rahman Ridoy HASIBUR.RIDOY@NORTHSOUTH.EDU

Department of Electrical and Computer Engineering, North South University, Dhaka 1229, Bangladesh

Abstract

Water is the most important vital source in the world. Industrial revolution introduced us with new sources of water pollution. Factories began to throw pollutants directly into water and later on Chemical waste released into water around the world.In 21st century it becomes total disastrous.Water quality testing in lab is expensive and time consuming. In this case if we can predict safe water for people with the help of technology can be used to prevent horrible diseases.With this vision of safe water our aim is to explore Supervised Machine Learning algorithms to predict safe and quality water.The overall prediction methods consist of 20 different chemical elements in a sample. Several classification methods achieves reasonable accuracy with a good precision and recall.

Keywords: Safe water prediction, water quality Data-set,Supervised Machine Learning,Classification.

1. Introduction

Water is the most important vital source.70 Percent of Earth surface is full of water. Water with poisonous chemical is a threat for all the animals that live in the Earth. According to WHO, Annually, safe water can prevent 1.4 million child deaths from diarrhoea, 500 000 deaths from malaria, 860 000 child deaths from malnutrition.
In addition, 5 million people can be protected from being seriously incapacitated from lymphatic filariasis and another 5 million from trachoma (WHO). Governments and international health organizations are trying their best to help people provide safe water. We are intended to do the same. Our goal is to predict safe water by looking at the ratio of other ingredients present in water. Water is safe until all the ingredients in water is as per limit. above that can be poisonous or dangerous. We have 7999 samples in our dataset with detail information of 20 ingredient present in water.
We have the numerical values of ingredients present in water samples.There is a standard limit for all the ingredients present in water. If the limit exceeds the water is consider as unsafe and dangerous which can cause various diseases. Limits of all ingredients are listed below,
aluminium - dangerous if greater than 2.8
ammonia - dangerous if greater than 32.5
arsenic - dangerous if greater than 0.01
barium - dangerous if greater than 2
cadmium - dangerous if greater than 0.005
Charmaine - dangerous if greater than 4
chromium - dangerous if greater than 0.1
copper - dangerous if greater than 1.3
fluoride - dangerous if greater than 1.5

bacteria - dangerous if greater than 0
viruses - dangerous if greater than 0
lead - dangerous if greater than 0.015
nitrates - dangerous if greater than 10
nitrites - dangerous if greater than 1
mercury - dangerous if greater than 0.002
Perchlorate - dangerous if greater than 56
radium - dangerous if greater than 5
selenium - dangerous if greater than 0.5
silver - dangerous if greater than 0.1
uranium - dangerous if greater than 0.3
The main contribution of the study was,
At first analysis was performed on the available data to understand the data. To check if there is any null value.and therefore few other data visualization techniques were applied to visualize the data.
After cleaning and processing the data properly a series of classification algorithms were applied. With different algorithms accuracy will be different. we have applied few algorithms and among them prioritize the best one.

2. Methods and Materials

The Dataset was obtained from open source Kaggle.It contains 7999 samples with 20 different ingredients present in water. This dataset was created to use only for educational purpose. We have performed different data visualization methods to understand the dataset like Dist-plot,Hist-plot,Scatter-plot,Heat-map,Box-plot. Then we have applied classification methods like, Decision Tree,KNN,SVC,Random Forest.
This study explores the methodologies that have been employed to help solve problems related to water quality.
We have used python because it is ideal programming language for Data Analysis. Alongside we have used Anaconda navigator and Jupyter notebook. As it was a team work we have used GitHub for code management and tracking.It is very much suitable for team work.

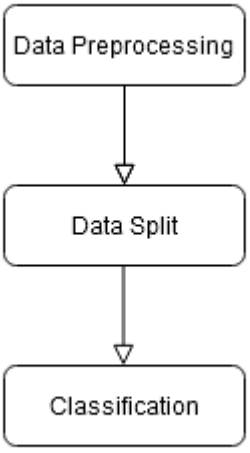


Figure 1: Work Flow

3. Data Prepossessing

Few parameters varied enough and were on the higher values . We perform Box-plot analysis on the data and normalize few of its parameters that have bigger values and convert them into decimals so that We get good result from KNN(Euclidean Distance). Among the 7999 instances in the dataset we checked if there is any null values or not . There were no null values in the dataset.
Everyone knows about arsenic problem in earlier days in Bangladesh. Still many other people of different countries are facing arsenic related diseases. According to standard arsenic level, Water is dangerous if it is more than 0.01 mg/l. We have visualize arsenic to understand how much it is present in our data. We can clearly see in the graph that we have high rate of instances vary from 0.00-0.1 .

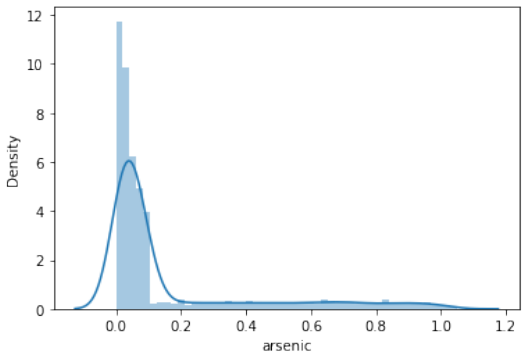


Figure 2: Arsenic

Bacteria and virus graphs are,

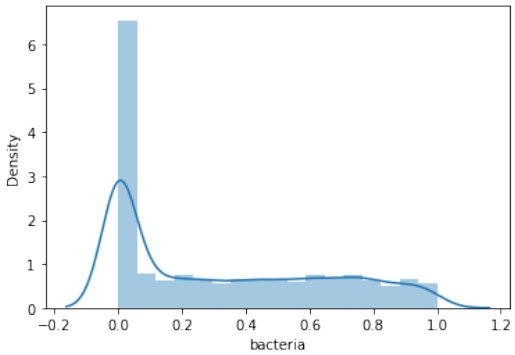


Figure 3: Bacteria

3.1. Data Analysis

After all the data cleaning and processing, for data analysis we have applied some data analysis methods to predict safe water. before applying Machine Learning algorithm there are few steps to prepare the data for applying Machine Learning algorithm like correlation analysis, data splitting.

Correlation: There are no correlation between the features. so we cant minimize any feature. There is no dependent variable. We applied Pearson correlation method to extract any possible relation or dependency between parameters but there were none more than 60 percent.

3.2. Data Splitting and Cross Validation

Now its time to split the data into 2 different part one is train part to build a model and test the other part and measure the accuracy. This way we establish models performance. We can apply different cross validation method.
First we will apply Holdout Validation. 70 percent train set and rest 30 percent test set. after splitting we will measure the performance.
But we can see that we have a imbalanced dataset.

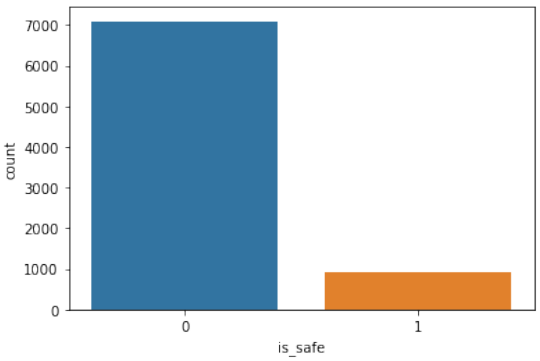


Figure 4: Imbalanced Dataset

So, Stratified K Fold Cross Validation will be best for this dataset.Here 1 represent safe water and 0 represent unsafe water. this is our target to predict by seeing all the ingredients that our water is safe or not.

4. Machine Learning algorithms

We have used Classification algorithms. Classification algorithms are used to predict the water quality. Samples are predefined with two quality class. Safe and not

safe. Our Classification algorithms will predict how accurate our model is behaving. We have used four classification algorithms.

4.1. Decision Tree

The Decision tree algorithms is a supervised machine learning algorithms. Here all rel event input parameters take part to make a decision. All the parameter decisions are arranged in a top down tree. then according to the tree decisions are made. Using Stratified K Fold Validation our anerage accuracy on Decision Tree is 0.8613652934363941 +/- 0.12910874383907842,

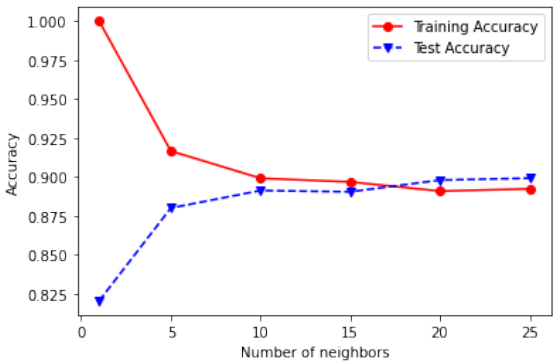
Dataset	Accuracy
SKFold1	0.55293088
SKFold2	0.90463692
SKFold3	0.9720035
SKFold4	0.88538933
SKFold5	0.93088364
SKFold6	0.90105079
SKFold7	0.882662

4.2. K Nearest Neighbor

K Nearest Neighbor algorithm is simple supervised machine learning algorithm. Can be used for both classification and regression problems. Here we have used for classification.

This algorithm classify by finding points nearest N neighbor. It computes the nearest neighbor each time. We have large values in our dataset that's why using Euclidean Distance rules we have normalized those features. It helps us to get better result with KNN. As distance values converted into decimals.

This is our result with KNN classification algorithm,



4.3. Random Forest

A random Forest Algorithm consist of many decision trees. We have used this model also to visualize the accuracy on our dataset. And we got better result than decision tree algorithm, Average result is, 0.8853617465942852 +/- 0.10962738528112591 as we have used stratified splitting method.

Dataset	Result
SKFold1	0.62729659
SKFold2	0.91951006
SKFold3	0.97287839
SKFold4	0.93350831
SKFold5	0.96500437
SKFold6	0.88791594
SKFold7	0.89141856

5. Results and Confusion Matrix

As our dataset is imbalanced we have calculate the confusion matrix to and calculated precision, recall and f1-Score. We have good precision with 100 percent Recall. That means there is no false negative.

$$\begin{bmatrix} 2140 & 0 \\ 260 & 0 \end{bmatrix}$$

Classification Report	
Accuracy	0.891
Precision	0.892
Recall	1.000
F1-Score	0.943

5.1. Accuracy

Accuracy is the correct number of prediction.made by the model.Accuracy is measured by this equation(1),

$$ACCURACY = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Here, TP refers to true Positive, TN refers to True Negative, FP FN refers to False Positive and False Negative.

5.2. Precision

Precision is the portion of correctly classified positive classes out of total positive predicted class. Formula for precision is in equation(2),

$$PRECISION = \frac{TP}{TP + FP} \quad (2)$$

5.3. Recall

Recall is the proportion of actual positive class that were classified correctly,We measure the error by measuring Recall that means if there is few instances that actually positive but predicted as Negative.The formula is,

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

5.4. F1-Score

Precision And recall does not cover all aspects of accuracy that's why we calculate F1-Score, the harmonic mean of Precision and Recall.

$$F1 - Score = \frac{2(Precision * Recall)}{Precision + Recall} \quad (4)$$

References

[1] Drinking Water Quality Guidelines By WHO(World Health Organization)

[2] Developing drinking-water quality regulations and standards,ISBN: 978-92-4-151394-4, 2018

[3] Water Quality - Dataset for water quality Classification

[4] Liaw, A.; Wiener, M. Classification and regression by randomForest. R News 2002, 2, 18–22.

[5] Water quality Parameters, Nayla Hassan Omer, October 16th 2019