Data Memo

PSTAT 131: STATISTICAL MACHINE LEARNING

Aarti Garaye

2025-04-12

Introduction to the Data Set

A healthy, avergage adult drinks between 2.7 liters to 3.7 liters of water every day. The exact amount of water needed can vary depending on factors like activity level, climate, and overall health. However, there is no doubt that water is one of the very essential resource to us. It is vital to know the quality of water we are consuming. This is why my final project will be built on the dataset that will help determine whether the water is potable or not.

Source

The author of the dataset is Laksika Tharmalingam who posted the Water Quality and Potability dataset on Kaggle. The data is sourced from GitHub.

I will be working on this dataset in R after seeing whether it has previously been cleaned or not. From the first glance, it looks pretty neat and the only missing values are in the pH column of the dataset.

Familiarizing with the Dataset

This dataset contains water quality measurements and assessments related to potability, which is the suitability of water for human consumption. Each row in the dataset represents a water sample with specific attributes, and the "Potability" column indicates whether the water is suitable for consumption.

Columns

pH: The pH level of the water.

Hardness: Water hardness, a measure of mineral content.

Solids: Total dissolved solids in the water.

Chloramines: Chloramines concentration in the water.

Sulfate: Sulfate concentration in the water.

Confuctivity: Electrical conductivity of the water.

Organic Carbon: Organic carbon content in the water.

Trihalomethanes: Trihalomethanes concentration in the water.

Turbidity: Turbidity level, a measure of water clarity.

Potability: Target variable; indicates water potability with values 1 (potable) and 0 (not potable).

Observations and Predictors

Let's load the dataset and check how many rowns does it have since that is the number of observations in that dataset. This will be an example of a classification problem since my response variable is whether the water is potable or not. So `Potability` is my response and that means every other column is a predictor variables. I have nine predictor variables" `ph`, `Hardness`, `Solids`, `Chloramines`, `Sulfate`, `Conductivity`, `Organic_carbon`, `Trihalomethanes`, and `Turbidity`.

```r
library(readxl)
library(knitr)
water <- read_excel("water_potability_excel.xlsx")
nrow(water)
```

```
## [1] 3276
```

Thus, there are 3276 observations in the dataset.

Types of Variables and Missing Data

Below is a preview of the uncleaned just loaded dataset that has the first five observations

```r
head(water, 5)
```

```
## # A tibble: 5 x 10
##   ph             Hardness Solids Chloramines Sulfate Conductivity Organic_carbon
##   <chr>          <chr>    <chr>  <chr>       <chr>   <chr>        <chr>
## 1 <NA>           204.890~ 20791~ 7.30021187~ 368.51~ 564.3086541~ 10.3797830780~
## 2 3.71608007538~ 129.422~ 18630~ 6.63524588~ <NA>    592.8853591~ 15.1800131163~
## 3 8.09912418929~ 224.236~ 19909~ 9.27588360~ <NA>    418.6062130~ 16.8686369295~
## 4 8.31676588421~ 214.373~ 22018~ 8.05933237~ 356.88~ 363.2665161~ 18.4365244954~
## 5 9.09222345629~ 181.101~ 17978~ 6.54659997~ 310.13~ 398.4108133~ 11.5582794434~
## # i 3 more variables: Trihalomethanes <chr>, Turbidity <chr>, Potability <dbl>
```

As we can see there are a few observations that have missing values in some of the columns in this dataset. In the future, I am planning on just eliminating the observations that have missing values, however, I have to be careful whether they are changing the entire dataset of not (meaning that they are influential points / outliers or not). Furthermore, there might be some observations where the missing values are of the variable that doesn't contribute much to the potability of the water and in that case we could include it.

As we saw in the preview all of the predictor variables are numeric and the response variable is a boolean which means it's either a success (potable) or a failure (not potable) i.e. it takes a true or false value, in this case it's 1 or 0.

Research Questions

The main objective of this dataset is to assess and predict water potability based on water quality attributes. It can be used for evaluating the safety and suitability of water sources for human consumption, making informed decisions about water treatment, and ensuring compliance with water quality standards.

This dataset is valuable for water quality assessment, water treatment planning, and ensuring the safety of drinking water supplies. It can be utilized by water treatment plants, environmental agencies, and researchers to make data-driven decisions regarding water quality and potability.

Predicting Data and Annwering Questions

The main research question guiding this project is "Is the water safe (potable)?" Some of the sub questions that will be answered along the way of this study are, "what makes water more potable?" or "Which attribute of the water affects the change in potability the most?" and we could try extending this project so that the users can put in their location (zip code) and check whether the water they get in their neighborhood is potable of not.

A Classification Problem

Just based on this dataset, this is a classification problem as my predictor variable, potability is answered as either "yes, the water is potable," or as "no, it's not potable." The outcome variable takes a qualitative value, even though in the dataset it is either 1 or 0 which are numerical values. It is important to know the context behind these numerical 1 or 0 as they represent a qualitative answer making this a classification problem.

Influential Predictors

Looking at the dataset some of the predictors that I think will be most helpful would be `pH`, `haredness`, `Chloramines`, `sulfates`, and `organic carbon`. We can make a correlation matrix to check, however, we have to make sure that when we import, the columns have a correct type i.e. numeric for basic correlation, even though we know that potability is not numeric.

```r
library(readxl)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
water <- read_excel("water_potability_excel.xlsx",
    col_types = c("numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric"))
water_wo_na <- na.omit(water)

cor(water_wo_na)[, "Potability"]
```

```
##              ph        Hardness          Solids     Chloramines         Sulfate
##      0.01453004     -0.00150502      0.04067418      0.02078361     -0.01530315
##    Conductivity  Organic_carbon Trihalomethanes       Turbidity      Potability
##     -0.01549572     -0.01556703      0.00924411      0.02268240      1.00000000
```

They all show very slight correlation to Potability. The most correlation is between Sulfates and Potability of almost 4.06%. Water with more solids is very slightly more likely to be potable.No variable shows strong linear correlation with Potability. So we need to move away from a logistical regression.

Since the correlation values with `Potability` are all very weak, selecting variables just by correlation might not be helpful alone. We might need to check the multicollinearity, or do something like a random forest since it doesn't use linearity.

Goal of the Model

The main objective of this dataset is to assess and predict water potability based on water quality attributes. It aims to be a predictive model or a combination of inferential and predictive model since we would want people to avoid areas where the water is not potable and for that we need to make inferences.

Timeline

By the end of following weeks I intend to finish these tasks. Note, that these are very tentative and are subjected to change.

Week 2: Finalize and familiarize myself with the raw data

Week 4: Finish cleaning the data and performing some EDA. Read more about randome forest and other non-regression (other than logistic regression) classification machine learning models.

Week 6: Try fitting the model to the data and learn from the output. Evaluate whether the model was a good fit.

Week 8: Fit the best possible model to the data and try making inferences and start on the final project report.

Week 10: Finish the report and finalize everything. Make sure to read over and keep editing. Submit the final Project.

Questions and Concerns

Is there any advantage of doing regression problem over a classification machine learning problem? Or vice-versa?

A lot of the datasets that I have been seeing on Kaggle and UCI Machine learning datasets have been classification problem and I was wondering if there's any particular reason for it.

Some problems that I anticipate are that the predictor variables not being correlated enough with the response. I wonder if none of the other methods work, then I would just have a model that doesn't have a very good performance accuracy.

I hope this dataset works for the final project for the class, I'm quite interested especially after seeing the correlation results. I think it would be an interesting problem to solve.