141C Project Proposal

Kyle Dinh, Jayoung Kim, Melanie Lue, and Daniel Momeni

Project topic/Questions

For our project we wish to determine whether or not drinking water is safe for consumption based on a variety of factors that could easily affect our drinking water infrastructure. This would give us a baseline idea of what factors lead to more dangerous water and how much of each substance can be neglected. The results of this case study could be helpful to government agencies who can better regulate these substances in our drinking water.

We will aim to answer the following questions:

1. Which variables/factors have a higher effect on determining water quality?
2. Are there variables that don't have an effect on water quality?
3. Is there any correlation between the variables?
4. How accurate is the classification method chosen?
5. Which classification method is best in determining water potability?

Data Source

Water quality dataset

1. ph: pH of 1. water (0 to 14).
2. Hardness: Capacity of water to precipitate soap in mg/L.
3. Solids: Total dissolved solids in ppm.
4. Chloramines: Amount of Chloramines in ppm.
5. Sulfate: Amount of Sulfates dissolved in mg/L.
6. Conductivity: Electrical conductivity of water in µS/cm.
7. Organic_carbon: Amount of organic carbon in ppm.
8. Trihalomethanes: Amount of Trihalomethanes in µg/L.
9. Turbidity: Measure of light emitting property of water in NTU.
10. Potability: Indicates if water is safe for human consumption. Potable -1 and Not potable -0

https://www.kaggle.com/datasets/adityakadiwal/water-potability

Methods

Since the goal is to categorize whether water is safe to drink or not based on different conditions, it makes sense to use classification methods to solve this problem. We will consider using some of the following:

- Logistic regression
- K Nearest Neighbors
- Decision tree
- Gradient descent

Challenge

- There are some missing values in the dataset which would require us to deal with that
- There may be some collinearity problems that exist in our models
- There will not be enough time to do some background research on how this dataset is collected.