Water Potability

Is your water safe to drink?

Who could use this?

Our stakeholders for this question are generally going to be scientific communities, those interested in studying water quality, government agencies (like the EPA) and potentially private businesses that have stake in making sure water is safe.

What is “Potability”?

“Potability” is used to describe whether or not water is safe for human consumption. Our goal will be to create a model that is capable of predicting that safety based on the given data.

Our Data

- Our data is derived from Kaggle
- This data consists of multiple metrics used for determining the safety of water
- Examples of our data:
    - Solids: The “Total Dissolved Solids” in water such as Potassium, Calcium, Chlorides and Sulfates: These produce diluted color and unwanted taste.
    - Chloramines: These are the major disinfectants used in public water. Most commonly formed when Ammonia is added to treat the water.
    - Sulfates: Naturally occurring substances found in minerals, soil, and rocks.
    - Conductivity: Pure water is not a good conductor of electricity. Therefore we study this as a representation of it’s impurities.
    - Organic Carbon: Organic Carbons are produced from natural decaying material, and some synthetic sources.

About our model

- After testing a couple different prediction models, I came to the conclusion that we would use a “KNearestNeighbors” model. This model makes predictions, basically, by examining the types of, and how many points are nearest our target.
- Strengths: This model type is flexible. It can be used for both binary classification like we see here, potable or non-potable, as well as “Multi-class” classification, so it’s possible with testing this model may perform well if science determines a new way to classify water. Our Model ended up with the ability to accurately predict potability 66% of the time. Of our other models it also was able to most reliably avoid “False Positive” predictions. Which considering the cost, is very important.
- Weaknesses: In this case, unfortunately, one of our model’s strengths is one of it’s weaknesses. Considering the subject matter, a 66% accuracy is not really something I’d feel comfortable putting into production. Additionally, when dealing with very large amounts of data these types of models can be very slow to make their predictions. With the need to test many samples of water across large bodies of land for the sake of drinking water for people, this may prove to be an issue.

Suggestions

This prediction model seems very promising in the respect that it is in its early stages. With more types of data, like maybe location, date-time data, or even additional data regarding the individual types of some of our available metrics (for instance sodium or magnesium deposit levels individually) I feel that it could be a contender, potentially, to be used in real world practice.

End

Thank you for all of your time!