Water Quality Predictions

# **Adonis Shareef Fall Semester https://github.com/adonisShareef/DataScience**

# Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.

My domain is the going to be on water quality which is something I was thinking about after watching the movie Signs when the little girl had an issue with the quality of the water, but the water was coming from the same sink. Living in the Pacific Northwest I live around pretty good water quality so seeing how water is determined at that quality I want to be able to predict the water quality.

1. <https://www.kaggle.com/adityakadiwal/water-potability>
2. <https://www.kaggle.com/mssmartypants/water-quality>
3. <https://www.hindawi.com/journals/abb/2020/6659314/>
4. <https://iwaponline.com/wqrj/article/53/1/3/38171/Water-quality-prediction-using-machine-learning>
5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7787777/>
6. <https://link.springer.com/article/10.1007/s11356-021-13875-w>
7. <https://www.youtube.com/watch?v=MWLUtTlHxpw>
8. <https://www.stantec.com/en/projects/united-states-projects/m/machine-learning-water-quality-predictions>
9. <https://thecleverprogrammer.com/2021/08/19/water-quality-analysis/>
10. <https://dl.acm.org/doi/abs/10.1145/3471287.3471308>

# Which Data?

What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

1. <https://www.kaggle.com/adityakadiwal/water-potability>

* This data has drinking water features and the ability to make predictions off of this data set will help me use features like ph levels to make these predictions.

1. <https://www.kaggle.com/mssmartypants/water-quality>

* This data actually has a column I can use to match against my predictions, so I know if the water is safe or not to drink

# Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.

I want to be able to predict from water samples if the quality is good enough for humans to drink from this water.

This is a fun project because it has serious real-world applications that can save lives and I am sure top researchers constantly do even in space when they are looking for life on other planets and also to see what the water on Earth is being impacted over time and if it is still safe for us to drink.

# What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

It looks like a good approach to this project will be using PCA and sklearn classifiers to make predictions and see analyze the data with plenty of descriptions because there is so much data here and a lot of cool looking features that I see on my dataset already to use.

# Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

An issue I can see is not using the correct model, but this project has been done before so going through some other projects can help me see how other have gotten higher accuracy when it comes to predicting the quality of water.

# Concluding Remarks

Tie it all together. Think of this section as your final report’s abstract.

I have some initial predictions that the accuracy will be within 98% because I should be able to tell if the water is not good for drinking well and the samples that aren’t safe for drinking should be easy to identify because having metals in the water is the first thing to come to mind. I am looking forward to starting this project and seeing what I can make out of this because this one is not as much for fun but running EDA on something that I think can really be used to help people in the world.