DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Milestone 2: Prediction of Water Quality
Author: Jake Meyer
Date: 09/11/2022

Introduction

Water is a key factor for human survival. Although water may seem abundant since it makes up roughly 75% of the Earth's surface, only 0.78% of this resource is accessible for human consumption (Misachi, 2018). According to the article posted by World Atlas (2018), an estimated 2.5% of water on Earth is non-saline, freshwater. Only a fraction of this freshwater, roughly 31%, is accessible. The remaining freshwater is currently frozen in glaciers, ice sheets, or ice caps. Two common sources for freshwater are groundwater (underground aquifers) and surface water such as rivers and lakes (BOSAQ, 2020). Other potential sources for potable water are rain/snow, wastewater, and saltwater processed through desalination systems. The rain/snow fall is essential for helping to replenish the groundwater and surface water, but there are also systems in place to harvest this rainfall/snowfall in a more efficient manner. The wastewater requires extensive filtration to remove contaminants, such as fecal matter, and may be used for irrigation purposes as well. The desalination performed on saltwater is currently high in cost, however this may be a suitable path due to the plethora of saltwater available on Earth. In the United States, a common source of water is considered tap water in which water is pulled from a centralized water supply (BOSAQ, 2020). Unfortunately, about 30% of the world population do not have access to potable water. This is most common in third world or developing countries.

The freshwater availability is currently limited. The costs associated with setting up and managing water filtration and sanitation systems differ depending on the water source. The objective of this project is to construct a model to predict whether water is potable based on water quality measurements. This predictive model is beneficial for multiple reasons. First, a water sanitation company could benefit by testing the potability of the water through certain stages of the sanitation process. By testing the water at different stages and inputting the data into this model, a company may find they are able to eliminate non-value-added steps within their process. Second, an organization such as the Environmental Protection Agency (EPA) may benefit from this model through the testing of raw water sources. Additional bodies of water or water collection systems may be identified and tested to distinguish water sources either approved for drinking or requiring minimal sanitation. Third, individuals will benefit from a potable water prediction model since it will help reduce the consumption of non-potable water. Some example diseases transferred from non-potable water are diarrhea, polio, typhoid, and cholera. According to the World Health Organization (2022), each year there are 829,000 individuals that die from diarrhea transmitted from unsafe drinking water or poor hand hygiene. Many of these individuals, roughly 35.8%, are children under the age of five. A predictive model for drinkable water will help to reduce this unfortunate statistic. Water quality is extremely important and needs to be a major focus worldwide.

The dataset for this model is from Aditya Kadiwal's post on Kaggle. The file format is Comma Separated Value (CSV) and consists of a consolidation of ten features measured across 3276 unique bodies of water. Unfortunately, there is no information available as to how the measurements were obtained, so the data will need to be explored and understood thoroughly. The ten features captured in this dataset are pH, Hardness, Total Dissolved Solids (TDS), Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability. The pH test will indicate how acidic (0-6.9), neutral (7), or basic (7.1-14) the sample measures. It serves as a useful measure for water quality since corrosive water can lead to contamination from pipes and appliances. The recommended pH value from the World Health Organization (2022) is between 6.5-8.5, but will vary depending on materials touching the surface of the water. Hardness does not contain any specific guidelines from the World Health Organization. Hardness is a measure, typically in milligrams of calcium per liter, of the volume required to react with soap. There are usually calcium or magnesium cations that contribute to the Hardness in water. Total Dissolved Solids (TSD) represent traces of organic matter and inorganic salts caused from environmental sources. The desired range for TDS is between 500-1000 milligrams per liter (Kadiwal, 2021). Chloramine is generated from chlorine reacting with ammonia and results in odd smells or taste within the water. The desired level of chlorine is below 4 milligrams per liter (Kadiwal, 2021). Sulfate in drinking water may result in stomach issues and contribute to undesired taste. According to Kadiwal, most freshwater sources have sulfate in the 3-30 milligram per liter range. Conductivity will help indicate the amount of dissolved solids in water by measuring how well the water conducts electrical current. The guidelines recommend conductivity below 400 micro-Siemens per centimeter (Kadiwal, 2021). Total Organic Carbon (TOC) results from decomposing natural organic matter and is recommended to be below 2 milligrams per liter for drinking water (Kadiwal, 2021). Trihalomethanes result from chlorine treated water and are recommended to be below 80 parts per million for drinking water (Kadiwal, 2021). Turbidity utilizes light with water to measure waste discharge. The recommended Turbidity for drinking water is below 5 Nephelometric Turbidity Units (NTU). Lastly, the Potability feature indicates whether the water sample is drinkable or not. This data will be useful to solve the problem because it includes nine features based on measurements for water samples across over 3200 bodies of water. The target variable, Potability, will be used to help evaluate how well the model performs.

There will be three models evaluated for this project. Logistic Regression, k-Nearest Neighbor, and Decision Tree models will be constructed, trained, and tested with the selected data. These models were chosen since the framework for the problem includes Supervised Learning with a binary target variable (Potable, non-Potable). Some of the evaluation metrics that will be considered will be accuracy, precision, recall, F1-Scores, and confusion matrices. The models will be compared based on these different metrics and the best model will be identified. A review of the best model will be discussed and a recommendation will be given whether it is ready for deployment. In terms of what I hope to learn, the high-level questions for this project are shown below:

- How balanced is the dataset for Potable vs. non-Potable water samples?
- Are there any insights for the feature distributions included in this dataset?
- What are the main features contributing to water Potability?
- Which model performs the best to predict water Potability?
- What are the evaluation metrics for the best performing model?
- Is the model recommended to be deployed?

There are a few risks associated with this project. The dataset chosen did not have any details regarding the way the water quality measurements were obtained or consolidated into the CSV file. The measurement method for the values within this dataset are assumed to be valid. In addition, the exploratory analysis performed on the dataset will require a thorough review of missing values, feature distributions, and outlier detection/handling. Another risk associated with this project is the inability to construct a deployable model with the training and test data available. The evaluation metrics will be illustrated to compare the three predictive models. In the event the project does not go as expected, the contingency plan will be to locate an alternate dataset pertaining to water quality. Depending on the dataset(s) available, the scope of the project may need to be modified. For example, the scope may need to be narrowed down to water quality in a particular region/country. The current plan is to utilize Python, specifically Jupyter Notebook, to perform the analysis. This project may serve as a template for domain experts in water quality, water treatment, or environmentalists to build from for additional insights.

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Milestone 2: Prediction of Water Quality
Author: Jake Meyer
Date: 09/11/2022

References

Kadiwal, Aditya. (2021, April 25). Water Quality: Drinking water potability. *Kaggle*. Water Quality | Kaggle

BOSAQ. (2020, March 18). Everything You Need To Know About Water Resources. *BOSAQ Blog.* Everything you need to know about water resources | Bosaq

Misachi, John. (2018, February). What Percentage of Earth's Water Is Drinkable? *WorldAtlas*. What Percentage of the Earth's Water Is Drinkable? - WorldAtlas

World Health Organization. (2022, March 21). Drinking-Water. *World Health Organization Newsroom.* Drinking-water (who.int)

World Health Organization. (2022, March 21). Drinking-water quality guidelines. *Water Sanitation and Health*. Water Sanitation and Health (who.int)

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Milestone 2: Prediction of Water Quality
Author: Jake Meyer
Date: 09/11/2022

Milestone 2 Requirements:

Introduction

- Problem statement
- Explain why the problem is important/interesting
- Who would be interested in solving this problem, i.e., who would you be trying to sell
- this project to?
- Where did you get your data?
- Why is this data useful to solve the problem?

Data selection and your project proposal are due this week. While you might decide to add additional data sources as the project progresses, you should have a good idea of your initial dataset by this milestone.

Milestone 2 should include the information outlined in the introduction above. Additional items to address are the following:

- What types of model or models do you plan to use and why?
- How do you plan to evaluate your results?
- What do you hope to learn?
- Assess any risks with your proposal.
- Identify a contingency plan if your original project plan does not work out.
- Include anything else you believe is important.

The proposal should be a minimum of three pages, double-spaced. You should treat this proposal as the start of your final project paper submission. But also remember this is only the initial proposal. Your findings might take you in a different direction for the final submission.

Please submit Milestone 2 in Blackboard under the group submission link.

This should be submitted through the group assignment submission regardless if it is an independent project or multi-person group.

Also, post your Milestone 2 in your Teams project folder for peer reviews.

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Milestone 2: Prediction of Water Quality
Author: Jake Meyer
Date: 09/11/2022

Water Quality Dataset

water_potability.csv

CSV Attachment -