DS 2500 Project Proposal – Tamara Hadzic

The motivation for this project came from a dilemma my friends and I were facing last year when planning a weekend trip – I was coming back to Boston after a semester abroad, and we wanted to visit one of the beaches along the coast of Massachusetts. However, due to the heavy rainfall Massachusetts had experienced, many beaches were shut down due to overgrowth of bacteria. At the time, the data on the Massachusetts government website was exclusively available as a downloadable CSV file for the current week. Since then, the website has added an interactive dashboard that displays the current beach postings on a map. However, as beaches are currently off season and past data is not displayed on the dashboard, users cannot easily look for trends in the data.

Since Massachusetts beaches are currently in their off-season, there is complete data available that can be used to track the patterns in beach closures due to bacteria for the entire 2023 season. Unfortunately, the data is not easily accessible, as the website archives past data, and does not appear to have a centralized location where it can be downloaded at once. For example, a CSV file for beach closures on 7/24 is available at https://www.mass.gov/files/csv/2023-07/BeachPostingTbl-07-24-23.csv. Tables for other dates can be found by simply changing the date in the link; however, tables were not posted daily. In this case, trying to download tables for 7/23 and 7/22 leads to a 404 error, while 7/21 is available under https://www.mass.gov/files/csv/2023-07/BeachPostingTbl-07-21-23.csv. The other challenge with accessing the data will likely be finding locations for each of the beaches. Although the downloadable data from the interactive dashboard currently lists all the beaches as "Off-season" and does not include historical data, it does include the latitude and longitude for each of the beaches, which will be extremely helpful in the data analysis.

(https://www.mass.gov/info-details/interactive-beach-water-quality-dashboard) Therefore, my first step in the project will be to find a way to easily access and load in all of the tables for the entire season, and then merge these with the full list of beaches downloaded to create a data table that includes each beach's status throughout the season.

The main data science algorithm I hope to use is a KNN classifier to predict whether beaches will be open or closed a week from the current posting. As features, I hope to use the latitude and longitude coordinates (since bacteria can likely move quicker to beaches geographically close to each other), and data from previous weeks. For this last feature, I would likely need to come up with a metric of my own, such as a score that would consider the beach's status over a set number of weeks. For example, a beach that has been closed for the past 3 weeks might be more likely to also close the following week than a beach that was closed 2 weeks ago but reopened a week ago. I would likely need to set a time limit on the amount of data considered (ex: 3 weeks) to make this less complicated.

The upside of having data available for the entire season is that the accuracy of the KNN algorithm can be evaluated by comparing the predicted beach closings to the actual beach closings. As a success metric, I will use the accuracy, precision, recall, and F1 score. I hope that through this project, I will be able to build a fairly reliable tool that would allow beachgoers to interpret trends in beach bacteria levels and make predictions about future trends.