STAT 426 **Statistical Modeling II** **Fall 2023**

Mini-Project 2

1 Instructions

Due Date: Friday Dec 8th, 11:59 pm on Canvas
Students must submit a report no longer than 10 pages with their main findings from this data analysis. Please make sure you submit your RMarkdown file and your knitted .html or .pdf file into the Canvas Website.
Please note this is a group assignment. It is highly recommended that you work with the same team members as for Mini-Project 1. PLEASE INCLUDE YOUR NAMES AND THE CONTRIBUTIONS FROM EACH TEAM MEMBER AT THE BEGINNING OF YOUR FINAL REPORT.

2 Problems

1. Data from Kaggle data repository: *Water Quality and Potability.*

   This data set contains a number of features representing water quality measurements. These measurements are associated with the suitability of water for human consumption. The primary objective of the analysis is to asses the water quality for the sampling locations, and the use of these attributes to determine the potability of water. Each row in the data set represents a water sample with specific attributes, and the "Potability" column indicates whether the water is suitable for consumption or not. A description of the data set columns from the Kaggle website follows:

   - pH: The pH level of the water.
   - Hardness: Water hardness, a measure of mineral content.
   - Solids: Total dissolved solids in the water.
   - Chloramines: Chloramines concentration in the water.
   - Sulfate: Sulfate concentration in the water.
   - Conductivity: Electrical conductivity of the water.
   - $Organic_carbon$: Organic carbon content in the water.
   - Trihalomethanes: Trihalomethanes concentration in the water.
   - Turbidity: Turbidity level, a measure of water clarity.
   - Potability: Target variable; indicates water potability with values 1 (potable) and 0 (not potable).

   Make sure you include the following analysis in your final report:

   (a) Include an exploratory analysis of this data set in order to evaluate the water quality attributes and their relationship with the drinking water status.

(b) Fit an appropriate model to this data set in order to make predictions about the potability of water given a set of measurements on water quality attributes.

(c) Select the best predictors determining the water potability. Make sure you take care about potential collinearity issues (if any), and you include appropriate model diagnostics taking into account potential unduly observations.

(d) Summarize the predictive power of the final model selected, using correlation based measures and likelihood based measures.

(e) Use classification tables to measure the predictive power of the best final model selected based on the results obtained in (c). Implement the leave-one-out cross-validation method for this step.

Note: You can use the heart disease dataset analysis posted in Canvas Module 6 as a guide for your analysis. Please note that this is not an extensive analysis but you can use the pieces of code that you need. More details about this example can be found in *Faraway, J.J. (2016): Extending the Linear Model with R, Chapter 2.* A good reference about model diagnostics can be found in *Dunn, P/K. and Smyth, G.K. (2018): Generalized Linear Models with Examples in R, Chapter 8.*