**CIND 820: Big Data Analytics Project**

*Abstract*

Sviatlana Shakibaei

501213480

Dr. Tamer Abdou

January 20, 2024

The lakes are one of the most important water resources and have been used as a source of water supply for human consumption and in general accounts for about 0.3% of the total surface water body sources. The conditions of lakes have been in constant deterioration due to increased anthropogenic activities surrounding them. In general, natural lakes are confined bodies of water lacking a strong flow for self-cleansing of its water and therefore leading to accumulation of various impurities. The determination of existing properties helps in determination of future trends of such pollutants and thereby the quality of the lake water in future scenario.

This project will research the Lake Water Quality at Drinking Water Intakes, specifically the All Ontario Great Lakes dataset. The data is obtained from Ontario Data Catalogue published by the Government of Ontario on November 30, 2020. URL: <https://files.ontario.ca/moe_mapping/downloads/2Water/GLIP/All_Lakes_GLIP.csv>

This dataset includes information on sampling locations, water chemistry and chlorophyll collected at 18 locations in the Great Lakes-St. Lawrence River and 4 locations in Lake Simcoe. In total the dataset contains 425011 rows and 14 attributes. Data range: January 1, 1976 – December 31, 2019. Last updated: November 30, 2020.

While acknowledging the significance of research across all the Great Lakes, it is evident that certain lakes may necessitate greater research emphasis due to unique challenges or environmental concerns. According to that, the dataset described above will be analyzed to answer the following question:

1. Are there significant differences in researching pollution levels, ecosystem health, and human impacts between Lake Ontario, Lake Erie, Lake Huron, and Lake Superior?

Answering this question, we will answer the following questions as well:

1.a. How many sampling locations are associated with each lake?

1.b. Analysis of the association between lake name and specific water parameter.

1.c. How many sampling locations are associated with each facility name by lake and how many facilities associated with each lake?

In addition, we will answer:

1. Identifying groups of sampling locations with similar water quality characteristics using Clustering algorithms such as K-means clustering.
2. Performing time series analysis to detect trends and patterns in the number of sampling locations over time for each lake.

To answer all the questions the following machine-learning approaches will be used: Classification machine learning such as Decision Trees, Clustering algorithms such as K-means clustering, Linear regression model. Also, the following statistical tools and techniques will be used: the chi-square test, analysis of variance(ANOVA), non-parametric tests, multivariate analysis technique (such as principal component analysis (PCA)) and visualization. All statistical analyses will be conducted by using R software.

**References**

1. Prachi Vasistha, Rajiv Ganguly, Water quality assessment of natural lakes and its importance: An overview, Materials Today: Proceedings, Volume 32, Part 4, 2020, Pages 544-552, ISSN 2214-7853, <https://doi.org/10.1016/j.matpr.2020.02.092>. (<https://www.sciencedirect.com/science/article/pii/S2214785320308373>)
2. Liu J, Zhang D, Tang Q, Xu H, Huang S, et al. (2021) Water quality assessment and source identification of the Shuangji River (China) using multivariate statistical methods. PLOS ONE 16(1): e0245525. <https://doi.org/10.1371/journal.pone.0245525>
3. Stefan G. Schreiber; Sanja Schreiber; Rajiv N. Tanna; David R. Roberts; Tim J. Arciszewski, Statistical tools for water quality assessment – a scoping review and recommendations for data analysis, Water Quality Research Journal (2022) 57 (1): 40–57. <https://doi.org/10.2166/wqrj.2022.028>