Wastewater treatment plants(WTP) are elaborate environmental techniques that are challenging to manage. As the WTP is complex, traditional laboratory methods and mathematical models have limitations in optimizing this type of operation. To overcome this issue, we will utilize machine learning and deep learning methods for WTP optimization.

For this project, we are utilizing the water treatment Dataset of the Town of Grand Falls company. It covers the years from 2009 to 2015 and 2019 to 2023. It includes data on compounds that can dissolve in water, such as aluminum and chlorine, and information about water properties like hardness, pH, and conductivity. Also, it includes measurements of each attribute at the Raw, Filter, Finish stages of water treatment. For example, pH-raw indicates level of acidity in the source water. In contrast, pH-Finish indicates level of acidity or alkalinity of treated water before it enters the urban water distribution network. According to the company's development manager, a normal value of 6.3 has been set for the pH level of treated water before it enters the distribution network.

Our team has conducted research and consulted with the company's development manager. Based on these efforts, we aim to identify the relationships between following features.

1- Color and Turbidity.

2- UVA and Turbidity.

3- UVT and Turbidity.

4- pH and Temperature.

We can use statistical methods like Pearson or Spearman rank correlation coefficients to find the correlation between the above attributes. We can also visualize the correlation using a scatter plot to see their relationship.

Furthermore, we can cluster the water samples based on their similar characteristics, which help us identify patterns/relationships for example by using K-Means and Hierarchical Clustering. It may be useful for deciding the water treatment process. K-Means can cluster the water samples based on their similar values for different attributes, such as pH, chlorine, temperature, etc. K-means clustering is also helpful in detecting anomalies, identifying trends, or making predictions. Also, Hierarchical Clustering can visualize the relationships between the different water samples in a dendrogram. The closer the samples are to each other, the more similar their attributes are.

Besides, we can classify the samples based on their attributes/characteristics to predict healthy and hazardous samples. For instans, if we want to predict based on the levels of specific attributes (e.g., pH, chlorine, temperature, color), decision-trees/random-forest may be appropriate. They can help identify important attributes for classification and optimal thresholds for healthy/hazardous water.

Otherwise, if the definition of healthy and hazardous water is complex(depends on multiple attributes), neural networks or support vector machines(SVM) may be better. They can capture intricate patterns/relationships among attributes and handle non-linear relationships between them and the class labels.