Wastewater treatment plants(WTP) are elaborate environmental techniques that are challenging to manage. It may vary slightly at different locations, depending on the technology of the plant and the water it needs to process, but the basic principles are essentially the same. 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 to provide various solutions for WTP optimization.

For this project, we are utilizing a dataset related to the water treatment of the Town of Grand Falls company, Windsor. This dataset covers the years from 2009 to 2015 and 2019 to 2023. It includes data on various compounds and chemicals that can dissolve in water, such as aluminum and chlorine, as well as information about water properties like hardness, pH, and conductivity. Also, this dataset includes measurements of each attribute at different stages of water treatment, including the Raw, Filter, Finish steps, etc.

As an example, the pH level of water in the raw stage indicates the level of acidity in the source water. In contrast, the pH level in the Finish stage indicates the level of acidity or alkalinity of the treated water before it enters the urban water distribution network. It is worth noting that a normal value of 6.3 has been set for the pH level of treated water before it enters the distribution network, according to the company's development manager.

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

1- The relationship between Color and Turbidity.

2- The relationship between UVA and Turbidity.

3-The relationship between UVT and Turbidity.

4-The relationship between pH and Temperature.

We can use statistical methods such as the Pearson correlation coefficient or the Spearman rank correlation coefficient 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. This can help us identify patterns or relationships in the data that may be useful for deciding the water treatment process. This can be done using clustering methods such as K-Means and Hierarchical Clustering. K-Means can be used to cluster the water samples based on their similar values for the different attributes, such as pH, chlorine, temperature, etc. K means clustering is also useful for detecting anomalies, identifying trends, or making predictions. Also, using Hierarchical Clustering, we 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 or characteristics to predict which sample is healthy and which is hazardous.

Different classification methods can be used for this purpose. For example, if we want to predict based on the levels of specific attributes (e.g., pH, chlorine, temperature, color), decision trees or random forest may be appropriate methods for classification. These methods can help identify the most important attributes for classification and the optimal thresholds for defining healthy and hazardous water.

On the other hand, if the definition of healthy and hazardous water is more complex and depends on multiple attributes, neural networks or support vector machines (SVM) may be more appropriate. These methods can capture complex patterns and relationships among attributes and handle non-linear relationships between the attributes and the class labels.