ChE 197/297: Intro to AI/ML for Chemical Engineers
Case Studies in ChemE

Instructions: Answer each problem then create a solution using Python code via Jupyter Notebook.

Problem: Unsupervised Learning in a Wastewater Treatment Plant Data Set

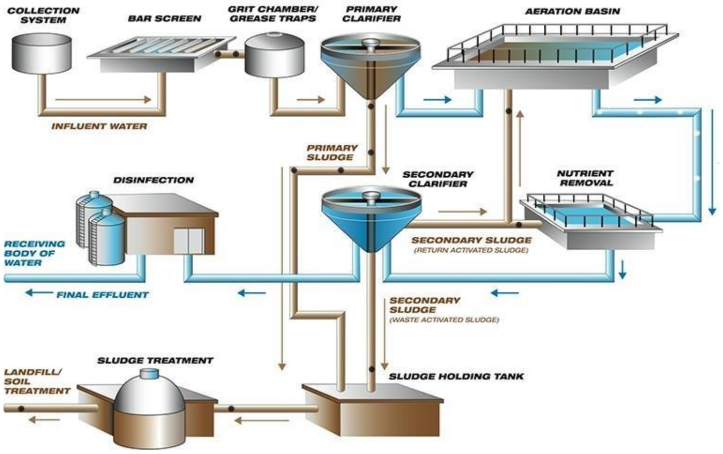

Figure 1. Typical process flow diagram of a wastewater treatment plant.

The data set for this problem contains multivariate time series collected from a wastewater treatment plant. The number of variables (features) is 38 and the number of samples is over 500, which are daily values from 1990 to 1991. The following are the names and descriptions of the variables:

Variable No.	Name	Description
1	Q-E	(input flow to plant)
2	ZN-E	(input Zinc to plant)
3	PH-E	(input pH to plant)
4	DBO-E	(input Biological demand of oxygen to plant)
5	DQO-E	(input chemical demand of oxygen to plant)
6	SS-E	(input suspended solids to plant)
7	SSV-E	(input volatile supended solids to plant)
8	SED-E	(input sediments to plant)
9	COND-E	(input conductivity to plant)
10	PH-P	(input pH to primary settler)
11	DBO-P	(input Biological demand of oxygen to primary settler)
12	SS-P	(input suspended solids to primary settler)
13	SSV-P	(input volatile supended solids to primary settler)
14	SED-P	(input sediments to primary settler)
15	COND-P	(input conductivity to primary settler)
16	PH-D	(input pH to secondary settler)
17	DBO-D	(input Biological demand of oxygen to secondary settler)
18	DQO-D	(input chemical demand of oxygen to secondary settler)
19	SS-D	(input suspended solids to secondary settler)
20	SSV-D	(input volatile supended solids to secondary settler)

21	SED-D	(input sediments to secondary settler)
22	COND-D	(input conductivity to secondary settler)
23	PH-S	(output pH)
24	DBO-S	(output Biological demand of oxygen)
25	DQO-S	(output chemical demand of oxygen)
26	SS-S	(output suspended solids)
27	SSV-S	(output volatile supended solids)
28	SED-S	(output sediments)
29	COND-S	(output conductivity)
30	RD-DBO-P	(performance input Biological demand of oxygen in primary settler)
31	RD-SS-P	(performance input suspended solids to primary settler)
32	RD-SED-P	(performance input sediments to primary settler)
33	RD-DBO-S	(performance input Biological demand of oxygen to secondary settler)
34	RD-DQO-S	(performance input chemical demand of oxygen to secondary settler)
35	RD-DBO-G	(global performance input Biological demand of oxygen)
36	RD-DQO-G	(global performance input chemical demand of oxygen)
37	RD-SS-G	(global performance input suspended solids)
38	RD-SED-G	(global performance input sediments)

Clearly, we need unsupervised learning techniques such as dimensionality reduction, clustering, density estimation, and anomaly detection to make sense of our large data set.

Do the following.

1. Remove rows with missing data, then split the data into training (1990) and testing (1991).
2. Normalize the training data to zero-mean and unit-variance. Apply the transform to the test data.
3. Use PCA to project the training data onto 2D space and 3D space, then apply the projection to the test data. What is the cumulative percent variance retained by 2D and 3D scores?
4. In the 2D PCA latent space, perform any clustering technique and make insights about the data.
5. Compute the Hotelling’s T2 statistic to the 2D PCA scores from the training data, then use KDE to calculate a detection limit at 95% confidence. Apply KDE for anomaly detection on the test data.

END OF EXERCISE