ME395 – Mechanistic Data Science for Engineering

Homework 2

Tuba Dolar

Access to safe drinking water is an essential human right and a must for health. Even though water quality is known to be improving, it is estimated that every year water approximately 1.1 million people get sick because of contaminants in drinking water (Lambertini et al., 2012). To prevent health problems and reduce the economic burden of heath care costs, it becomes crucial to define some water quality metrics and assess the water potability according to these.

Determining water potability with respect to some quality metrics starts with collecting data for different water samples. This problem relies on experimental analysis of water samples and would be considered as a type 1 problem as it is purely data driven. An open source dataset has already been identified which contains assessment of potability with respect to measurements of pH value, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity (Kadiwal, 2021). As listed, dataset includes many features as it is. Next step should be defining the features within the dataset that will be used for the water potability analysis since it is likely that the available raw data has some irrelevant portions. Based on the extracted mechanistic features, dimension reduction is necessary for narrowing data down to a few of meaningful features because there is a possibility that not all water quality metrics has an effect on the potability. After obtaining the reduced dataset, next step would be defining a reduced order surrogate model of the complex real-life problem based on scientific principles. However, water potability is assessed with experiments only and there are no scientific principles governing that problem. Thus, this step might be skipped for that purely data driven problem. Following that, regression analysis will be performed to define the relationship between water potability and selected meaningful water quality metrics. In the end, obtained regression model will be used to identify potability of new water samples based on measurements of selected quality metrics which are the inputs of the regression model (Liu et al., 2020).

References

Kadiwal, A. (2021, April 25). *Water quality*. Kaggle. Retrieved October 2, 2021, from https://www.kaggle.com/adityakadiwal/water-potability.

Lambertini, E., Borchardt, M. A., Kieke Jr, B. A., Spencer, S. K., & Loge, F. J. (2012). Risk of viral acute gastrointestinal illness from nondisinfected drinking water distribution systems. *Environmental science & technology*, *46*(17), 9299-9307.

Liu, W. K., Gan, Z., & Fleming, M. (2020). Mechanistic Data Science for Engineering Draft: An

Introduction.