**Predicting and Analysing Urban Water Quality Using IBM Watson Machine Learning Service**

**Group Members:**

**Khushi Mattu 20BCE1189**

**Anushka Bisani 20BCE1885**

**P Karthikeya 18MIS7063**

**S. Kalyan Sai 18MIS7065**

**V. Akhil 18MIS7199**

**1. Introduction:**

**a. Overview:**

In this section, we present the motivation of our method by analysing the correlations of urban water quality with its influential factors. There is an interrelationship between various influential factors for the urban water quality. Generally, those factors can be categorized as direct and indirect contributing factors. Those that have intermediate impacts on the water quality are direct contributing factors. For example, hydraulic characteristics and pipe network have direct influence on the water quality since water from the service reservoirs is distributed to users via extensive networks of water mains and pipe corrosion or flow velocities can cause water quality to deteriorate in a direct way. Hence, hydraulic characteristics and pipe network are main influencing factors for the water quality. Additionally we have the D.O levels, B.O.D levels, pH levels the affect the quality of water too. Besides direct factors, those that do not have direct influences on the water quality are considered as indirect contributing factors. Take time as an example, time plays an important role in the prediction of water quality, even if it is not directly related to it. One reason is that time is a key factor that affect human behaviour, which has an effect on the water quality by influencing the water usage patterns. Similarly, meteorology, road network, POIs are indirect contributing factors for the water quality.

**b. Purpose:**

Predicting urban water quality is very challenging due to the following reasons. First, the water quality of an area is affected by multiple complex factors, including spatial factors (e.g., pipe attributes) and temporal factors (e.g., flow and pressure). Capturing these complex factors as well as the spatio-temporal heterogeneity simultaneously is a tough challenge. Existing hydraulic models-based approaches try to model water quality from physical and chemical perspective, but such hydraulic models can hardly capture all of those complex factors. Moreover, the parameters in models are hard to get, which makes it difficult to extend to other water distribution systems. Second, as all the stations are connected through the pipeline system, the water quality among different stations are mutually correlated by several complex factors, such as attributes in pipe networks and distribution of Points of Interests (POIs). Therefore, characterizing such relatedness globally is another challenge. Traditional hydraulic models-based approaches build hydraulic models for each station and ignore their spatial correlations, and thus their performance is far from satisfactory.

To address the aforementioned issues, in this paper, we predict the water quality of an area through a data-driven perspective using a variety of data sets, including water quality data, hydraulic data, meteorology data, pipeline networks data, road networks data, and POIs.

We summarize the contributions as follows:

• We present a novel data-driven approach to co-predict the future water quality among different stations with data from multiple domains. Additionally, the approach is not restricted to urban water quality prediction, but also can be applied to other multi-locations based co-prediction problem in many other urban applications.

**Literature survey:**

1. **Existing problem:**

* Low DO: Low dissolved oxygen (DO) primarily results from excessive algae growth caused by phosphorus. Nitrogen is another nutrient that can contribute to algae growth. As the algae die and decompose, the process consumes dissolved oxygen. This can result in insufficient amounts of dissolved oxygen available for fish and other aquatic life. Die-off and decomposition of submerged plants also contributes to low dissolved oxygen. In urban areas, sources of phosphorus include discharges from municipal and private wastewater treatment, cropland and urban storm water runoff, and natural decay of vegetation. Direct discharge of pollutants from point source and nonpoint sources into a river segment add to its CBOD loadings, creating an oxygen demand that may depress DO below acceptable concentrations. Nutrient levels can in certain rivers occasionally cause sufficient eutrophication to generate CBOD loads from decaying algae. This may not occur locally, but instead farther downstream in pools where the current slows and algae collect.
* Acidic water: There are many causes of low pH water that don’t just occur in urban areas. These include natural causes like acidic rain. Soil microbes, tree roots, and some rock formations can also generate acids that cause nearby water to become acidic.
* High BOD: High biochemical oxygen demand in urban areas can be caused by:

High levels of organic pollution, caused usually by poorly treated wastewater;

High nitrate levels, which trigger high plant growth.

* E coli: The presence of faecal coliform bacteria in aquatic environments indicates that the water has been contaminated with the faecal material of man or other animals. At the time this occurred, the source water may have been contaminated by pathogens or disease producing bacteria or viruses which can also exist in faecal material. Some waterborne pathogenic diseases include typhoid fever, viral and bacterial gastroenteritis and hepatitis A. The presence of faecal contamination is an indicator that a potential health risk exists for individuals exposed to this water. In urban areas, faecal coliform bacteria may occur in ambient water as a result of the overflow of domestic sewage or nonpoint sources of human and animal waste.

1. **Proposed Solution:**

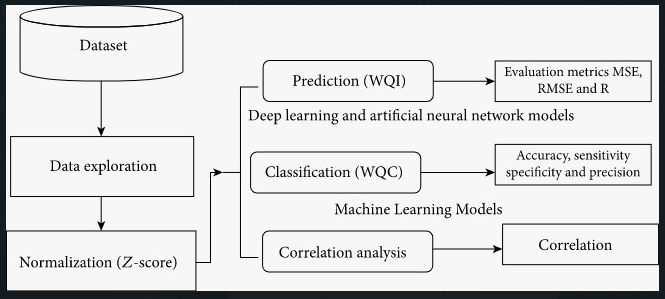
Solutions involve finding sustainable ways for the urban area to reduce both its dependence on pollutants and the amount of pollutants it produces, and to properly recycle or dispose of pollutants before they contaminate soil, water, or air. See the discussion below under "Lakes that Face this Problem" for more detailed solutions that have been tried at various lakes.

Preventing pollution in urban areas is often largely a public relations task. People need to be educated about proper ways to dispose of waste. Showing each other where waste goes and the problems it can create in our watersheds is an effective way to get the message across.

Of course, regulations are often necessary to reduce the amount of pollutants contaminating our watersheds, and the Lake Biwa Ordinance is an example of regulatory measures (such as prohibiting synthetic detergents) making a big difference.

**Theoretical analysis:**

1. **Block diagram:**

****

1. **Hardware/ Software designing:**

Minimum configuration required to complete this project:

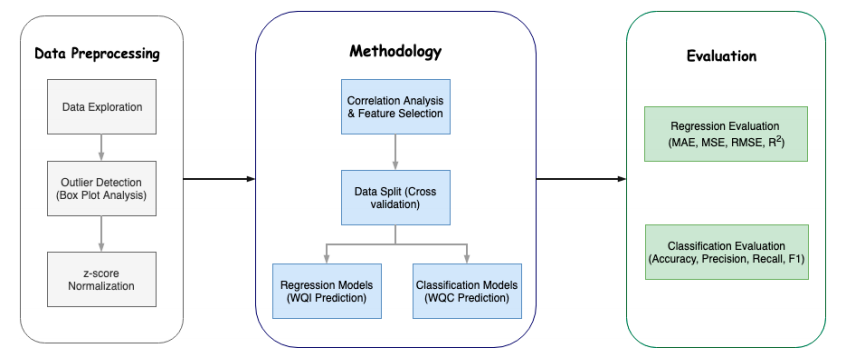
CPU: 2 x 64-bit, 2.8 GHz, 8.00 GT/s CPUs or better. Memory: minimum RAM size of 32 GB, or 16 GB RAM with 1600 MHz DDR3 installed.

In the project we make use of anaconda IDE, jupyter notebook, spyder and IBM cloud.

**Experimental investigation:**

We applied various operations on the dataset. First we checked if it has missing values. Then we changed some of the object datatypes to float datatype and then again checked for missing values. We then replaced the missing values with mean or mode depending on the distribution. We then dropped similar columns. We then calculated ph, dissolved oxygen, total coliform, B.D.O, electrical conductivity, nitra and then applied the necessary formulas. We then found the wqi using mean. We then visualised the data with the help of boxplot and scatterplot and countplot. We then fit the model using random forest and applied feature scaling. We then split the data into train and test. We then applied random forest regressor and found the score and y\_pred. We then analysed and evaluated the model. We then saved the model by importing pickle. We then build the HTML and python codes. After building the codes we apply the process of flasking via anaconda prompt. Then we move on to creating an IBM account. After creating the IBM account we train the ML model on IBM by following the steps given.

**Flow Chart:**



**Advantages and disadvantages**

**A. Advantages:**

In situ sampling must be undertaken, usually on research vessels that require substantial time, effort, and financial support in order to reach the area of interest.

Precision calibration equipment (stable reagents and standards), supporting infrastructure, such as flow systems, and interval frequency consistency are among the major issues that affect in situ monitoring.

Nitrate sensors particularly require consistent cleaning to remove biofilm or unwanted contaminants.

**B. Disadvantages:**

A microwave water level sensor has several advantages over acoustic sensors, including a higher reflectivity and lower sensitivity to variations in the air temperature and humidity.

Recently, the use of a chip less radio-frequency identification sensor, which exhibits a reliable detection with relatively low cost for water level monitoring was suggested.

In addition, wireless sensors are one of the most efficient methods for collecting field data.

Despite the prominence and widespread use of in situ sensors, there are many disadvantages that stem from the manual handling and instrumentation required to collect the samples. Both filtered and unfiltered water samples are typically collected, frozen, and analysed in the lab, using photochemical methods for total nitrogen (TN) and total phosphorus (TP).

CDT probes and plug-in devices are expensive and bulky, limiting their roles in sensitive sensor networks.

**Conclusions:**

Modelling and prediction of water quality are very important for the protection of the environment.

Developing a model by using advanced artificial intelligence algorithms can be used to measure the future water quality.

In this proposed methodology, the advanced artificial intelligence algorithms were used. The proposed models were evaluated and examined by some statistical parameters.

However, the algorithm has achieved the highest accuracy of the prediction of the WQC After examining the robustness and efficiency of the proposed model for predicting the WQI, in future work, the developed models will be implemented to predict the water quality.

Source Code:

<https://github.com/smartinternz02/SI-GuidedProject-4881-1627461954>

**References:**

Water Quality Research Journal (2018) 53 (1): 3–13.

<https://doi.org/10.2166/wqrj.2018.025>

Finlay, J. C., Small, G. E. & Sterner, R. W. Human Influences on Nitrogen Removal in Lakes. *Science* **342**, 247–250 (2013).

Lindberg, R. H., Östman, M., Olofsson, U., Grabic, R. & Fick, J. Occurrence and behaviour of 105 active pharmaceutical ingredients in sewage waters of a municipal sewer collection system. *Water Res.* **58**(3), 221–229 (2014).

Appendix

