IS509: Introduction to Data Science Term Project

Submission Deadline of Phase I Report: 04.01.2022

Submission Deadline of Final Report: 30.01.2022

Instructions – Deliverables:

- You may work in groups on this project. Each project group may comprise at most three people.
- You are required to inform us about your group information latest by **26.12.2021 in ODTUClass**.
- The requirements of the **Final Report** is specified at the end of this document.
- You are required to submit your Phase I report with your python code files in ipynb format latest by **04.01.2022 in ODTUClass**. This submission constitutes 30% of your project grade. Please keep in mind that you can make changes in the later stages of the term project.
- At the end of the term project, you have to submit a report (in pdf or Docx format) and your Python codes(in ipynb format).
- The report should be a maximum of 7 pages long and should follow the IEEE conference format.
- Your project report: Must include all the analyses you performed in detail, justifications, and comparisons. **You must explain the reasoning behind your studies.**
- Your results **must be reobtainable**. Therefore, please ensure that you used seeds or/and your codes are executable in your ipynb files.
- You are required to upload your files to the ODTU-Class with a compressed file named as "**YourID_Project.rar**".
- One submission per group is sufficient.

In Figure 1, the illustration of a part of the production process in a hypothetical chemical plant is provided. In this chemical plant, various fluid products are derived from a raw material (fluid) which consists of a set of compound substances. These compound substances have distinct densities and specific heat values (these properties affect the pressure applied by fluid flowing through the pipes and the amount of heat exchanged in heat exchangers).

The raw materials are received from five different sources (A, B, C, D, and E). The processing of a batch of the material obtained from one of the sources takes approximately two days, on average. The process involves heating the input fluid to a specific temperature and applying a series of chemical processes to derive the final products and waste from the input. It is also necessary to cool final products before storing them in tanks. Hence, heat exchangers (shell and tube type) are employed heat input and cool output products to achieve energy efficiency.

The input usually contains one or more substances that cause corrosion of the equipment used in the production. In order to minimize corrosion in the equipment, two kinds of chemicals are used: Caustic soda (referred to as "caustic") and corrosion inhibitor (referred to as "cinh"). Caustic is used to regulate the pH value of the liquid flowing in the equipment. Corrosion inhibitor (cinh) is used to prevent corrosion of the equipment. In every 2 hours (at 8:00 am, 10:00 am, 12:00 pm, 2:00 pm, …), a sample is taken from the fluid at a sampling point very close to the end of the process to measure pH and iron levels. Lab processing takes 30 minutes, and the caustic and cinh injection ratios are adjusted according to the measured pH and iron levels. The goal is to keep the pH value between 5.5 and 6.5 and the iron value less than 1 to minimize corrosion.

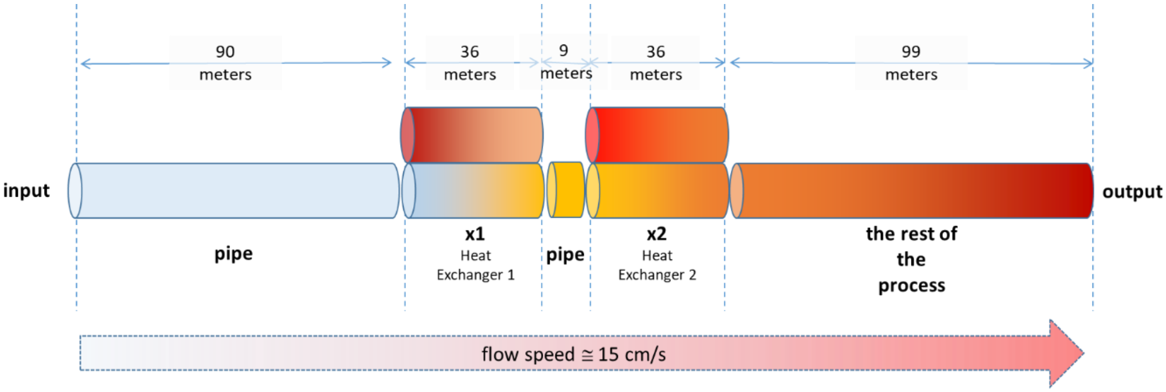

Figure 1 A part of the process and equipment used

In the current process, a simple policy outlined in Figure 2 is employed to determine the caustic and cinh injection ratios for corrosion control purposes.

Suppose

- the currently applied caustic and cinh ratios are $r_{caustic\_current}$ and $r_{cinh\_current}$, respectively.
- the measured pH and iron levels are `ph_lab` and `fe_lab`, respectively.
- the caustic and cinh ratios to apply thereafter are $r_{caustic\_next}$ and r_{cinh_next}, respectively.

Then,

```
if $5.6 \leq ph\_lab \leq 6.4$ then $r_{caustic\_next} = r_{caustic\_current}$
if $ph\_lab < 5.6$ then $r_{caustic\_next} = 1.1 * r_{caustic\_current}$
if $ph\_lab > 6.4$ then $r_{caustic\_next} = 0.9 * r_{caustic\_current}$
if $0.80 \leq fe\_lab \leq 0.95$ then $r_{cinh\_next} = r_{cinh\_current}$
if $fe\_lab < 0.80$ then $r_{cinh\_next} = 1.05 * r_{cinh\_current}$
if $fe\_lab > 0.95$ then $r_{cinh\_next} = 0.95 * r_{cinh\_current}$
```

Figure 2 The corrosion control method

The laboratory periodically measures the pH and iron levels by using the samples taken from the process. The locations of some of these sensors are shown in Figure 3. The sensors with names ending with “_t” are temperature sensors, “_p” are pressure sensors, “_f” are flow sensors, and “_v” are speed sensors.

In addition to these, pH and iron levels are measured periodically by the laboratory using the samples taken from the process. The sampling locations are indicated by the ph_lab and fe_lab labels in Figure 3. Caustic and cinh injection is performed at the source as shown in Figure 3.

The measurement units (cgs) for these are as follows:

- t: fluid temperature – °C
- p: fluid pressure – g/cm^2
- v: fluid speed – cm/s
- f: fluid flow – cm^3/s
- caustic: ratio (i.e., x/1 of the input will be caustic)
- cinh: ratio (i.e., x/1 of the input will be cinh)
- pH: varies from 0 to 14
- fe: iron level in the fluid ≥0

The sensor readings, caustic/cinh injection ratios, and pH/iron measurements at different times are available in the attached files with the names "<sensor_name | injected_chemical | lab_measurement>.csv". In addition, the source from which the raw input is received is provided in "source.csv" file.

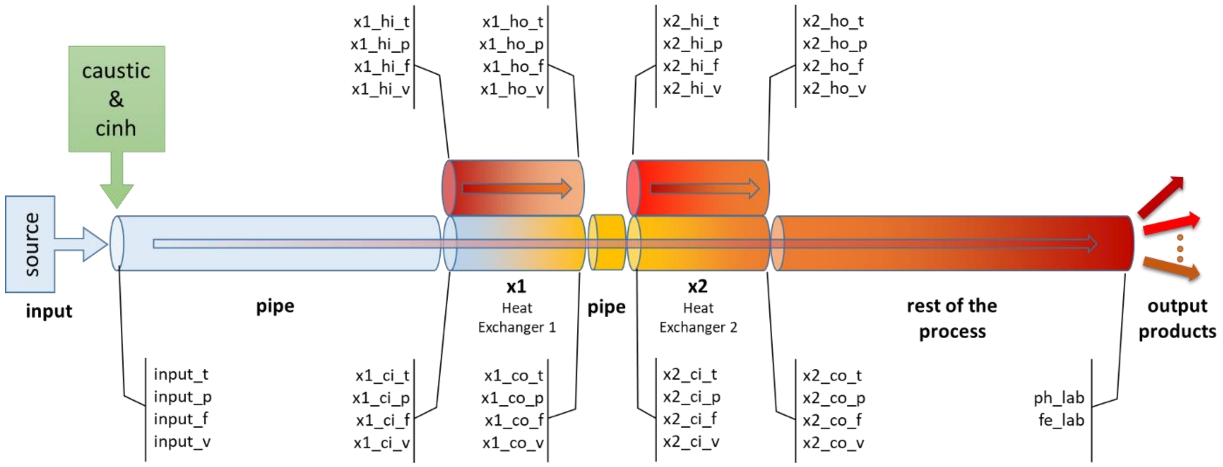

Figure 3 The input, coustic/cinh injection, sensor, lab sampling and the output locations

ci: cold in	co: cold out	hi: hot in	ho: hot out	→ hot fluid is cooled down while the cool fluid is heated.

Project – Phase I - Requirements

1. Read the sensor measurements, lab measurements, and other relevant data provided as separate files, integrate and form a dataset that will be used for further analyses. Please do not manually integrate data; instead use the python libraries pandas/numpy to construct the dataset from the input files.
2. Describe the data available in terms of descriptive statistics and other observations made via visualizations. How fast do the sensor readings, lab measurements, and other relevant data change in time? Do the changes in successive values occur rapidly or slowly?
3. Train two separate models to predict ph and iron levels in terms of sensor readings and costic/cinh injection ratios and evaluate the model performances. You are not allowed to use the provided "source" information in your models. Moreover, you are not allowed to use any time-series or similar (e.g., recurrent neural networks) analysis method. On the other hand, you might consider the time lag between the measurements and apply hyperparameter tuning when needed.

Project – Phase II – Requirements

1. A new dataset is provided in this phase, and the provided sensor readings contain some noise/errors. Read the sensor measurements, lab measurements, and other relevant data provided as separate files, integrate and form a dataset used for further analyses. Please do not manually integrate data; instead, use the python libraries pandas/numpy to construct the dataset from the input files. Do not use the data provided in the first phase in any analysis in Phase II.
2. Describe the data available using descriptive statistics and other observations made via visualizations. Characterize missing data and outliers, if there are any.
3. Preprocess data to prepare it for modeling.
4. Train two separate models to predict ph and iron levels in terms of sensor readings and coustic/cinh injection ratios and evaluate the model performances. You are not allowed to use the provided "source" information in your models. Moreover, you are not allowed to use any time-series or similar (e.g., recurrent neural networks) analysis method. On the other hand, you might consider the time lag between the measurements and apply hyperparameter tuning when needed.
5. Suppose the "source" information is not available, and even the number of sources from which raw material is received is unknown. Try to discover the number of sources and sources using the available information. Using appropriate metrics and the provided "source" information as the ground truth, evaluate your findings.

6. Assume that the unit cost of cinh is equal to twice the unit cost of caustic (i.e., $C_{cinh}=2*C_{caustic}$). Propose a new corrosion control method to replace the procedure specified in Figure 2. The new method should minimize the total cost of chemicals injected for corrosion control (i.e., the monetary cost of cinh and caustic used) while ensuring the pH value is between 5.5 and 6.5 and the iron value is less than 1. Compare the performance of the proposed method with that of the method given in Figure 2.
(Hint: Trained ph and iron level prediction models may be used to evaluate the feasibility and the cost of various injection ratios)