**2022-1 CSI4117 Data Mining: Individual Project I**

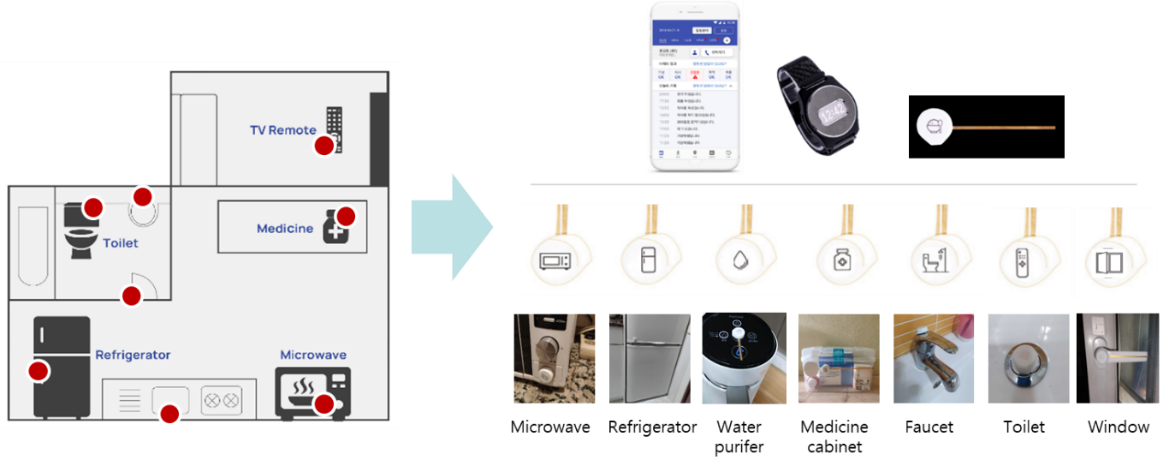
**Submission:** Upload your report at the Assignment page called ‘Data Mining Project I’ in the LearnUs course site. The report has the title of **‘DM-P1-[STUDENT ID]-[NAME].pdf‘**

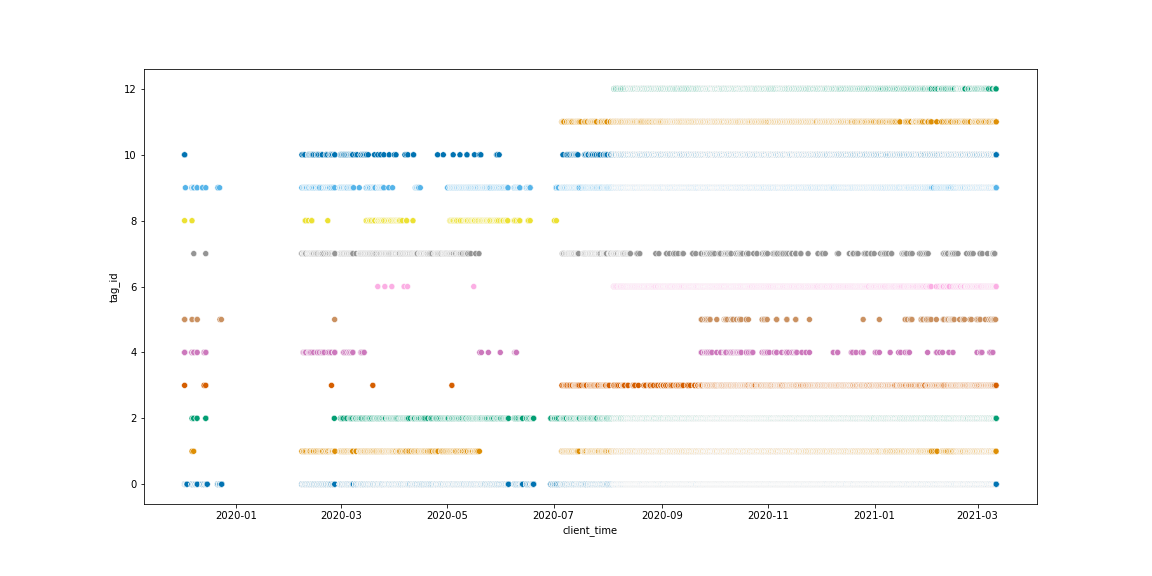
**Due Date:** 2022.04.06 (by 11:59pm)

**Dataset:** Available on the Assignment page

**0. Background**

From 2019.01.01. to 2021.03.11 (800 days), a company specializing in IoT devices has developed several sensors to collect everyday activities of 67 users for lifelogging; small sensors (tags) are installed in the house of elderly citizens who live alone, collecting various data using smartphones and smartwatches as shown in Fig. 1. Based on the lifelogging record, we encourage you to construct an automated service to support elderly citizens.

  
**Fig. 1. Sensor devices installed in a home for lifelogging.**

  
**Fig. 2. Example visualization of a user’s activities over 800 days (x-axis: time/y-axis: sensors)**

**1. Project Goals**

This project is preliminary steps to explore and preprocess the given lifelogging data for future prediction as well as identify each user’s life patterns. The given data is a lifelogging record collected from the daily life of the elderly person living alone. You need to explore raw data to propose various services and preprocess data to analyze users’ behavior or patterns from user’s daily life.

**2. Assignments**

1. Take samples to explore what types of users' lifelogging tags have. What type of data should be for each attribute? Give some sampling examples to explore the given data.

2. There are some cases that are incomplete or uncertain since it is raw data recorded in daily life. In this case, data values may not be provided. Then, how do you solve these missing values? Explain how to handle this problem with justification.

3. There are some cases that are out of normal range due to misleading tagging sensors and errors while recording users’ lifelogging. Could you find which records have to do with these cases? If it exists, explain how to handle these records with justification.

4. Tagging information is the most basic attribute that can express a user’s daily life. However, some data may be overlapped with others or be unnecessary (out of range) to predict user activities. How are you going to select the necessary variable over the data?

5. Using the inputs from multiple tags as-is can lead to the input dimension becoming too large. This can cause the models to become too complex and may hinder in extracting the information you want. How can we reduce the dimension of the input? What benefits are there in reducing the input dimension?

6. Visualize the records corresponding to users’ daily life from the preprocessed data. How could tagging information correspond to the expected daily routine of the user?

7. Visualize lifelogging patterns for each user from the preprocessed data. Do you agree that it is possible to make a new service based on various patterns you discover over data? If possible do you need any additional information to develop your idea?

**3. Data Description**

1. device\_uplink.csv (1,976,032 x 8): Payload data from each upload from sensor devices.
2. uplink\_id: relevant id with the movement dataset
3. owner\_id: id of user who owns the mobile device.
4. client\_time: when the action was recorded.
5. tag\_id: tagging device/space id.

1: refrigerator, 2: toilet, 3: bathroom, 4: front door, 5: bedroom, 6: medicine, 7: water purifier, 8: remote controller, 9: kitchen, 10: microwave, 11: window, 12: sink, 13: fitness equipment, 14: desk, 15: living room, 16: vacuum cleaner, 17: washer**~~, 18: home, 19: on-line, 20: off-line~~**

**(Note: tag 18, 19 & 20 are reserved tags intended for sensors detecting interaction between users. Since there were no valid sensors installed during the testing period, data points with these tags should be considered invalid.)**

1. step: how many steps did the user take?
2. battery\_level: remaining device battery level
3. is\_charge: is the device charged? (0: False, 1: True)
4. tag\_battery\_low: Status of tag battery life (0: high, 1: low)
5. user\_information.csv (52 x 6)
6. id: participant’s id
7. birth year
8. age: how old they are?
9. sex: gender
10. **~~depression level: depression level based on survey (missing)~~**
11. etc.: additional information of participants (medicine history, medical records, or personal information etc.)