

# Purpose

In this project, you will use a training dataset to train and test a machine model. The purpose is to distinguish between meal and no meal time series data.

# Objectives

Learners will be able to:

- Develop code to train a machine model.
- Assess the accuracy of a machine model.

# Technology Requirements

- Python 3.6 to 3.8 (do not use 3.9).
- scikit-learn==0.21.2
- pandas==0.25.1
- Python pickle

# Project Description

In this project, you will train a machine model to assess whether a person has eaten a meal or not each time a data set is provided.

## Directions

Meal data can be extracted as follows:

- From the InsulinData.csv file, search the column Y for a non NAN non zero value. This time indicates consumption time  $t_m$ . Meal data comprises a 2hr 30 min stretch of CGM data that starts from  $t_m$  to  $t_m+2\text{hrs}$ .
- No meal data comprises 2 hrs of raw data that does not have meal intake.

## Extraction: Meal data

Start of a meal can be obtained from InsulinData.csv. Search column Y for a non NAN non zero value to find the start of a meal. There can be three conditions:

1. There is no meal from time  $t_m$  to time  $t_m+2\text{hrs}$ . Then use this stretch as meal data
2. There is a meal at some time  $t_p$  in between  $t_p > t_m$  and  $t_p < t_m+2\text{hrs}$ . Ignore the meal data at time  $t_p$  and use the meal at time  $t_m$  instead.
3. There is a meal at time  $t_m+2\text{hrs}$ , then consider the stretch from  $t_m+1\text{hr } 30\text{min}$  to  $t_m+4\text{hrs}$  as meal data

## Extraction: No Meal data

Start of no meal is at time  $t_m+2\text{hrs}$  where  $t_m$  is the start of some meal. We need to obtain a 2 hr stretch of no meal data. you need to find all 2 hr stretches in a day that have no meal and do not fall within 2 hrs of the start of a meal.

## Handling missing data:

You have to carefully handle missing data. This is an important data mining step that is required for many applications. Here there are several approaches:

1. Ignore the meal or no meal data stretch if the number of missing data points in that stretch is greater than a certain threshold.
2. Use linear interpolation (not a good idea for meal data but maybe for no meal data)
3. Use polynomial regression to fill up missing data (untested in this domain).

Choose wisely.

# Feature Extraction and Selection:

You have to carefully select features from the meal time series that are discriminatory between meal

## Test Data:

The test data will be a matrix of size  $N \times 24$ , where  $N$  is the total number of tests and 24 is the size of  $N$  will have some distribution of meal and no meal data.

Note here that for meal data you are asked to obtain a 2 hr 30 min time series data, while for no meal data you are asked to obtain a 2 hr 30 min time series data. However, a machine will not take data with different lengths. Hence, in the feature extraction step, you will extract features from both meal and no meal data and they will have the same length.

## Output format:

You have to output an  $N \times 1$  vector of 1s and 0s, where if a row is determined to be meal data, then the entry will be 1, and if determined to be no meal, the corresponding entry will be 0.

- This vector should be saved in a “**Result.csv**” file.

## Given:

- Meal Data and No Meal Data of subjects 1 and 2
- Ground truth labels of Meal and No Meal for subjects 1 and 2

Using Python, train a machine model to recognize whether a sample in the training data set represents eaten (Meal), or not eaten (No Meal). The training data set contains ground truth labels of Meal and No Meal subjects.

You will need to perform the following tasks:

1. Extract features from Meal and No Meal training data set.
2. Make sure that the features are discriminatory.
3. Train a machine to recognize Meal or No Meal data.
4. Use k fold cross validation on the training data to evaluate your recognition system.
5. Write a function that takes a single test sample as input, and outputs 1 if it predicts the test sample as Meal, and 0 if it predicts test sample as No meal.

## **Submission Directions for Project Deliverables**

### **Deliverables:**

- Two python files: 1) train.py and 2) test.py
- The train.py reads CGMData.csv, CGM\_patient2.csv and InsulinData.csv, Insulin\_patient2.csv, extracts meal data, extracts features, trains your machine to recognize meal and no-meal classes, and stores the model as a pickle file (Python API pickle).
- The test.py reads test.csv which has the N x 24 matrix and outputs a Result.csv file which has N x 1 matrix of 0s and 1s, where 1 denotes meal, 0 denotes no meal.

- Assume that CGMData.csv, CGM\_patient2.csv and InsulinData.csv, Insulin\_patient2.csv files are in the current directory and execution folder. Avoid using static paths.