Project Description

For this project, you will write a program using Python that takes a dataset and performs clustering. U
training data set you will perform cluster validation to determine the amount of carbohydrates in each

There are two main parts to the process:

1. Extract features from Meal data
2. Cluster Meal data based on the amount of carbohydrates in each meal

Data:

Use the Project 1 data files:

- CGMData.csv

- InsulinData.csv

Extracting Ground Truth:

Derive the max and min value of meal intake amount from the Y column of the Insulin data. Discretiz
bins of size 20. Consider each row in the meal data matrix that you generated in Project 2. Put them
according to their meal amount label.

In total, you should have n = (max-min)/20 bins.

Performing clustering:

Use the features in your Project 2 to cluster the meal data into n clusters. Using P*30 matrix from prc
with 17 columns having velocity (min,max,mean), Acceleration(min,meax,mean), iqr, entropy, fft(1 to

1. **velocity** - 3 columns will be the "**first_differential**" and consider min, max and mean

2. **acceleration** - 3 columns will be the "**second_differential**" and consider min, max and avg

3. **Entropy** - 1 column - use the inbuilt from scipy.stats import entropy

4. **iqr** -1 column - use inbuilt function scipy.stats.iqr

5. **fft** - 6 columns - apply fft and consider the top 6 max values resulting in 6 columns.

6 **psd -** 3 columns- you can use inbuilt function scipy.signal.periodogram and take the mean of first (
psd2 10:16 for psd3

Report your accuracy of clustering based on SSE, entropy, and purity metrics.

Expected Output:

A Result.csv file which contains a 1 X 6 vector. The vector should have the following format:

SSE for Kmeans	SSE for DBSCAN	Entropy for KMeans	Entropy for DBSCAN	Purity for KMeans	Purity for DBSCAN

The Result.csv file should not have any headers, just the six values in six columns.