

General Instructions

Dataset Preparation

1. You should combine same-level datasets for different diseases. For example, you should combine **CAD/ClassCAD_train.csv**, **CKD/ClassCKD_train.csv**, **IBD/ClassIBD_train.csv**, and **T2D/ClassT2D_train.csv** files into one file.
2. You should merge datasets by row and not columns. That means if the first dataset has dimension **m1xn** and the second one has dimensions **m2xn**, the merged dataset should have dimensions **(m1+m2)xn**.
3. When merging datasets, the value *0* (0 = healthy) should be unchanged. However, you have to change the value *1* in the label column of each dataset according to the following guideline:
 - 1 = Type 2 Diabetes (T2D)
 - 2 = Irritated Bowel Disease (IBD)
 - 3 = Coronary Artery Disease (CAD)
 - 4 = Chronic Kidney Disease (CKD)

Example

Merged dataset for *Class* should look like this:

sample_ID	Bacteria;Abditibacteriota;Abditibacteria	Bacteria;Acidobacteriota;Aminicenantia	Bacteria;Acidobacteriota;Blastocatellia	...	label
H26709	0	0	0	...	0
H26781	0	0	0	...	3

As you can see, the first *sample_ID* does not have any disease, while the second *sample_ID* has CAD. The same procedure should be followed for *Family*, *Genus*, *Order*, and *Species* levels. **You can decide which level you want to use for training.** When submitting, you will tell us which level you used so that we can evaluate your model against that level.

ML Part

1. Exclude the *sample_ID* column for training. You can use all other columns.
2. You can use any model from the *scikit-learn* or the *keras* library.
3. When models are trained and ready to be evaluated by us, follow this procedure:
 1. Save the model
 1. If your model comes from the *scikit-learn* library use `joblib.dump(model_variable, 'model_name.joblib')` where *model_name* is consisted of the first names of your team. For example, if John, Anne, and Marie are in the same team, your export command should look like `joblib.dump(model_variable, 'JohnAnneMarie.joblib')`. You can find more information about the *joblib* package here: https://scikit-learn.org/stable/model_persistence.html.
 2. If your model comes from the *keras* library use `keras.Model.save('model_name.h5')` where *model_name* is consisted of the first names of your team. For example, if John, Anne, and Marie are in the same team, your export command should look like `keras.Model.save('JohnAnneMarie.h5')`. You can find more information about the *keras.Model.save* method here: https://keras.io/api/models/model_saving_apis/#save-method.
 2. Send the model
 1. You should send us the saved model (*joblib*, or *.h5*).
 2. You should send us the level you used to train your model. Please don't send us the complete dataset you created, just the name of the level. Therefore, when you write an e-mail, you should attach the exported model, and in the e-mail text body you should write one of the following *Class*, *Family*, *Genus*, *Order*, or *Species*.

General Info

The deadline for submission is the end of the week after the lectures are finished. Therefore, if lectures are completed on Wednesday, the deadline will be Friday next week at 23:59.