Explanation of approach

Task description : If a certain number of profiles are given, the model needs to classify if all profiles belong to the same patient (label='0') or not (label='1'). The dataset contains profiles of participants. But it is not well suited for the task.

Outline of my work:

1. **Directly approach**: We model the problem as multiclass classification. Perform training and testing using each profile as an individual datapoint and label as participantID. The results are not good as shown in **svm.ipynb**.
2. **Data Analysis:** As shown in **eda.ipynb,** if we aggregate features for each participant, there is strong correlation between participants. This explains why it is difficult to distinguish between participant profiles. (Additional observations are also present in the **eda.ipynb** notebook)
3. **Dataset remodeling:** In the setting of the given task, we will have multiple profiles at once and we will assign 0/1 labels. I am creating a new dataset appending multiple profiles to a single datapoint and assigning a new label based on task description.

   First split original dataset into train (80%) and test(20%). So that same profiles are not present in train and test both. For a fixed number of datapoints, we append a random number of profile features together (also selected randomly) and assign label = 0 if all are from the same participant and 1 otherwise. We produce **train_new.csv** and **test_new.csv.** Please refer to **format_dataset.ipynb** for specific details and code. The problem of data imbalance also came, which was handled by additional points with all the same participant profiles.

4. **Final method:** Since the dataset is now modeled according to the task. I consider this as a binary classification problem. After experimenting with some algorithms, I selected **gradient boosted trees** and implemented them using *xgboost* python package. For training, **logistic loss** (logloss) was used. As an evaluation metric, **F1-score** was used on the test set. Since the problem is binary classification, aforementioned selections were made.

   Confusion matrix are shown for both train and test dataset, which help us to get better insight on model learning. Finally, we receive **0.738** F1-score which is adequate for such a dataset. Please refer to **xgboost.ipynb**