Assignment-X
Deadline: 6/4/2023 at midnight

Predicting the Diagnosis/Occurrence of the Disease

Instructions. Submit the code and report. Also, you can only use this dataset for this assignment only. You cannot use this dataset for any other purposes and share with any others. Please consult the instructor if you have any questions.

1. Download the training & test dataset from here.

2. Task
   In this assignment, your task is to predict/classify the diagnosis/occurrence of the disease from the given dataset, where the column "timestamp(day)" will be the label. For example: If the input data has a label ("timestamp(day)") as "-2", the model has to predict that the data of a patient is from 2 days before the disease is diagnosed. There will be a total of 4 classes: 0, -1, -2, and (-3, -4, -5). Here you have to consider -3,-4,-5 as a single class.

3. Dataset Descriptions and Details
   a. Medical data of patients in time series format
   b. There are 37 columns (features) in the data frame.
   c. "original": 1 if there is at least one feature value at that particular timestep, otherwise NaN (You should remove the rows if original != 1).
   d. "timestamp(day)" and "timestamp(hr)": how many days(hours) before the diagnosis of the disease. "**timestamp(day)**" will be the (ground-truth) **label** for your model.
   e. Feature "1" is a gender and feature "2" is an age. The rest of the features are **continuous values**. (male=0, female=1)
   f. Columns "original" and "timestep(hrs)" are dummy (auxiliary) features; in other words, **metadata**. They can be used in your favor at your own will.
   g. The data is ordered with each patient's data **sequentially**. One patient's data follows another patient's data row-wise and so on.

(when both timestamp(hr) and timestamp(day) is 0, it is the end of a patient's record.)

4. Preprocessing
    a. **Imputation** (or interpolation) is recommended. However, be careful on how and what you impute.
    b. **Data balancing** (such as data augmentation, upsampling, or downsampling) is recommended, as these samples are not balanced (e.g. SMOTE).
    c. (Optional) However, be careful to not **leak information** from the train set to the validation set when performing cross-validation.

Part 1.

5. **EDA (Due 5/26 midnight)**

    After the preprocessing step, explain the key insights about the data and provide the answers to the following questions

    1) How many patients are there?
    2) What are the mean and median value of the label (day)?
    3) Perform EDA and calculate the statistics of the dataset: mean, std, correlations among features, etc. (e.g.There are 34 features and you have to find the correlations among each feature (34 by 34 correlation matrix)).
    4) Explain the key insight from your observation above.
    5) Perform feature engineering/selection/importance (you may remove no more than 5 features). You can employ any well-known feature engineering methods (PCA, clustering, etc.). Justify your choices and processes.

Part 2.

6. **Task formulation/Machine Learning (Due June 4 midnight)**
    a. Choose the best models to predict/classify the diagnosis/occurrence of the disease from the given dataset. The column "timestep(days)" will be the label. For example: If the input data has a label ("timestep(days)") as "-2", the model has to predict that the data of a patient is from **2 days before**

the disease is diagnosed. There will be a total of **4 classes**: 0, -1, -2, and (-3, -4, -5). Here you have to consider -3,-4,-5 as a single class

7. **Result reporting**
   Report the Accuracy, F1 scores, Recall and Precision of your approach in the table below. Plot confusion matrix and ROC curve as well (For example, the following is the baseline performance with and without SMOTE method.)

	Accuracy	**F1 Score (macro)**	Precision	Recall
LightGBM (W/O SMOTE)	**0.44**	**0.20**	0.25	0.25
LightGBM (W SMOTE)	**0.39**	**0.43**	0.51	0.39
Your Approach				

8. **Performance Requirement**
   Your performance has to be better than the baseline result (macro F1)

9. **Final Submission**
   a. **Code:** You should follow this template to write your code. Write down the proper explanation and details in each text cell. Explain the experiments and results.
   b. **Report**: Explain all the steps above.
   c. You must make the notebook runnable sequentially and without error end-to-end.

Note: General Suggestions/Recommendations
   d. Deep neural networks are not recommended. Stick with ML algorithms as much as possible
   e. You may use multiple models and apply the ensemble methods to improve the performance.

f. Preprocessing is crucial. While imputation/balancing is important, they should preserve the data's original characteristics, which can be easily corrupted or overfit.
g. Hyperparameter tuning is very important so one can apply search.
h. Do not change the order of the columns or their name.
i. **Extra credit:** If you can predict the hr (regression) in addition to predicting day (classification) with minimum MAE and perform more reliably compared to the baseline classification model, an extra credit + pizza will be given.