Structured Data Assignment Report

NOTE:

Due to insufficient time and tight deadline I was not able to complete the 3rd problem statement for this assignment but I have completed the 1st problem statement and the report has been made below. Refer the flowchart in the 2nd page.

Problem Statement:

So, the Structured Data Assignment consists of two files 'Train.parquet' and 'Test.parquet'.

1) Train.parquet - Dataset to be used for training and to build ML model
2) Test.parquet - Dataset to be used for testing and to make predictions.

Brief Description of the Dataset:

The dataset in question contains a comprehensive collection of electronic health records belonging to patients who have been diagnosed with a specific disease. These health records comprise a detailed log of every aspect of the patients' medical history, including all diagnoses, symptoms, prescribed drug treatments, and medical tests that they have undergone. Each row represents a healthcare record/medical event for a patient and it includes a timestamp for each entry/event, thereby allowing for a chronological view of the patient's medical history.

The Data has mainly three columns:
1) Patient-Uid - Unique Alphanumeric Identifier for a patient
2) Date - Date when patient encountered the event.
3) Incident - This columns describes which event occurred on the day.

Problem 1 - The development of drugs is critical in providing therapeutic options for patients suffering from chronic and terminal illnesses. "Target Drug", in particular, is designed to enhance the patient's health and well-being without causing dependence on other medications that could potentially lead to severe and life-threatening side effects. These drugs are specifically tailored to treat a particular disease or condition, offering a more focused and effective approach to treatment, while minimising the risk of harmful reactions.
The objective in this assignment is to develop a predictive model which will predict whether a patient will be eligible*** for "Target Drug" or not in next 30 days. Knowing if the patient is eligible or not will help physician treating the patient make informed decision on the which treatments to give.
*** - A patient is considered eligible for a particular drug when they have taken their first prescription for that drug

FLOWCHART WHICH DESCRIBES MY SOLUTION FOR THIS ASSIGNMENT:

Loaded the 'Train.parquet'

- Grouped the dataset based on unique 'patient-uid' and recorded the **number of times** that patient was diagnosed with each drugs along with the TARGET DRUG.
- Record the difference of days between the first normal drug (first incident) and the last normal drug (last incident) i.e before the first prescription of TARGET DRUG for some patients.
- Recorded the **difference of days** between the last normal drug and the the first prescription of TARGET DRUG and if TARGET DRUG has not been taken by some patients then the value I recorded is 0.

Stored the resulting data in '.csv' format in a file named **'train2.csv'**

Followed the above same 3 steps in the 'Test.parquet' file and stored the results in 'test2.csv'

Did all this processes in **'001.1.ipynb'** and found that this problem statement is a **binary classification task** aimed to predict whether the patient will be eligible to take the TARGET DRUG in the next 30days.

Loaded the **'train2.csv'** in a new **jupyter notebook** named **001.ipynb** . Splitted the dataset into train and test sets, built the model using Logistic regression, Random forest classifier and XGBoost classifier.

XGBoost classifier gave the best AUC_ROC score and Accuracy. So I used this model to make predictions for the **'Patients '** in the '**test2.csv'** file.

All **11k** patients in the **'test2.csv'** file is eligible for the **'TARGET DRUG'** in the next **30 days** according to my model. I don't know whether the prediction is good!

KINDLY REFER THE JUPYTER NOTEBOOKS!!!