STRUCTURED DATA ASSIGNMENT - 001 DOCUMENTATION

DATA PREPROCESSING :

- The data is loaded with the help of a pyarrow engine.
- All the columns in the dataframe are checked for counts and unique values.
- One unique patient uid is first taken to devise the steps for preprocessing.
- The ‘Date’ column is converted as datetime datatype to enable sorting.
- Then sorted by date putting the old incident in the first and latest incident in the last.
- In the ‘Incident’ columns except for the ‘TARGET DRUG’ value, others are replaced as ‘NO TARGET DRUG’.
- Then the data frame is sorted by the Incident column such that ‘NO TARGET DRUG’ is at the top and ‘TARGET DRUG’ is at the bottom (alphabetical order).
  It is stored in a new variable to prevent any loss of data or mishappenings.
- Duplicates are dropped in the dataframe by Patient Uid and the Incident, and parameter is given as keep = last, such that only one unique Patient Uid is there for both ‘NO TARGET DRUG’ and ‘TARGET DRUG’ values.
- Again duplicates are dropped in the dataframe by the Incident, and parameter is given as keep = last, such that only ‘TARGET DRUG’ values are present.
  (For patients who have not taken the target drug, only the ‘NO TARGET DRUG’ row remains).
- Positive and negative sets are created based on the condition and stored in a list. This list is added to the dataframe.
- ALL THE ABOVE STEPS ARE APPLIED TO THE WHOLE DATAFRAME.

- The data frame is sorted based on the index to maintain the original order of the dataframe.
- The ‘Date’ column is split into year ,month, week, day, day of week columns and added to the dataframe.
- The ‘Patient-Uid’ and ‘Date’ columns are dropped.
- The ‘Incident’ column is encoded by mapping.

SPLITTING:

- The data frame is made into X and y such that X is the features and y is the target.
- It is then split into training and validation sets by train_test_split function keeping the validation set size as 0.2 of dataframe.

BALANCING:

- The target variable is checked for imbalances.
- The data was imbalanced with more than 65% being 0.

- The training data is balanced by the combination of both oversampling and undersampling.
  (Oversampling = SMOTE, Undersampling = Tomek links).

MODELLING:

- The training data is fitted on a Decision tree model.
  (This was done after the comparison of models in pycaret and the best model was chosen).

- The model is evaluated with F1 score as well as the AUROC score.

TEST DATA PREDICTION:

- The test data is loaded and preprocessed as mentioned in the above steps.
- The predictions are made and downloaded in the csv format and entered in the final submission