STRUCTURED DATA ASSIGNMENT

PROBLEM STATEMENT-1

Data Pre-processing

1.Data Jar

1.1 Data Cleaning

- Imported necessary libraries
- Load the dataset
- Checked data types and Structure of data
- Removed duplicate values
- Checked missing values

1.2 Creating Positive set and Negative set data

1.2.1 Positive Set

- If patient taken the first “Target Drug” is eligible for the next target drug based on that criteria data are split from the main dataset

1.2.2 Negative Set

- If Patient not taken the “Target Drug” is not eligible for the next target drug so, consider it as negative data set

1.2.3 Combine both positive and negative set

- Concatenated both positive set and negative set.
- Created new Target column and added 1 for positive set data and 0 for negative set data.

1.3 Splitting

- Split 70-80% of data for training data
- Split 30-20% of data for test data

1.4 Imbalance check

- Checked the data is imbalanced or not but data is balanced data so no need of under sampling and over sampling

1.5 Scaling

- Scaling is done to maintain same range for all the column

2.Task Jar

- It’s a supervised learning problem
- Target type is binary classification type
- So we can use all type of classification model
- I have used Extreme Gradient Boosting

3.Model Jar

- By using Extreme Gradient Boosting model has been developed
- Find the best learning rate by using hit & trail and cross validation
- Based on best learning model was developed

4.Evaluation Jar

- Using F1 score and Accuracy model was evaluated
- F1 score value = 0.72
- Accuracy value = 0.81

Importing Test Data

Data Pre-processing

1.Data Jar

- Test data is loaded
- Data Cleaning is done
- Made data into structure format

2.Prediction

- Based on the trained model the test data was predicted whether patient will take the Target Drug within next 30 days.

3.File Handling

- The Predicted data is formed as data frame as per the sample file then based on that file predicted data is stored as csv file