Group Details:

Group Name	SoleDS
Name	Anju Paul
Email	anjupaul14@yahoo.co.in
Country	UK
College	Anglia Ruskin University
Specialization	Data Science

Problem Description:

To understand the persistency of a drug as per the prescription given by the physician is an important question faced by the pharmaceutical companies. The problem here is to build a classification model to understand the persistency (persistent or not) of a drug for the given dataset.

Data Cleansing and transformation:

The data available in excel sheet with 69 columns and 3424 rows with a data type of each column as object. There are only 2 columns with numerical data, but the rest are available as categorical data. This was converted back to numerical by using the label encoding technique.

There are no null values present in the data, hence it does not require any special treatment for that.

Skewness:

The gender data is highly skewed towards the Male. The race data is majorly for cacussians. Ethnicity towards non-Hispanic. The age group >75 data is available more when compared to other groups. Likewise, almost all the groups have the data skewed.

Outliers:

The data in the range of 25 to 75 are considered as the normal data, and the rest of the others is considered as outliers and is removed. Ie, Datapoints outside 25% and 75% Quarters are outliers and will be removed. After that the data has been reduced to the shape 2956 rows × 69 columns

The different methods for removing the outliers are explained in the notebook and the selected method is applied on the numerical data. Also, 2 different methods to convert the data from categorical to numerical is also explained(factorize and encoding).