GROUP NAME: Birmingham DS

TEAM MEMBER DETAILS:

1. Name: Meghna Ijjapureddy
   Email: ijjapureddym@gmail.com
   Country: United Kingdom
   Collage/Company: N/A
   Specialization: Data Science
2. Name: Yashwanth Reddy Bomma Reddy
   Email: yashwanth245967@gmail.com
   Country: United Kingdom
   Collage/Company: N/A
   Specialization: Data Science

PROBLEM DESCRIPTION:

Pharmaceutical companies strive to understand how consistently patients follow the prescribed dosage and duration of a drug as recommended by physicians.

The objective is to gather insights on the factors influencing persistency and build a classification model using the provided dataset.

By assessing persistency, Pharma companies can evaluate the effectiveness of their medications and identify any issues or barriers that may hinder patients from adhering to the prescribed treatment. This information is crucial for pharmaceutical companies to develop strategies and interventions that promote better patient compliance and improve treatment outcomes.

DATA UNDERSTANDING:

File Type = .xlsx

Number of Columns = 69

Number of Rows = 3424

There are 67 Non-numerical columns and 2 numerical columns.

There are no Null values in the dataset.

Data has “unknown” values in the columns: Race, Ethnicity, Region, Ntm_Speciality, Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score and Change_Risk_Segment.

PROBLEMS IN THE DATA:

1. Number of **NA** values = Nil
2. Number of **duplicates** = Nil
3. **Outliers** are present in both the numerical columns (Dex_Freq_During_Rx and Count_Of_Risk) has a positive skewed data

APPROACHES TO THE PROBLEM:

1. Outliers will be dropped by using Inter-Quartile-Range method
2. Skewed data applying the min-max and log transformation to the numerical features