WEEK 8

Name: Eric Mpofu

Email: erickdmpofu@gmail.com

Country: South Africa

University: Nelson Mandela University

Specialization: Data Science

Problem Description

Pharmaceutical companies are encountering challenges in understanding the persistency of a drug based on physicians' prescriptions and determining the number of patients who continue with the prescribed treatment. This involves tracking various prescription details, such as the duration the drug should be taken, the interval between doses, and specific instructions like whether the drug should be taken before or after meals.

Data Understanding

1. The dataset includes 2 features with discrete values, 1 feature with continuous values, and 66 features with categorical values, including the target feature, which is the consistency flag.
2. The data contains outliers within the **Dexa_Freq_During_Rx** column. Random forest algorithm and decision trees may be used to deal with outliers effectively.
3. The count of risks column is right skewed.
4. The dataset contains missing values in the following columns: **Race**, **Ethnicity**, **Region**, **Risk_Segment_During_Rx**, **Ntm_Speciality**, **Tscore_Bucket_During_Rx**, **Change_T_Score and Change_Risk_Segment**.
5. In order to resolve the missing value problem we may drop all the missing value columns, however this will result in a lot of data being lost. Perhaps a better strategy will be to use machine learning in to generate the missing values. Another approach may be to fill in the missing data with the most popular label.