Week 8: Deliverables

Group Name: Medical Data Science

Name: Peter Abban

Email: abbanpeter12@gmail.com

Country: Hungary

Specialization: Data Science

❖ Problem Description

A critical challenge for pharmaceutical companies is **understanding and monitoring drug persistency**, which reflects the extent to which patients adhere to prescribed medication schedules over time. Poor adherence compromises treatment efficacy, increases healthcare costs, and negatively impacts patient outcomes.

To address this challenge, **ABC Pharmaceutical Company engaged an analytics firm** to develop an automated system based on some feature variables of the patients

By leveraging advanced analytics, the solution aims to enhance **patient care, support physicians in monitoring compliance, and inform strategic decisions for the company's pharmaceutical portfolio** by making predictions on whether the patient is persistent or non-persistent

❖ Dataset type for analysis: Healthcare dataset - Pharmaceutical

❖ Data Understanding:

The dataset consists of **3,424 unique patients**, each identified by a distinct patient identifier (Ptid). For every patient, there are **69 feature variables** capturing *Unique Row Id, Target Variable, Demographic, Provider Attributes, Clinical, Factors and Treatment- Factor/Disease* related attributes.

The **target variable** is *Persistence*, categorized into two classes:

- **Non-Persistent** (most frequent, *n = 2,135*)
- **Persistent**

Some Feature Contributions:

- **Race**: Four categories — *Caucasian, Other/Unknown, African American, Asian*. *Caucasians* dominate the distribution with *3,148 individuals*.
- **Ethnicity**: Two categories — *Non-Hispanic* and *Hispanic*. *Non-Hispanics* are the majority group with *3,235 individuals*.
- **Region**: Five categories — *Midwest, South, West, Northeast, Other/Unknown*. The *Midwest* has the largest representation (*1,383 individuals*).
- **Age Group (Age_Bucket)**: The most common age bracket is *75+ years*, comprising *1,439 individuals*.
- **Gender**: The dataset is highly imbalanced toward *Female* patients (*3,230 individuals*).
- **Ntm_Specialty (prescriber specialty)**: The most frequent specialty associated with treatment initiation is *General Practitioner*, accounting for *1,535 occurrences*.

❖ Data Quality Issues

The dataset exhibits several data quality challenges that require consideration prior to analysis and modeling. These include **missing values (NA), outliers, and skewness** in the distribution of certain variables.

Missing Values:
Several feature variables contain missing observations, with varying degrees of impact:

- **Ethnicity**: 91 missing values
- **Ntm_Speciality**: 310 missing values
- **Change_T_Score**: 1,497 missing values
- **Tscore_Bucket_During_Rx**: 1,497 missing values
- **Risk_Segment_During_Rx**: 1,497 missing values
- **Change_Risk_Segment**: 2,220 missing values

❖ Handling of Missing Values

Missing values in the dataset will be addressed based on the **type and proportion of missingness** for each feature:

- **Categorical variables**:
    - *Ethnicity* and *Ntm_Speciality* have a relatively small number of missing values. These will be imputed using the **mode**, representing the most frequent category, which preserves the categorical distribution without introducing bias.
- **Numerical/clinical variables**:

    Change_T_Score, Tscore_Bucket_During_Rx, Risk_Segment_During_Rx, and *Change_Risk_Segment* have a higher proportion of missing values and require more sophisticated imputation. After ensuring that these features are appropriately transformed to numerical representations (where necessary), **K-Nearest Neighbors (KNN) imputation** will be applied.

    This method predicts missing values based on the most similar instances in the dataset, leveraging patterns in the available data to provide accurate and consistent imputations. This approach balances simplicity and accuracy by applying **central tendency imputation for low-missing categorical features** and **predictive modeling techniques for higher-missing clinical variables**, ensuring minimal distortion of the dataset's overall distribution.

❖ Outliers and Skewness

Analysis of the numerical variables — **Dexa_Freq_During_Rx, Count_Of_Risks, and Age_Bucket** — reveals the presence of both outliers and skewed distributions.

Outliers

Outlier detection will be conducted using **box plot analysis**. Identified outliers will be subsequently addressed through appropriate transformation techniques to minimize their impact on the overall distribution while retaining clinically meaningful observations.

Skewness

The skewness of all numerical variables will be evaluated to assess distributional symmetry.

- For **right-skewed variables**, a **logarithmic transformation** will be applied to approximate normality.
- For **left-skewed variables**, a **squared transformation** will also be used to achieve a more balanced distribution.

These corrective measures will be implemented to reduce distributional bias, enhance statistical validity, and improve the performance of predictive models when finally being applied.