Name: Saif Amer, saifamer67@gmail.com , United States, Ohio State University, Data Science

Problem

ABC Pharma has provided a real-world dataset involving patient persistence information, demographic factors, comorbidities, risk factors, and treatment indicators. The objective is to prepare the dataset for downstream analysis and modeling by performing systematic data cleansing and transformation.

Key tasks include:

- Handling missing values in both categorical and numeric variables
- Identifying and treating outliers
- Creating derived features to support machine learning models

Missing Value Imputation

- Age_Bucket, Risk_Segment_During_Ntm, and Tscore_Bucket_During_Rx had missing values imputed using the mode, as they are categorical.
- Change_Risk_Segment contained frequent “Unknown” entries, which were replaced with a placeholder value “No Change” to maintain categorical consistency.

Outlier Detection and Treatment

- Count_Of_Risks was analyzed using the Interquartile Range (IQR) method to detect outliers. Entries outside 1.5×IQR from Q1 and Q3 were removed to reduce skewness.
- A Z-score method was applied as a secondary check for extreme values.

Feature Engineering

- The Persistency_Flag was mapped to numeric format (1 for “Persistent”, 0 for “Non-Persistent”).

- A comorbidity count was created by summing up all “Y” values across Comorb_ columns, providing a quantitative measure of patient health complexity.

- Binary columns (e.g., Gender, Region, Age_Bucket) were one-hot encoded to prepare them for modeling.

- A new feature was derived to calculate risk density, reflecting the total number of risk factors flagged across Risk_ columns.

Column Standardization

- Text fields were normalized to ensure consistent casing.

- Flag columns containing “Y”/“N” were checked for typos or mislabeling and standardized.