Name: Data Science :: Healthcare - Persistency of a drug:: Group Project

Report date: 24th April 2024

Internship Batch: LISUM30:30 Jan24 - 30 Apr 24

Version: 1.0

Data intake by: David Paul Nalumenya

Group name: None

Specialization: Data science

Email: nalumenyad@gmail.com

Company: Makerere University

Data intake reviewer: None

Data storage location:

1.0 Exploratory Data Analysis (EDA)

1.1 Unveiling Patient Characteristics

The initial EDA revealed a dataset free of significant skew or outliers, potentially due to the cleaning and transformation steps mentioned earlier. This is a positive finding, as it avoids introducing biases during model training.

To gain deeper insights, the data was strategically divided into clusters based on meaningful categories: demographics, specialties, comorbidities, and risk factors. This approach, known as feature clustering, allows for targeted analysis within each group, potentially revealing unique patterns related to drug persistency.

2.0 Descriptive Statistics

2.1 Delving into Demographics

Descriptive statistics, often obtained using functions like .describe() in Python, were employed to explore each cluster. For the demographic data (demo_data), the analysis identified a higher frequency of:

- Females
- Caucasian ethnicity
- Non-Hispanic origin

- Midwestern residence
- Age above 75 years

Interestingly, this demographic group also exhibited a higher prevalence of non-persistence. This suggests a potential association between these demographic factors and adherence behavior, warranting further investigation.

2.2 Risk Factors, Comorbidities, and Medication Patterns

Similar analysis for risk factors revealed "No" as the most frequent response, indicating a low prevalence of reported risks in the year preceding the initial NTM reaction. This finding might be further explored to understand if the absence of reported risks translates to actual low-risk patients or potential under-reporting.

Comorbidities analysis also highlighted a frequent lack of reported comorbidities associated with the specific therapy. This could be due to genuine absence of comorbidities or limitations in data capture. Investigating correlations with other features might shed light on this aspect.

General Practitioners (GPs) emerged as the most frequent prescribers, followed by the broader category of OB/GYN/Others/PCP/Unknown. This reflects the potential involvement of various specialties in managing the condition.

The analysis also revealed frequent non-use of glucocorticoids before and during NTM reactions, with similar patterns observed for prior DEXA scans, fragility fractures, and risk segment changes. These findings warrant further exploration to understand their potential influence on drug persistency.

3.0 Univariate and Bivariate Analysis

3.1 Visualizing the Data

The EDA employed univariate analysis techniques like histograms to visualize the distribution of variables within each cluster, confirming the observations from descriptive statistics. Bivariate analysis then examined the relationships between individual features and the dependent variable, "persistency_flag." The consistent observation of non-persistence in over 80% of patients across different clusters suggests a potential overall bias towards non-adherence in the data. This could

be due to inherent characteristics of the patient population or limitations in data collection regarding persistence.

4.0 Recommendations:

- Investigate potential reasons behind the high prevalence of non-persistence observed in the data.
- Explore relationships between features across different clusters to identify potential interactions influencing drug persistency.
- Consider using more advanced techniques like correlation analysis or hypothesis testing to solidify the findings from EDA.

By employing a structured EDA approach, we were able to gain valuable insights into patient characteristics, medication patterns, and potential factors associated with drug persistency. These findings will guide further feature engineering and model development in the machine learning pipeline.