Data Intake Report

Name: Data Science Project: Healthcare – Persistency of the drug
Report date: 6th May 2023
Internship Batch: LISUM19
Version:
Data intake by: HIRA FAHIM
Data intake reviewer
Data storage location: DataGlacier/week13 at hirafahim-patch-2 · hirafahim/DataGlacier (github.com)

Tabular data details: Healthcare_dataset

Total number of observations	236,256
Total number of files	1
Total number of features	69
Base format of the file	.xlsx
Size of the data	888 KB

Proposed Approach:

- There are no "Null values" in the dataset.
- Drop "Ptid" column as all values are unique.
- There are Outliers in the dataset, but we keep them to better understand the dataset in the big picture.
- Chi-Square test of independence is performed for categorical features selection.
- Dataset was imbalanced, so SMOTE technique is used to balance the data.