Group Project Deliverables Week 8

Project Name: Healthcare - Persistency of a drug
Report date: 26.09.2023
Internship Batch: LISUM24
Group: HealthData Collective
Data storage location: https://github.com/venusflytrapfairy/HealthDataCollective-Group

- **Team member's details**

Group: HealthData Collective
Group members:

Name	email	country	College/Company	Specialization
Selena	selena.elensto@gmail.com	Macedonia	-/-	Data Science
Anusha Asim	anushaasim21@yahoo.com	UAE	Regent Middle East	Data Science
Sukurat	salamsukurat@gmail.com	United Kingdom	Teesside University	Data Science
Khushboo	Khushboomasih193@gmail.com	United Kingdom	University of Aberdeen	Data Science

- **Problem description**

The problem to be solved is related to pharmaceutical companies' need to understand the persistence of drug usage among patients based on physician prescriptions. Specifically, the challenge is to automate the process of identifying whether a patient is persistent in following their prescribed therapy regimen. The primary problem is to build a classification model using machine learning to determine the persistency of patients based on various features and factors.

- **Data understanding**
  - **Data type:** tabular, 69 features: 2 numeric, 16 categorical, 50 boolean and 1 text

  - **Problems in dataset (number of NA values, outliers, skewed, etc.)**

- There are no missing values in the dataset.
- There are 22 imbalanced features, including the target feature.
- There are 6 features with "Unknown" values. In 4 of these features (Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score, and Change_Risk_Segment), 1497 values are "Unknown" - the "Unknown" in the 'Change_...' features come because of those in the 'Risk...' and 'Tscore'.
- The first column, the unique identifier for each patient, would be dropped because this does not have any mathematical interpretation and cannot contribute to the model.

- **Approach to overcome problems and why?**

- *Resampling*: Increase (oversampling) or decrease (under-sampling) the number of instances in the minority class by duplicating or generating synthetic data points. This will be used to address the problem of class imbalance in the dataset.
- *Use Different Evaluation Metrics* when evaluating ML models: Instead of accuracy that might be biased to the minority class, evaluation metrics that are more informative for imbalanced datasets are to be used. Examples: precision, recall, F1-score, ROC-AUC, or area under the precision-recall curve (AUC-PR).
- *Data Augmentation*: Augment the minority class by applying transformations or adding noise to existing data points. This can help create more diverse examples.