# *Data Science Internship*

## *Final Project:* Predicting the persistency of a drug (Healthcare)

## ***Week 8 deliverables***

**Group name:** Gold Standard Team
**Name:** Mario Rodrigues Peres
**Email:** mariorope@hotmail.com
**Country:** Brazil
**Specialization:** Data Science
**Report date:** 2nd of May 2023
**Internship Batch:** LISUM19
**Project home:** GitHub

## Contents:

# 1. Problem Description

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

With an objective to gather insights on the factors that are impacting the persistency, a classification model will be built for the given dataset.

# 2. Project Timeline

| Activity | Section | Deadline |
|---|---|---|
| Problem description, project timeline, data intake report and GitHub repository link | Week 7 | 19 Apr 2023 |
| Understanding the data and checking for problems | Week 8 | 26 Apr 2023 |
| Data cleansing and transformation | Week 9 | 2 May 2023 |
| Exploratory data analysis (EDA) and recommendations | Week 10 | 9 May 2023 |
| EDA presentation and proposed modeling technique | Week 11 | 16 May 2023 |
| Model selection and model building | Week 12 | 23 May 2023 |
| Final project report and code | Week 13 | 30 May 2023 |

# 3. Data Understanding

The dataset presents 3424 registers and 69 features. The features include the target, 2 numerical and 66 categorical variables. There is no duplicated register nor missing values.

The 'Persistency_Flag' variable is the target variable and all the others are predictors. This variable presents two categories: Persistent and Non-Persistent. There are more registers for the Non-Persistent category (62.4%).

# 4. Dataset Cleansing and Transformation

### 4.1. Inspection of NULL Values and Duplicate Rows

Inspection of Null values is done using an Auto EDA library called Pandas Profiling where we got no null values present in the data and also same goes with duplicate rows there are no duplicate rows present in it.

### 4.2. Outlier Detection and Resolution

Both numerical variables (‘Count_Of_Risks’ and ‘Dexa_Freq_During_Rx’) seems to be strongly correlated with the target, however, these variables are skewed and present outliers that need to be treated before proceeding with the modeling steps.

The Outliers were found for both numerical variables through box plot and each variable is divided into two types based on Dependent Variable Persistency_Flag(‘Persistent‘ ,’Non-Persistent’).

IQR was used for each category of each variable to deal with outliers by finding out lower and upper bounds for all four categories in two variables. To solve the presence of outliers and skewness of numerical variables, the outliers were treated according to the approaches shown below.

Approach 1: Replace outlier values by the median value(Approach by: Harshith Sakala Santhosh)

The outliers were replaced by the median value of the distribution, considering each target class. That resulted in removal of outliers in all three categories except this category (‘Dexa_Freq_During_Rx’ under Persistent category of output) could not completely avoid outliers but the value of outliers has a variance of 3 to 4 and the number of outliers compared to previous number there was a lot reduction observed.

Approach 2: Replace outlier values by the upper limit threshold(Approach by: Mario Rodrigues Peres)

The upper limit threshold for outliers was calculated considering 1.5 times of the interquartile range (IQR) above the upper quartile (Q3). In this cases there were not outliers left in the new transformed dataset.

Approach 3: Remove outliers(Approach by: Alexis Michael-Igbokwe)

Finally, as there were not too many outliers, another approach adopted by our group was related to the removal of outliers, which generated a dataset with lower number of registers.

### 4.3. Encoding of Categorical Variables

As mentioned previously, There are totally 66 categorical variables present out of 68 variables and most of them are flag based categories and others has less categories present in them. By observing all these, It was decided to use Label Encoding for all categories instead of going for one hot encoding.