Group Name: **Attack on Data**

Team Members:

Name	Email	Country	College/Company	Specialization
Armin Khayati	Northatlas@gmail.com	UAE	-	Data Science
Ezzuldin Zaky	Ezzulding.zaky@gmail.com	UAE	American University of Sharjah	Data Science
Orcun Sami Tandogan	tandogan.orcun@metu.edu.tr	Turkey	Middle East Technical University	Data Science

Problem description

One of the challenges 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.

Data understanding

What type of data you have got for analysis?

Our dataset has 69 columns, 67 of them are categorical and two of them are numerical.

What are the problems in the data (number of NA values, outliers , skewed etc)?

There are no nan values in the dataset. The main problem of it is that some columns have rare labels and values for some of the variables and maybe some outliers in the numerical variables.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

The main problem of it is that some columns have rare labels and values and therefore for columns with multiple categories we have to combine low frequency categories into one larger category and for columns with binary values like *Injectable_Experience_During_Rx*, one solution is over sampling.

Rare values also happen for the numerical variable in our dataset and one solution is binning the variable to gain more values for each bin and another is considering them as outliers and remove them.

Github Repo:

https://github.com/Arminkhayati/dataglacier_internship