



**Data Glacier**

Your Deep Learning Partner

# Data Science Intern at Data Glacier

## Week 8 Deliverables

Healthcare - Persistency of a drug

LISUM12



Name: Mariam Ali Bashandy

Individual: Solo

Email: [mariam.a.bashandy.1@gmail.com](mailto:mariam.a.bashandy.1@gmail.com)

Country: Egypt

College: German International University of Applied Science

Specialization: Data Science

Date : 19/9/2022

Batch code: LISUM12

# Table of Contents



**Data Glacier**

Your Deep Learning Partner

- 1- Project Plan
- 2- Problem statement and business understanding
- 3- Data Collection
- 4- Data Understanding
  - 4.1- Types of data
  - 4.2- Data Cleaning
    - 4.2.1- Check Missing Values
    - 4.2.2- Check Outliers
    - 4.2.3- Check skew for numeric

# Project Lifecycle along with deadlines



**Data Glacier**

Your Deep Learning Partner

Weeks	Date	Plan
Week 07	19/9/2022	Problem statement, Data Collection and Data Report
Week 08	26/9/2022	Data preprocessing
Week 09	2/10/2022	Feature engineering
Week 10	9/10/2022	Build the model
Week 11	16/10/2022	Model Result Evaluation
Week 12	23/10/2022	Flask deployment
Week 13	20/10/2022	Final submission



One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription and ABC pharma company has the same challenge and to solve this problem wants to automate the process of identification.

Objective to gather insights on the factors that are impacting the persistency and build a classification for the given dataset to automate the process.

For Business Understanding needs:

- Know drug persistency and how it calculated
- Define the business problem
- Define objective and criteria to solve it



The Data is about Healthcare which contain 69 number of features and 3424 number of observations and it was used to detect Persistent vs Non-Persistent

<b>Total number of observations</b>	3424
<b>Total number of files</b>	1
<b>Total number of features</b>	69
<b>Base Format of the file</b>	.CSV
<b>Size of the data</b>	898KB



In this part, we will explain data understanding

## Types of data

The dataset contains 69 features, we found 67 out of 69 data type are categorical and only two features were numerical (names of numeric: Dexa\_Freq\_During\_RX and Count\_Of\_Risks)

## Missing Values

No missing values found in this dataset

## Outliers

We found outliers in both two numerical attributes and may have solve them by quartile or Simple Imputer in the next step.

## Skew

We found that both numeric have skew greater than zero so will own more weight in the left