**[Project :](https://canvas.instructure.com/courses/10363240/modules/items/117435121" \o "Data Science:: Healthcare - Persistency of a drug:: Group Project)** [**[Healthcare - Persistency of a drug](https://canvas.instructure.com/courses/10363240/modules/items/117435121" \o "Data Science:: Healthcare - Persistency of a drug:: Group Project)**](https://canvas.instructure.com/courses/10363240/modules/items/117435121)

**Week 8 Deliverable**

**Batch: LISUM38**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Team: Salus AI Data Avengers** | | | | |
| **Name** | **Email** | **Country** | **Company** | **Specialisation** |
| **Amrin Shaikh** | **amrin02@gmail.com** | **United Kingdom** | **Freelance** | **Data Science** |
| **Keilor Fallas Prado** | **kfallasprado@gmail.com** | **Costa Rica** | **Immune Technology Institute Madrid** | **Data Science** |
| **Hyejoon Lee** | **candy0543@gmail.com** | **United Kingdom** | **--** | **Data Science** |
| **Raina singh** | **Rainasinghh1@gmail.com** | **United States of America** | **The Ohio State University** | **Data Analytics Specialization** |

**1. Description of the problem**

The dataset is related to the persistence of patients in medical treatments.

The objective is to analyze patterns related to treatment adherence, identifying how specific factors affect retention or persistence.

**2. Understanding the data**

* **Data types:**
  + The dataset contains 3424 rows and 69 columns.
  + Of the 69 columns there are only two numerical columns and the others are categorical.
  + The two numerical columns are: Dexa\_Freq\_During\_Rx, Count\_Of\_Risks.

**3. Data issues**

* **Missing values (NA):**
  + The dataset does not contain null values with respect to the EDA analysis that was performed.
* **Outliers:**
  + Outliers were identified using the boxplots in the numerical variables.
* **Ignorance of data:**
  + There is a lack of knowledge of some data so it can affect a more accurate analysis.
* **Unbalanced distribution:**
  + Some categorical variables have very unbalanced distributions, such as Race and Ethnicity, which could reduce their predictive importance.
* **Categorical variable coding:**   
  Categorical variables need to be transformed using techniques such as One Hot Encoding to allow their use in machine learning models.

**4. Approaches to overcoming problems**

* **Management of outliers:**  
   The limits of 1.5 \* IQR were used to identify and eliminate extreme values in the numerical variables. This ensures that the models are not biased by outlier data.
* **Categorical Data Coding:**   
  One-Hot Encoding will be implemented for categorical variables, ensuring that the model can process them correctly.
* **Evaluation of the target variable:**  
   In case of imbalance in the classes, algorithm techniques that handle class weights will be applied to balance the data set.
* **Elimination of categorical variables with low variability, which contribute little to the model.**