Deliverables – Week 08

Team member's details:

Group Name: Future shapers

Names: Mohamed Mohamed – Krishna Haripur.

Email: Mohamed.hussien155@yahoo.com ,

Country: USA & Canda.

College/Company: Houston Community College.

Specialization (Data Science, NLP, Data Analyst) : Data Science.

Problem Description:

One of the challenges for all pharmaceutical companies is to understand the persistency of drug as per the physician prescription.

Business understanding:

The health sector is one of the biggest areas that are interested in data and processing it because it is directly related to the lives of patients, whether current or future, so this data must be dealt with and analyzed in a very accurate way and not neglect any of them to get the maximum possible benefit and reduce any risk rates.

What type of data you have got for analysis:

I got an Excel file with (3424 rows × 69 columns), all the features type are object except two of them (*Dexa_Freq_During_Rx* and *Count_Of_Risks*) are integers.

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

- NA Values : No missing values.
- Outliers : No outliers.
- Duplicates : No duplicates.
- Skewed : unbalanced data. The data skewed toward a specific object in the different categories like, the number of non-persistent is highly more than the number of persistent, more than 90% of cases are females, more than 85% of cases are Caucasians,

85% of cases are non-Hispanic, most of the cases age are above 75 years. All those reasons make the data unbalanced.

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

- I'm going to use some techniques to overcome the imbalanced data like Decision tree and oversampling to get more accurate results.

Github Repo link

Midohussien/Data-Science-Healthcare---Persistency-of-a-drug-: understanding the persistency of drug as per the physician prescription, and gather insights on the factors that are impacting the persistency, build a classification for the given dataset. (github.com)