Deliverables – Week 10

Team member's details:

Group Name: Future shapers

Names: Mohamed Mohamed.

Email: Mohamed.hussien155@yahoo.com ,

Country: USA.

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.

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)

EDA Summary:

- The data in Drug_presistent.csv is tabular (2-dimensional).
- The original data set has 3424 rows and 69 columns.
- The total number of values was 236256.
- There were no duplicate rows or missing values in the original data set.
- all columns are object (string) data types, except two columns are numeric integer data
- we have four races: Caucasian, Asian, Other/Unknown, or African American.
- we have three different Ethnicity categories: Not Hispanic, Hispanic, or Unknown.
- The patients were from 5 different regions: West, Midwest, South, Northeast, or Unknown.
- The patients were set in 4 different age categories: >75, 55-65, 65-75, or < 55.
- The given data was imbalanced and skewed towards some categories:
  - 1 - the number of non-persistent is higher than the number of persistent.
  - 2 - more than 90% of cases are females.
  - 3 - more than 85% of cases are Caucasians.
  - 4 - 85% of cases are non-Hispanic.
  - 5 - most of the cases, the age are above 75 years.

All those reasons make the data imbalanced.