Healthcare - Persistency of a drug
Data Science Final Project Report

Group Name: Data Science Warrior
Name: Ugur Selim Ozen
Email: ugur_ozen58@hotmail.com
Country: Turkey
Collage/Company : Yıldız Technical University
Specialization : Data Science
Submission date: 11.15.2021
Github Repo Link:
https://github.com/UGURSELIMOZEN/Data_Glacier_DS_Internship/tree/main/DataScience_Healthcare_Final_Project

1.Problem Description

One of the challenge 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.

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

2.Data Understanding

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	.xlsx
Size of the data	1.8 MB
Null/NA Values	0

3. Data Cleaning and Transformation

3.a. Null / NA Values Problem in Dataset

Firstly, by checking null values in all features we can see there is **no null values** but now we need to focus on much more to observe any missing or unknown value assigned for null values.

3.b. Null / Missing Values Handling Approach

Secondly, this is a **classification problem** so we can **impute null or missing values generally with two approaches**; first one is **filling with most recurring value (mode)** and second one is we can **categorize the missing values with some value like 'missing' or 'unknown'**. In this dataset null or missing values were filled 'unknown' value therefore we can apply first method which is filling with mode.

Finally, I will make **filling with mode operation for only 4 columns**; **Race, Ethnicity, Region and Ntm_Speciality** because in other columns, ratio of 'Unknown' is more than %50 that means 'Unknown' itself is mode in column so it can be meaningless and not correct operation for other columns.