Team Member's Details

Name: Jessica Gallo

Email: jessica.gallo522@gmail.com

Country: United States of America

Company: Wagner College

Specialization Data Science

Problem Description

One of the challenges 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.

Data Understanding

The data given is personal data about someone's health and their demographics. The data mostly consists of Yes and No answers about specific health problems.

What type of data you have got for analysis?

The data is mostly object datatypes and two int64 datatypes (Dexa_Freq_During_Rx and Count_Of_Risks. The data is pharmaceutical data and personal data about the patient.

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

There are no NA values in this dataset. There are also no outliers in this dataset because this dataset is all categorical. The dataset also is not skewed since, again, this is all categorical data.

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

There weren't any problems in my dataset to overcome. If I had to overcome an NA problem, I could have remedied the problem by either using the mean, median or row above it to fill in the dataset. If my dataset had more numerical data, I could have used box and whisker plots and Inter Quartile Range to identify the outliers, however, depending on the data, outlier may not be an issue.

Github Repo Link

https://github.com/Gallo13/DataGlacierProject/tree/main/Week%208