Week 8 Project Deliverables

Group Name: Decent Healthcare Analytics

Name: Tolga Yaz

E-Mail: tolgayaz1991@gmail.com

Country: Turkey

Specialization: Data Science

Problem Description and Business Understanding:

- One of the challenges for pharmaceutical companies is to understand the persistency of drugs as per the physician prescription.
- To solve this problem a pharma company wants to automate this process of identification via using analytics.
- **Main Goal:** Predicting drug persistency of patients to provide better healthcare.
- **Main Objective:** Building a classification model to predict drug persistency of patients via using provided data and data science methods.

Data Understanding:

- The data provided are mostly composed of categorical features (67 of 69 features are categorical with data type of object). And only 2 features have data type of int64.
- When the data were examined, no dominant outlier problem was seen.
- When the data were examined, it was seen that there are some values named “Unknown”.
- The “Unknown” values were changed into NaN values to easily examine them via Pandas.
- The number of NaN values and the column names they are found are as below:
  ('Ethnicity', 91),
  ('Ntm_Speciality', 310),
  ('Risk_Segment_During_Rx', 1497),
  ('Tscore_Bucket_During_Rx', 1497),
  ('Change_T_Score', 1497),
  ('Change_Risk_Segment', 2229)
- We plan to use the columns with NaN values by dropping the rows because we cannot change them with other values like most-frequent due to sensitivity of the data.

	Dexa_Freq_During_Rx	Count_Of_Risks
count	3424.000000	3424.000000
mean	3.016063	1.239486
std	8.136545	1.094914
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	1.000000
75%	3.000000	2.000000
max	146.000000	7.000000

Fig. 1. Main Statistics of Numerical Variables of the Data Provided

	Dexa_Freq_During_Rx	Count_Of_Risks
Dexa_Freq_During_Rx	1.000000	0.013964
Count_Of_Risks	0.013964	1.000000

Fig. 2. Correlation of Numerical Variables of the Data Provided