Week 9 Project Deliverables

Group Name: Decent Healthcare Analytics

Name: Tolga Yaz

E-Mail: tolgayaz1991@gmail.com

Country: Turkey

Specialization: Data Science

Problem Description:

- 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.

Github Repo Link: https://github.com/tolgayaz1991/DgInternshipProjects

Data Cleansing and Transformation

Some Facts about Data Provided:

- 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.
- No necessary outliers were detected.
- There are some values named “Unknown”.
- The “Unknown” values were changed into NaN to easily examine them via Pandas.
- The number of NaN values and the column names they are found in 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)

Dealing with NaN Data:

- Since we have many columns and the data are sensitive, we dropped the columns with >200 missing values and got a new dataframe without these missing values.
- For the "Ethnicity" column, we used mode value to fill missing values.