Week 8: Data Overview

Team Name:			SAAN		
SL. No	Name	Email	Country	College / Company	Specialization
1	Aman Niyaz	amanniyaz.waplog@hotmail.com	Ireland	University of Limerick	Data Science
2	Aftab Ara	ara.aftab@gmail.com	Saudi Arabia	NA	Data Science
3	Brian Orina	brayohorina@gmail.com	Kenya	NA	Data Science
4	Niyusha Baghayi	niyusha.baghayi@gmail.com	Iran	University of Tehran	Data Science

Problem Description:

ABC is a pharmaceutical company that wants to undertsand the persistency of a drug as per the physician's prescription for a patient. This company has approached an Analytics company to automate this process of identification. This Analytics company has given responsibility to *Team SAAN* and has asked to come up with a solution to automate the persistency of a drug for the client ABC.

Data Understanding:

The healthcare dataset is considerble size data having 69 columns altogether. The label is the Persistency_Flag which is a binary data having value as True or False depending on the other features. The first column being the unique id of the patient is of no use for us as this is not going to help us in training the model. Hence the first thing we do is drop that column. We have analyzed the dataset and upon analysis we found that there are only few numerical data column and rest are either binary or string value.

Exploratory Data Analysis:

We analyzed and dataset and found that there are many binary data having "Y" and "N" values. We mapped each if the data to 1 and 0 respectively. Also, we find no null values in the dataset and hence there was no need of handling it. We checked the numerical columns and found that one of the feature is having some outliears and we handled it using log transformation.

Github Link: https://github.com/aniyaz/Healthcare-DataScience2021

Submitted by: *Aman Niyaz*

Date: *12[h] April 2021*