Week 8: Deliverables

Team member's details :

Group Name - Chandni

Name – Chandni

Email – cpatel521351@hotmail.com

Country – England

College/Company – NA

Specialization - Data Science

Problem description - ABC is a pharma company who wants to understand the persistency of drug as per the physician prescription for the patient. To solve this problem ABC pharma company approached us to automate the process of identification for the drug by building a classification machine learning model. Using a classification machine learning model, we will be able to classify the dataset and find the variables that affect the target variables "Persistency Flag".

ABC pharma company has provided us with a dataset in xlsx format which needs to be cleaned so that the machine learning model is accurate and therefore can be used to gain actionable insight.

Data understanding & what type of data do we have for analysis - ABC pharma provided us with a healthcare dataset. The have 3424 observations and 69 features. The data is labelled and the “Persistency_Flag” feature will assist as the label for the machine learning model. All the features expect “Count_Of_Risk” and “Dexa_Freq_During_Rx”) are categorial data. (“Count_Of_Risk” and “Dexa_Freq_During_Rx”) feature is numerical data). The categorial data will be converted in to numerical for the machine learning model. Features which are of no use to us such as “Patient ID” will be dropped.

What are the problems in the data – Firstly the categorial data will need to converted to integer values to create a machine learning model. There are no null values or duplicate values hence there is no need to deal with this. Count_Of_Risk and Dexa_Freq_During_Rx both have outliers and are screwed to the left.

How we will overcome the problems

Firstly, the categorial data can be converted to integer values using one hot encoding method. Outlier values can be replaced by the mean value or upper bound value. Data screwed to the left can be normalised using MinMaxScaler library.