Name: Yusuf Aishat
Email: aishatyusuf12@gmail.com
Country: Nigeria
Specialization: Data Science

PROBLEM DESCRIPTION

ABC Pharma wants to automate the identification of drug persistence as per the physician's prescription. The goal is to build a classification model to predict whether a patient will persist with their treatment and to identify the factors that influence drug persistence.

GITHUB REPOSITORY

https://github.com/aishatyusuf/drug_persistence_abc_pharma

DATA CLEANSING AND TRANSFORMATION

- **Race**: ~2.83% of entries are labeled Other/Unknown. These entries were relabeled as “Other” because there are other races outside of Caucasian, African American and Asian

- **Ethnicity**: ~2.65% were entered as Unknown. These were replaced by the mode (“Not Hispanic”) which is about 94.5%

- **Region**: ~1.75% are labeled as “Other/Unknown”. These were also replaced by the mode (“Midwest”)

- **NTM Speciality**: An entry labeled as “Obstetrics & Obstetrics & Gynecology & Obstetrics & Gynecology” seemed to be a data entry error and it was added to the “Obstetrics and Gynecology” category.

- **Risk_Segment_During_Rx, Change_Risk_Segment, Tscore_Bucket_During_Rx, Change_T_Score**:
  These columns have >40% of their entries as Unknown, as they were not adding much information to the data, they were dropped.

At the end of the cleaning process, the cleaned dataset was saved as a .csv file.