Data Glacier Team: The Data Dynamos
Batch Code: `LISUM04`
Remote Internship Cohort: September to December 2021

1. Team member's details:

Shari Arangi
Shariarangi504@gmail.com
Ph +254 0742164535
We! Hub Victoria LTD, Kenya
Data Science

Ateeb Aqil Aziz
Atteeb.aqil@gmail.com
Ph +92 3044693262
Data Science

Jennifer Turley
Jennifer.turley@ucdconnect.ie
Ph +353 83 065 7252
University College Dublin, Ireland
Data Science

Queen Echerenwa
Quin_codes@outlooks.com
+234 806 352 8847
Data Science

Github repo link:

https://github.com/DataDynamosTeam/DrugPersistencyProject

Data Cleansing and Transformation:

There are 3424 observations (each associated with a unique patient) and 69 features, with a majority of them categorical (many are 'Y' or 'N').

There are no missing values of any kind, no NAs or blanks.

One column had a problematic column name which included a comma. ''Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx' needed to have the comma removed for some other code to work.

One observation belonging in speciality OBSTETRICS & GYNECOLOGY was miscoded as OBSTETRICS & GYNECOLOGY plus other garbage within the speciality label; this had to be corrected so that observation would be properly included when the ML classification is applied.

Data cleansing required correcting the above specific problems involving Comorbid Encounters and coding of the one OB & GYN observation.

For data transformation, the categorical values were converted to numerical (Y/N to 1/0). Also, flags were created for entries that were not Y/N. For age_bucket, a new, ordered category was created.

Bar charts for each independent (potential predictor) variable against Persistency_Flag permitted visualization of which variables might be useful predictors. There is generally an issue with too many apparently weak predictors in this data: identifying useless predictors (so they can be excluded) will facilitate the classification process, especially where it involves fitting a linear model (logistic regression). Predictors that are potentially useful similarly need to have their usefulness maximized. For this purpose, some new columns were created: such as one for Asian race (which appears correlated with persistency) and having certain Ntm_Specialty (some are correlated with persistency and some with non-persistency). The provided Ntm_Specialty buckets from the original data set were unimpressively correlated with persistency.

Using the bar charts from Week 8 code, obviously unhelpful variables were then eliminated from a created list ‘features’ with the plan that early runs of classification using this list of features might allow for further refinement of the features list.

Finally, in the week 9 .ipynb file, a csv of the chosen subset of features plus the Persistency_Flag was created for purposes of facilitating model building by the team.