GROUP NAME: OpenML

MEMBER'S DETAILS:

Name	Email	Country	College/Company	Specialization
Juan Carlos	juanca.gutierrez@outlook.com	Spain	Everis	Data Science
Laith Adi	Laith_adi@hotmail.com	Canada	Laurier University	Data Science
Gerson Orihuela	yovanni.orihuela@gmail.com	Peru	Inspira IT	Data Science
Walquer Valles	wx.vr@outlook.com	Peru	KeepCoding	Data Science

PROBLEM DESCRIPTION: ABC Pharma contacted OpenML to carry out an analysis in order to have an understanding on the persistence of taking of a drug they released to market. The aim is to know if a patient, based on his/her information, will follow the prescription of the physician and continue taking the drug for all the treatment time. We have been provided with a dataset with patients' details.

GITHUB REPO LINK: https://github.com/jaycee-ds/Drug_Persistency_ABC_Pharma

DATA CLEANSING AND TRANSFORMATION

Race variable - missing values

- use the mode as an imputer. Two reasons why:
    1. only 2.83% (97 instances out of 3424) are "Other/Unknown". So, it feels safe
    to use the mode to fill in the values for now.
    2. The mode accounts for 91.94% of the data. And if we were to group the data
    by ethnicity, the mode accounts for 93.45% (3023 instances out of 3235) for
    "Not Hispanic" and 61.22% (60 instances out of 98) for "Hispanic".
- For those reasons, it's safe to assume that it is likely that the "Other/Unknown" values
can be treated as the mode.

Region variable - missing values

- use the Region mode for "Not Hispanic" Ethnicity group. Reasons:
    - 100% of "Other/Unknown" values in the Region variable, the instances Ethnicity falls under "Not Hispanic"

Ethnicity variable - missing values

- use the mode as an imputer. Reason:
  - the mode accounts for 94.48% (3235 instances out of 3424) of the values for
    Ethnicity.
  - There are only 2.66% of missing values so the number is not alarmingly large
    to reconsider what we use for the missing values. The mode should be safe/good enough.

Ntm_Speciality variable - missing values

- We will keep unknowns as a category and see how it relates to other variables.
- Also the categories that accounts for less than 0.01 of the number of observations will be treated as 'OTHER'.

NTM - Injectable Experience, Risk Factors, Comorbidity, Concomitancy and Frag_Frac_During_Rx (group of variables) - handling categorical data.

- "Y" will be replaced with 1 and "N" with 0

Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score and Change_Risk_Segment missing values

- These variables have more that 40% missing values, consequently they'll be eliminated.

Tscore_Bucket_Prior_Ntm - handling categorical data.

- ">-2.5" will be replaced with 1 and "<=-2.5" with 0

Risk_Segment_Prior_Ntm - handling categorical data.

- "VLR_LR" will be replaced with 1 and "HR_VHR" with 0

CODING:

Code by	Reviewed by
Laith	Gerson
Walquer	Juan Carlos
Gerson	Laith Adi

Juan Carlos	Gerson