Project – Phase 2 (Data Preprocessing)

Aim

The purpose of the project phase 2 is to preprocess the dataset with which you can conduct data analyses for the term project. Please note that this should be group work. Basically, you are expected to 1. select 10 to 15 variables related to your medical condition 2. find any indiscriminative (not meaningful) variable(s) and remove them, 3. process missing values, 4. remove outliers, 5. create a new attribute by using existing attributes as necessary, and 6. normalize data. Basically, what you did for assignment 2 can be applied to the project phase 2. More details will be explained below.

## Dataset

You need to use a dataset you created for the project phase 1 report. It is important for you to discuss what variables you are going to use for the project. Each of the group member need to find 5 to 7 variables independently. Then, please combine them as a group. As there are some duplicated variable, you may end up with 10 to 15 variable. You may want to include some new variables to practice with removing redundancy. For instance, If you only had "RACEX" in the dataset, it would be good to include "RACETHNX" to examine whether they are redundant or not.

## What to Do for the Project Phase 2

1. Do practice with the weeks 6 and 7 example code (data processing & data preprocessing)
2. Find indiscriminative variables
   * You may find or not indiscriminative variables. Please provide your rationale why they are discriminative or indiscriminative.
3. Process missing data
   * Some of your variables would have negative values. For some reason, those data are missed. Please process (e.g., ignore or replace with new values) these missing data and provide your rationale for selecting the method your group chose. Also, please explain some pros and cons of your approach.
4. Find outliers and remove them
   * Please explain what thresholds you used
5. Create new attributes as necessary (e.g., merging "BMINDX53" and"CHBMIX42"). You can skip this step if you don't need to create new attributes.
6. Check redundancy of attributes by using correlations tests and remove redundant attributes.
   * Please provide test results (e.g., Person's correlation coefficient for numeric data and Pearson’s χ2 correlation coefficient, degree of freedom, and critical values for categorical data) to support your decision.
   * You may not find redundant attributes which are totally fine. In this case, you have to provide evidence (i.e., test results).
7. Normalize data

## How to write

For the project phase 2, I will not put constraints on the format. However, you need to explain what you have done for data processing in high-level. Also, please provide rationale for your decision and supporting evidence (e.g., correlation coefficients or descriptive statistics) (refer to What to Do for the Project Phase 2). Up to this part, you can write the report in a **MS word file**. You also need to submit your **Jupyter notebook** you used for data preprocessing and the final data file (csv).