SI 670 Proposal, Team 18

Team members: ***Chen Xie, Xun Wang, Xinye Jiang***

Motivation:

Whether those who assume themselves healthy actually have perfect health conditions? In fact, there is a gap between self-reported and actual health situations. Therefore, we intend to identify people's real health conditions by using clustering methods, depending on measurable data. Furthermore, we will analyze the relationship between real health condition clusters and self-reported general health.

Methods:

Pre-processing: missing value imputation, one-hot encoding, normalization, etc
Clustering: k-means, hierarchical clustering, etc
Classification: logistic regression, SVM, decision tree, cross-validation, etc.

Datasets: We are going to use the National Health and Nutrition Examination Survey data from 2015 to 2016 in our project. The relevant datasets are listed below:

1. Demographic Variables and Sample Weights Dataset: 9971 records, 37 variables, basic info
https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2015
2. Current Health Status Dataset: 9165 records, 9 variables, info about health assessment
https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Questionnaire&CycleBeginYear=2015
3. Examination & Laboratory Data (same resource as before, info omitted due to limited space)

Before applying any ML techniques, we need to join these datasets by their shared respondent id, select variables of interest and handle missing values. We also need to use the one-hot encoding transformation for some specific categorical variables.

Evaluation:

For clustering, we may use ANOVA or chi-squared test to verify the effectiveness of grouping. If there exists a difference between clusters by general health report, then we could encode clusters into categorical variables and process to classification.

For classification, we will compare prediction metrics, such as the accuracy and confusion matrix, of the baseline classifier (e.g. Naive Bayes classifier). If the prediction metrics are improved significantly by using advanced classifiers, we could define it as "success".

Computing: The computing resources we're likely to use are Python Jupyter notebook on our own laptops and probably the school's computing resource.

Existing Work:

Samieri C, Jutand MA, Féart C, Capuron L, Letenneur L, Barberger-Gateau P. Dietary patterns derived by hybrid clustering method in older people: association with cognition, mood, and self-rated health. *J Am Diet Assoc*. 2008;108(9):1461-147118755318

Duty: Xinye is responsible for data preprocessing, Chen is in charge of data visualization, Xun will take charge of model fitting. We would help each other if anyone has difficulty. The report will be accomplished together.