Capstone 1 Data Wrangling

Predicting Prediabetes and Diabetes with Machine Learning

In order to predict prediabetes or diabetes with machine learning, we must first procure data to analyze. This document outlines the steps taken to obtain data and get it ready for analysis.

1. Choose NHANES data for analysis. NHANES stands for National Health and Nutrition Examination. NHANES is a long term health survey conducted by the National Center for Health Statistics, which is a part of the Centers for Disease Control and Prevention (CDC). Start with preprocessed NHANES data on Kaggle at: https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey.
2. Determine kaggle data is not enough for a good analysis. Additional data is on the CDC website. Determine a time frame for analysis that will yield a better quantity of data for analysis is 10 years. Choose the five most recent two year survey cycles: 2007-2016.
3. Determine which features are available for this entire time period. Many survey questions, lab tests, and measurements were dropped or added during this time frame. The list of features chosen is in the document titled Capstone 1 Data.
4. Download all available feature files from the CDC website. These files are in SAS transport format.
5. Download and install Python xport package, which reads SAS transport format. The most current release of xport for Microsoft Windows is not available through Anaconda, so install xport in a separate virtual environment called capstone1_data with pip. Keep this separate in case we need to install further packages with conda for analyzing data as it is not recommended to install packages with conda after pip. Conda doens't understand pip dependencies.
6. Process all downloaded SAS transport format files (.XPT) with xport and convert to csv format for processing in capstone1 conda environment.
7. Start Jupyter notebook, read in 314 csv files by NHANES file prefix. NHANES data is organized by survey cycle, and broken into many small files to ease downloading. There are 62 file prefixes times 5 survey cycles each, and 2 extra files (insulin and total arsenic) times 2 survey cycles each.
8. Create a dataframe for each NHANES category, which generally corresponds to one prefix. Concatenate all five survey cycles into one dataframe.
9. Piece together dataframes requiring special processing:
    a. merge insulin back into glucose dataframe for the 2 survey cycles it was pulled out; and
    b. merge total arsenic back into arsenic for the 2 survey cycles it was pulled out.
10. Create special processing for a few categories that had file prefix and format changes across the 5 survey cycles:

    a. Personal Care and Consumer Product Chemicals and Metabolites/Environmental Phenols/Environmental Phenols & Parabens (EPH/EPHPP);
    b. Mercury: Inorganic, Ethyl and Methyl – Blood/ Mercury, inorganic (IHGEM/IHG);
    c. Perfluoroalkyl and Polyfluoroalkyl Substances/ Polyfluoroalkyl Chemicals (PFAS/PFC); and
    d. Metals – Urine (UHM/UM).
11. Hand edit three RXQ_RX prefix files due to commas in text strings that were not surrounded by quotes. There were one or two instances in each of first three cycles. Apparently this is a bug in the xport package .XPT to .CSV conversion function.
12. Investigate options in xport package due to thousands of unquoted strings containing commas in the last two cycles of NHANES prescription data. Cannot find an option to wrap strings in quotes. Download SAS Universal Viewer, open SAS transport files manually, use SAS Universal Viewer option to save as csv format file. Get UTF-8 encoding error on reading these two files with pandas read_csv function. Try Latin1 encoding (encoding = "ISO-8859-1"). Read in last two RXQ_RX files with Latin1 encoding without error.
13. Create a heatmap of null data for each dataframe. Determine which columns were added or dropped during the 10 year analysis period. Drop all columns that do not span the entire 10 year period as there will be too many nulls to analyze. Remap null data with heatmap.
14. Merge data into dataframes by categories as was done on Kaggle as it seems to make sense for analysis and is easier to manage 6 dataframes than 62. These categories match the NHANES categories with the exception of Prescription data. Prescription data comes out of the NHANES Questionnaire category. The 6 categories are:
    a. Demographics data;
    b. Dietary survey data;
    c. Physical examination data;
    d. Prescription data;
    e. Laboratory test data; and
    f. Questionnaire data.
15. Find rows with all null data except keys in dataframes that were built from one or two NHANES categories (demographics, diet, exam, prescription). Drop these rows. Lab and questionnaire dataframes are created by merging data from nearly 30 category files each, so it would not be useful to drop null rows. We will be creating more null data by merging these features together into the laboratory and questionnaire dataframes as we use an outer merge to avoid data loss.
16. Save the 6 dataframes as csv files for further analysis.