Data Wrangling and Cleaning Steps Utilized to Create the df_final.csv Dataset

The data analysis project, Predicting the Likelihood of Medical Devices Being Recalled, is looking to determine predictive factors that significantly influence medical devices being recalled. Thus, the df_final.csv file was merged and cleaned from three CSV files (devices-1562662526.csv, events-1562662544.csv, and manufacturers-1562662522.csv) for the purpose of this project. Below describes the data wrangling and cleaning steps and methods used to create the final df_final.csv file.

Using IPython Notebook, devices-1562662526.csv, events-1562662544.csv, and manufacturers-1562662522.csv were read into dataframes called devices, events, and manufacturers respectfully. The devices and manufacturers dataframes were merged (in an outer join fashion on manufacturer_id) to create the df dataframe. Then, the df dataframe was merged (in an outer join fashion on device_id) with the events dataframe to create the df_final dataframe.

Columns in the df_final dataframe were removed based on a number of factors. By a boolean array, columns with a missing (null) percentage of 60% or more were removed. Additionally, based on common sense judgement, columns that have no relation to the outcome column, type (Event Type), were removed. These columns included the id numbers used to merge the original dataframes, device numbers, address of manufacturers, urls to respective authorities, dates the investigative events were initiated by the firms, UIDs, UID hashes, urls to the specific investigative events, and dates each entry in the original dataframes were created and updated. Word frequency count was used to determine whether or not to drop certain columns with lengthy texts. These columns included description (descriptions of the devices), action (descriptions of actions taken concerning the devices and their manufacturers), data_notes (additional notes concerning the events in which how the devices were dealt), icij_notes (notes by the ICIJ), and reason (reasons why the devices were reported and put under investigation). To determine and explore the word frequency of these columns, a copy (denoted as df_final_copy) of the df_final dataframe was created. Then, the texts of the columns in the copy dataframe were converted to lowercase, and the nulls of the columns were removed. Finally, a Python tool called Counter from the collections library was used to count the five most frequent words of each of the textual columns. The description, action, and reason columns were dropped from the df_final dataframe as they contain stop words and symbols as the top five most frequent words.

The values of the action_classification (Event Risk Class) column and the outcome column, type (Event Type), were reorganized. The values of the action_classification column were denoted into numeric values: 0 - No classification; 1 - Class 1; 2 - Class 2; 3 - Class 3. As there were only four outlier values of “Unclassified Correction” and “Voluntary recall” under the action_classification column, the rows containing them were removed from the df_final dataframe since their removal won’t affect the analysis results of this project. The values of the type column were changed into numeric binary values: 1 – recall; 0 - no recall. As the type column is the outcome column of the project, the rows containing the null values of the type column were removed from the df_final dataframe. Finally, the action_classification and the type columns were converted from object types into integer types.

As the rest of the remaining columns of the df_final dataframe are object types, the rest of the null values in the dataframe were changed to the string value, “None.”

The final df_final dataframe was exported as a CSV file named, df_final.csv.