**MIDTERM PROJECT**

**Data PreProcessing**

Data preprocessing is one of the most important steps before applying any algorithm to a given dataset. It is very important because a clean data will ensure less effort in application of any machine learning algorithm. It would result in less load on the system and a much better result.

Like any Data Science project, we have also performed Data Preprocessing Steps. They can be classified as the following

**Data Cleansing**

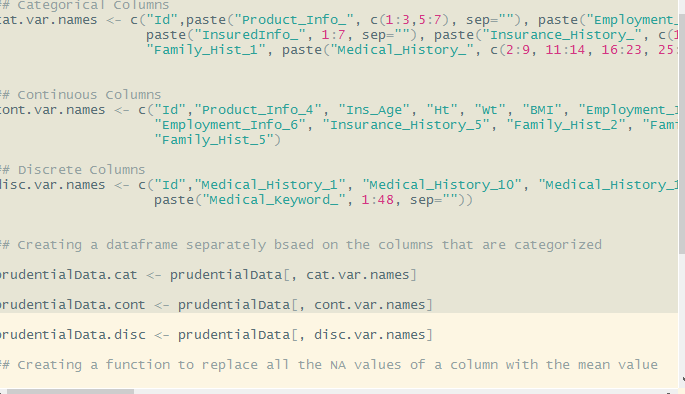
Data cleansing is to remove unwanted, junk and garbage values and eventually create a new dataset to be fed into the system.The following steps were performed as part of our Data Cleansing process

Step 1:

Splitting of the main dataset into three key data frames based on type of attributes – categorical, discrete and continuous.

This was done to split the task equally and to b able to clean the data without any value acting as a predominant value for the other.

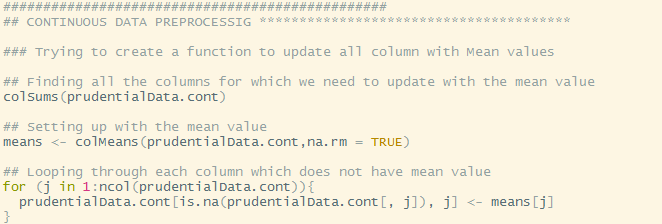
For Example: Value 1 holds importance in categorical but can be rejected in a continuous dataset.

Code: 

Step 2:

For continuous variables, all the missing (NA) values were replaced with the Mean value for that column. This was to ensure the total model still manages to classify if that particular attribute is important for classification and regression or not.

Code:



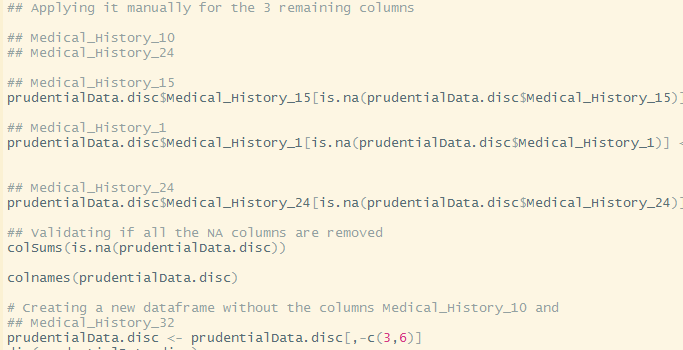
Step 3:

For Discrete variables, we have replaced all the NA or NULL values with the median value of the column. This is because each discrete value is treated as a whole number, and in case we put a mean, then it would be marked in decimals.

Also it is important to note that 2 values have their values almost 95% as NULL. So we are discarding those values.

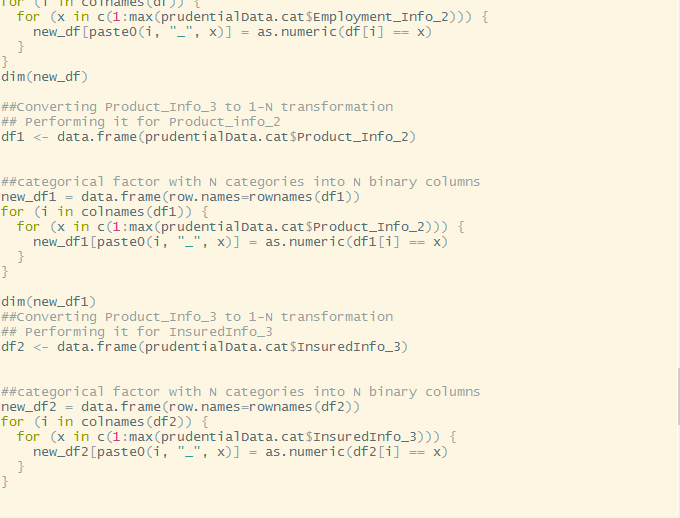
Columns are Medical\_History\_10 and Medical\_History\_24

Code:



Step 4:

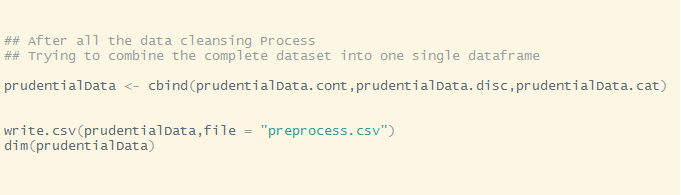
In case of categorical values, for the columns which have more than 3 values we have performed 1-N split. There was also a column, Product\_info\_2 which had “factor” as the type. For that column we have converted it to numeric and split it into N possible values



Step 5:

After separate dataframes are created, we combine all of them, with the final dependent variable “Response” and get one single CSV file.

Code :



**LINEAR REGRESSION**