**UCI Diabetes data**

The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of subjects diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. The goal is to determine the early readmission of the subject within 30 days of discharge.

1. Originally there are 101766 observations on 50 variables with the variables being described in the above link. (UCI repository link for full data and variable description: <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>.)
2. After removing some data, we store 69169 data points in the UCIdiabetes.csv file on the same features. You will be working with this dataset.
3. For details on features (variables) look at the Diabetes\_variables.xlsx or the above link. Details on variables are in Diabetes\_variables sheet. The variables “diag\_1”, “diag\_2” and “diag\_3” from the original data have been recoded to “diag\_1\_name”, “diag\_2\_name” and “diag\_3\_name” respectively to provide you with more meaning.
4. The IDS\_mapping worksheet in Diabetes\_variables.xlsx contains more information on variables “admission\_type\_id”, “discharge\_disposition\_id” and “admission\_source\_id”. If appropriate, you may recategorize certain variables: for example, “admission\_type\_id” = 1,2,3,4,7 as known and unknown elsewhere.

With the above in mind, create a presentation exploring the following:

Analysis questions:

1. Starting with an exploratory analysis of the features, do you find anything surprising/noteworthy regarding any of the variables? Explain. (For example, amongst others, consider “num\_medications”; are patients being given too many medications?).
2. If there are any missing data, explore the same. How do you tackle the missingness if you were to fit a predictive model? Please provide a logical explanation.
3. How would you suggest we analyze the impact of HbA1c (use variable “A1Cresult”) on hospital readmission (use variable “readmitted”)? Is there a model that you can specify? You may recategorize “readmitted” to be a binary outcome as Yes/No. Recall that early readmission is defined as readmission within 30 days of discharge. You may focus only on primary diagnosis (diag\_1\_name) only.
4. It is important that we know how diabetes medication changes (variable “change”) were being done under different scenarios of HbA1c measurements. Suggestions?
5. Are there any issues of multi-collinearity? If so, what might those variables be and how can we tackle?
6. Ultimately you may do a validation to justify performance of your model.

Reference:

Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J.