CIND 119: Introduction to Big Data Analytics Assignment 1 (15% of the final grade)

Perform K-Means clustering on a dataset and analyze the results with using SAS

Dataset: The dataset for this assignment is "heart.csv" attached to the assignment. Details about the dataset can be found at dataset link.

The dataset contains information about various cardiovascular disease indicators for patients, with 13 numerical/categorical attributes and 1 binary target attribute indicating the presence or absence of heart disease.

Download the heart.csv from your D2L Assignment 1 link. Complete the following tasks (15 points):

1. Read the file in SAS and display the contents using the **PROC IMPORT** and **PROC PRINT** procedures, print only the **first 10** observations. (3 points)
2. Perform basic Data analysis using **PROC Means** (2 points).
3. Apply standardization to the numerical attributes using **stdize** procedure and print the data (obs=10) (2 points).
4. Apply k-means clustering using **fastclus** procedure of SAS use your standardized dataset. Scatter plot your cluster labels (use y=chol and x=age) to visualize and compare with the original data labels. Assuming that you do not know the exact number of clusters in the dataset, try k=2, 3, 4, 5 and evaluate the solutions. Choose the best K value based on the RMS Std. Deviation. (8 points)

End of CIND119 Assignment 1