**Emory Big Data Program – Capstone Project**

This capstone project is designed and intended to give you an opportunity to apply what you have learned in this program in a holistic manner. Throughout this project you need to think through and decide what are the best data mining methodologies and techniques in answering various questions. As much as you will get the chance to master what you have learned in this program with this project, you will also get the chance to strengthen your analytical thinking and self-learning skills. You may be faced with questions that you may not be able to answer directly using the methodologies and techniques that you have learned in the program, however, you have also learned to use different ways to answer questions, such as using the Internet and other available resources.

Please remember that most of the questions can be answered with a variety of data modeling techniques methodologies. Which methodology you use solely depends on your own personal preference, modeling technique limitations and feature, data set limitations and features, and of course the goal of analysis.

For this project you are given a subset of World Development Indicators data and you are asked to perform various tasks and analysis using RapidMiner.

1. Are there any missing data or inconsistencies in this dataset?
   1. Lots of missing values for many attributes
2. Explain your strategies for dealing with missing data for each variable with a missing value. You need to explain why you chose your strategy versus other possible strategies.
   1. By attribute:
      1. Birth rate: avg
3. Which region has the most amount of overall CO2 emission?
4. What are the top 3 countries producing CO2?
5. Is there any relationship between CO2 emission and the GDP?
   1. Ans: run correlation between the 2 attrributes
   2. NOTE: use [+/-] 0.7 as cutoff where > 0.7 means highly correlated
6. What are the highly correlated attributes in this dataset?
   1. Ans: run correlation matrix across entire data set
7. What are the moderately correlated attributes in this dataset?
   1. Moderate correlation: 0.3 < corr < 0.6 [+/-]
8. We want to group countries into specific categories based on their GDP? How many categories do we need and why?
   1. Based on distribution can run K-cluster analysis
9. Excluding categorical attributes, run a k-means clustering analysis with k=4, can you identify any pattern in the cluster formations?
10. What are the most important factors impacting GDP?
11. What do you think would be the best modeling technique (e.g. regression, decision trees, etc.) in predicting GDP and why? What would be the second best?
12. Build the models according to your answers to the above questions and calculate and report model performances. Did the models perform the way you theorized they would? Are there any other models that can outperform your suggested models?