**Project: Milestone 1, 2 & 3**

Name: Soukhna Wade   
Date: 05/26/2020  
Title: DSC550-part 1, 2 &3

**Dataset Name:** students-performance.csv

**Source:** https://www.kaggle.com/spscientist/students-performance-in-exams

**Case Study:**

Analyze the data to understand the influence of the parents’ background, test preparation, etc. on students’ performance dataset.

**Data Analysis and Visualization Using Pandas:**

Importing and cleaning data by performing exploratory data analysis (EDA). And construct questions that lead to a more profound analysis of a visual representation of the dataset, in the form of graphs.

1. Load the data from the “StudentsPerformance.csv” file into a DataFrame.
2. Display the dimensions of the data to have a good idea of the amount of dataset.
3. Display the first five rows of data so you can see the column headings and the type of data for each column.
4. Check for any missing values or null values (NA or NaN)
5. Initialize the required values by setting the minimum mark
6. To understand the necessary information of the data, like min, max, mean, and standard deviation
7. Think about some questions that might help to predict the students’ performance dataset
8. Determine the type of variables (numerical or categorical dataset)
   1. If they are numerical, what are their distribution;
   2. if they are categorical, how many are they in different categories?
   3. Are the numerical variables correlated?
   4. Are the distributions of numerical variables the same or different?
9. Look at summary information about the dataset (total, mean, min, max,, etc.) Does this present any more questions?
10. Make some histograms of the dataset and describe the result
11. Make some boxplots for numerical variables
12. Showing if the data is correlated (using Pearson Ranking charts)
13. Use count plot to plot the student’s grades obtained in an order
14. Check the percentage of missing data in each column present in the data
15. Working on Dimensionality and Feature Reduction
16. Convert your categorical data into numbers
17. Training - Split your data into two sets: Training and Testing
18. Evaluation - There are many algorithms that could be used but we’re going to use logistic regression.

a. Metrics for the evaluation:

i. Confusion Matrix (you should get 84% - pretty good)

ii. Precision, Recall & F1 score (all 3 were very good)

iii. ROC curve (the dotted line is the randomly guessed so anything above that is good metric)