Assignment 7: Data Analytics (Fall 2018) (20% written)
Due: FRIDAY Nov. 30, 2018 (by 5pm ET)

Submission method: written by LMS
Please use the following file naming for electronic submission:
DataAnalytics2018Fall_A7_YOURFIRSTNAME_YOURLASTNAME.xxx, etc.

Late submission policy: **If you are more than 10 days late it is likely that you will not have your grade for this assignment included in your final grade before they need to be submitted.**

Note: Your assignment should be the result of your own individual work. Take care to avoid plagiarism ("copying"), and include references to all web resources, texts, and class presentations. You may discuss the project with other students, but do not take written notes during these discussions, and do not share your written assignment or presentation before the class they are presented in.

General assignment: Predictive and Prescriptive data analytics. You should develop and validate predictive *models* (regression, classification, clustering – using one or more of the methods covered in class to date or one of your choosing) for ***two*** of the seven (the Wine Quality contains red wine and white wine datasets) datasets below and apply them for decision purposes. Use the section numbering below for your written submission for this assignment. Include references – websites, papers, packages, data refs...
http://archive.ics.uci.edu/ml/datasets/News+Popularity+in+Multiple+Social+Media+Platforms
http://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT
http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
http://archive.ics.uci.edu/ml/datasets/Bank+Marketing,
http://archive.ics.uci.edu/ml/datasets/Wine+Quality,
http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

1. Exploratory Data Analysis (3%)
   Explore the statistical aspects of both datasets. Analyze the distributions and provide summaries of the relevant statistics. Perform any cleaning, transformations, interpolations, smoothing, outlier detection/ removal, etc. required on the data. Include figures and descriptions of this exploration and a short description of what you concluded (e.g. nature of distribution, indication of suitable model approaches you would try, etc.). Min.1 page text + graphics (required).

2. Model Development, Validation, Optimization and Tuning (14%)
   Choose two (4000-level*) or three (6000-level) or more different *models* (e.g. a model with a different set/ number of variables/ features in a regression, or classification, etc. does NOT count as a different model). Explain why you chose them. Construct the *models*, test/ validate them. Explain the validation approach. You can use any method(s) covered in the course. Include your code in your submission. Compare model results if applicable. Report the results of the model (fits, coefficients, graphs, trees, other

measures of fit/ importance, etc.), predictors, and summary statistics. Min. 4 pages of text + graphics (required). * 4000-level will receive extra credit for 6000-level responses.

3. Decisions (3%)
   Describe your conclusions in regard to the *model* fit, predictions and how well (or not) it could be used for decisions and why. Min. 1 page of text + graphics.