Assignment 7: Data Analytics (Fall 2024) (15% written)

Due: Tuesday, December 10th 2024 by 11:59pm EDT. (12/10/2024 by 11:59 pm EDT)
Submission method: written by email (eleisa2@rpi.edu)

Please use the following file naming for electronic submission:
DataAnalytics2024Fall_A7_YOURFIRSTNAME_YOURLASTNAME.xxx, etc.

Late submission policy: If you are more than 5 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. First time with valid reason – no penalty, otherwise 20% of score deducted each late day.

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 **one** of the ten datasets below. Use the section numbering below for your written submission for this assignment.

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/Communities+and+Crime

https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk

https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

https://archive.ics.uci.edu/dataset/890/aids+clinical+trials+group+study+175

https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation

https://archive.ics.uci.edu/dataset/20/census+income

Conduct the following analysis for the dataset:

1. Exploratory Data Analysis (3%) Explore the statistical aspects of the dataset. 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 and Optimization (10%) Choose two (4000-level*) or three (6000-level) or more different models (the models must include at least 2 objectives from regression, classification & clustering). The choice of independent and response variables is up to you.

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, trees, other measures of fit/ importance, etc., predictors and summary statistics). Min. 2 pages of text + graphics (required).

3. Decisions (2%) Describe your conclusions from the model fits, predictions and how well (or not) it could be used for decisions and why. Min. 1/2 page of text + graphics.