Assignment 6: Data Analytics (Spring 2026) (15% written)

Due: Friday, December 11th 2026 by 11:59pm ET.

Submission method: LMS

Please use the following file naming for electronic submission: DataAnalytics2026Fall_A6_YOURFIRSTNAME_YOURLASTNAME.xxx, etc.

Assignment: Predictive analytics, feature selection, model development and evaluation. You should develop and validate predictive models (regression, classification, clustering) for **one** of the datasets below. Use the section numbering below for your written submission for this assignment.

Detection of IOT botnet attacks: https://doi.org/10.24432/C5RC8J

Census Income: https://doi.org/10.24432/C5GP7S

Online Retail II: https://doi.org/10.24432/C5CG6D

Conduct the following analysis with the dataset:

1. **Exploratory Data Analysis (3%)** Explore the statistical aspects of the dataset. Analyze the distributions and provide summaries of the relevant statistics. Begin with a GGpairs plot. Identify interesting variables/ variable associations. Perform any cleaning, transformations, outlier detection/ removal, etc. required on the variables to obtain informative plots. Include figures and descriptions of this exploration and a short description of what you concluded (e.g. nature of distributions, indication of suitable model approaches you would try, etc.)

Min.1 page text + graphics (required).

2. **Model Development, Validation and Optimization (10% 4000-level / 7% 6000-level)** Develop and evaluate three (4000-level) or four (6000-level) (or more) models. **If possible**, these models should cover more than one objective, i.e. regression, classification, clustering, where possible. Consider the effect of dimension reduction of the dataset on model performance. The choice of independent and response variables is up to you.

Construct the models, test/ validate them using an appropriate cross-validation strategy. Report the results of the models (summary statistics, coefficients, measures of

accuracy/error, etc.) and compare model results if applicable. List and explain your takeaways from the predictive analysis.

Min. 2 pages of text + graphics (required).

3. **Decisions (2% 4000-level / 5% 6000-level)** 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.

Notes:

1) Submit a word, pdf, etc. doc containing your responses, tables, figures, etc. + your R code.
2) Make sure all your plots are clear/legible/informative.
3) Make sure to caption all tables/figures.
4) Some of these datasets are very large, make a reasonable effort to evaluate your models robustly but do not fry your laptops.

EOF