

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

Due: Tuesday, April 28th 2026 by 11:59pm ET.

Submission method: LMS

Please use the following file naming for electronic submission:

DataAnalytics2026Spring_A6_YOURFIRSTNAME_YOURLASTNAME.xxx, etc.

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_Balot

<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 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. 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% 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. Consider the effect of dimension reduction of the dataset on model performance. The choice of independent and response variables is up to you.

Explain why you chose them. Construct the models, test/ validate them. Briefly 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 (summary statistics, coefficients, measures of accuracy/error, etc.).

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.

EOF