Assignment 5: Data Analytics (10% of overall credit score)

Due: Friday December 12, 2025 (by 10 pm EDT)

Submission method: LMS

Please use the following file naming for electronic submission for any individual documents: DataScience_2025_A5_YOURFIRSTNAME_YOURLASTNAME.xxx

Late submission policy: If you are more than 2 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 report for this assignment should be the result of individual work. Take care to avoid plagiarism ("copying"), including all web resources, texts, and class presentations. You may meet to discuss the tasks for this assignment with other members of the class but must produce the materials for the final assignment themselves, citing all outside sources.

Assignment Instructions: Select a dataset from those listed below and conduct a data analytics exercise. Perform exploratory data analysis by examining a subset of the features in the dataset, visualizing their distributions (boxplots, histograms, etc.) and the relationships between them (scatter plots, heatmaps, etc.) Use what you learn from exploration to develop a predictive model that attempts to solve the problem for which the dataset was intended.

Datasets:

1) Air Quality https://doi.org/10.24432/C59K5F
2) Seoul Bike Sharing Demand https://doi.org/10.24432/C5F62R
3) Magic Gamma Telescope https://doi.org/10.24432/C52C8B
4) Dry Bean https://doi.org/10.24432/C50S4B
5) Online Shoppers Purchasing Intention Dataset https://doi.org/10.24432/C5F88Q

Use the following numbering to organize your written responses.

1. Exploratory Data Analysis (**4000** 5% / **6000** 3%)

- Examine a subset of features in the dataset both separately and in pairs. This may involve cleaning the dataset, for example removing missing values, applying transformations such as log scale and/or taking subsets. You have complete freedom in the selection of features to explore and any necessary filtering. Make sure to use suitable plots to examine variable distributions as well as pairwise relationships between variables. Consider how the presence of outliers affects the

plots and consider removing them if they have a significant effect. Explain what you learned about the dataset in 5-6 sentences.

2. Predictive Modeling (5%)

- Decide on a problem to solve by developing predictive models using the data. This could be the same problem for which the dataset was intended or a different problem that you believe can be addressed with this dataset. The solution could involve predicting a continuous variable (regression) or a categorical variable (classification). Consider what you learned about the variables during exploratory analysis to decide which features should be used as inputs to the models. The response variable (output of the model) is usually clear and is dictated by the problem being addressed. Train and evaluate **two** models and compare their results. The two models should utilize different algorithms (e.g. kNN and Random Forest) but the same set of input variables. Classification models should be evaluated and compared using a confusion matrix and Precision/Recall measurements. Regression models should be evaluated using Mean Squared Error. Explain how models performed and comment on the suitability of the dataset to solve the problem you chose in 5-6 sentences.

3. **(6000-level** 2%**)** Meta-analysis

- Discuss how the secondary aspects of the dataset helped or hindered the analysis. Consider the effects of dataset structure, file format, metadata, contextual information, etc. on your analytics plan. These should roughly correspond to some of the areas of data management covered earlier in the course. If there are deficiencies in the dataset, provide recommendations to fix them. Three items should be sufficient.

NOTE:

- Use the code files in my box folder as a reference for the types of functions you will need (in R and Python): https://rpi.box.com/s/c9vpi24e5jizjaw9ef49s1nblym7xq8i
- You may use any programming language you prefer.

THE END