# Data Science (Eng) 774/874
# Post-block Assignment 3

10 March 2021

Kindly complete the following assignment and submit it as an electronic file in PDF format on SUNLearn by **23:50 on 30 April 2021.**

## 1. Data Science (Eng) 874 AND 774 (Postgraduate Diploma) students [40]

In this assignment, you may use any dataset of your choice except the ones used in other assignments. It is not allowed to reuse any example provided in this course. You should provide a reason why you selected a specific dataset. The final outputs of this assignment are a data visualization and a supporting report combined in an electronic file in PDF format.

The CRISP-DM methodology must be at the center of your data visualization strategy. In your report (where applicable), you should clearly outline how each of the phases of the CRISP-DM was applied to your selected dataset.

- Students in the Data Science 774 module are exempted from explaining the modelling and evaluation phases.
- Students in the Data Science 874 module should explain in detail how they applied the modelling and evaluation phases of the CRISP-DM approach.

Note: You should use a dataset that addresses a pertinent question, and your visualization step should be able to tell a story about the data.

## 2. ONLY Data Science (Eng) 874 students [30]

2.1. In addition to the requirements outlined in Section 1 of this assignment, you need to use at least 3 Machine Learning (ML) algorithms for predictions from the following methods: k-Nearest Neighbor, Logistic Regression, Linear Regression, Random Forest, Neural Networks (Feed forward, do not use more than two hidden layers) and Naïve Bayes. You need to provide a reason why you selected a particular ML method.

For each model, conduct predictions for 10 records within the testing set.

**Note:** Use 80% of your dataset for training and 20% for testing.

2.2. Demonstrate your ability to evaluate the selected models by using the following performance metrics: Accuracy, Precision, and Recall. Compare the results.

2.3. Compute the Confusion Matrices of your selected models and discuss the findings.

### 3. Evaluation rubric

Your visualization and predictions will be marked according to the following rubric:

| **Criteria** | | **Excellent - above average** | | | **Not at desired level of competency** |
|---|---|---|---|---|---|
| | **Weighting** | **4** | **3** | **2** | **1** |
| **Excellent choice of dataset.**<br>**The dataset assists in answering a clear question** | 10 | Visualization has a concise and clearly defined topic that addresses one question | Topic is well defined, but visualization addresses too many questions to be useful. | Topic is somewhat defined, and the visualization addresses multiple questions. | Poorly defined topic that addresses too many questions to be useful. |
| **Adherence to the application of the CRISP-DM methodology** | 20 | Accurate, concise description of how each relevant phase of CRISP-DM was executed in the assignment. | Only minor errors in the application of CRISP-DM | Major flaws in the application of CRISP-DM | No reference to the use of CRISP-DM |
| **Compliance with Tufte's Visualization Aesthetic** | 5 | Outstanding application of Tufte's visualization aesthetic | Generally competent application of Tufte's visualization aesthetic. Minor errors may exist. | Major violations of Tufte's visualization aesthetic. | No consideration of Tufte's visualization aesthetic. |
| **Level of originality and innovation** | 5 | Excellent and innovative visualization. | Fair, creative use of visualization tools. | Limited effort made to select and use data visualization tools. | Poor visualization. No effort was made to show innovation. |
| **ONLY Data Science (Eng) 874 students** | | | | | |
| **Adequate and competent use of the selected ML algorithms.** | 15 | Correct explanation of the choice of ML methods**.**<br>Excellent use of the selected ML methods. | Generally competent application of prediction algorithms. Minor errors may exist. | Major errors in the use of prediction algorithms. | No or highly flawed prediction of future data points. |
| **Successful models comparison using the specified metrics.** | 15 | Model comparison correctly done using all performance metrics and confusion matrix. Excellent discussion of the results. | Error in the implementation of 1 performance metric. | Error in the implementation of 2 performance metrics. | Error in the implementation of 3 performance metric<br>**OR**<br>No performance metric was implemented. |

## 4. Examples of data visualization tools

There exist several data visualization tools and techniques that can assist you in this assignment. The following links contain some resources that can be used as a starting point:

- https://blog.udacity.com/2015/01/15-data-visualizations-will-blow-mind.html
- https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97fbed
- https://medium.datadriveninvestor.com/artificial-intelligence-series-part-4-data-visualization-in-python-da457ff3a70b
- https://www.kdnuggets.com/2019/08/activestate-exploratory-data-analysis-python.html