Honours Assignment for DSA822S Data Science & Analytics: 2024
30% Contribution: Deadline Friday, October 11, 2024

BASIC REQUIREMENTS

- □ This assignment is not an essay assignment, but your iPython notebook should be well commented with both text, equations, and explanations.
- □ Your submitted notebook should be able to run to completion successfully.

PART 1 - NEURAL NETWORKS

IDE Notebook Code completion

The biomedical industry uses machinery to capture images and other data and also software that provides the specialist with information to make decisions. These software applications often use vast data sets, such as the mammogram data set you used in the previous activity.
Complete the notebook by reading the data in the **Mamo-data.csv** file and filling in the missing code in order to answer the questions at the end.

Questions

1. Consider the medical data and the context in which it is used. Why is machine learning, and specifically neural networks, an appropriate method for prediction in these circumstances? (Max. 200 words)

2. What is the purpose of a response curve in the context of neural networks? What insights can be gained from the response curves you generated in this IDE notebook?

MARK SCHEME: Part 1 - Neural Networks

IDE Notebook Code completion	**Question 1**	**Question 2**	**TOTAL**
(ex 30)	(ex 10)	(ex 10)	(ex 40)

PART 2: K-MEANS CLUSTERING

Using the "**online_shoppers_intention.csv**" data file provided, complete the IDE notebook code where you have to build the K-means model, and lastly analyse the model output generated by the K-means algorithm. Justify the use of the K-means algorithm on a given data set.

Insight Questions from Part 2 Notebook

1.1 At the end of the Part 2 notebook, you created the pair-wise plots. Consider the behaviour of customers in each cluster based on the distribution and scatterplots related to the variable ***ProductRelated_Duration***.

2.1 Discuss the quality of the clustering solution by referring to the number of observations in each cluster and the distribution of the data along each dimension. How does this affect the way you can interpret the scatterplots?

3.1 Discuss whether the K-means algorithm was an appropriate method for identifying clusters in this particular dataset by referring to at least two elements of the dataset that contributed to your decision.
The following are some elements that you could take into consideration to justify your answer:

- Discuss the applicability of the K-means algorithm for the number of observations in the data set. Compare this data set with one that has substantially more observations (e.g. 900,000 vs the present 12,000).
- There were some outliers in the data set. How do you think this can influence the predictive capability of the analysis?

NOTE: Remember to set the `random_state=1` for reproducibility.

MARK SCHEME: K-Means Clustering.

Code completion	**Insights, Question 1.1**	**Insights Question 1.2**	**Insights, Question 2**	**Style of Presentation & Insights from the K-means Model**	**TOTAL**
(ex 25)	(ex 10)	(ex 10)	(ex 10)	(ex 5)	(ex 60)

The annotated notebook is worth 30% of the total of 100% credits for your CAS, and therefore evidence for a significant amount of work, over and above the repackaging of lecture material, is required. It should demonstrate your own calculations/estimates of relevant quantities. The level should be suitable for reading by a fellow honours student. You should properly reference your academic sources, which may include websites.

A PDF OF THE NOTEBOOK* INCLUDING A LINK TO THE EXECUTABLE VERSION ON BINDER** MUST BE SUBMITTED **VIA EMAIL** BY **23:59 10th October, FRIDAY 2024**

* https://nbconvert.readthedocs.io/en/latest/usage.html

** https://mybinder.org