School of Computer Sciences

Academic Session 2022/2023

Semester I

CDS501 – Principles & Practices of Data Science & Analytics

Assignment 2 (7.5%)

1. Part A (Choosing and Evaluating Methods): Apply at least **two (2)** different cross-validation methods (e.g., k-fold, repeated k-fold, LOOCV etc.) to your selected dataset from your group project. Evaluate each method (give advantages and disadvantages) and select one that best suits to solve your business problem.
   Note: For the group with text/review dataset: You can use the dataset from your project (if applicable) or any datasets from Kaggle or UCI.

2. Part B (Unsupervised Modelling): Apply one unsupervised modelling algorithm (**e.g. clustering, market basket analysis, topic modelling etc)** to your selected dataset. You can use the dataset from your project or any datasets from Kaggle or UCI under this domain:
   a) clustering for customer segmentation
   b) clustering in healthcare and medicine
   c) clustering in cybersecurity
   d) clustering in social media analysis
   e) clustering in business analysis
   ***please provide the link to your dataset*

   Show the process of your work, evaluate and describe your results/output.

3. The expected outcome of this assignment is not more than 10 pages write-up of the process you conducted. Please also include your R file/script in your submission

Submission:

Online submission: Saturday (14 January 2023), before 11.59 p.m.

Late submission will be penalized.

Formatting

1. Cover page: Course code and course title, Title, Full name of group members and metric numbers, and Submission date.
2. File format: **pdf and r script/rmarkdown/knitr HTML.**

3. Please name your file in this order: <groupno_assignment2>
4. Page limit: The document should be not more than 10 pages.
5. Font size: 11. Spacing: 1.5

Read Me: USM Policy

- All assignments and lab exercises MUST be submitted **before or on** the specified date. Late submissions without any reasons and without permission from the lecturer(s) will not be accepted. The grade for late submission (even with permission) will be reduced as determined by the lecturer(s).
- Students who copied or **plagiarized** other's work or let their work be copied or plagiarized will be given F grade for the work, test or the whole coursework component as determined by the lecturer(s). The said student may be barred from sitting for final exam and reported to the university's disciplinary board.
- **Plagiarism** (using other people's ideas and text without proper acknowledgment and using them as your own) is a serious academic offence. The consequences for plagiarism are severe.

Reference: https://www.kdnuggets.com/2020/03/trends-machine-learning-2020.html