CPS803 / CPS8318 Assignment 4 (Due: 30-Nov-2025)

Choose a dataset (as opposed to the example ones we used in class), **requiring pre-processing**, with a reasonable size, to solve a practical **clustering** problem. You may select a dataset from one of the following sources (other sources are also possible, e.g., Kaggle):

- UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.php.

- KDD Cup challenges, http://www.kdd.org/kdd-cup.

You can use any appropriate libraries.

Marking: 50% for the writeup and 50% for the results. In the write-up, cite the sources of your data and ideas, and use your own words to express your thoughts. If you have to use someone else's words or close to them, use quotes and a citation. The citation is a number in brackets (like [1]) that refers to a similar number in the References section at the end of your paper or in a footnote, where the source is given as an author, title, URL or journal/conference/book reference. Grammar is important. Concerning the 50% for results, elaborate on what (if any) manipulations you did, what are the results for the algorithms you tried, what else you tried.

Submit the document on the D2L site. Please submit a zipped file including the report (PDF file) and the python script (.py file(s)). If the dataset is not in the public domain, you also need to submit the data file. Submit the data you used, your code (.py file), and your report (pdf file). Ensure you named your documents appropriately:
report_FirstnameLastName.pdf
script_ FirstnameLastName.py

(1) Background (~1/4 to 1/2 page)
Describe your data, how you obtained it, and why it is interesting. Introduce all important concepts and background information.
(2) Methods (~1/2 to 3/4 page)
Describe your methodology: give flowcharts, diagrams or formulas where appropriate. Describe evaluation strategy.
(3) Results (~1/4 to 1/2 page)
Organize and present all your results and findings. Make sure it is easy to understand what your results are.
(4) Conclusions (~1/8 to 1/4 page)
Summarize what the conclusions are and how they derive from the results.
(5) References
List books, scientific papers, web sites etc. that you referenced in the body of your manuscript.

The above-mentioned paragraph lengths are approximative. As long as this is reasonable, you will have the flexibility to write slightly more text (but not less). You also have the flexibility to format your report as you wish, as long as this is reasonable. For example, you could use 1-inch margins from each side, use 11pt or 12pt font size, use standard font types such as Times New Roman.