# CPS844 Assignment 1 (Due: 29-Feb-2024)

Choose a practical dataset (as opposed to the example ones we used in class) with a reasonable size 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.

Download the data, read the description, and try various approaches to solve a classification problem as best as you can. Write up a report of approximately 5 pages (absolute maximum is 10 pages), double spaced, in which you briefly describe the dataset (e.g., the size – number of instances and number of attributes, what type of data, source), the problem, the approaches that you tried and the results. You can use any appropriate libraries. You can work in teams of two (or alone). If you work with someone, each of you should submit your documents on D2L. You need to make it clear with whom you worked with (e.g. name your documents with both of your names: report_FirstnameLastName1_FirstnameLastName2.pdf)
If you fail to do so, then your mark will at most be 60%.

Your tasks are:

1. to use at least 5 different classification methods to see which one does the best job (present your comparison).
2. to report on which attributes are most important for your classifier.
3. to report on anything else inventive you can think to do, but the above 3 tasks would probably be enough.

**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.
Name your documents appropriately:
report_FirstnameLastName1_FirstnameLastName2.pdf
script_ FirstnameLastName1_FirstnameLastName2.py