CPS844 Project (Due: 4/5/2025)

Choose a practical dataset (as opposed to the small example ones we used in class) with a reasonable size (at least 1k instances) from one of the following sources (other sources are also possible):

- UCI Machine Learning Repository, https://archive.ics.uci.edu/datasets
- Kaggle datasets, https://www.kaggle.com/datasets

Download the data, read the description, and try various approaches to solve the problem as best as you can. Write up a report of 10 to 15 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 need to write Python code using appropriate libraries. You can work in teams of two (or alone).

What actually is the "problem"?
The problem for each dataset is usually to predict the class. It could be to predict a numeric value, to find associations, or to find nature groupings, too.

Your tasks are:

1. to perform some explorative analysis on the dataset and decide whether to do any pre-processing on the data;
2. to try at least 5 different algorithms (if possible, each from a different category) to see which one does the best job (present your comparison);
3. to try at least 1 feature selection algorithm and report on which attributes are most important for the prediction;
4. to compare the accuracy of all the data mining algorithms with or without the feature selection (i.e., with all features vs. with selected features) and report on whether feature selection helps;
5. 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 writeup, 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. With regard to the 50% for results, show how you got the data the proper format, what (if any) pre-processing you did, what are the cross-validation results for algorithms you tried, what feature selection methods you tried. Please only include the evaluation metrics which are appropriate for the problem. Don't include the unnecessary metrics.

Submit the report and the source code to D2L. The report should be named as cps844w25_yourname.pdf. One group only needs to do one submission, just making sure you have group members' names included in the report. If the dataset is not in a public domain, you also need to submit the data file (if it is big, upload to Google drive and share a link).