CPS844 Assignment (Due: 4/9/2021)

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. Some of the recent ones have huge datasets which are probably beyond our capacity, but earlier ones are manageable, for example, the KDD Cup 2004. The 2011 data is manageable, but requires expanding the cache in Weka.

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 can either use Weka to perform all the tasks (no programming) or write programs using appropriate libraries. You can work in teams of two (or alone).

What actually is the "problem"?
The problem in 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 try at least 5 different algorithms (if possible, each from a different category) to see which one does the best job (present your comparison);
2. to try at least 2 attribute selection algorithms and report on which attributes are most important for the prediction;
3. to compare the accuracy of the data mining algorithms (either all from task 1 or the best one from task 1) with or without the attribute selection (i.e., with all attributes vs. with selected attributes) and report on whether attribute selection helps;
4. 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 into the Weka format, what (if any) manipulations you did, what are the cross-validation results for algorithms you tried, what attribute selection methods you tried. Please only include the evaluation metrics which are appropriate for the problem. Don't include the unnecessary metrics and the details of the learned model (e.g., the decision tree).

Submit the document on the D2L site. The document should be named as cps844w21_yourname. The PDF file is required. If the dataset is not in a public domain, you also need to submit the data file, and in this case, a zipped file should be submitted. In case that you write programs, please also submit a zipped file including the report and the source code. If you choose to use Jupyter notebook, please make sure that the document part covers everything required.