Description

The class project is due at the end of finals. You can choose to do essentially anything inside the area of data mining, so long as I approve it in advance. You can build on material we do in class, or you can use it as an opportunity to learn about subject matter we don't have time to cover. Perhaps you'd like to find an algorithm described in the textbook or in a paper that we haven't talked about, and implement it. If you go this route, you should turn in a short description of what you implemented, a reference to what you've done, and your working code with data. Alternatively, perhaps you can find a dataset of interest, and use either your own code or R or Weka to heavily analyze it and learn what you can. There are a number of data mining contests out there: perhaps you want to pick a current one and try it. If you go the route of analyzing a dataset, you should turn in a paper describing what you've learned: tell me what patterns you've found in the dataset, and interpret the results (tie it back to reality).
Here are some potential sources of data:

1. UCI KDD Archive
2. UCI Machine Learning Repository
3. Kaggle. Kaggle is often running a variety of contests you may be interested in trying.

Most online data comes with some kind of license. Make sure that you can appropriately use the data as you intend to.
You are strongly encouraged to work in pairs. There are a lot of seniors in the class, and the college gives me essentially one day to have everything graded for them. If you wish to work individually, you must turn in your project early.

Proposal

Submit a one page proposal, including the names of both people involved (if it is a team project), describing what it is you want to do for your project.

Deadline for working in pairs

If working in a pair, your deadline is the end of the last final exam. That will be on Monday, May 10, at 9:30 pm. This is not a midnight deadline: the end of the last exam is 9:30 pm.
If you miss this deadline, I am by forbidden by college policy to extended it, and I cannot grade it. You will need to contact your class dean for permission to have late work graded.

Deadline for working individually

If you choose to work individually, your deadline is Wednesday, June 5, at 11:55 pm. That's the end of the last day of classes. (Note that you have an exam on the same day.)
If you miss this deadline, the standard course late policy will apply; see the syllabus, since the reason this option is structured this way is to make sure that I get individual projects in early enough so that I can grade them.

Submitting your data

If you are using a large dataset, you shouldn't submit it via Moodle. There are submission limits, and Moodle's disk capacity is likely limited. If your dataset is large, make sure to work with me before the deadline so you can get me the data via some other means.