Course Project

You have two options:

(a). use any programming language / software and datasets to do the course project (e.g., python, your own mobile phone data, etc)

(b). use the given data and Weka to do the course project.

Project steps:

1. Datasets:
   1. If you will use the given data, here are the links to download them (link1 (weather.arff), link2(soybean.arff), [link3](http://kdd.ics.uci.edu/databases/tic/tic.html) (use only the ticdata2000.txt Training data. (1M).))
2. Experiments:
   1. If you will use Weka,
      * Translate the dataset into the arff format if needed.
      * Open the dataset in Weka.
      * Preprocess the dataset attributes using Weka's filters. In particular,
        1. explore different ways of discretizing continuous attributes. That is, convert numeric attributes into "nominal" ones by binning numeric values into intervals - See the weka.filter.DiscretizeFilter in Weka. Play with the filter and read the Java code implementing it.
        2. explore different ways of removing missing values. Missing values in arff files are represented with the character "?". See the weka.filter.ReplaceMissingValuesFilter in Weka. Play with the filter and read the Java code implementing it.
   2. Use the "ZeroR" classifier under the "Classify" tab. Use different ways of testing your results. That is, explore the following alternatives offered by the Weka system:
      * Testing your results over the training data.
      * Splitting your input file into two parts one for training and one for testing.
      * Using n-fold cross validation. Play with different values for n.

Analyze the results obtained (i.e. interpret the meaning of the output produced by Weka). Read to the extent possible the Java code implementing the ZeroR classifier.

* 1. Run several experiments with your data and the system varying the parameters so that you gain familiarity with the system.

1. Write a 2 to 4-page report of your analyses.
   1. Data: Describe the datasets that you used in terms of the attributes present in the data, the number of instances, missing values, and other relevant characteristics.
   2. Code Description: Describe to the extent possible any observations you made when looking at the Weka code implementing the filters you used and the ZeroR function.
   3. Experiments: For each experiment you ran describe:
   4. Instances: What data did you use for the experiments? That is, did you use the entire dataset of just a subset of it?
   5. Any pre-processing done to the data. That is, did you remove any attributes? Did you discretize any continuous attribute? If so, what strategy did you use to bin the values? Did you replace missing values? If so, what strategy did you use to select a replacement of the missing values?
   6. Your system parameters.
   7. For the ZeroR function, analysis of results of the experiments you ran using different ways of testing the classifier (crossvalidation, etc.).
   8. Summary of Results
   9. Discuss the strengths and the weaknesses of your project.