1. **Note**: HW8 data is the same as HW7:

* [original training data](https://s3-us-west-2.amazonaws.com/syr-mac/prod/IST+565+Data+Mining/PDFs/Data+and+Code+/Kaggle-digit-train.csv) (77MB)
* [original test data](https://s3-us-west-2.amazonaws.com/syr-mac/prod/IST+565+Data+Mining/PDFs/Data+and+Code+/Kaggle-digit-test.csv) (51MB)
* CSV: [small sample of training data](https://s3-us-west-2.amazonaws.com/syr-mac/prod/IST+565+Data+Mining/PDFs/Data+and+Code+/Kaggle-digit-train-sample-small-1400.csv)
* ARFF: [small sample of training data](https://s3-us-west-2.amazonaws.com/syr-mac/prod/IST+565+Data+Mining/PDFs/Data+and+Code+/Kaggle-digit-train-sample-small.arff)
* CSV: [small sample of test data](https://s3-us-west-2.amazonaws.com/syr-mac/prod/IST+565+Data+Mining/PDFs/Data+and+Code+/Kaggle-digit-test-sample1000.csv)
* ARFF: [small sample of test data](https://s3-us-west-2.amazonaws.com/syr-mac/prod/IST+565+Data+Mining/PDFs/Data+and+Code+/Kaggle-digit-test-sample1000.arff)

2. Sample code for calling multiple classification models in RWeka:

* [digit-multi-model-sample-code](https://s3-us-west-2.amazonaws.com/syr-mac/prod/IST+565+Data+Mining/PDFs/Data+and+Code+/digit-multi-model-sample-code.R)

3. **Instruction on increasing java heapsize for Weka**: For both windows and mac, go to the folder where your weka.jar is located, then use the following command to launch your weka GUI. In this command the memory size is set to 1g, you can choose to increase or decrease based on your computer config.

* "java -Xmx1g -jar weka.jar"

If you have difficulty finding your weka.jar on mac, see if the following tip helps. On my mac I don't have the VMOptions, but I do find weka.jar under the same folder. You can then use the "Get Info" to find the entire path.

* Go to Application==> Left click on specific application (like Weka)==> Show Package Content==> Contents==>Info.Plist==>Java==> VMOptions==> increase the size of memory as you want (in my mac the default was 1 Gb, I increased to 6 Gb)

**FAQs**

**Q**: Why don’t the HW8 test examples have labels?

**A**: You are right the test data does not have label. It is to show that you can cut the large test data into this smaller chunks if your Weka complains out of memory. Then you need to combine them back together in order to submit the complete answer to Kaggle.

If you did not get to submit to Kaggle because of the confusion, it's ok for HW8 as long as you understand how to do it. It can take a while to run all test examples through Weka. Using R for this task is definitely faster.

**Q**: Where can I find “numTrees” in random forest in the new Weka version?

**A**: In the new Weka version, numTrees becomes numIterations, and the default value 10 trees becomes 100 trees. If your data is not large, consider lowering the number to 10, 20 or 50. If you are working on the entire digit recognition data set, then 100 trees might be ok. Can tune this parameter to find out.

**Q**: Where can I find “ChiSquareAttributeEval” in Weka?

**A**: Note that “ChiSquareAttributeEval” has been removed from the basic Weka 3.8 package. You have to use Package Manager to install chiSquareAttributeEval, and then it will show up in "Select attributes" choices.