Tutorial Notes

1. 38,160 instances (rows)
2. 20 attributes
3. 1 class/target feature (column contam\_code)
4. obscuredA.arff and obscuredB.arff are identical except the final column has been randomly shuffled
5. Confusion matrix can be used to see whether your model is biased toward certain classes.

How can you tell there is a signal?

1. Train a model (e.g. decision tree, neural net etc) that is more accurate than just looking at the majority class
2. Look at the confusion matrix and seeing if the model just classifies everything under the majority class
3. If you build a decision tree and find that the attributes are being used

Techniques:

1. Majority class as a baseline
2. Use 10-fold cross validation
3. Confusion matrix comparison

Confusion Matrix Comparison