**CORE METHODS IN EDUCATIONAL DATA MINING
FALL 2014
BASIC ASSIGNMENT 1: CLASSIFIER**

In this assignment, you need to build a set of detectors for data file CogSci-Godwinetal-2013-3.csv. This data set is very similar (but not identical) to the data set used in

Godwin, K.E., Almeda, M.V., Petroccia, M., Baker, R.S., Fisher, A.V. (2013) Classroom activities and off-task behavior in elementary school children. Poster paper. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2428-2433.

This paper can be found at http://www.columbia.edu/~rsb2162/Godwinetal\_v12.pdf

A description of many of the variables can be found in that paper

You should complete this assignment in RapidMiner 5.3.

This assignment is due at noon, September 15, 2014.

Question 1:

Build a decision tree (using operator W-J48 from the Weka Extension Pack) on the entire data set, to predict the variable “ONTASK”. What is the non-cross-validated kappa?

Question 2:

The kappa value you just obtained is artificially high – the model is over-fitting to which student it is. What is the non-cross-validated kappa, if you build the model (using the same operator), excluding student?

Question 3: Exclude school, class, and coder from the data set, as these features will make your model overly specific to the current data set. What is the non-cross-validated kappa, if you build the W-J48 decision tree model (using the same operator), excluding student and the variables from Question 3?

Question 4: What is the non-cross-validated kappa, for the same set of variables you used for question 3, if you use Naïve Bayes?

Question 5: What is the non-cross-validated kappa, for the same set of variables you used for question 3, if you use W-JRip?

Question 6: What is the non-cross-validated kappa, for the same set of variables you used for question 3, if you use Logistic Regression? (Hint: You will need to transform some variables to make this work; RapidMiner will tell you what to do)

Question 7: What is the non-cross-validated kappa, for the same set of variables you used for question 3, if you use Step Regression (called Linear Regression in RapidMiner)?

Question 8: What is the non-cross-validated kappa, for the same set of variables you used for question 3, if you use k-NN instead of W-J48?

Question 9: What is the kappa, for the same set of variables you used for question 3, if you use W-J48, and conduct 10-fold stratified-sample cross-validation?

Correct answer: 0.106

Question 10: Why is the kappa lower for question 9 (cross-validation) than question 3 (no cross-validation?)

Question 11: What is the kappa, for the same set of variables you used for question 3, if you use k-NN, and conduct 10-fold stratified-sample cross-validation?

Question 12: k-NN and W-J48 got almost the same Kappa when compared using cross-validation. But the kappa for k-NN was much higher (1.000) when cross-validation wasn’t used. Why is that?