1. Description

**The Problem**

Final exams in certain courses consist of multiple-choice questions and are often administered electronically. Thus, each student’s answers can be represented as a line:

012345678, 2333132442444444124125534252552141d351141414125321144331222343431134154431441244221425112.....

The line consists of the student ID, a separator, and a vector of answers. The numbers 1, ..., 5 map to answers A, ..., E respectively.

Correspondingly, the answer key can be represented the same way:

answerkey,  22313224424434441243355515153223312333541314511132153221554142342221234322422231221425....

The instructor loads the answer key into grading software, which then evaluates all of the exams and returns a grade for each student.  If cheating is witnessed, corroborating evidence can be obtained from the exam data. If a student copies the answers from someone sitting nearby, the two answer vectors may appear suspiciously similar.

1. Instructions

**Your Job:**

In this course, we have so far encountered the following algorithms:  Simple Decision Trees, Naive Bayes, k Nearest Neighbor, Random Forest, Neural Networks/ Deep Learning, and the Generalized Linear Model.  Using the classification algorithm given to you in the question below on the attached sample dataset [Week8\_cheating.csv](https://blackboard.valpo.edu/bbcswebdav/pid-440744-dt-content-rid-2539394_1/xid-2539394_1)  of 10 exam questions, design a solution that could be implemented in a new cheating detection software.  You will need Excel, R, or h2o (or all of these) to preprocess the data, run your algorithm, and test its effectiveness. Then pick a second algorithm from the list and rerun your analysis.

Answer the following questions (2 pts each):

* 1. What preprocessing do you have to do to make the data work for your assigned algorithm?  Are there some solutions that obviously show the student has cheated?  How did you determine that?
  2. What is your class?  Did you add one?
  3. How does the algorithm you are given process the dataset to detect cheating?  What model does it build, and how reliable is the model (if you get error percentages, list these)?
  4. How does the (second) algorithm you picked from the list process the dataset to detect cheating?  What model does it build, and how reliable is the model (if you get error percentages, list these)?
  5. Given your results, which algorithm would you suggest should be integrated into the cheating detection software?

As always, show your R code and your h2o screenshots.

1. Force Completion

This test can be saved and resumed later.

1. Multiple Attempts

This test allows multiple attempts.

1. Due Date

This Test is due on March 3, 2017 11:59:00 PM CST.

Click **Continue** to continue: Quiz 8: Advanced Classification with Random Forest, Deep Learning, and GLM. Click **Cancel** to go back.

**2.  Submit**

*Click Begin to start. Click Cancel to quit.*

Your first assigned algorithm is the **Random Forest** algorithm.