**CS 4319 – Data Mining & Warehouses**

**Programming Assignment 2**

You are provided with a dataset that includes data of people who are interviewed for the Data Scientist job. The data is below and the fields are; each candidate’s level, their preferred language, whether they are active on Twitter and whether he/she has a PhD. The class label (last attribute) is either Yes (the candidate interviewed well) or No (the candidate interviewed poorly.

inputs = [

({'level':'Senior','lang':'Java','tweets':'no','phd':'no'}, False),

({'level':'Senior','lang':'Java','tweets':'no','phd':'yes'}, False),

({'level':'Mid','lang':'Python','tweets':'no','phd':'no'}, True),

({'level':'Junior','lang':'Python','tweets':'no','phd':'no'}, True),

({'level':'Junior','lang':'R','tweets':'yes','phd':'no'}, True),

({'level':'Junior','lang':'R','tweets':'yes','phd':'yes'}, False),

({'level':'Mid','lang':'R','tweets':'yes','phd':'yes'}, True),

({'level':'Senior','lang':'Python','tweets':'no','phd':'no'}, False),

({'level':'Senior','lang':'R','tweets':'yes','phd':'no'}, True),

({'level':'Junior','lang':'Python','tweets':'yes','phd':'no'}, True),

({'level':'Senior','lang':'Python','tweets':'yes','phd':'yes'},True),

({'level':'Mid','lang':'Python','tweets':'no','phd':'yes'}, True),

({'level':'Mid','lang':'Java','tweets':'yes','phd':'no'}, True),

({'level':'Junior','lang':'Python','tweets':'no','phd':'yes'},False)

]

You are asked using this data to build a model identifying which candidates will interview well? (so that your boss does not have to waste his time interviewing candidates!)

HINTS:

* Use decision trees, and ID3 decision tree
* Use Entropy to decide whether or not split a node, or on which attribute to split (unlike in class, where we used mostly GINI)