**Problem 1** Given the following weather dataset:

**Day Outlook Temp Humidity Windy Play**

D1 Overcast Hot Normal False Yes

D2 Overcast Mild High True Yes

D3 Sunny Mild Normal True Yes

D4 Rainy Mild Normal False Yes

D5 Sunny Cool Normal False Yes

D6 Sunny Mild High False No

D7 Overcast Cool Normal True Yes

D8 Rainy Cool Normal True No

D9 Rainy Cool Normal False Yes

D10 Rainy Mild High False Yes

D11 Sunny Hot High False No

D12 Rainy Cool High True No

1. **[10 points]** Calculate the Gain(S,A1) and Gain(S,A2) for the dataset S and attributes A1 and A2**. Show all your work.** (Show each step of the calculation. )

**A1 = Outlook**

**Outlook = Overcast: info(3,0) = entropy(1,0) = -1log(1) – 0log(0) =   
 0 bits**

**Outlook = Sunny: info(2,2) = entropy(1/2,1/2) = -1/2log(1/2) – 1/2log(1/2) = 0.301 bits**

**Outlook = Rainy: info(3,2) = entropy(3/5,2/5) = -3/5log(3/5) – 2/5log(2/5) = 0.292 bits**

**Expected: info([3,0],[2,2],[3,2]) = (3/12)\*0 + (4/12)\*0.301 + (5/12)\*0.292 = 0.222 bits**

**gain(Outlook) = info([8,4]) – 0.222 = 0.276 – 0.222 = 0.054 bits**

**A2 = Temperature**

**Temperature = Hot: info(1,1) = entropy(1,1) = -1/2log(1/2) – 1/2log(1/2) = 0.301 bits**

**Temperature = Mild: info(4,1) = entropy(4,1) = -4/5log(4/5) – 1/5log(1/5) = 0.217 bits**

**Temperature = Cool: info(3,2) = entropy(3,2) = -3/5log(3/5) – 2/5log(2/5) = 0.292**

**Expected: info([1,1],[4,1],[3,2]) = (2/12)\*0.301 + (5/12)\*0.217 + (5/12)\*0.292 = 0.262**

**gain(Temperature) = info([8,4]) – 0.262 = 0.276 – 0.262 = 0.014 bits**

1. **[5 points]** Calculate the Gain\_ratio(S,A1) and Gain\_ratio(S,A2) for the dataset S and attributes A1 and A2**. Show all your work.** (Show each step of the calculation. )If you use ID3 algorithm, which of these two attributes you will chose for a root node of a decision tree? **Explain your answer!   
     
   Gain ratio(S, A1): info([3,4,5]) = 0.468 bits. Ratio = 0.045/0.468 = 0.096 bits  
     
   Gain ratio(S, A2): info([2,5,5]) = 0.447 bits. Ratio = 0.014/0.447 = 0.031 bits  
     
   If I were to use ID3 I would choose the Outlook attribute because it has the greatest gain ratio and therefore the greatest chance at producing accurate predictions.**
2. **[10 points]** Construct classification rule for class=”Yes” (Play=”Yes”) using Simple Covering Algorithm. **Show all stages/steps of rule construction and all your calculations!  
     
   If ? then Play = Yes  
     
   Outlook = Overcast 3/12  
   Outlook = Sunny 2/12  
   Outlook = Rainy 3/12  
   Temperature = Hot 1/12  
   Temperature = Mild 4/12  
   Temperature = Cool 3/12  
   Humidity = Normal 6/12  
   Humidity = High 2/12  
   Windy = True 3/12**

**Windy = False 5/12  
  
 If Humidity = Normal and ? then Play = Yes  
  
 Outlook = Overcast 3/12  
 Outlook = Sunny 2/12  
 Outlook = Rainy 3/12  
 Temperature = Hot 1/12  
 Temperature = Mild 4/12  
 Temperature = Cool 3/12  
 Windy = True 3/12**

**Windy = False 5/12  
  
 If Humidity = Normal and Windy = False and ? then Play = Yes  
  
 Outlook = Overcast 3/12  
 Outlook = Sunny 2/12  
 Outlook = Rainy 3/12  
 Temperature = Hot 1/12  
 Temperature = Mild 4/12  
 Temperature = Cool 3/12  
  
 If Humidity = Normal and Windy = False and Temperature = Mild and Outlook = Overcast then Play = Yes**

**Problem 2 [25 points] – Programming - Using WEKA.**

1. **Data Mining Technique(s):** We will run experiment using the following decision trees techniques:
   * ID3, and
   * J4.8 (given that J4.8 is able to handle numeric attributes and missing values directly, make sure to run some experiments with no pre-processing and some experiments with pre-processing, and compare your results).
2. **Dataset(s):** In this project, we will use two datasets:
   * The [census-income dataset](ftp://ftp.ics.uci.edu/pub/machine-learning-databases/census-income/) from the US Census Bureau which is available at the <http://archive.ics.uci.edu/ml/datasets.html> ([Univ. of California Irvine Repository](http://www.ics.uci.edu/~mlearn/MLRepository.html)).  
     The census-income dataset contains census information for 48,842 people. It has 14 attributes for each person (age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country) and a boolean attribute *class* classifying the input of the person as belonging to one of two categories >50K, <=50K.
   * A dataset that you choose depending on your own insterests. It should contain enough instances (at least 200 instances) and several attributes (at least 10). Ideally it should contain a good mix of numeric and nominal attributes.   
     I include below some links to Data Repositories containing multiple datasets to choose from:
     + [Univ. of California Irvine KDD Data Repository](http://kdd.ics.uci.edu/).
     + [Univ. of California Irvine Machine Learning Data Repository](http://www.ics.uci.edu/~mlearn/MLRepository.html).
     + [Time Series Data Library](http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/)
     + [Datasets for Data Mining](http://www.kdnuggets.com/datasets.html)
     + [CMU's StatLib-Datasets Archive](http://lib.stat.cmu.edu/datasets/)
     + Miscellaneous
       - [Intrusion Detection Data - Sequences](http://iris.cs.uml.edu:8080/network.html)

**THIS DATASET CANNOT BE ONE OF THOSE INCLUDED IN THE WEKA SYSTEM.**

1. **Performance Metric(s):**
   * Use (1) classification accuracy, (2) size of the tree, and (3) readability of the tree, as separate measures to evaluate the "goodness" of your models.
   * Compare each accuracy you obtained against those of benchmarking techniques as ZeroR and OneR over the same (sub-)set of data instances you used in the corresponding experiment (the same training set).
2. Your Problem 2 report should include:

* Description of your second dataset – link to it, size, attributes, attributes type;
* Training set size (in % of the entire dataset) and testing set size (in % of the entire dataset);
* Performance metrics (accuracy, size of the tree, readability of the tree) using ID3, J4.8;
* Performance metrics (accuracy) for both your datasets using ZeroR and OneR algorithms;
* Accuracy comparison for ID3, J4.8 and ZeroR and OneR classifiers.

**Submit your Problem 1. and Problem 2. to me as a hard copy on Monday, March 21, at the beginning of class.**