# CS 7265- Homework 2

In HW2, we will implement Decision Trees, where we train a decision tree model with training data and classify with new data (test data).

You are given two data sets: car.training.csv and car.test.csv, where "car.training.csv" is a training data that you will induce a decision tree model from, and "car.test.csv" is a test data that you classify with the trained decision tree model.

You can find the data description at https://archive.ics.uci.edu/ml/datasets/Car+Evaluation, but have to use the given simplified data sets (all classes are equally distributed).

For this homework assignment, please follow the procedure:

1. Make a decision tree by using one of GINI index, Information Gain (or Gain Ratio), or misclassification errors. You can also compare the performances of the three approaches (strongly recommended, but not required).
2. Show **WHAT FEATURES** are selected in order for each node. See Figure 2.
3. Show your decision tree model as a figure. You can manually draw the tree, but recommended to (automatically) construct the decision tree using the data structure of TREE in the script language.
4. Show the accuracy of your decision tree. For accuracy, you have to use the given test data. The test data includes class labels (acc vs unacc), so you check whether your prediction is correct or not. Then you can compute accuracy by:

$$accuracy = \frac{\#\ of\ your\ predictions\ which\ are\ correct}{\#\ of\ total\ test\ data}$$

5. You can convert the ordinal data to numerical data for simplicity's sake. (but it may not give the optimal performance)
6. **You CANNOT use library or built-in functions of decision trees. You have to implement it.**

You have to submit the followings to D2L:

1. MS word file
   - Describe what you have done for the homework assignment.
   - Include the decision tree model as a figure.
   - Include the accuracy that you got with your decision tree.
2. Source code file(s)
   - Any languages, but recommend R or Python
   - Must be well organized (comments, indentation, …)
   - You need to upload the "original R or python file (*.r or *.py)" to text files E.g., "*.r.txt" or ".py.txt".

You have to submit the files SEPERATELY. DO NOT compress into a ZIP file. If you fail to provide all required information or files, you may be given zero score without grading.

**Deadline:**

The deadline is **11:59pm Sunday, February 17, 2019**. Late assignments will not be accepted.

This is an example of the outcome. Fig. 1 shows the message how the model constructs during the training process, and Fig. 2 is the result with the test data. Fig. 3 illustrates the optimal decision tree trained with the training data.

```
*****************************************
** Decision Tree for [data/car.training] **
*****************************************

Decision: buying, Information Gain=0.4932452372049694
-- Attribute: high
---- Decision: maint, Information Gain=1.0
------ Attribute: vhigh
-------- Class: unacc
------ Attribute: high or low or med
-------- Class: acc
-- Attribute: vhigh
---- Decision: maint, Information Gain=0.5408520829727552
------ Attribute: low
-------- Decision: safety, Information Gain=0.37093224580041206
---------- Attribute: high
------------ Class: acc
---------- Attribute: low or med
------------ Class: unacc
------ Attribute: med
-------- Decision: safety, Information Gain=0.37093224580041206
---------- Attribute: high
------------ Class: acc
---------- Attribute: low or med
```

```
*************************************
** Test Results for [data/car.test] **
*************************************

training_mode information_gain
matches 168
test_rows 200
overall 0.84
```

*Figure 2. Testing*

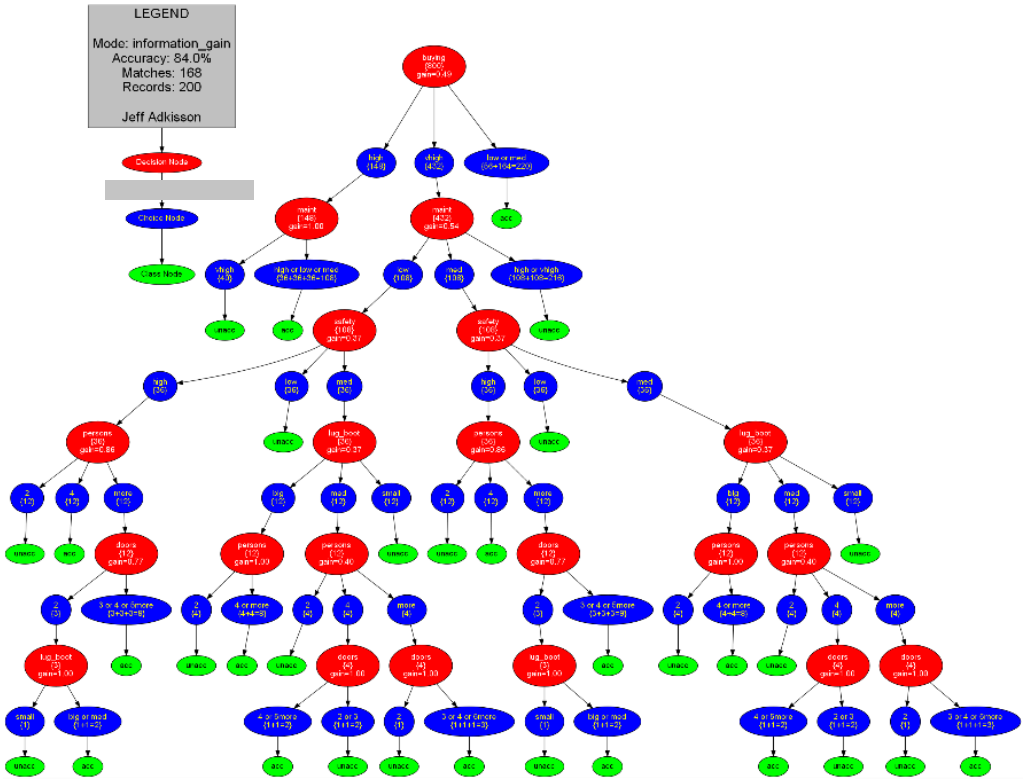


*Figure 3. Decision Tree*