# Task 1: Exploring Dataset w/ Entropy & Gini

To demonstrate your understanding of Decision Trees, start by calculating a few measure with the dataset:

* General Information: How many examples, How many positives (survived), How many negatives (did not survive)?
  + There were 891 recorded victims (examples?) in the Titanic data provided. Out of the 891 victims, 342 survived (positive) and 549 did not survive (negative).
* What is the Entropy and Gini for the training set data?

The following code will be used to determine the Entropy, Gini, and Information Gain of a sample:

**def** entropy(parameters):  
 sum = 0.0  
 **for** i **in** parameters:  
 sum -= i\*float(log(i,2))  
 **return** sum  
  
**def** gini(parameters):  
 sum = 0.0  
 **for** i **in** parameters:  
 sum += i\*(1-i)  
 **return** sum  
  
**def** infoGain(newResult, oldResult):  
 **return** oldResult- newResult

* + We know that according to the and . So, using the entropy function above, we get the following: 0.9607
  + We know that the equation for gini is the following:. We also know that according to the and . So, using the entropy function above, we get the following: 0.4730
* What is the information gain using both metrics for the following features: sex (male/female), pclass (1/2/3), embarked (C/Q/S).
  + So, using the sex metrics, we will have the following probabilities:

Then, the new entropy is: 0. 7441

The information gain is:

* + So, using the pclass metrics, we will have the following probabilities:

And the new entropy is: 0.8769

The information gain is:

* + So, using the embarked metrics, we will have the following probabilities:

And the new entropy is: : 0. 9367

The information gain is:

# Task 2: Build Simple Decision Tree`

Using either entropy or gini, build a two level tree for predicting survival using only the features sex, pclass or embarked (you may or may not use all 3). Then determine how well your tree performs on the test data.

**Sex**

IG =

**Class**

IG =

**Class**

IG =

**1**

Not

Survived

**1**

Survived

**2**

Survived

**3**

Split?

**2**

Not

Survived

**3**

Not

Survived

?

# 

# The tree shows that all men would have died in the Titanic accident. That means that a total of 577 males would have died out of 577. The data shows that only 468 males died. That means that the tree predicted 109 incorrect outcomes. One of the tree options is that all of the females would have survived the accident. That means that a total of 0 females would have died. The data, however, shows that only 81 out of 342 females died. That means that the tree predicted another 81 outcomes incorrectly. Overall, the tree predicted 190 incorrect outcomes, and 701 correct outcomes. That means the tree has a success rate of .

# Task 3: Explore ML Tool

Using an ML Tool (Weka if no other choice) build a decision tree using the training data. Document the tree and its performance, and contrast and compare with your simple tree from task 2. Weka can be downloaded for multiple platforms from the URL below:

<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

