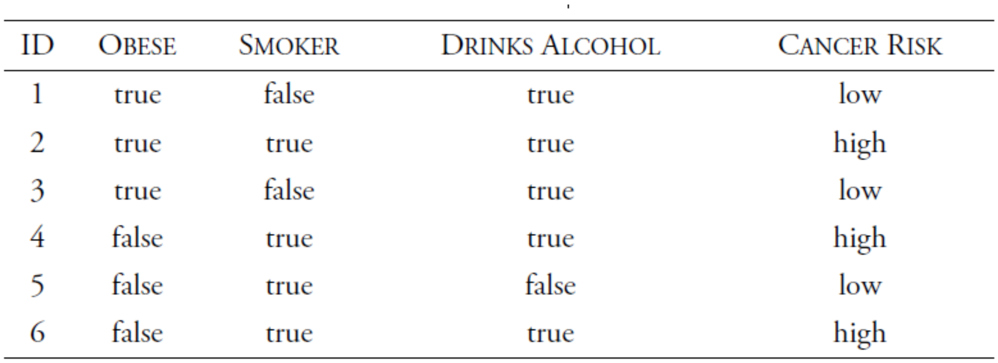
Lab 4: Decision Tree

CPSC429/529 Machine Learning

1. **Hand written part**

   The following table lists a dataset containing the details of six patients. Each patient is described in terms of three binary descriptive features (OBESE, SMOKER, and DRINKS ALCOHOL) and a target feature (CANCER RISK). You will use ID3 algorithm discussed in class to find out the feature for the root node of the decision tree. Particularly, you will need to do the followings:



ID	OBESE	SMOKER	DRINKS ALCOHOL	CANCER RISK
1	true	false	true	low
2	true	true	true	high
3	true	false	true	low
4	false	true	true	high
5	false	true	false	low
6	false	true	true	high

   (a) Calculate the overall entropy $Entropy(S)$, where S is the whole dataset.

   (b) Calculate the information gain for the feature of OBESE. Note: you need to calculate each entropy $Entropy(S_f)$ first, where S_f is the subdataset split by the feature, and then weighted entropy, and finally the information gain.

   (c) Calculate the information gain for the feature of SMOKER. Note: you need to calculate each entropy $Entropy(S_f)$ first, where S_f is the subdataset split by the feature, and then weighted entropy, and finally the information gain.

   (d) Calculate the information gain for the feature of DRINKS ALCOHOL. Note: you need to calculate each entropy $Entropy(S_f)$ first, where S_f is the subdataset split by the feature, and then weighted entropy, and finally the information gain.

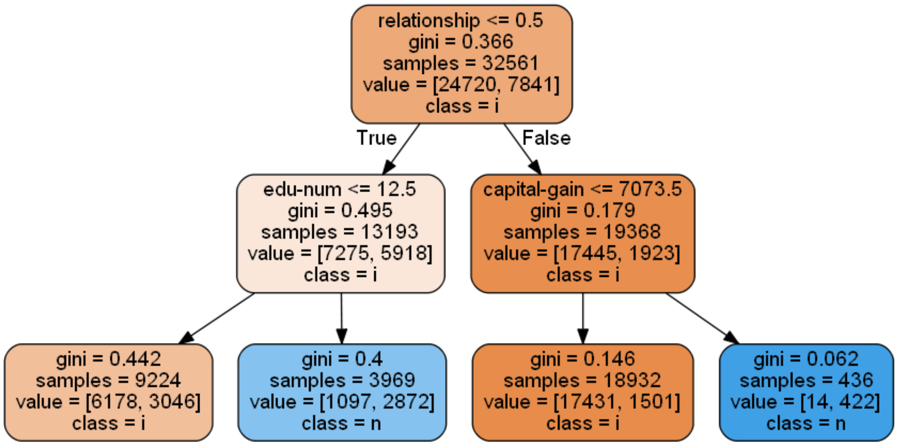
   (e) Draw the decision tree after the data has been split using your best feature you found.

   Submission instruction: Take a picture of your hand written answer, save it as `lab4_1.png` or `lab4_1.pdf`, upload it to D2L and hand-in the original copy to me.

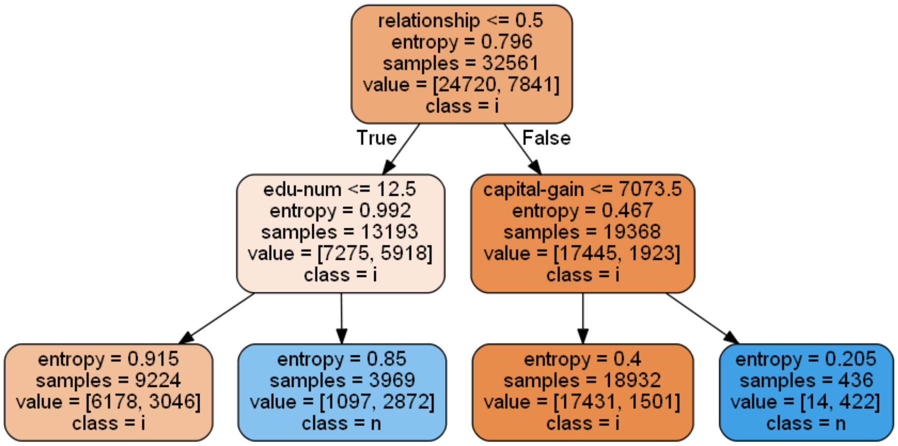
2. **Programming part**

You are given a adult dataset (`adult.data`) (`https://archive.ics.uci.edu/ml/datasets/adult`). The data preprocessing step is done for you already. You are asked to do the followings:

(a) Use dataset (**X, y**) build your decision tree model (criterion='gini', max_depth=2). generate the decision tree image corresponding to the model, and display it on Jupyter notebook (See below as reference).



(b) Use dataset (**X, y**) build your decision tree model (criterion='entropy', max_depth=2). generate the decision tree image corresponding to the model, and display it on Jupyter notebook (See below as reference).



The jupyter notebook skeleton of lab 3 (`Lab4_2_DecisionTree.ipynb`) is given to you, so you can complete the remaining parts (**Hints:** Refer lecture slides).

Submission instruction: Submit this jupyter notebook (`Lab4_2_DecisionTree.ipynb`) to D2L, and demo it to me.