EE6435 Homework 2

**Points**: 60.

**Out**: September 18, 2020 (Friday)

**Due**: **11:59PM**, September 30, 2020 (Wed.). **No late homework will be accepted**.

**Handin method and requirement**: name your notebook file (.ipynb) as **yourlastname-firstname-studentID-hw2.ipynb**. For example, if your name is Amy Zhang, the file should be named as zhang-amy-5678910-hw2.ipynb. Also, attach an html file (generated by the notebook file) with your notebook using the naming rule: **yourlastname-firstname-studentID-hw2.html**.

======================================================================

**Homework overview: Implement the decision tree construction using entropy as the pureness function. As this is your first programming homework, we will be gentle in the requirement. You don’t need to output the whole tree. Instead, just output **two top layers**. See the picture below:**

**Yes No**

**Data:**

Canvas🡪files/homework 🡪 hw2-decision-tree-input.txt

The training data is used to decide the risk of heart attack.

The first line shows the attributes of the training data (all binary to simplify your programming). The last attribute is the target attribute, also known as the label.

Then you will see the training data.

The six binary attributes are:

Chest pain, Male, Smoke, Drink, Exercise, Heart attack

It is followed by training data from previous patients/healthy individuals. For example, if there is an individual who has chest pain, is male, does not smoke, drinks, does not exercise, and had a heart attack, the corresponding representation in the file is:

1,1,0,1,0,1

**Requirement and grading:**

1. **Your python program must take a file as input. Don’t hardcode the training data. (10 pts)**
2. **Test your program using the given training data. (20 pts)**
3. **Test your program using a different training data file (same format, but different number of binary attributes, different number of samples) (20 pts)**
4. **Output format: you can output the tree using text description. If so, use the following format (10 pts):** 
   1. **The root node is attribute \_\_\_. Its left edge has label \_\_\_. Its right edge has label \_\_\_. Its left child node’s attribute is \_\_\_\_. Its right child node’s attribute is \_\_\_\_\_.**
   2. **If any child node is already the target attribute, just put the target attribute (i.e. label). It should be either “heart attack” or “no heart attack”.**

To get partial credit, I suggest that you output some intermediate information. For example, if your code can read the file correctly, output the training data, which will give you 10 pts.