CS 472/572, Spring 2024
Homework 2 (Programming): Decision Trees
DUE DATE: May 1st at 11:59pm.

1 Requirement

In this assignment, you will implement the ID3 algorithm for learning decision trees. You may assume that the class label and all attributes are binary (only 2 values). Please use the provided skeleton code in Python to implement the algorithm.

The ID3 algorithm is similar to what we discussed in class: Start with an empty tree and build it recursively. Use the information gain to select the attribute to split on. (Do not divide by split information.) Use a threshold on the information gain to determine when to stop. The full algorithm is described in this paper: http://dept.cs.williams.edu/ andrea/cs374/Articles/Quinlan.pdf

You may look at open-source reference implementations, such as WEKA, but please **do not copy code from open-source projects**. Your code must be your own. Undergraduates may complete the assignment in a team of 2. Graduates must complete the assignment alone.

2 Starter code

The starter code is provided at `https://ix.cs.uoregon.edu/~swalton2/cis472/hw2-starter-code.zip`. You should write your code in python 3. The code should run from the command line and accept the following arguments:

```
python3 id3.py <train> <test> <model>
```

where train/test are the paths to files containing training data and testing data; model is a path to a file where you will save the model for the decision tree. If you don't have an Unix-like environment (e.g. MacOS, Linux, WSL), you can use the ix.cs.uoregon.edu server; ask **systems@cs.uoregon.edu** for any issue while accessing this server.

The data files are in CSV format. The first line lists the names of the attributes. The last attribute is the class label.

We are providing skeleton code in Python that handles input, output, and some of the internal data structures. Please use it as the starting point because we'd be using that API to grade.

For saving model files, please use the following format:

```
wesley = 0 :
| honor = 0 :
```

```
| | barclay = 0 : 1
| | barclay = 1 : 0
| honor = 1 :
| | tea = 0 : 0
| | tea = 1 : 1
wesley = 1 : 0
```

According to this tree, if wesley = 0 and honor = 0 and barclay = 0, then the class value of the corresponding instance should be 1. In other words, the value appearing before a colon is an attribute value, and the value appearing after a colon is a class value.

Once you are done with coding, you may want to check your code with this autograder: `https://ix.cs.uoregon.edu/~swalton2/cis472/hw2/upload.html`. You should submit a single file named **id3.py**. The testing may take upto 500 seconds per submission.

3 Turn in

You should submit a single file **id3.py** through canvas.