Programming Assignment 2

Write a Python3 program (you can do that in a group of 3 to 5 people, but everybody has to submit individually!) that implements a decision tree using the ID3 algorithm presented in the lecture. Use the following entropy calculation:

$$entropy(S) = -\sum_{i=1}^{C} p_i log_C(p_i)$$

where p_i is the proportion of class i (with C being all classes in the data set). Use Information Gain as your decision measure and treat all features as discrete multinomial distributions.
Given are the three data sets (on Moodle) named *car*, *breast-cancer* and *nursery* as csv files. Your program should be able to read those data sets and treat the last value of each line as the class. Your task is to correctly implement the ID3 algorithm and return the nodes of the final tree without stopping early. In the event, that no more attribute can be split upon, use the majority class to decide on the output of the leaf. Of course, you would still stop earlier, if entropy is 0! The output of your algorithm (again on console only) should look like the example solution given for the *car* data set. With that, you can check the correctness of your solution. All features are unnamed on purpose, please number them according to the column starting from 0 (e.g. att0). I.e. the meaning of each line is as follows:
`depth,attname=value,entropy,class`
Here, `depth` indicates the depth of the tree (root node is depth 1), `attname` the name of the attribute, which is split upon (e.g. att1), `value` the corresponding of value of that attribute for the branch of that decision, `entropy` the entropy value of the node after filtering by that attribute and `class` the class value to which this branch leads to. Therefore, if the branch leads to a leaf node, it should have the actual class as an output. If the branch leads to another decision node, then it has the special value `no_leaf`. Please note, there is a special node at the beginning, which has depth 0, is called root and has the entropy of the whole data set. (All following nodes only have the entropy of the splits, that's why this is necessary.)
For each data set, you can acquire one point, if the solution of your program returns correct results. If the program fails, the data format is incorrect or nodes are wrong or missing, you will get zero points. Machine learning libraries are not allowed. You can use numpy and ElementTree though, if you want.
Name your program `student.py`. Your program must accept the following parameters:

1. **data** - The location of the data file (e.g. car.csv).

Your program on the server will therefore be executed as follows:

```
python3 student.py --data car.csv
```

The final program code must be uploaded to Moodle until Wednesday, the 9th of December, 1 am.

2 points