

# Decision Trees




- Explain the advantages and disadvantages of writing a program on your own vs using a pre-created suite such as WEKA.

The main advantage is that we can decide how to implement our program and do whatever we want to, from being a lab just to practice and learn the ID3 algorithm to have the ability to change the behavior of how the tree is built and made a custom algorithm depending on the dataset that we have. However, WEKA is a full suite developed by not just students but a University so their implementation could be more complete and may be faster and less resource consuming, giving teachers and researchers a perfect tool to use with an easy to use interface and graphics.

- Explain what criteria you followed to choose the datasets for your tree and the WEKA tests.

We started looking for datasets on the page UCI Machine Learning but we had problems with them, the ones we “randomly” choose had overfit because they had many attributes even some of them had missing values, to actually retrieve functional datasets we filter the data sets by the following, resulting on the image below.

- Default Task: Classification
- Attribute Type: Categorical
- Attributes: Less than 10
- Instances: Less than 10

 <u>Balloons</u>	Multivariate	Classification	Categorical	16	4	
 <u>Lenses</u>	Multivariate	Classification	Categorical	24	4	1990
 <u>Shuttle Landing Control</u>	Multivariate	Classification	Categorical	15	6	1988

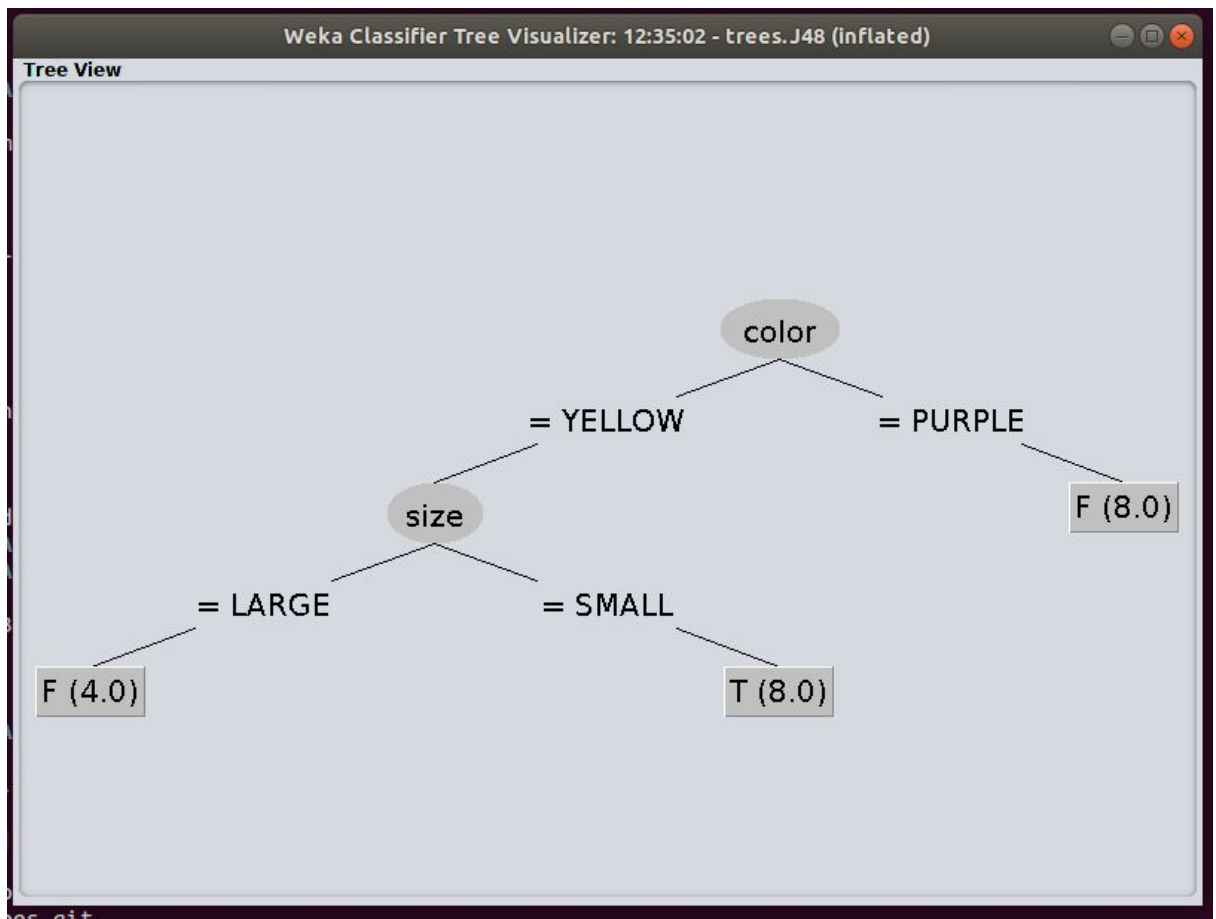
The data set Lenses is already tested on the assignment by Alphagrader but the others aren't, we tried to test the Shuttle Landing Control dataset but the problem was that there are several “don't care” attributes and we did not make it work, luckily for us we had the Balloons dataset left which did not gave us problems and we just needed to format to the arff style.

- Include the graphics of the trees or part of the trees you generated in WEKA and your own program. Are they different, and if so, why?

The trees generated for the Ballon dataset with Weka and our implementation were the same, we suppose it's because we choose the less possible attributes and instances so the complexity of the tree did not increase; also there was no improvement between ID3 and the J48 algorithm. Unlike the first dataset we test (Car Evaluation, we decided not to

include it) which gave us overfit and a complex tree to read with this, the trees were a little different because there were so much attributes and branches pruned.

```
(env) crcz@crcz-VirtualBox:~/Documents/AI/DecisionTrees$ python trees.py < balloons.arff
color: YELLOW
  size: LARGE
    ANSWER: F
  size: SMALL
    ANSWER: T
color: PURPLE
  ANSWER: F
```



- Based in what you have learned so far where would you use decision trees?

Decision trees are useful because they often mimic the human thinking so they are quite simple to understand and make good interpretations, for classification problems we could predict an output given certain features. we could use the ID3 with nominal values or the C4.5 with numerical values. An interesting application we thought about was medical diagnosis, there was a dataset called Breast Cancer it would help a lot to have a tool which helps to identify diseases and its possible treatment.