INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Artificial Intelligence

Report Lab 5

Equipo:

Alan Kuri García A01204805

Shara Teresa González Mena A01205254

Due Date:

October 30, 2018

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

By implementing the ID3 algorithm by hand, we understood deeply how the algorithm works, how to compute the entropy of a given dataset, how to compute the information gain and how to build a decision tree by using those concepts. In the other side, using only the WEKA tool and entering some given input, does not give us the same knowledge obtained by writing your own program because you just give and input and watch the magic happen.

Despite this,WEKA lets you understand the structure of the decision tree' by offering visual elements like the binary trees and showing additional information specifying details of the tree, aspect that an own implementation lacks of. The most gráphical representation that our program does is printing the nodes of the tree with certain tabulation to differentiate between the parents and the childs. This representation is not as easy to understand at the beginning, specially if you are learning about decision trees.

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

At the beginning we filtered the examples in the machine learning repository by the attribute type "Categorical" because the tests in alpha grader had categorical attributes and we wanted to keep trying with similar types of data sets. After this filter, we picked 2 data sets called: Arrythmia and Breast Cancer.

One of the examples had missing data and wasn't adapted for our algorithm and neither for WEKA because it was not possible to build the three by analyzing that data due to its incompleteness. The second data set, after being computed by weka, presented a clear overfit which we identified by visualizing a really messy tree that had a big number of nodes, and in our program the dataset was computed but resulted in a really big tree, different than WEKA's result, but also represented an overfit.

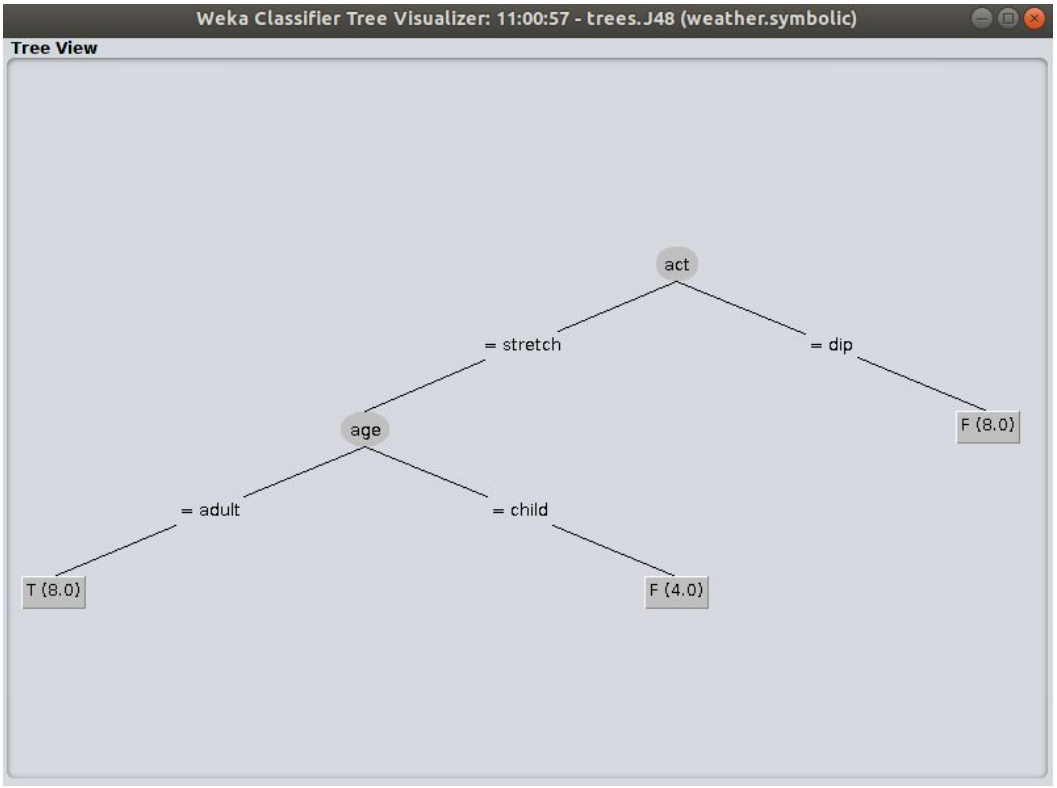
After several attempts to fix those problems, we did achieved a change in the representation of both trees but the overfit was present anyway so we decided to look for other examples that could let us compare and analyze the performance of both WEKA and our implementation. We picked a data set to model whether a balloon is inflated or not given a child or an adult have it, is stretch or dip and its color.

The criteria to pick this example was first to be categorical and second to have less than 15 attributes so that both programs could work with the information and we could get to appreciate both performances.

- **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?**

```
act:  stretch
  age:  adult
    ANSWER:  T
  age:  child
    ANSWER:  F
act:  dip
  ANSWER:  F
```

Our program answer.



WEKA answer.

In this case they have the same output in both tools, this is because the complexity of the tree is simple, in the case that the tree gets much larger the difference will be noticeable. There are no differences in this tree between ID3 and J48.

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

When there is a decision to make and that decision is affected by different aspects or attributes. There is where a decision tree, that models how the decision is taken, is needed by using a real data set of how that decision has been behaved in time given the different aspects that affect it.