Decision Trees Lab

Artificial Intelligence

Luis Carlos Acosta
Eduardo Emilio Tavarez

Introduction

For this lab we had to implement the ID3 decision tree algorithm and show the results it threw in form of a tree, this was a bit challenging since we had to understand the algorithm perfectly before starting to implement it, we had to understand the concepts of information gain and entropy in the first place and have a clear understanding of recursion. As a second part of the lab we got to play a little bit with the tool WEKA which contains features for pre-processing, classification, regression and visualization of data. On the next sections we will try to explain the advantages and disadvantages of implementing your own version of ID3 against using a pre-created suite for data processing and visualization such as WEKA.

Advantages/Disadvantages

Advantages

- ❏ We got to fully understand the algorithm. As in every programming problem we had to do some exercises and get familiar with the algorithm in order to start coding.
- ❏ You can make your own version of the algorithm, with modifications if desired.
- ❏ You get the chance to optimize the algorithm, which can be pretty challenging but it is possible.
- ❏ You strengthen your programming skills in general and in the chosen language.
- ❏ Last but not least, it’s more fun :)

Disadvantages

- ❏ It’s more time consuming. Since we had to do all the implementation from scratch, it took more time to understand how it works and then implement it, rather than just learning how to use the tool and looking for the data set.
- ❏ You depend on a pre-created suite to work with the algorithm.
- ❏ You don’t get to fully understand the algorithm and this may get a bit confusing when trying to understand the reasons of the results.
- ❏ You can find a lot of forums and information about how to use the tool in different ways.
- ❏ It would take a lot of time in order to implement all the functionalities the tool has.
- ❏ There’s more reliability for pre-created tools rather than student code.

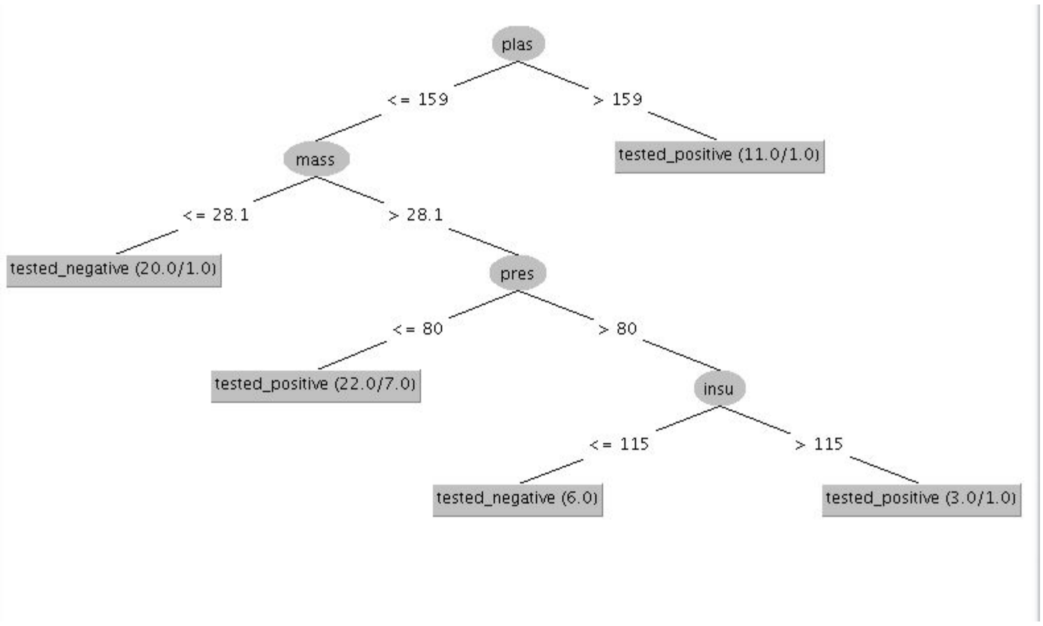
Criteria followed to choose the data set

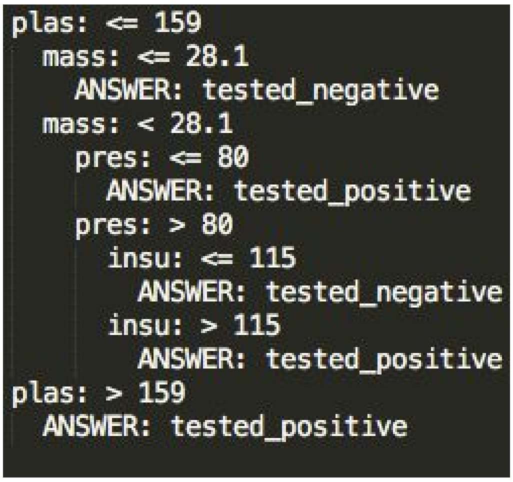
We chose a data set that aims to give a diagnosis to people that take a diabetes study, the binary variable investigated whether the patient shows signs of diabetes according to World Health Organization criteria. The attributes are the following:

@attribute 'preg' real: Number of times pregnant
@attribute 'plas' real: Plasma glucose concentration
@attribute 'pres' real: Diastolic blood pressure
@attribute 'skin' real: Triceps skin fold thickness
@attribute 'insu' real: 2-hour serum insulin
@attribute 'mass' real: Body mass index
@attribute 'pedi' real: Diabetes pedigree function
@attribute 'age' real: Age (years)
@attribute 'class' { tested_negative, tested_positive}: Class variable (0 or 1)

We were interested on looking for a data set that could help doctors predict if a patient has a disease or not, since diabetes is becoming a more common disease and is known for being letal we thought that decision trees could play an important role.

Comparison (program and WEKA tool)


plas
<= 159
> 159
tested_positive (11.0/1.0)
mass
<= 28.1
> 28.1
tested_negative (20.0/1.0)
pres
<= 80
> 80
tested_positive (22.0/7.0)
insu
<= 115
> 115
tested_negative (6.0)
tested_positive (3.0/1.0)



```
plas: <= 159
  mass: <= 28.1
    ANSWER: tested_negative
  mass: < 28.1
    pres: <= 80
      ANSWER: tested_positive
    pres: > 80
      insu: <= 115
        ANSWER: tested_negative
      insu: > 115
        ANSWER: tested_positive
plas: > 159
  ANSWER: tested_positive
```

As we can see the results above are the same

Conclusion

The concept of decision trees pretty interesting and we think it can be applied to almost any type of problem that you can gather data from and classify it and it can give valuable predictions, from simple and mundane problems such as what to have for dinner tonight to predict diseases and certain behaviours in stock markets, an interesting but challenging topic would also be psychiatry since it could help prevent some mental diseases. Decision trees are a powerful tool for classification and prediction and is our work as engineers to exploit its advantages.