Lindsay Cagarli and Anna Kurchenko

CSCI 420- Intro to Data Mining

HW 05

1. Who did what roles during the assignment?

It was a collaborative effort to complete this assignment but in particular Anna had more say in the actual building of the decision tree and Lindsay was responsible for more of handling data and documentation.

2. What was the maximum call depth of your final classifier?  
Did it actually use the maximum number of levels?

The maximum depth that we used is 8 however, depending on what data set is being run not all the levels were consistently used.

3. Describe the decisions of your final trained classifier program (the resulting classifier).  
What were the most important attributes (the ones at the top)?  
Inspecting it, what does it tell you about the relative importance of the attributes? What is the  
most important attribute?

When observing all the attributes, it is evident that there is some noise as the feature-based attributes such as if a car has bumper stickers is not an ideal

4. What was your choice of a minimum number of records in a node?

Minimum records in a node is set to 5.

5. What was your choice for a maximum splitting depth for the code?

Maximum splitting depth is set to 8.

6. What was your choice for a maximum node purity to stop at?

Maximum node purity is when a node reaches 90% of purity

7. What features did you create, if any?

We did not add any additional features given the time and complexity of the assignment as is.

8. Did you discover any features which could be ignored?

We found that

9. Generate a confusion matrix for the original training data:  
How many aggressive were classified as aggressive?

How many aggressive were classified as non-aggressive?

How many non-aggressive were classified as non-aggressive?

How many non-aggressive were classified as aggressive?

10. What was the accuracy of your resulting classifier, on the training data?

11. What was the hardest part of getting all this working? Did anything go wrong? Did anything go very well?

The most difficult aspect of this assignment was the initial getting started.

Although gone in depth in class of the expectations and even in the write up giving tips of where to start at the end of the day this was still a hard thing to do (lol)

12. Conclusions: How did you clean your data? What pre-processing did you do? Did you create any new features to use? What did you discover? Anything else you discovered along the way? Did you run into the accuracy paradox? Was the data completely separable? Write full paragraphs, and full sentences. Show strong evidence of learning.

The data given was genuinely overwhelming as a whole.