Aaron Morgenegg A02072659

CS 5665 Homework 4

Tan 3.2

a. Gini = 0.5
b. Gini = 0.95
c. Gini = 0.48
d. Gini = 0.524
e. Gini = 0.63
f. Gender is the best attribute to use between gender, car type, and shirt size.
g. Because using Customer ID does not generalize well – it just memorizes the data set.

Tan 3.3

a. a1 entropy = 0.61, a2 entropy = 0.66
b. a1 gain = , a2 gain =
c.
d.
e.
f.

Tan 3.6

a.
b.
c.

Tan 3.8

a.
b.
c.
d.
e.

Tan 3.12

a. I would expect T_{10} to work better on unseen instances, as T_{100} may be overfitted to the training data.
b. I would choose T_{10}, as the accuracy is close between both trees, while T_{10} is likely to be more generalized for other data, and is more consistent between accuracy of A and B.

Tan 4.1

a. No, depending on the dataset it is very possible for rules to overlap.
b. No, there is no rule that covers rust.
c. Yes, for example, if the mileage → low, air conditioner → working, and engine → bad, the value could be high or low depending on the ordering between mileage and the other rules.
d. Not if the rules were ordered.

Tan 4.4

a. Accuracy: R1 = 0.8, R2 = 0.75, R3 = 0.53. R1 is the best according to Accuracy.

Aaron Morgenegg A02072659

b. FOIL: R1 = 5.55, R2 = 39.65, R3 = 96.76. R3 is the best according to FOIL.