CSCE 587 – Big Data Analytics
Class / Homework Assignment 5: Decision Trees

Deliverables: You will upload to the assignment in Blackboard

- a Word document OR PDF with answers to the questions in part one OR you will turn in this piece of paper with your written answers during class on 2/27
- an R script and PDF(s) with the solutions for part two

Part 1 – Decision Trees – Write answers in Word or PDF document

Rewards Member	Coupon	Purchase Size	Expedited Shipping
Yes	None	Small	No
Yes	15% off	Small	No
No	Buy One Get One Free	Large	No
No	15% off	Medium	No
Yes	None	Large	Yes
No	None	Small	Yes
Yes	Buy One Get One Free	Medium	No
Yes	None	Medium	No
Yes	15% off	Large	Yes
No	Buy One Get One Free	Large	Yes

All Students

We plan to use the data in the table above to create a decision tree to predict the value of Expedited Shipping given values for Rewards Member, Coupon, and Purchase Size. You can use technology to compute the numeric values, if you like. You must write the expressions (using the formulae for entropy and gain) with the correct fractions from the training sample shown above in order to receive credit.

a. Entropy of Purchase Size = Small

b. Entropy of Purchase Size = Medium

c. Entropy of Purchase Size = Large

d. Gain of Purchase Size for this training sample

Part 2 – Decision Trees in R – Write solutions in R Script and create PDFs

Preparation:

You can download and save the csv file attached to the assignment in Blackboard OR you can use wget at the terminal to download the file from my CSE site to your virtual machine for use in part 2.

wget https://cse.sc.edu/~bhipp/587/penguins.csv --no-check-certificate

All Students

Load the rpart library in R. Use the rpart function (as we did in the lab) to predict the penguins' species using the island, flipper length, culmen length, culmen depth, and body mass. For the control, use a minsplit of 2 and cp of 0.001.

a. Write the R command you used to call rpart.
b. Plot the decision tree with the text labels to a PDF. This PDF should be submitted to the assignment in Blackboard.
c. What species does the decision tree predict for a penguin with the following values: island = Biscoe, flipper length = 180, culmen length = 43, culmen depth = 18, body mass = 4000?

Graduate Students Only

d. Use the rpart function (as we did in the lab) to predict the penguins species using the island, flipper length, culmen length, culmen depth, and body mass. For the control, use a minsplit of 5 and cp of 0.01. Plot the decision tree with the text labels to a PDF. This PDF should be submitted to the assignment in Blackboard.
e. Using the values in the decision tree, what is the specificity for Chinstrap? Write your answer as a comment in the R script.