1.SAS:

**proc** **import**

out= cancer

datafile ="V:\breast-cancer-dataset.csv"

dbms=csv replace;

getnames=yes;

**proc** **print** data = cancer;

title "Breast Cancer Dataset";

**run**;

ods graphics on;

**proc** **hpsplit** data=cancer seed=**1**;

class class age menopause tumor\_size inv\_nodes node\_caps breast breast\_quad irradiat;

model class = age menopause tumor\_size inv\_nodes node\_caps deg\_malig breast breast\_quad irradiat;

grow entropy;

prune costcomplexity;

**run**;

2. Using the confusion matrix, compute the following assessment metrics accuracy, recall, and precision (see lecture for formulas).

Grow - Entrophy:

True positive (f++) = 194

True negative (f--) = 23

False negative (f+-) = 7

False positive (f-+) = 62

T = 286

Accuracy = (f++ + f-- ) /T = (194+23)/286 = 0.758741259 or about 75.87%

Recall (TPR) = f++ / (f++ + f+-) = 194/(194+7) = 0.965174129 or about 96.52%

Precision = f++ / (f-+ + f++) = 194/(62+194) = 0.7578125 or about 75.78%

3. Change the grow algorithm to “gini” and recompute the metrics from question 2.

Grow – GINI:

f++ = 191

f-- = 27

f+- = 10

f-+ = 58

T = 286

Accuracy = (f++ + f-- ) /T = (191+27)/286 = 0.762237762 or about 76.22%

Recall (TPR) = f++ / (f++ + f+-) = 191/(191+10) = 0.950248756 or about 95.02%

Precision = f++ / (f-+ + f++) = 191/(58+191) = 0.767068273 or about 76.71%

Does entropy build a more accurate classifier or gini?

No, GINI has higher accuracy and higher precision.