Ajay Herod

500764104

CIND 119 D20

/\* Read the file in SAS and display the contents using the import and print procedures. \*/

**proc** **import**

out= breast\_cancer

datafile="V:\CIND 119\Assignment 1\breast-cancer-dataset.csv"

dbms=csv replace;

getnames=yes;

**proc** **print** data=breast\_cancer;

**run**;

/\* Develop a decision tree-based classification model using the hpsplit procedure of SAS. \*/

ods graphics on;

**proc** **hpsplit** data=breast\_cancer;

class class menopause deg\_malig breast irradiat;

model irradiat = class menopause deg\_malig breast;

grow entropy;

prune costcomplexity;

**run**;

/\* Navigate the contents of Results View by clicking on HPSplit breastcancer-dataset, and then by selecting Model Assessment.

Examine the confusion matrix, fit statistics, and variable importance. \*/

Based on the confusion matrix there is a low error rate for no, but a very high error rate on ye.

This means that the classification tree can predict for no well but often incorrectly predicts ye wrong

In the fit statistics we can see that there are 10 levels in the tree.

We can also see that the miss class is about 20% which means the accuracy is about 80%.

The AUC is the area under ROC, having a 0.74 tell us that the accuracy is acceptable.

The variable importance tells us the relative and importance of each variable in the tree.

Each variable shows a importance between 2.44 and 1.69 which means they each are important to the tree.

/\* Using the confusion matrix, compute the following assessment metrics accuracy, recall, and precision \*/

Accuracy = (TP + TN)/T, Recall = TP/(TP + FN), Precision = 208/(208 + 48)

Accuracy = (208 + 20)/286, Recall = 208/(208 + 10), Precision = 208/(208 + 48)

Accuracy = 79%, Recall = 95%, Precision = 81%

/\* Change the grow algorithm to “gini” and recompute the metrics from question 2.

Does entropy build a more accurate classifier or gini? \*/

**proc** **hpsplit** data=breast\_cancer;

class class menopause deg\_malig breast irradiat;

model irradiat = class menopause deg\_malig breast;

grow gini;

prune costcomplexity;

**run**;

Using the gini growth algorithm we can see in that gini confusion matrix predicts more ye than the entropy algorithm.

Because the difference is two and split between the actuals the accuracy is still equal.

Therefore, both algorithms are equally accurate for this decision tree classification.