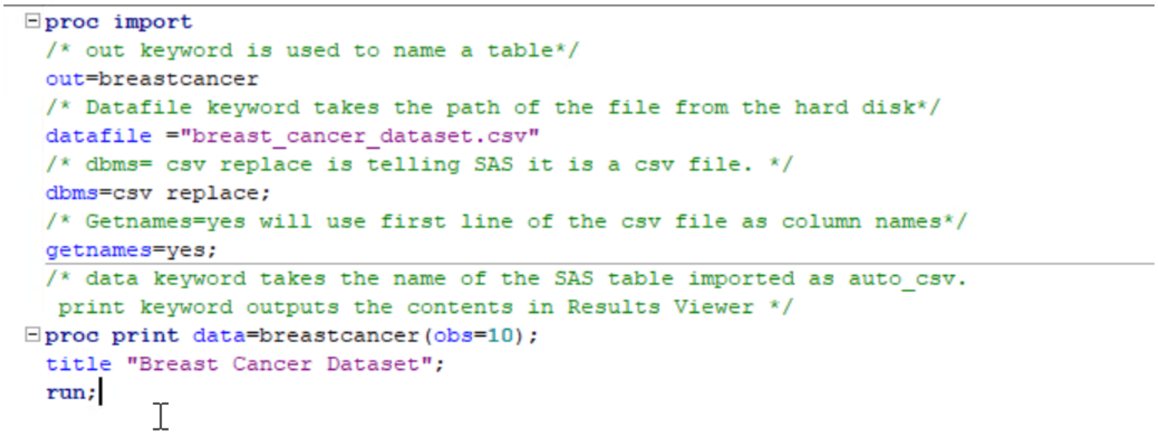
Big Data Analytics
Supervised Learning Using SAS

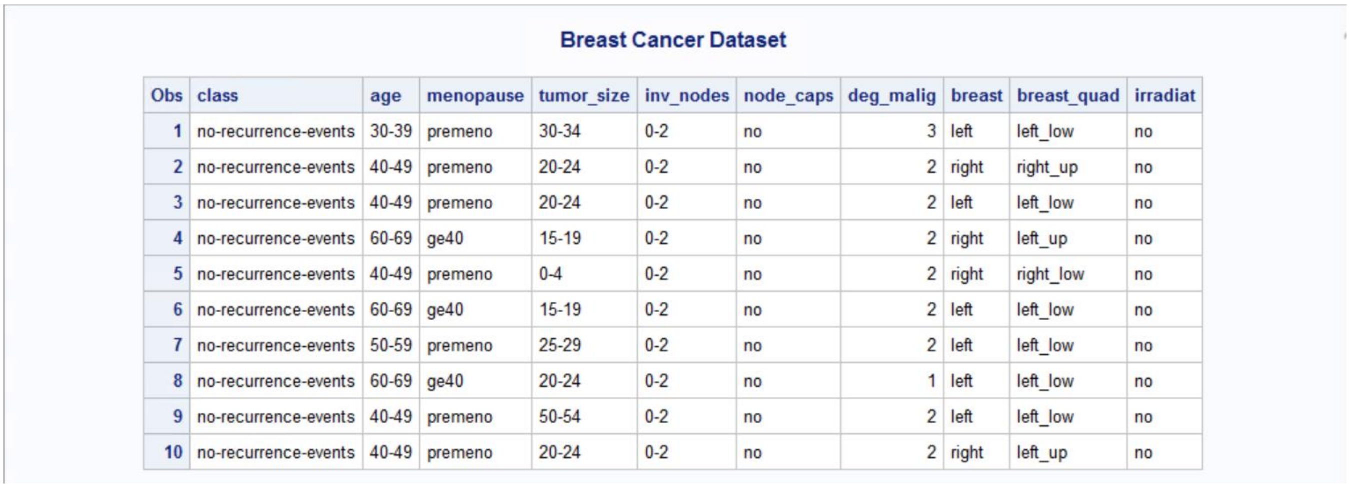
1.Download the breast-cancer-dataset.csv from your D2L Assignment 1 link. Complete the following tasks

a. Read the file in SAS and display the contents using the import and print procedures.

```
proc import
/* out keyword is used to name a table*/
out= breastcancer
/* Datafile keyword takes the path of the file from the hard disk*/
datafile ="breast_cancer_dataset.csv"
/* dbms= csv replace is telling SAS it is a csv file. */
dbms=csv replace;
/* Getnames=yes will use first line of the csv file as column names*/
getnames=yes;
/* data keyword takes the name of the SAS table imported as auto_csv.
 print keyword outputs the contents in Results Viewer */
proc print data= breastcancer (obs=10);
title "Breast Cancer Dataset";
run;
```



```
proc import
 /* out keyword is used to name a table*/
 out=breastcancer
 /* Datafile keyword takes the path of the file from the hard disk*/
 datafile ="breast_cancer_dataset.csv"
 /* dbms= csv replace is telling SAS it is a csv file. */
 dbms=csv replace;
 /* Getnames=yes will use first line of the csv file as column names*/
 getnames=yes;
 /* data keyword takes the name of the SAS table imported as auto_csv.
  print keyword outputs the contents in Results Viewer */
proc print data=breastcancer(obs=10);
 title "Breast Cancer Dataset";
 run;
```

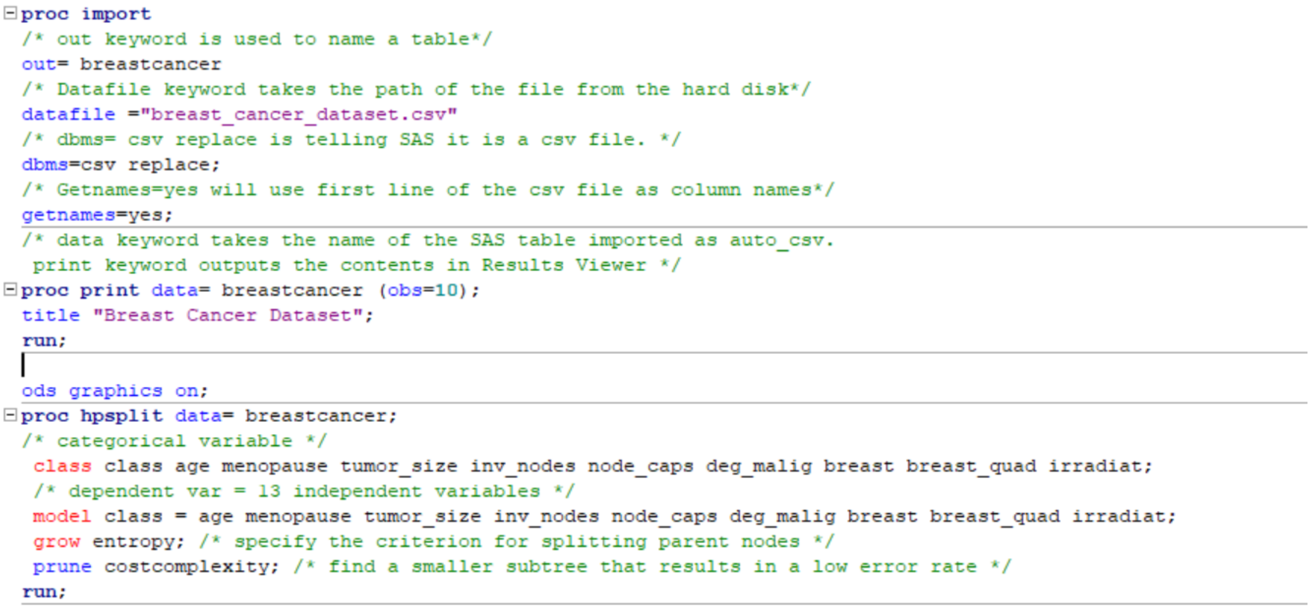


Breast Cancer Dataset

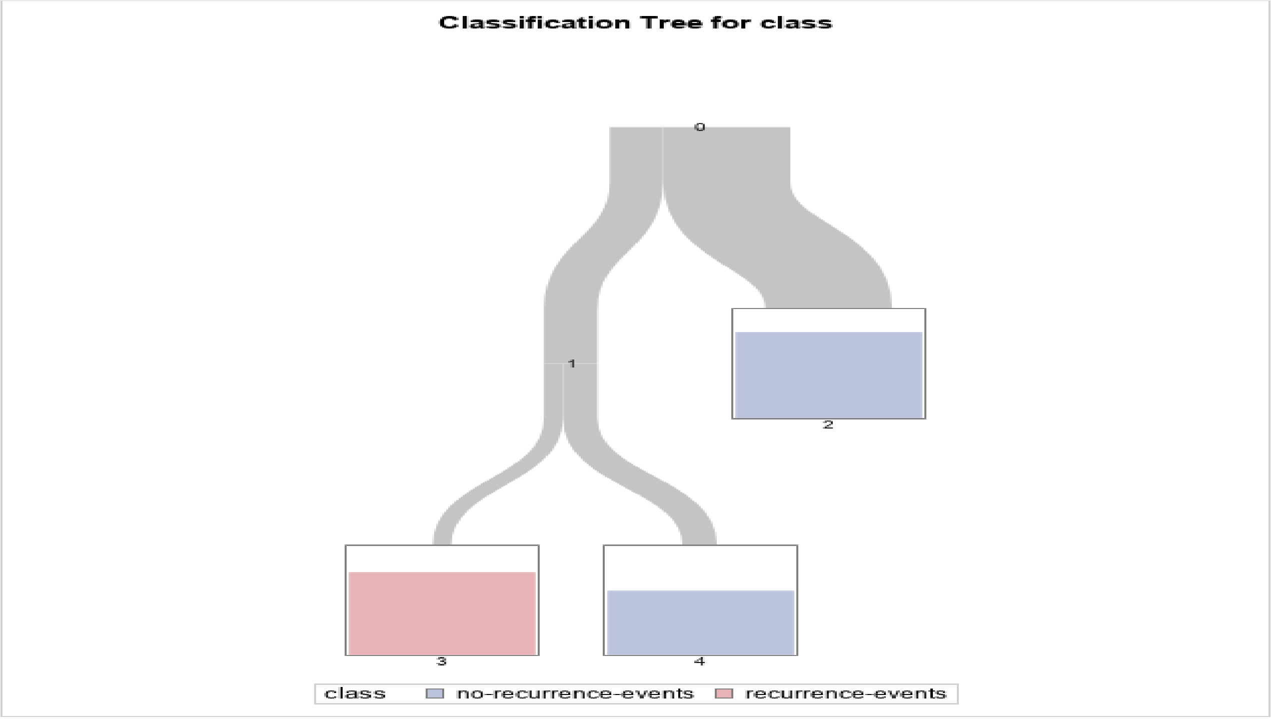
Obs	class	age	menopause	tumor_size	inv_nodes	node_caps	deg_malig	breast	breast_quad	irradiat
1	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
3	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
4	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
5	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
6	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	left	left_low	no
7	no-recurrence-events	50-59	premeno	25-29	0-2	no	2	left	left_low	no
8	no-recurrence-events	60-69	ge40	20-24	0-2	no	1	left	left_low	no
9	no-recurrence-events	40-49	premeno	50-54	0-2	no	2	left	left_low	no
10	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	left_up	no

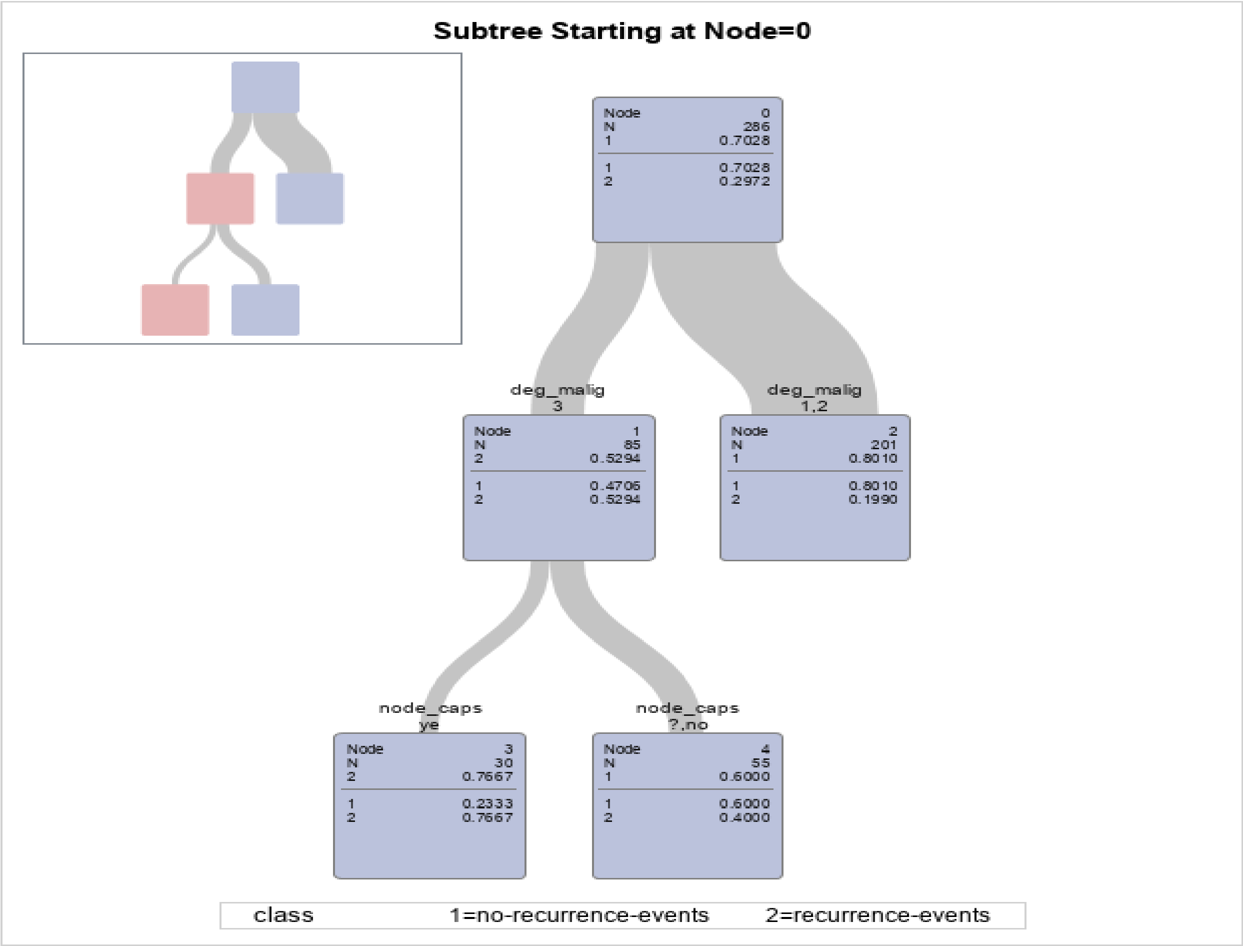
b. Develop a decision tree-based classification model using the hpsplit procedure of SAS.

```
proc import
/* out keyword is used to name a table*/
out= breastcancer
/* Datafile keyword takes the path of the file from the hard disk*/
datafile ="breast_cancer_dataset.csv"
/* dbms= csv replace is telling SAS it is a csv file. */
dbms=csv replace;
/* Getnames=yes will use first line of the csv file as column names*/
getnames=yes;
/* data keyword takes the name of the SAS table imported as auto_csv.
 print keyword outputs the contents in Results Viewer */
proc print data= breastcancer (obs=10);
title "Breast Cancer Dataset";
run;

ods graphics on;
proc hpsplit data= breastcancer;
/* categorical variable */
 class class age menopause tumor_size inv_nodes node_caps deg_malig breast
breast_quad irradiat;
 /* dependent var = 13 independent variables */
 model class = age menopause tumor_size inv_nodes node_caps deg_malig breast
breast_quad irradiat;
 grow entropy; /* specify the criterion for splitting parent nodes */
 prune costcomplexity; /* find a smaller subtree that results in a low error rate
*/
run;
```

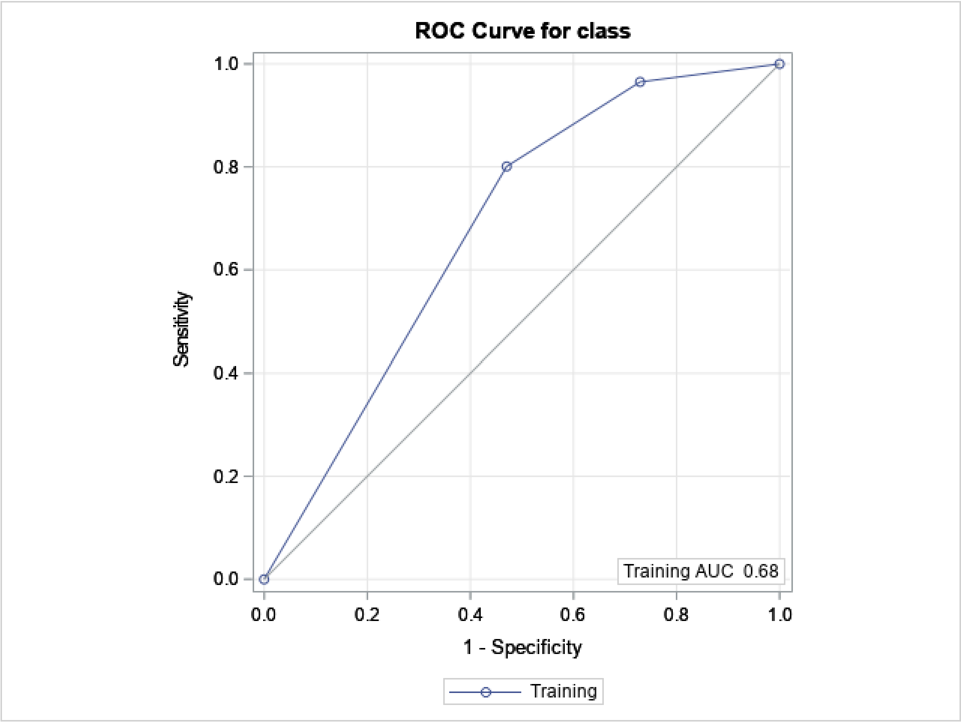


```
proc import
/* out keyword is used to name a table*/
out= breastcancer
/* Datafile keyword takes the path of the file from the hard disk*/
datafile ="breast_cancer_dataset.csv"
/* dbms= csv replace is telling SAS it is a csv file. */
dbms=csv replace;
/* Getnames=yes will use first line of the csv file as column names*/
getnames=yes;
/* data keyword takes the name of the SAS table imported as auto_csv.
 print keyword outputs the contents in Results Viewer */
proc print data= breastcancer (obs=10);
title "Breast Cancer Dataset";
run;

ods graphics on;
proc hpsplit data= breastcancer;
/* categorical variable */
 class class age menopause tumor_size inv_nodes node_caps deg_malig breast breast_quad irradiat;
 /* dependent var = 13 independent variables */
 model class = age menopause tumor_size inv_nodes node_caps deg_malig breast breast_quad irradiat;
 grow entropy; /* specify the criterion for splitting parent nodes */
 prune costcomplexity; /* find a smaller subtree that results in a low error rate */
run;
```

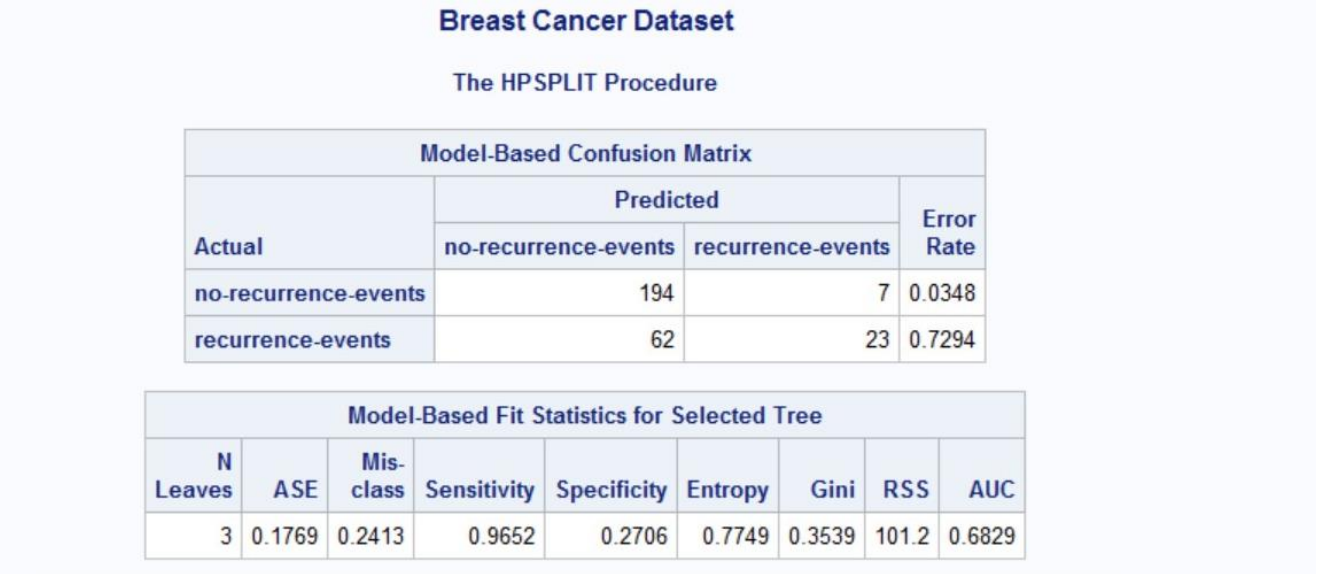




The resulting decision tree has 286 examples at the root node. Each decision node in the tree is labeled with the corresponding independent variable name and split value. The leaf nodes show the classification decision.



c. Navigate the contents of Results View by clicking on HPSplit breast-cancer-dataset, and then by selecting Model Assessment. Examine the confusion matrix, fit statistics, and variable importance.



Breast Cancer Dataset

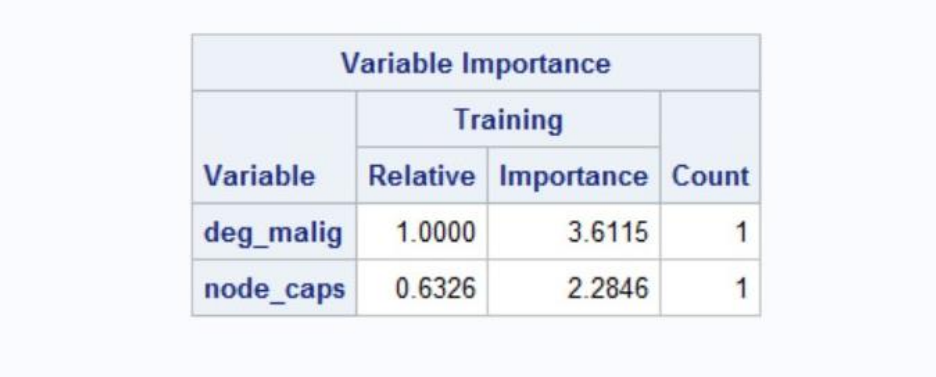
The HPSPLIT Procedure

Model-Based Confusion Matrix

Actual	Predicted no-recurrence-events	 recurrence-events	Error Rate
no-recurrence-events	194	7	0.0348
recurrence-events	62	23	0.7294

Model-Based Fit Statistics for Selected Tree

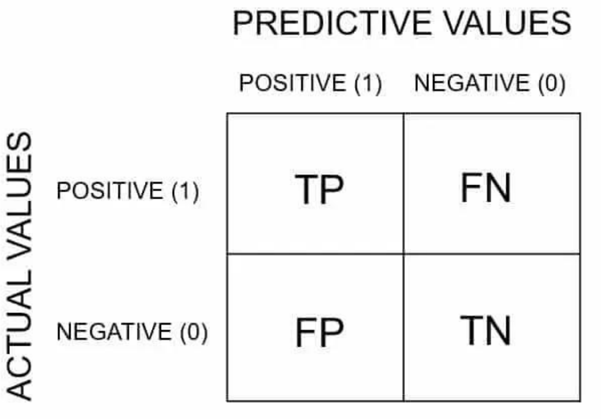
N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
3	0.1769	0.2413	0.9652	0.2706	0.7749	0.3539	101.2	0.6829



Variable Importance

Variable	Training Relative	 Importance	Count
deg_malig	1.0000	3.6115	1
node_caps	0.6326	2.2846	1

2. Using the confusion matrix, compute the following assessment metrics accuracy, recall, and precision (see lecture for formulas and state your assumption in bold which class in the confusion matrix you want to consider positive--i.e., recurrence-event or not-recurrence-event). (5 points)



		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

Condition for marks: 3 points for accuracy, 1 point for precision, and 1 point for recall.

If we consider Not -recurrence-event is considered as positive.

- Accuracy = TP+TN / TP+TN +FP+FN
  Accuracy = 194+23/ 194+7+62+23
  =217/286
  = 0.759

- Recall = TP/TP+FN
  Recall = 194/194+7
  =194/201
  =0.965

- Precision = TP/TP+FP
  Precision=194/194+62
  = 194/256
  = 0.757

If we considered recurrence-event as positive.

- Accuracy = TP+TN / TP+TN +FP+FN
  Accuracy = 23+ 194 /194+7+62+23
  = 217/286
  = 0.759

- Recall = TP/TP+FN
  Recall = 23/23+62
  = 23/85
  = 0.2706

- Precision = TP/TP+FP

  Precision= 23/23+7

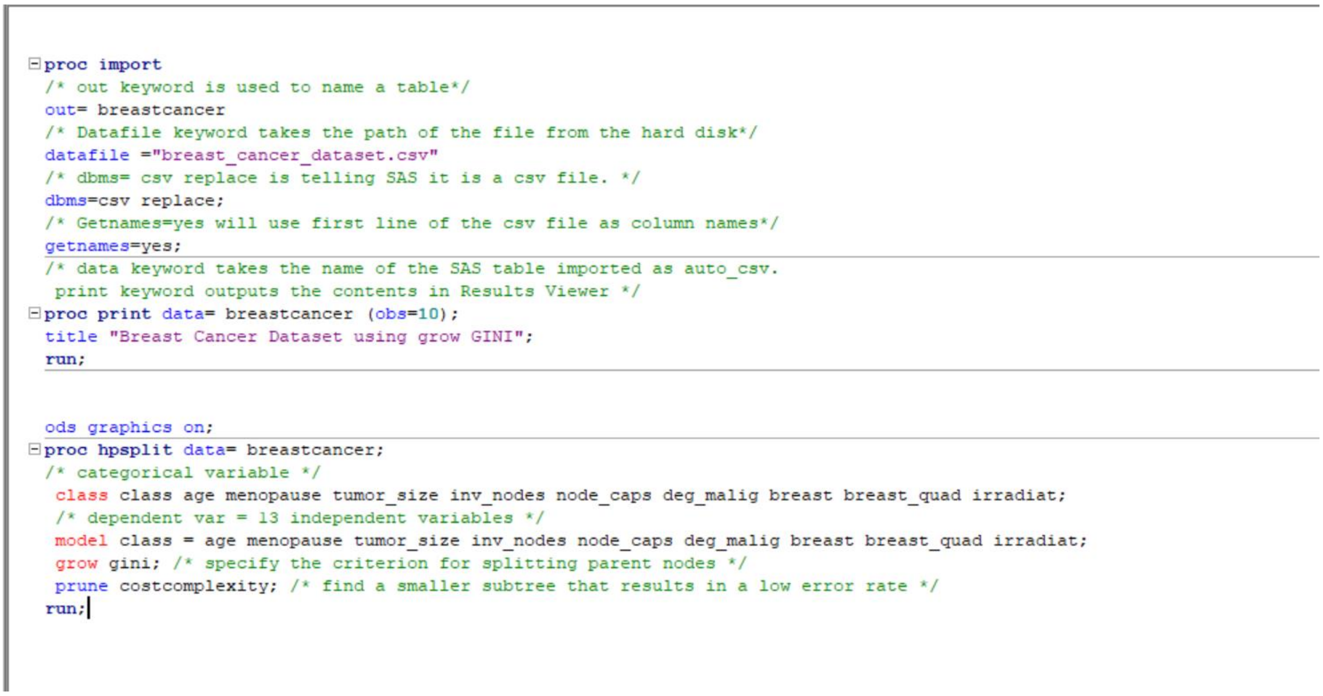
  = 23/30

  = 0.7667

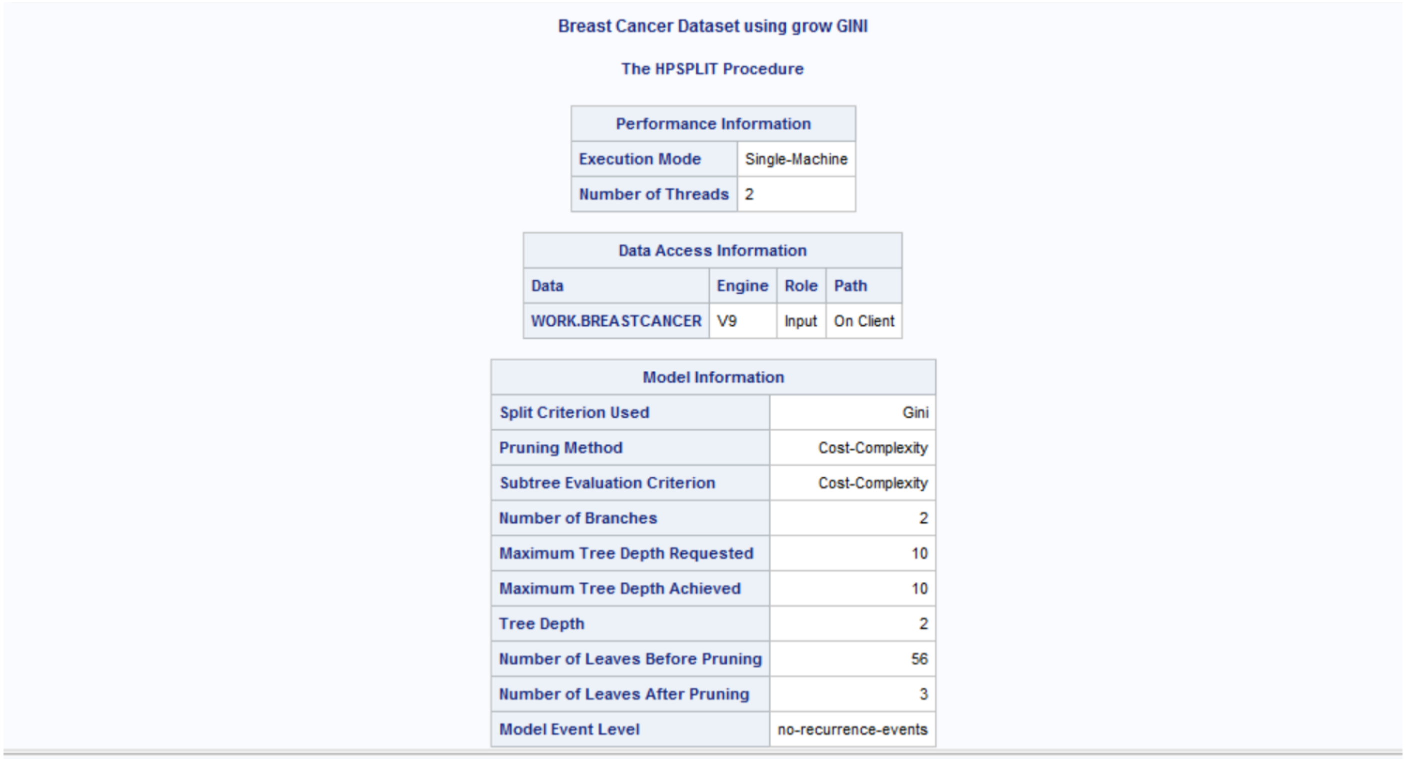
I will consider recurrence-event is position, because it is important to recurrence -event have less negative (FN). There ML algorithm missed 62 people who has tumors.

3. Change the grow algorithm to “gini” and recompute the metrics from question 2. Does entropy build a more accurate classifier or gini?

```
proc import
/* out keyword is used to name a table*/
out= breastcancer
/* Datafile keyword takes the path of the file from the hard disk*/
datafile ="breast_cancer_dataset.csv"
/* dbms= csv replace is telling SAS it is a csv file. */
dbms=csv replace;
/* Getnames=yes will use first line of the csv file as column names*/
getnames=yes;
/* data keyword takes the name of the SAS table imported as auto_csv.
 print keyword outputs the contents in Results Viewer */
proc print data= breastcancer (obs=10);
title "Breast Cancer Dataset using grow GINI";
run;


ods graphics on;
proc hpsplit data= breastcancer;
/* categorical variable */
 class class age menopause tumor_size inv_nodes node_caps deg_malig breast
breast_quad irradiat;
 /* dependent var = 13 independent variables */
 model class = age menopause tumor_size inv_nodes node_caps deg_malig breast
breast_quad irradiat;
 grow gini; /* specify the criterion for splitting parent nodes */
 prune costcomplexity; /* find a smaller subtree that results in a low error rate
*/
run;
```



```
proc import
/* out keyword is used to name a table*/
out= breastcancer
/* Datafile keyword takes the path of the file from the hard disk*/
datafile ="breast_cancer_dataset.csv"
/* dbms= csv replace is telling SAS it is a csv file. */
dbms=csv replace;
/* Getnames=yes will use first line of the csv file as column names*/
getnames=yes;
/* data keyword takes the name of the SAS table imported as auto_csv.
 print keyword outputs the contents in Results Viewer */
proc print data= breastcancer (obs=10);
title "Breast Cancer Dataset using grow GINI";
run;


ods graphics on;
proc hpsplit data= breastcancer;
/* categorical variable */
 class class age menopause tumor_size inv_nodes node_caps deg_malig breast breast_quad irradiat;
 /* dependent var = 13 independent variables */
 model class = age menopause tumor_size inv_nodes node_caps deg_malig breast breast_quad irradiat;
 grow gini; /* specify the criterion for splitting parent nodes */
 prune costcomplexity; /* find a smaller subtree that results in a low error rate */
run;
```



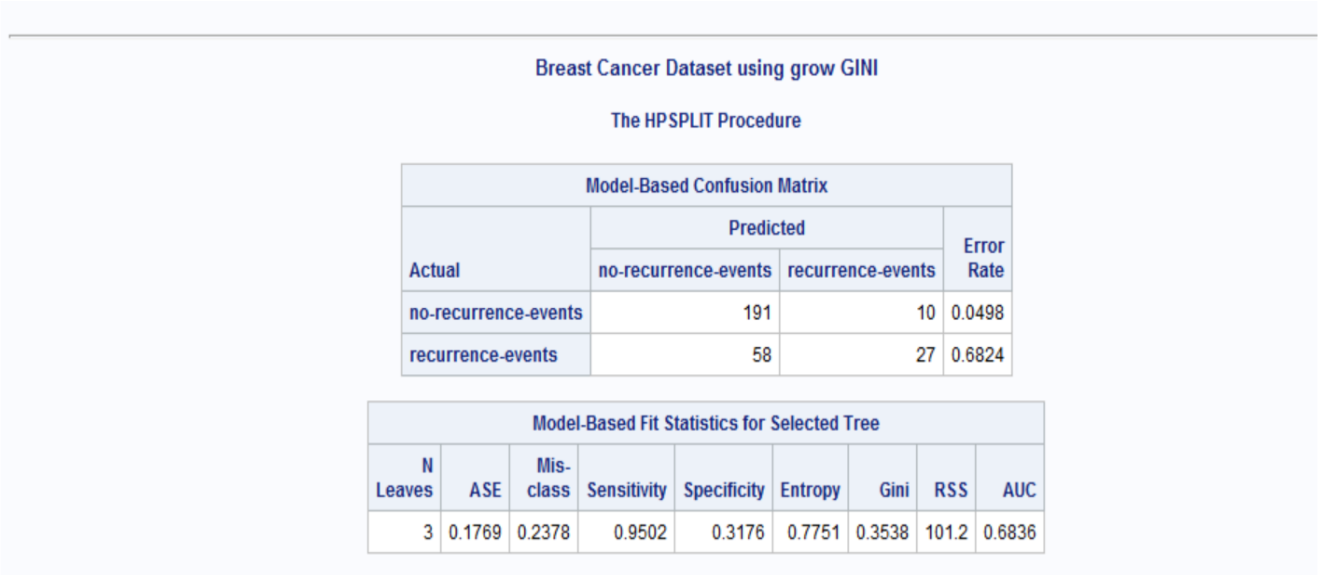
Breast Cancer Dataset using grow GINI

The HPSPLIT Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.BREASTCANCER	V9	Input	On Client

Model Information	
Split Criterion Used	Gini
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	2
Number of Leaves Before Pruning	56
Number of Leaves After Pruning	3
Model Event Level	no-recurrence-events



Breast Cancer Dataset using grow GINI

The HPSPLIT Procedure

Model-Based Confusion Matrix

Actual	Predicted no-recurrence-events	 recurrence-events	Error Rate
no-recurrence-events	191	10	0.0498
recurrence-events	58	27	0.6824

Model-Based Fit Statistics for Selected Tree

N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
3	0.1769	0.2378	0.9502	0.3176	0.7751	0.3538	101.2	0.6836

For entropy   Accuracy = TP+TN / TP+TN +FP+FN
Accuracy = 194+23/ 194+7+62+23
=217/286
= 0.759

For Gini Accuracy = TP+TN / TP+TN +FP+FN
Accuracy = 191+27/ 191+10+58+27
=218 /286
= 0.762

As result, Gini was more accuracy than entropy and also it give us high Ture recurrence-event .