**CIND 119: Introduction to Big Data Analytics**

**Assignment 1 (15% of the final grade)**

**Supervised Learning Using SAS**

**Sharlin Kahlon**

**501124232**

* 1. 1. Download the breast-cancer-dataset.csv from your D2L Assignment 1 link. Complete the following tasks (5 points):
  2. a. Read the file in SAS and display the contents using the import and print procedures. (1 point)

**Code:**

**proc** **import**

out=breast\_cancer\_dataset

datafile ="V:\Assignment\breast\_cancer\_dataset.csv"

dbms = csv replace;

getnames=yes;

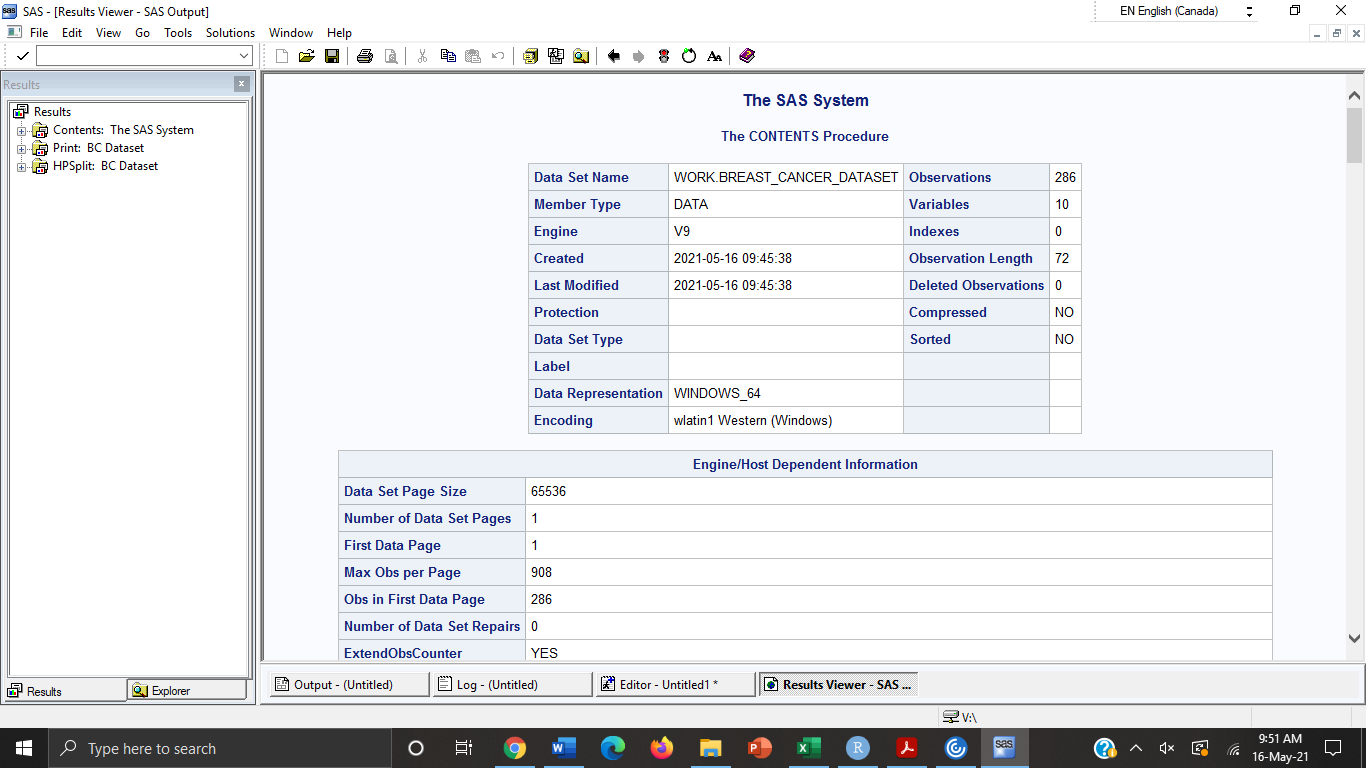
**proc** **contents** data=breast\_cancer\_dataset;

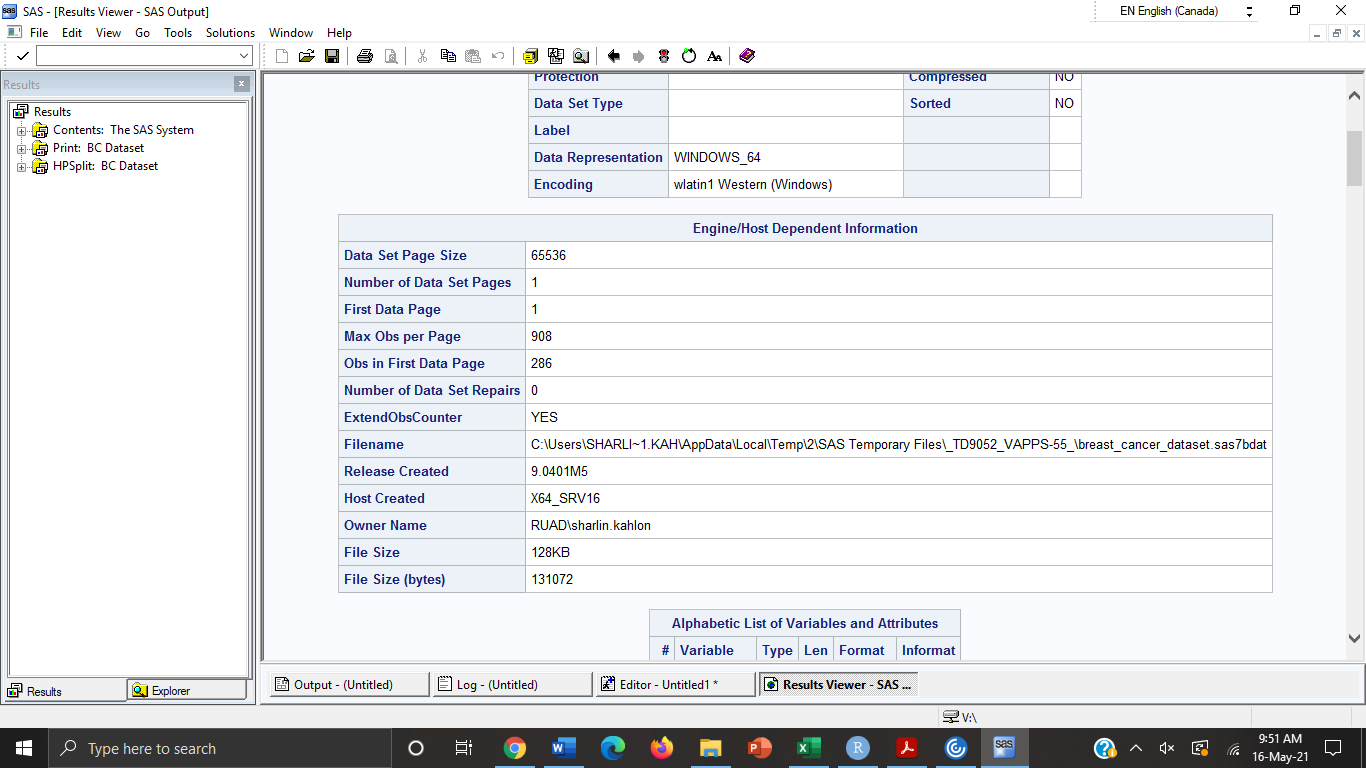
**run**;

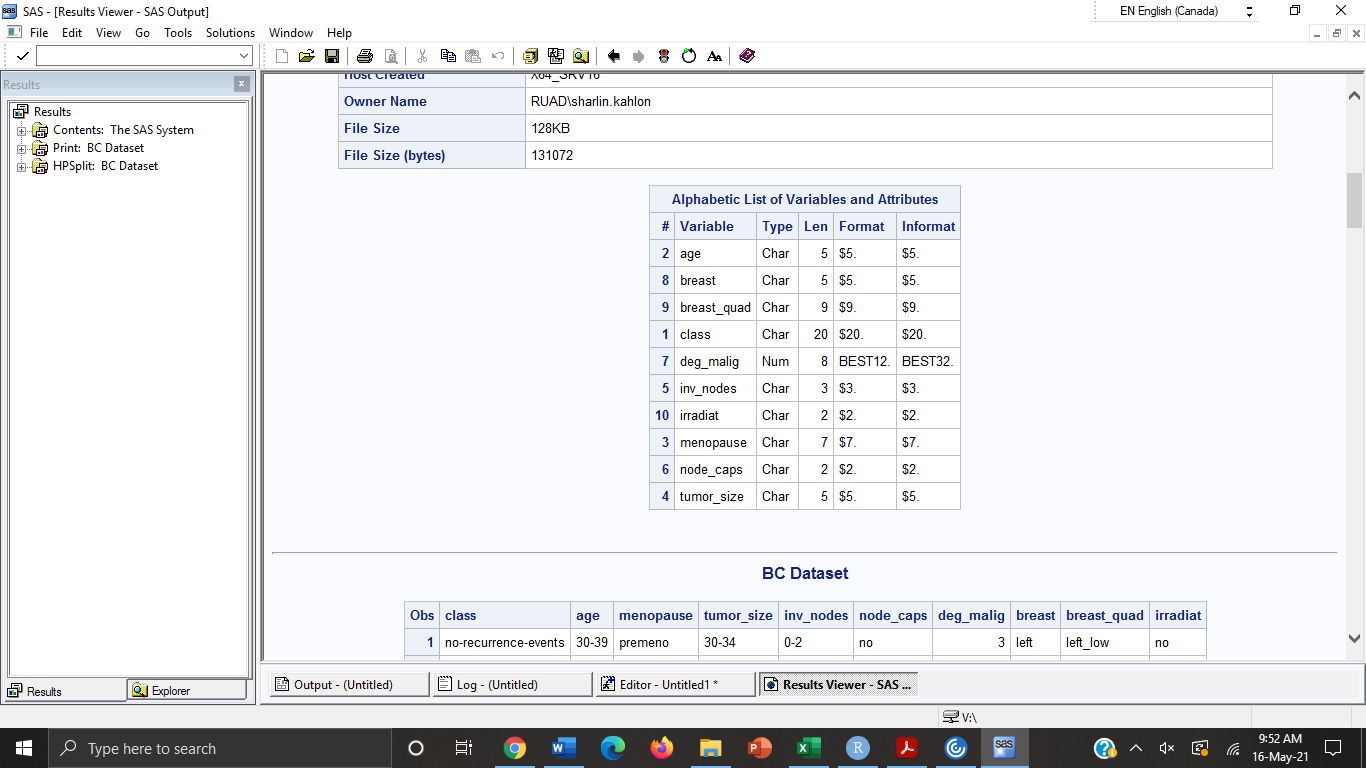
**proc** **print** data=breast\_cancer\_dataset(obs=**10**);

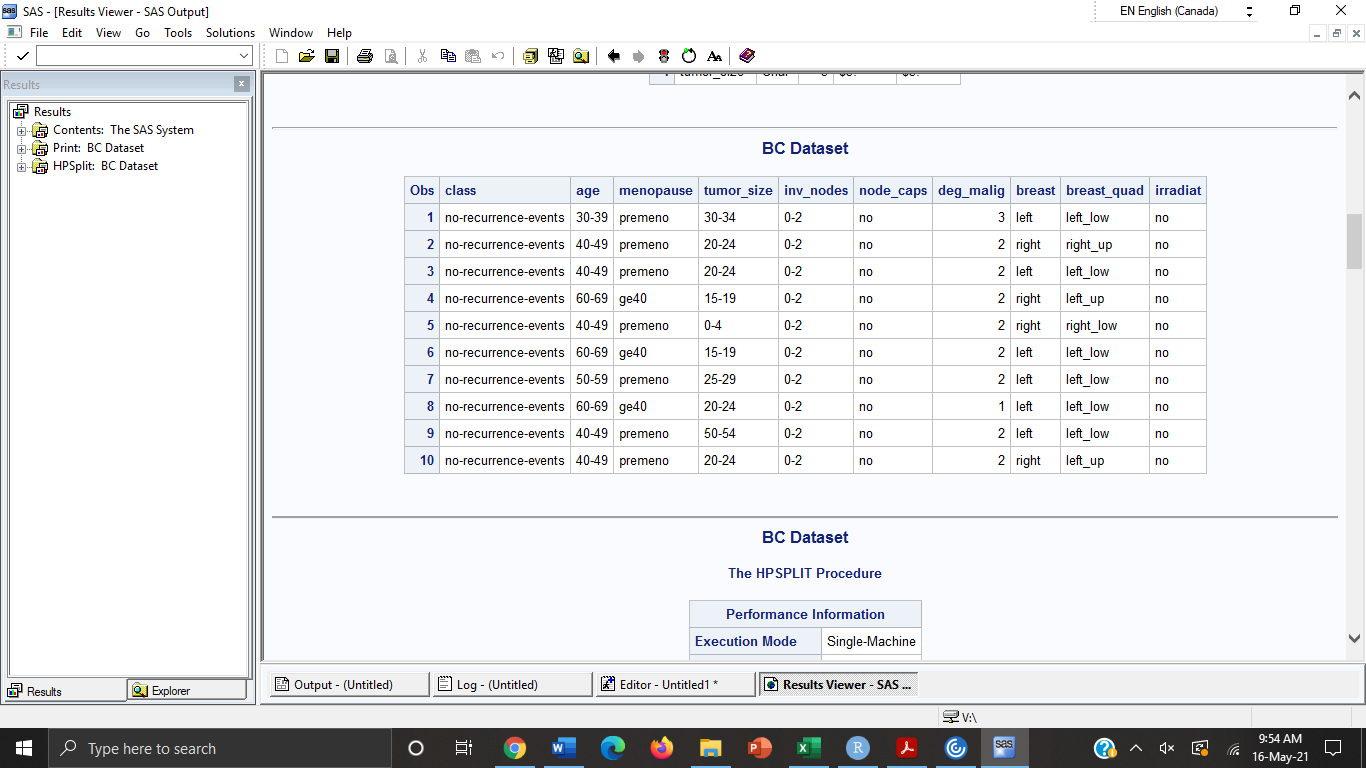
title "BC Dataset";

**run**;







* 1. 
  2. b. Develop a decision tree-based classification model using the hpsplit procedure of SAS. (2 points)
  3. 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. (2 points)
  4. Code:

proc hpsplit data=breast\_cancer\_dataset;

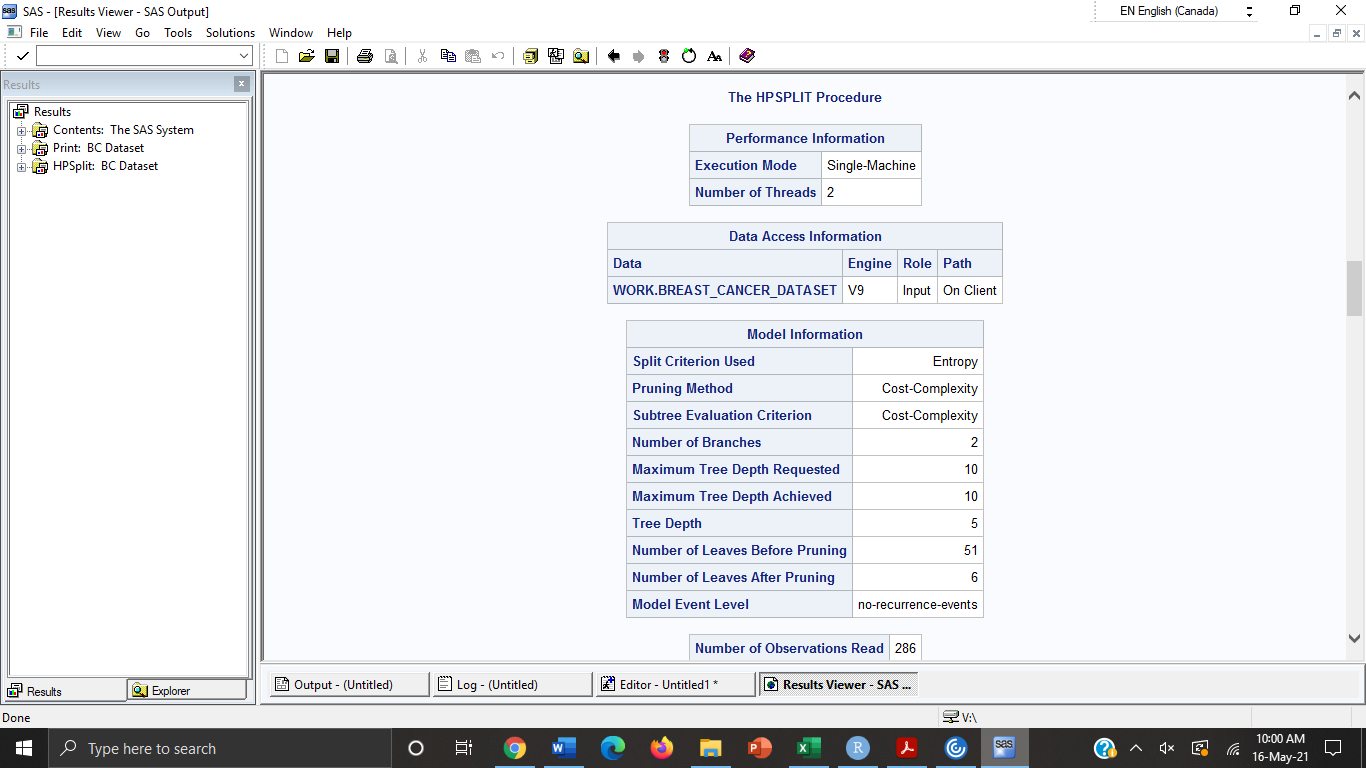
class class age menopause tumor\_size inv\_nodes node\_caps deg\_malig breast breast\_quad irradiat;

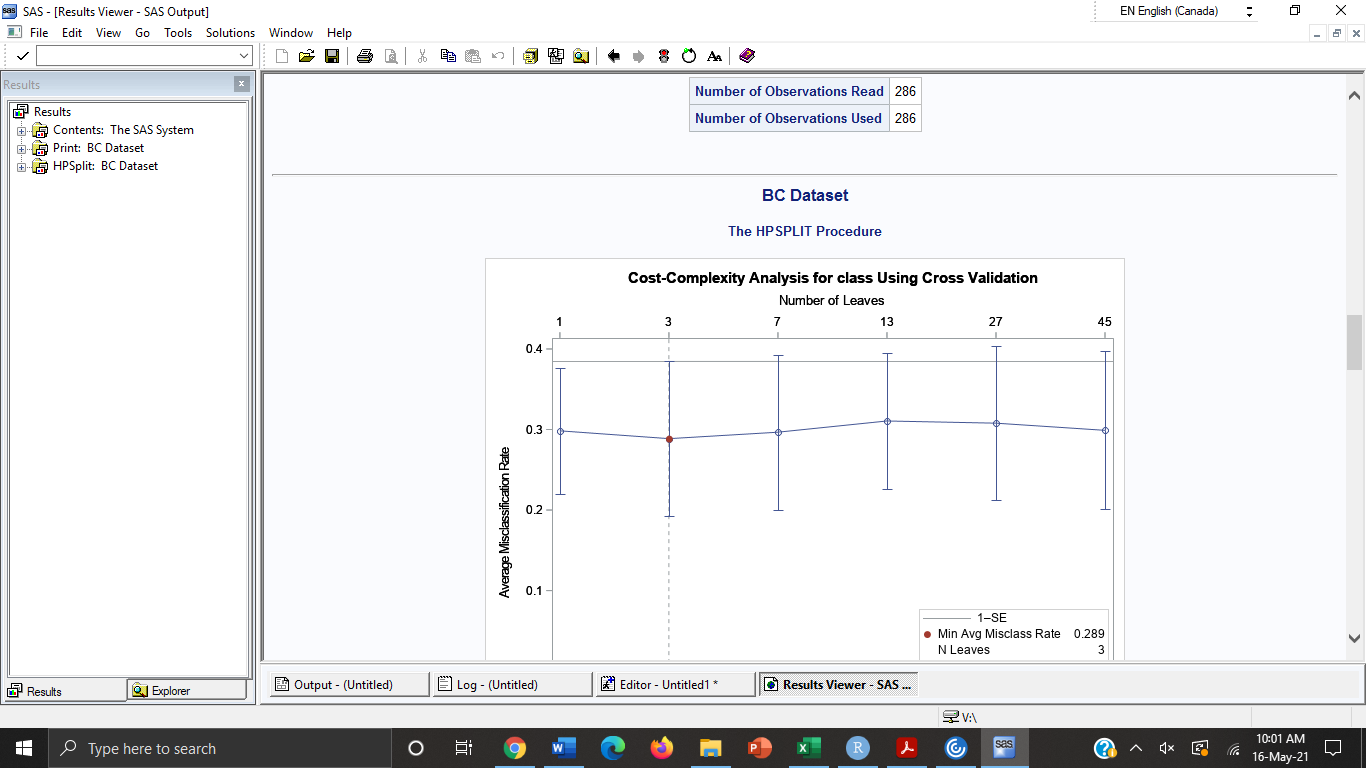
model class=age menopause tumor\_size inv\_nodes node\_caps deg\_malig breast breast\_quad irradiat;

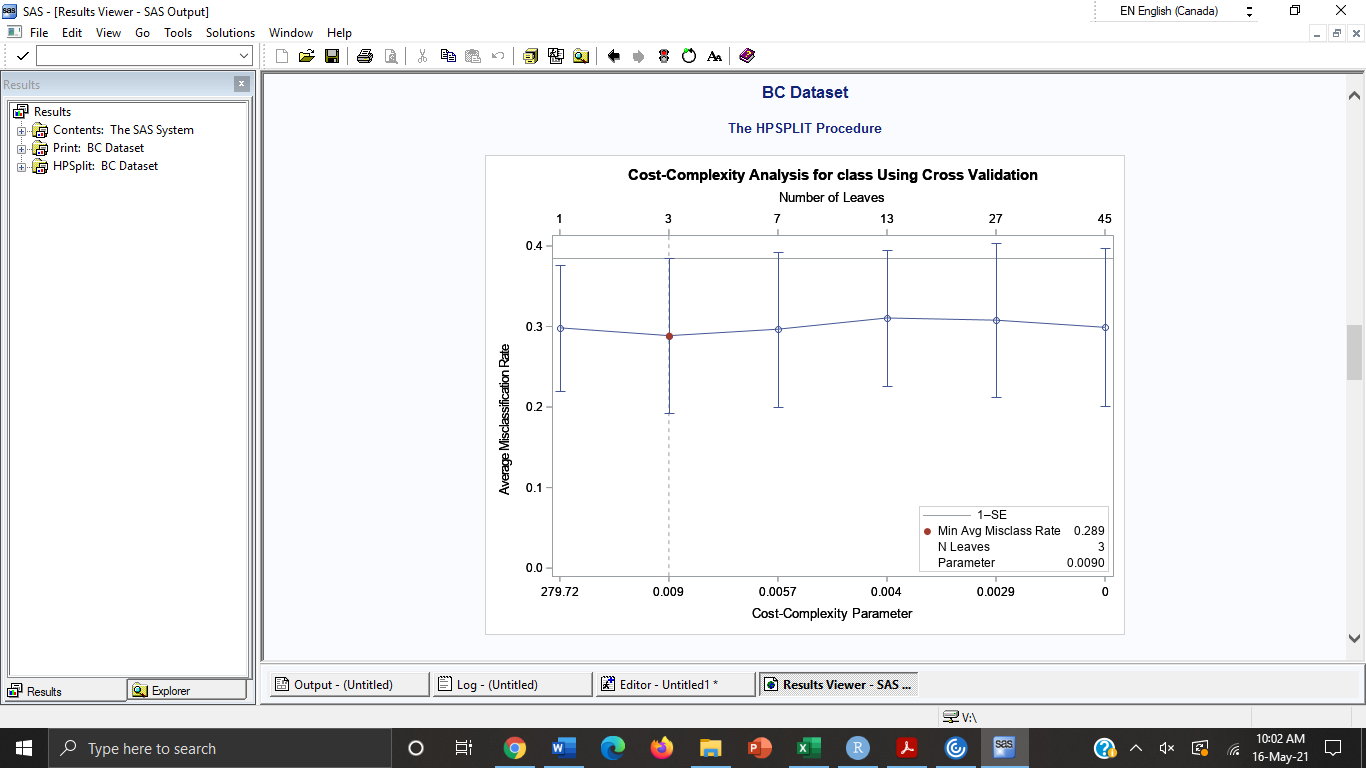
grow entropy;

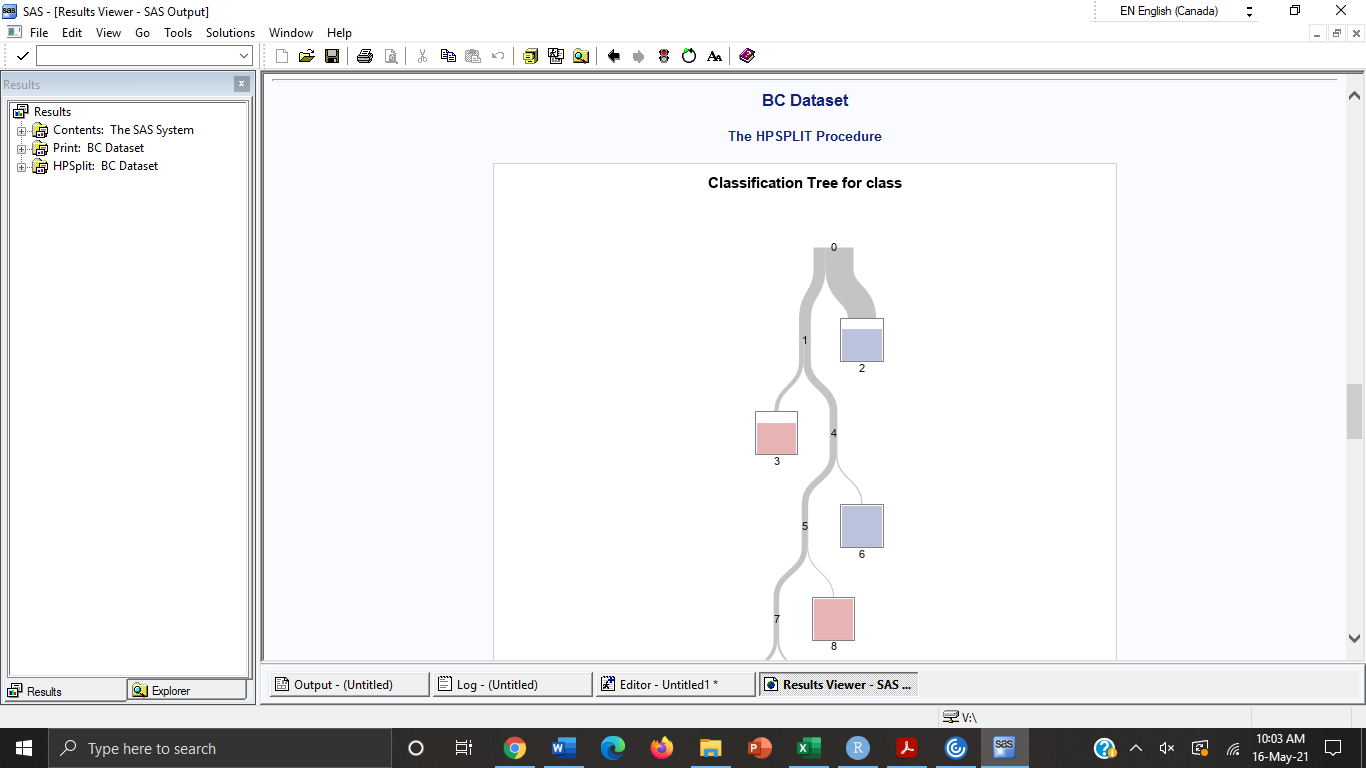
prune costcomplexity;

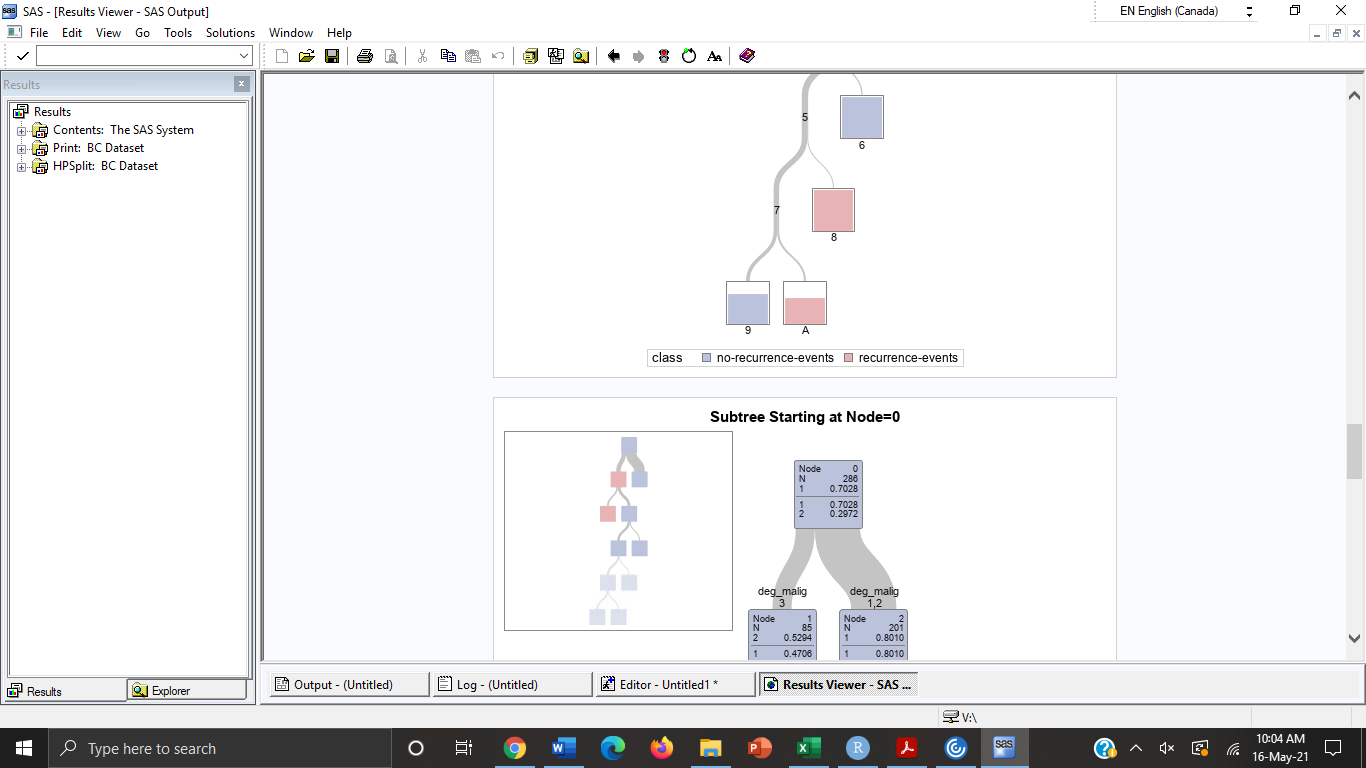
* 1. **run**;

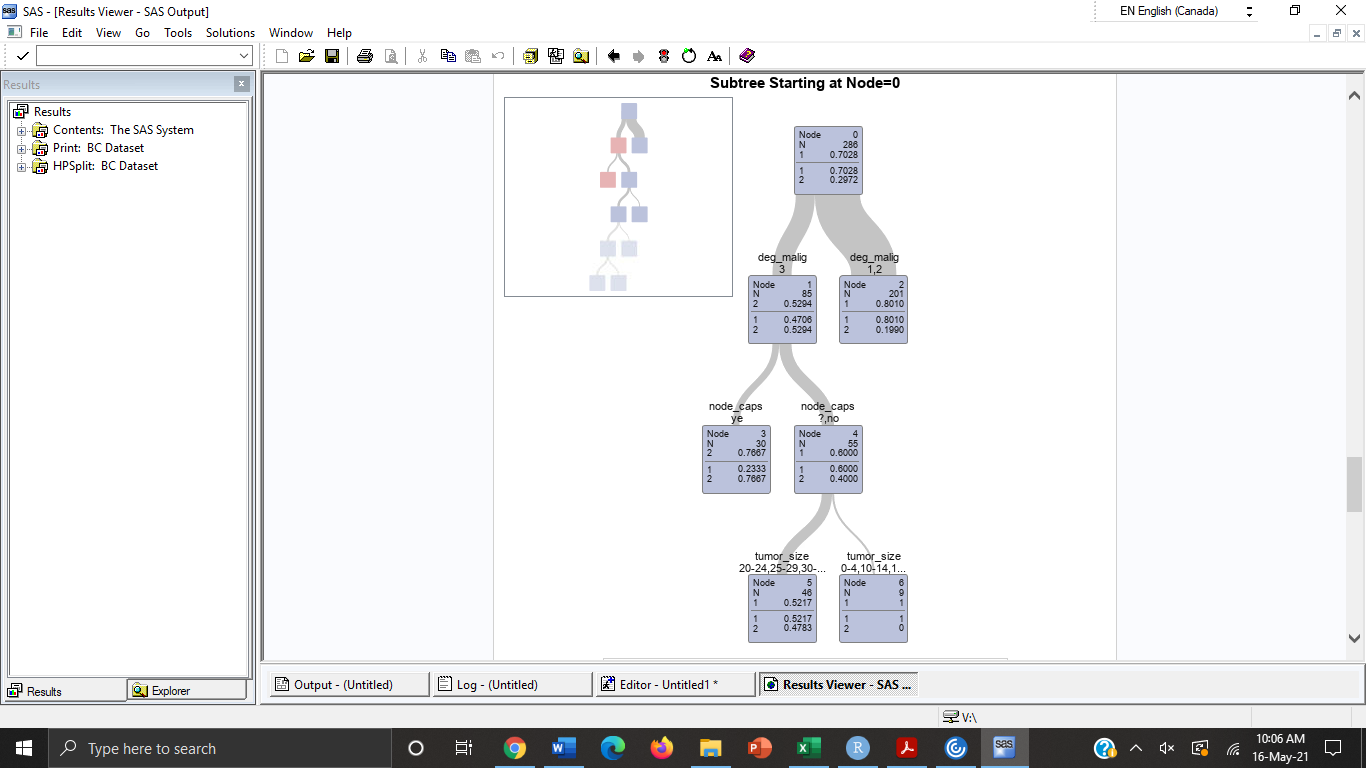
1. 

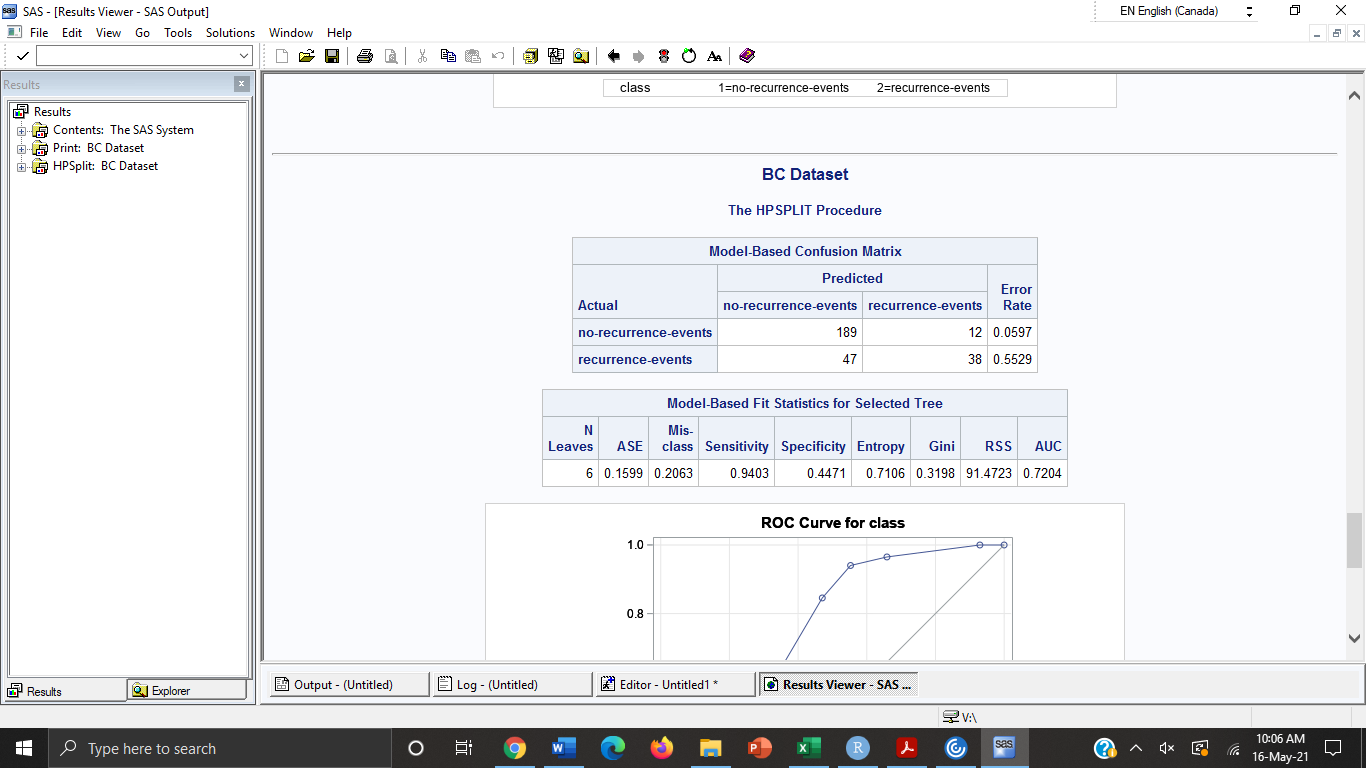


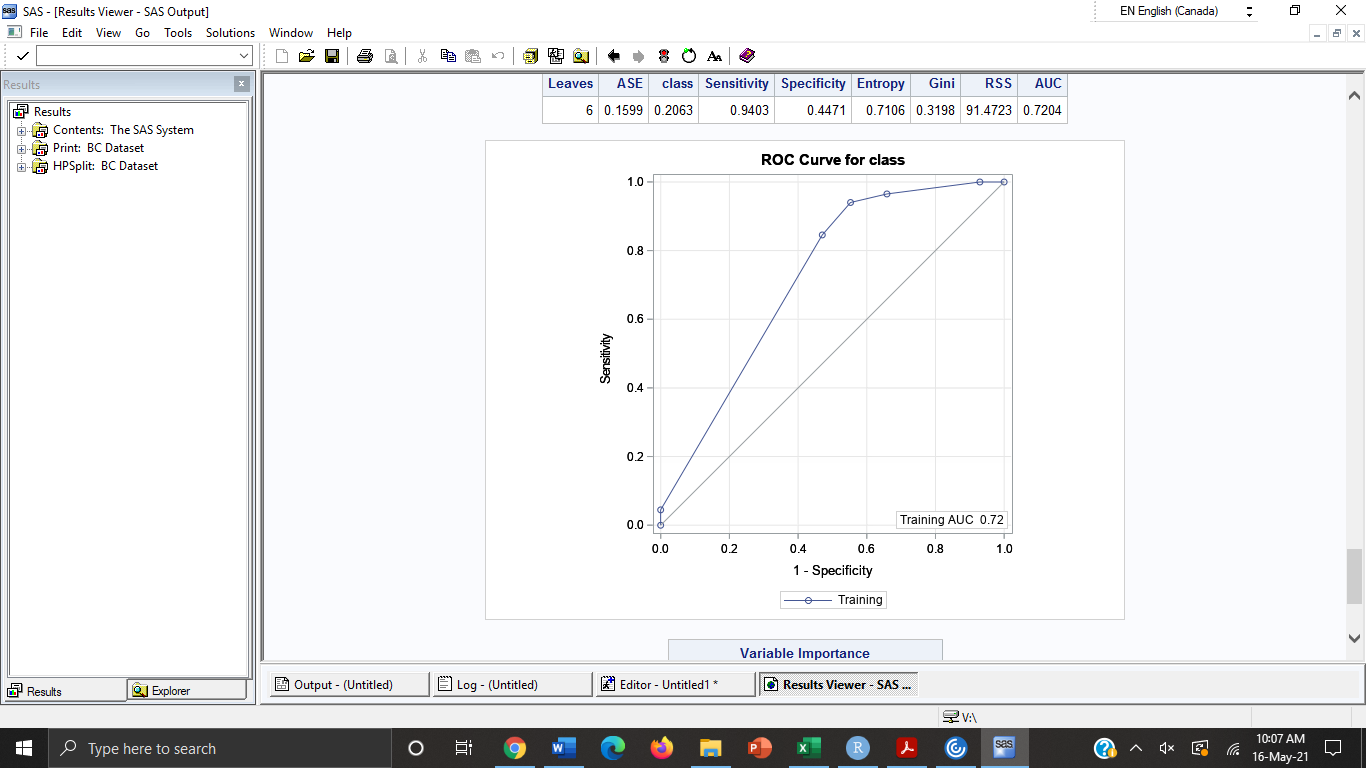


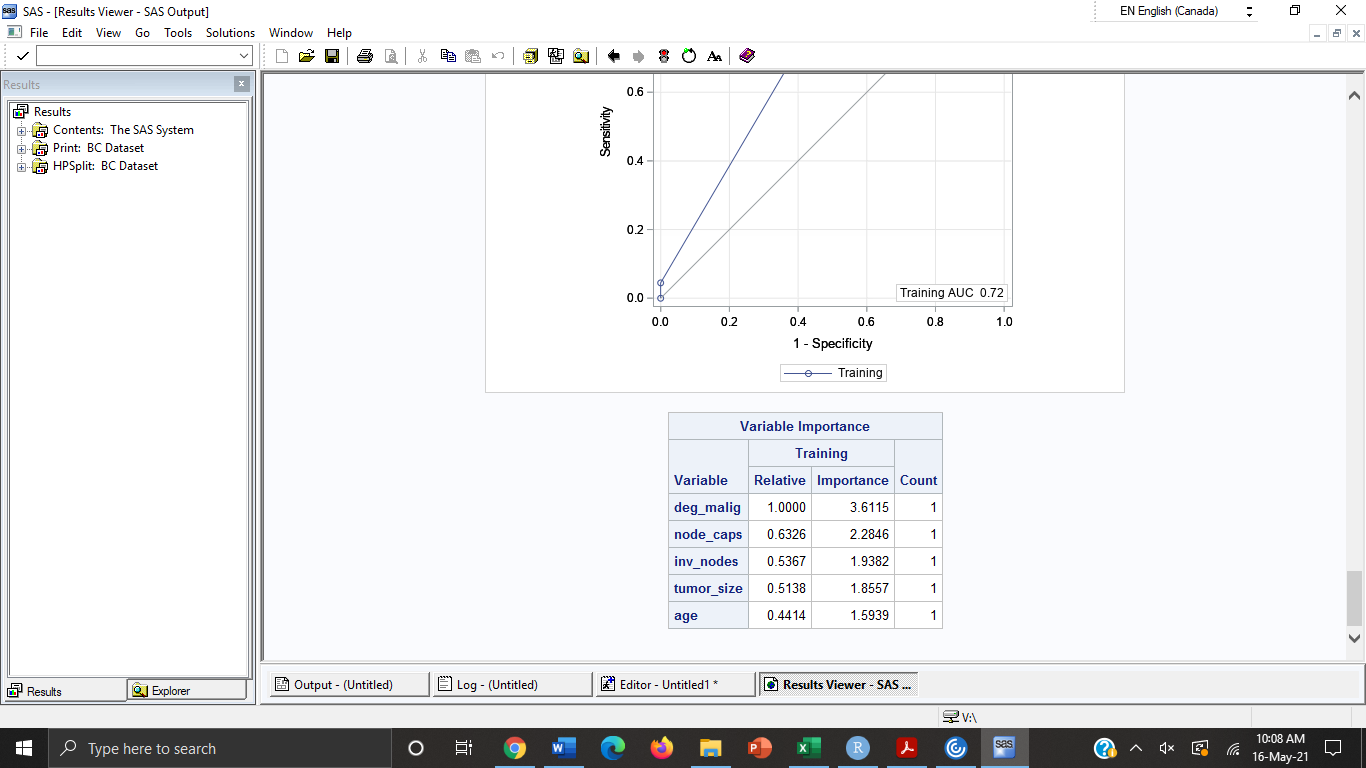






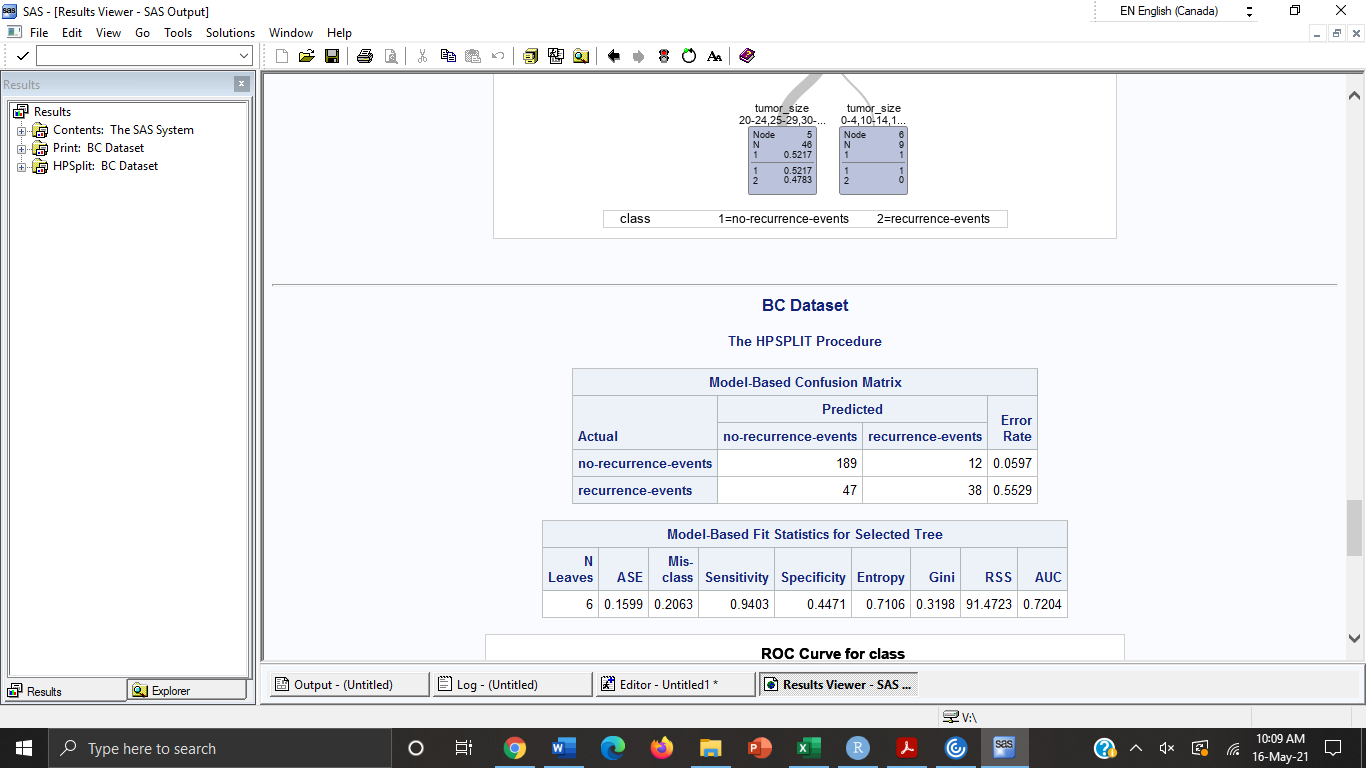






* 1. 2. Using the confusion matrix, compute the following assessment metrics accuracy, recall, and precision (see lecture for formulas). (5 points)

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



|  |  |  |
| --- | --- | --- |
|  | **+** | **-** |
| **+** | ++ | +- |
| **-** | -+ | -- |

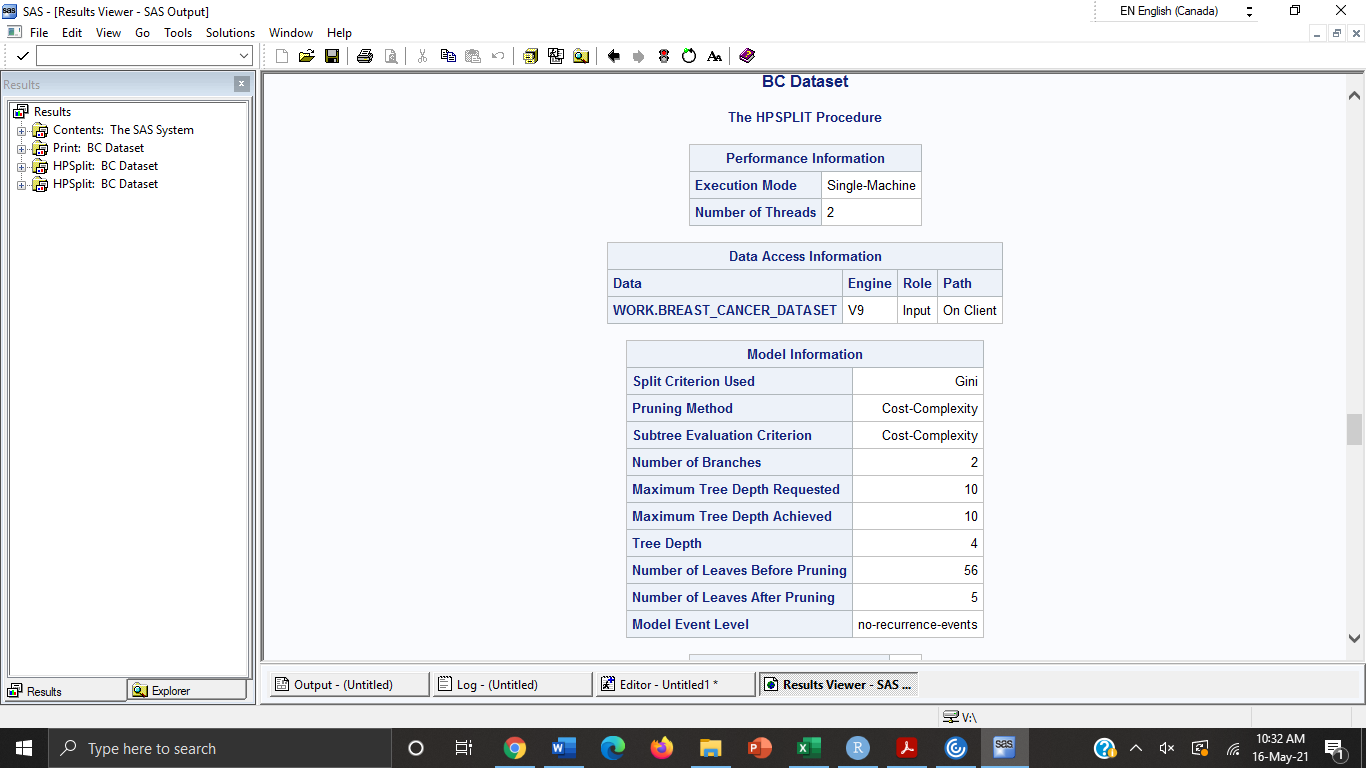
Accuracy= (f++ + f- -) /T= 189+38/286= 0.7937

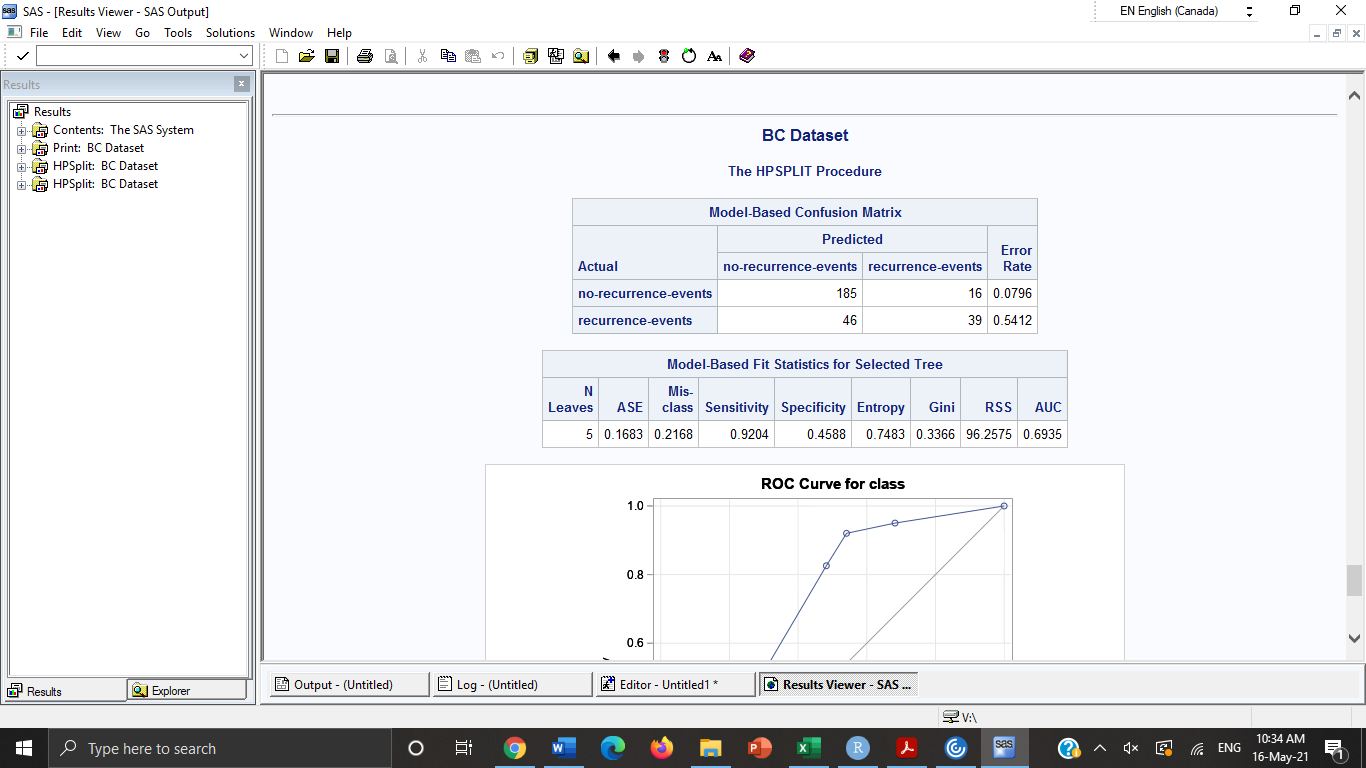
Recall= f++/ (f++ + f+-) = 189/ 189+12= 0.9402

Precision= f++/( f-+ + f++) = 189/ 47+189= 0.8008

.

3. Change the grow algorithm to “gini” and recompute the metrics from question 2. Does entropy build a more accurate classifier or gini? (5 points)

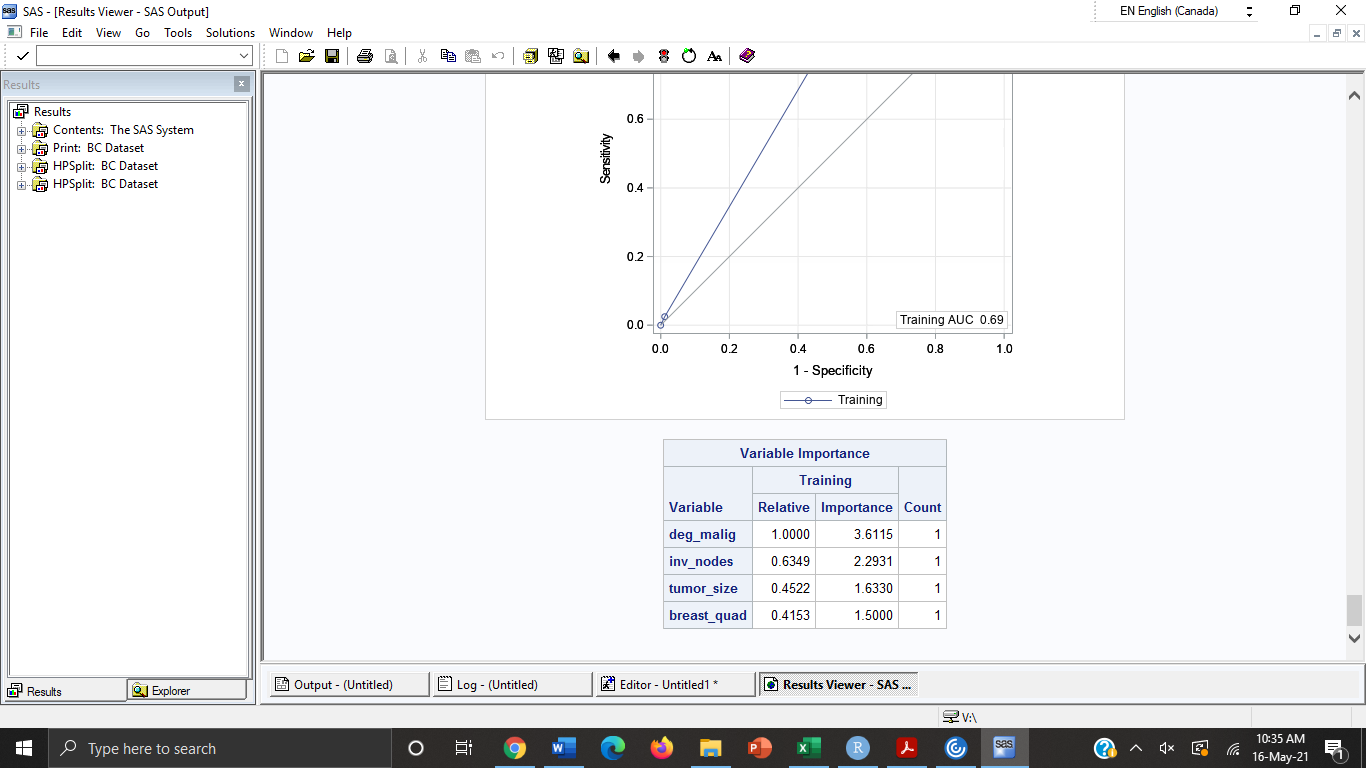




Accuracy= f++ + f- - /T= 185+39/286= 0.7832

Recall= f++/ f++ + f+- = 185/ 185+16= 0.9203

Precision= f++/ f-+ + f++ = 185/ 46+185= 0.8008



**Gini computes faster than entropy classifier based on the CPU time. However, based on metrics after calculating accuracy, recall and precision- entropy is better than Gini.**

