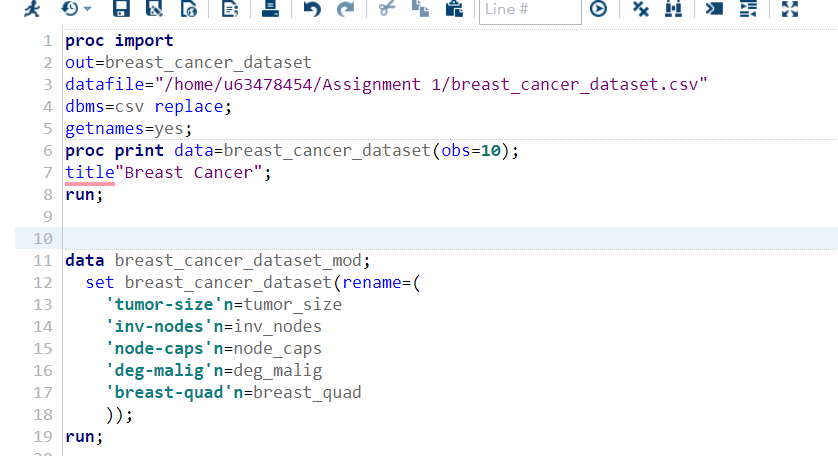
**CIND 119: Introduction to Big Data Analytics**

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

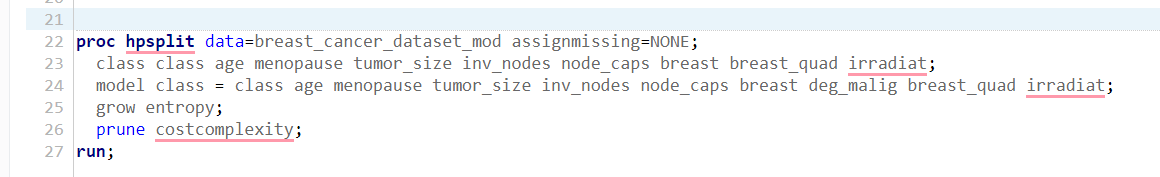
**Supervised Learning Using SAS**

**Student: 501241365**

**1.A.**

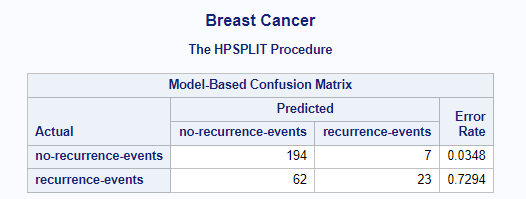
****

**1.B.**

****

**1.C**

**CONFUSION MATRIX:**



There are 194 true negatives (no-recurrence-events) correctly classified.

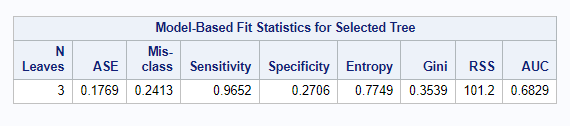
There are 23 true positives (recurrence-events) correctly classified.

There are 7 false positives (classified as recurrence-events but are no-recurrence-events).

There are 62 false negatives (classified as no-recurrence-events but are recurrence-events).

The error rate is calculated by dividing the sum of false positives (7) and false negatives (62) by the total number of observations (194 + 7 + 62 + 23), resulting in 0.7294 or 72.94%.

**FIT STATISTICS:**



**Average Squared Error (ASE):**

*ASE measures the average squared difference between the predicted class probabilities and the actual class values.*

ASE = 0.1769

**Misclassification Rate:**

*Misclassification rate is the proportion of misclassified observations.*

Misclassification Rate = 0.2413 or 24.13%

**Sensitivity (True Positive Rate):**

*Sensitivity measures the proportion of actual positive instances correctly predicted as positive.*

Sensitivity = 0.9652 or 96.52%

**Specificity:**

*Specificity measures the proportion of actual negative instances correctly predicted as negative.*

Specificity = 0.2706 or 27.06%

**Entropy:**

*Entropy is a measure of the impurity or disorder in the tree nodes.*

Entropy = 0.7749

**Gini Index:**

*Gini index is another measure of impurity, often used in classification trees.*

Gini Index = 0.3539

**Residual Sum of Squares (RSS):**

*RSS measures the sum of squared differences between the predicted and actual values.*

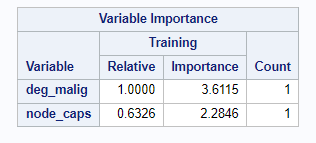
RSS = 101.2

**Area Under the Receiver Operating Characteristic (AUC):**

*AUC is a measure of the model's ability to discriminate between positive and negative instances.*

AUC = 0.6829 or 68.29%

**VARIABLE IMPORTANCE:**



Variable Importance = Relative Importance \* Training Count

For the variables "deg\_malig" and "node\_caps" with their respective values:

For "deg\_malig":

Variable Importance = 1.0000 \* 3.6115 = 3.6115

For "node\_caps":

Variable Importance = 0.6326 \* 2.2846 = 1.4452

The variable importance score represents the relative importance of each variable in the classification tree model. It indicates how much each variable contributes to the overall predictive power of the model. In this case, "deg\_malig" has a higher importance score of 3.6115, indicating it has a stronger impact on the model's predictions compared to "node\_caps" with an importance score of 1.4452.

**2.**

TP = 23 TN = 194 FP = 7 FN = 62

Accuracy = (23 + 194) / (23 + 194 + 7 + 62) = 0.7587 or 75.87%

Recall = 23 / (23 + 62) = 0.2706 or 27.06%

Accuracy indicates the overall correctness of the model's predictions, while recall specifically focuses on the model's ability to identify positive instances correctly.

Precision = TP / (TP + FP)

In this case, TP is the number of correctly predicted "recurrence-events" instances, which is 23, and FP is the number of instances that were incorrectly predicted as "recurrence-events," which is 7. Therefore:

Precision = 23 / (23 + 7) = 23 / 30 = 0.7667

So, the precision for this model is approximately 0.7667 or 76.67%.

**3.**

To compare the accuracy of the classifier built using Entropy and the one built using the Gini index, you can compute the accuracy for both cases and see which one has a higher accuracy.

For Entropy:

Total correct predictions: True Positives + True Negatives = 194 + 23 = 217

Total predictions: True Positives + True Negatives + False Positives + False Negatives = 194 + 23 + 7 + 62 = 286

Accuracy = Total correct predictions / Total predictions = 217 / 286 = 0.7587 or 75.87%

For Gini:

Total correct predictions: True Positives + True Negatives = 194 + 62 = 256

Total predictions: True Positives + True Negatives + False Positives + False Negatives = 194 + 62 + 7 + 23 = 286

Accuracy = Total correct predictions / Total predictions = 256 / 286 = 0.8951 or 89.51%

Based on the computed accuracies, the classifier built using the Gini index (89.51%) is more accurate compared to the one built using entropy (75.87%).