17/05/2021 **Mohd. Aftab Alam**
02013302717

EXPERIMENT 3

Aim :

Estimate the Accuracy of Decision Classifier on Breast Cancer Dataset using 5-fold Cross Validation (You need to choose the appropriate options for missing values).

Introduction :

There are various approaches to determine the performance of classifiers. It can most simply be measured by counting the proportion of correctly predicted examples in a test dataset. This value is the ***classification accuracy,*** which is also ***1-ErrorRate.*** Both terms are used in literature.

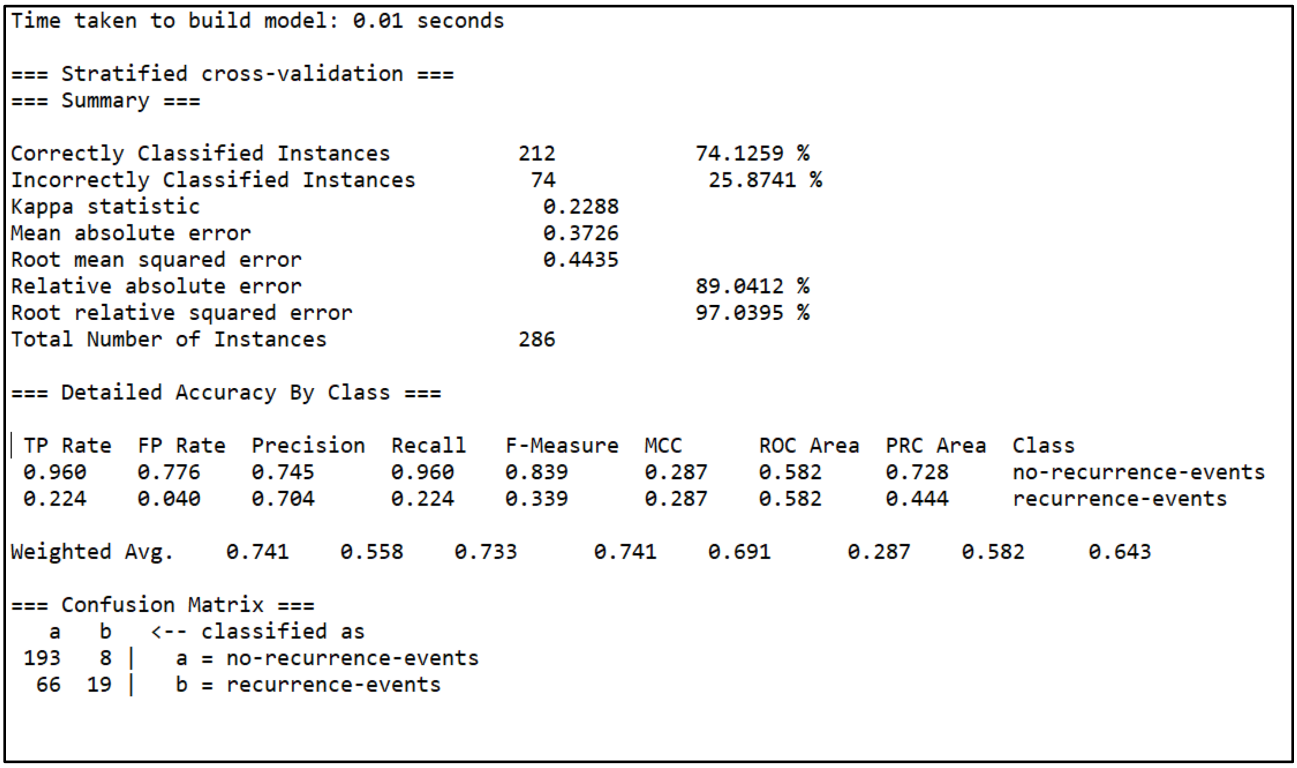
The simplest case for evaluation is when we use a training set and a test set which are mutually independent. This is referred to as hold-out estimate. To estimate variance in these performance estimates, hold-out estimates may be computed by repeatedly by resampling the same dataset -- i.e., randomly shuffling it and then splitting it into training and test sets with a specific proportion of the examples, collecting all estimates on the test sets and computing average and standard deviation of accuracy.

A more elaborate method is ***k***-fold cross-validation. Here, a number of folds ***k*** is specified. The dataset is randomly shuffled and then split into ***k*** folds of equal size. In each iteration, one-fold is used for testing and the other ***k-1*** folds are used for training the classifier. The test results are collected and pooled (or averaged) over all folds. This gives the cross-validation estimate of accuracy. The folds can be purely random or slightly modified to create the same class distributions in each fold as in the complete dataset. In the latter case the cross-validation is called ***stratified.***

Leave-one-out (loo) cross-validation signifies that ***k*** is equal to the number of examples. Out of necessity, loo cv has to be non-stratified, i.e., the class distributions in the test sets are not the same as those in the training data. Therefore, loo cv can produce misleading results in rare cases. However, it is still quite useful in dealing with small datasets since it utilizes the greatest amount of training data from the dataset.

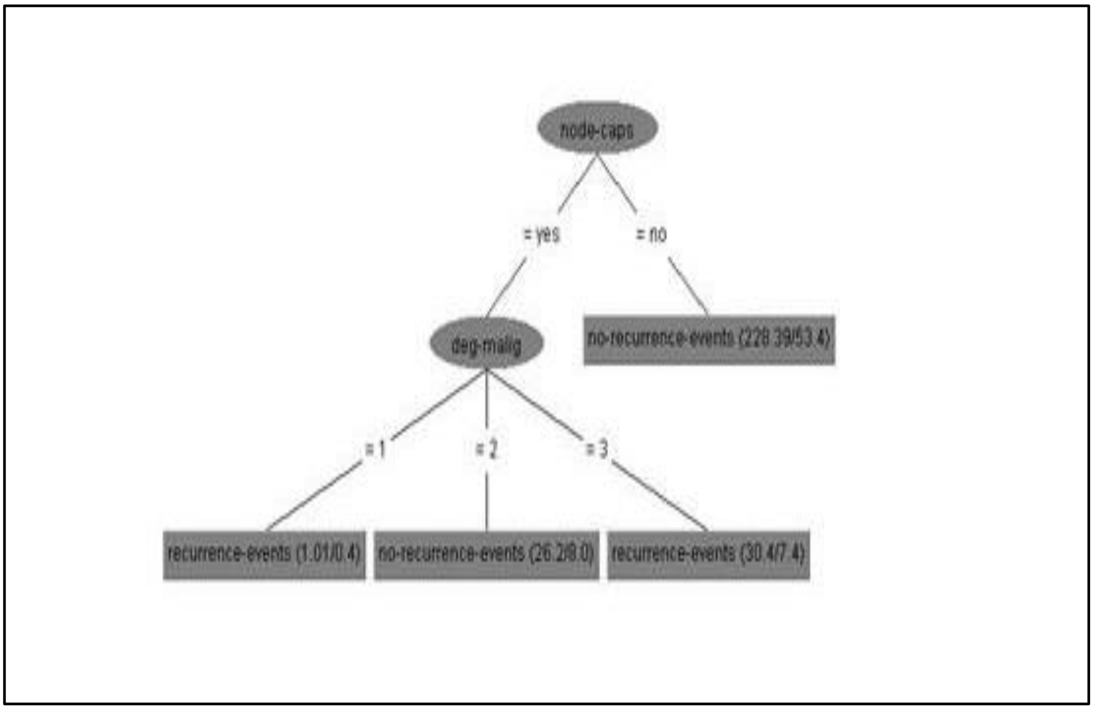
Implementation :

Classify using J48 (using 5-fold Cross Validation).



```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         212               74.1259 %
Incorrectly Classified Instances        74               25.8741 %
Kappa statistic                          0.2288
Mean absolute error                      0.3726
Root mean squared error                  0.4435
Relative absolute error                 89.0412 %
Root relative squared error             97.0395 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
 0.960    0.776    0.745      0.960    0.839      0.287    0.582     0.728     no-recurrence-events
 0.224    0.040    0.704      0.224    0.339      0.287    0.582     0.444     recurrence-events

Weighted Avg.    0.741    0.558    0.733      0.741    0.691      0.287    0.582     0.643

=== Confusion Matrix ===
   a   b   <-- classified as
 193   8 |   a = no-recurrence-events
  66  19 |   b = recurrence-events
```

Visualization Tree :


node-caps
= yes
= no
no-recurrence-events (228.39/53.4)
deg-malig
= 1
= 2
= 3
recurrence-events (1.01/0.4)
no-recurrence-events (26.2/8.0)
recurrence-events (30.4/7.4)