24/05/2021 **Mohd. Aftab Alam**
02013302717

EXPERIMENT 4

Aim :

Estimate the Precision, Recall, Accuracy and F-Measure of the Decision Tree Classifier on the Text Classification task for each of the 10 categories using 10-fold Cross Validation.

Introduction :

Text classification is one of the key techniques in text mining to categorize the documents in a supervised manner. The processing of text classification involves two main problems are the extraction of feature terms that become effective keywords in the training phase and then the actual classification of the document using these feature terms in the test phase. This text classification task has numerous applications such as automated indexing of scientific articles according to predefined thesauri of technical terms, routing of customer email in a customer service department, filing patents into patent directories, automated population of hierarchical catalogues of Web resources, selective dissemination of information to consumers, identification of document genre, or detection and identification of criminal activities for military, police, or secretes service environments and so on.

- TP = true positives: number of examples predicted positive that are actually positive
- FP = false positives: number of examples predicted positive that are actually negative
- TN = true negatives: number of examples predicted negative that are actually negative
- FN = false negatives: number of examples predicted negative that are actually positive

Recall is referred to as the true positive rate or sensitivity.

The Precision is the proportion of the examples which truly have class x among all those which were classified as class x.

$$\boldsymbol{Recall = \frac{tp}{tp + fn}}$$

$$\boldsymbol{Precision = \frac{tp}{tp + fn}}$$

$$\boldsymbol{Accuracy = \frac{tp + tn}{tp + tn + fp + fn}}$$

The F-Measure is simply ***2 * Precision * Recall / (Precision + Recall),*** a combined measure for precision and recall.

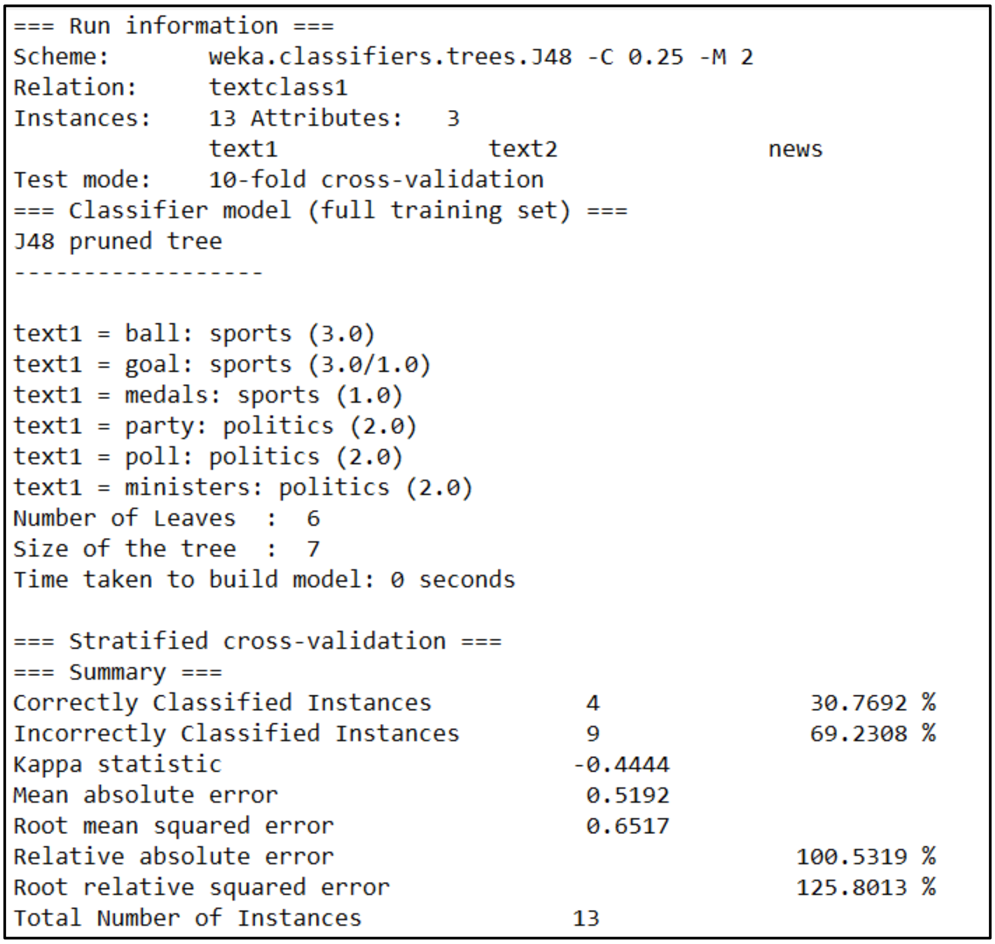
These measures are useful for comparing classifiers.

24/05/2021 **Mohd. Aftab Alam**
02013302717

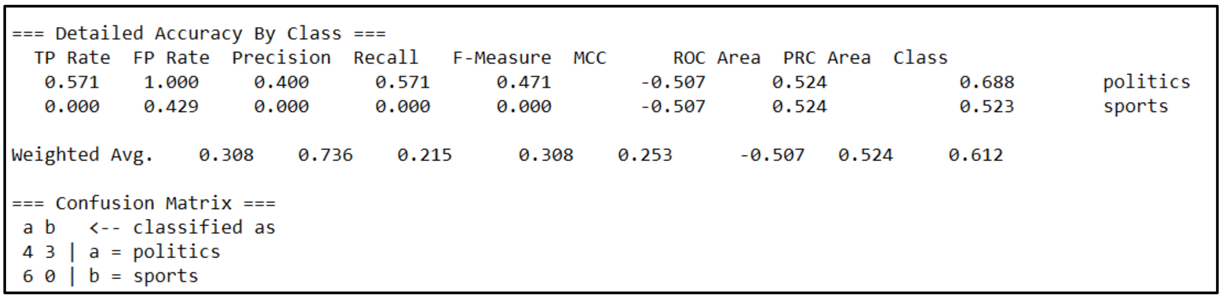
DataFile

```
@relation textclass1

@attribute text1 {ball, goal, medals, party, poll, ministers}

@attribute text2 {wicket, ball, poll, election, performance, party}

@attribute news {politics, sports}

@data

ball, wicket, sports

goal, ball, sports

party, poll, politics

poll, election, politics

ministers, election, politics

medals, performance, sports

ball, party, sports

goal, wicket, sports

ministers, party, politics

party, election, politics

goal, election, politics

poll, performance, politics

ball, performance, sports
```

Implementation :

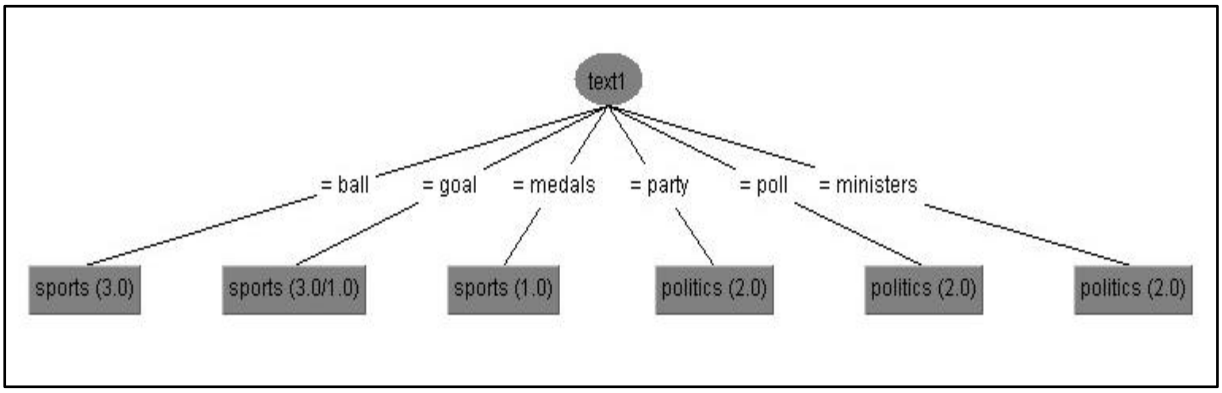


```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     textclass1
Instances:    13 Attributes:   3
              text1               text2               news
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------

text1 = ball: sports (3.0)
text1 = goal: sports (3.0/1.0)
text1 = medals: sports (1.0)
text1 = party: politics (2.0)
text1 = poll: politics (2.0)
text1 = ministers: politics (2.0)
Number of Leaves  :  6
Size of the tree  :  7
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances           4              30.7692 %
Incorrectly Classified Instances         9              69.2308 %
Kappa statistic                         -0.4444
Mean absolute error                      0.5192
Root mean squared error                  0.6517
Relative absolute error                                100.5319 %
Root relative squared error                            125.8013 %
Total Number of Instances               13
```



```
=== Detailed Accuracy By Class ===
   TP Rate   FP Rate   Precision   Recall   F-Measure  MCC      ROC Area  PRC Area  Class
    0.571     1.000     0.400      0.571     0.471          -0.507     0.524            0.688        politics
    0.000     0.429     0.000      0.000     0.000          -0.507     0.524            0.523        sports

Weighted Avg.    0.308     0.736     0.215      0.308     0.253      -0.507    0.524     0.612

=== Confusion Matrix ===
 a b   <-- classified as
 4 3 | a = politics
 6 0 | b = sports
```

Visualization Tree :


text1
= ball
= goal
= medals
= party
= poll
= ministers
sports (3.0)
sports (3.0/1.0)
sports (1.0)
politics (2.0)
politics (2.0)
politics (2.0)