10/05/2021 **Mohd. Aftab Alam**
02013302717

EXPERIMENT 2

Aim :

Study and Implement the Decision Tree Learner using WEKA (Breast Cancer Dataset).

Introduction :

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. In 2011, authors of the Weka machine learning software described the C4.5 algorithm as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date".

It became quite popular after ranking #1 in the *Top 10 Algorithms in Data Mining* pre-eminent paper published by Springer LNCS in 2008.

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $\mathbf{S = s_1, s_2, \ldots}$ of already classified samples. Each sample $\mathbf{s_i}$ consists of a p-dimensional vector $\mathbf{(x_{1,i}, x_{2,i}, \ldots, x_{p,i})}$ where the $\mathbf{x_j}$ represent attribute values or features of the sample, as well as the class in which $\mathbf{s_i}$ falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the partitioned sub lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Implementation :

J48 is an open-source Java implementation of the C4.5 algorithm in the WEKA data mining tool.

Classify using J48.

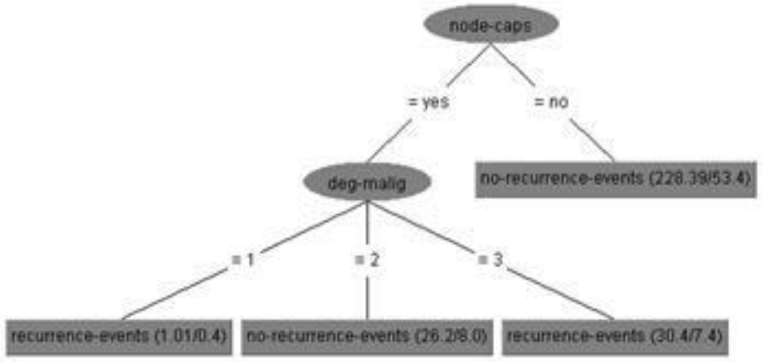
```
Time taken to build model: 0.02 seconds
=== Summary ===
Correctly Classified Instances          203              70.979  %
Incorrectly Classified Instances        83               29.021  %
Kappa statistic                         0.2055
Mean absolute error                     0.3767
Root mean squared error                 0.4586
Relative absolute error                                   89.9901 %
Root relative squared error                              100.335  %
Total Number of Instances               286

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
0.886    0.706    0.748      0.886  |0.811      0.220    0.609     0.755     no-recurrence-events
0.294    0.114    0.521      0.294   0.376      0.220    0.609     0.410     recurrence-events

Weighted Avg.    0.710    0.530    0.680      0.710    0.682      0.220    0.609     0.653
=== Confusion Matrix ===
   a   b   <-- classified as
 178  23 |    a = no-recurrence-events
  60  25 |    b = recurrence-events
```

Decision Tree Visualization :


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