# Συστήματα Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

# Εργασία 01 - HW1 - Decision Trees


## Ονοματεπώνυμο: Παναγιώτης Γιαννουτάκης

## ΑΜ: 12/38


### (α)

GINI(parent) = 1 – (4/9)^2 – (5/9)^2 = 1 – 16/81 – 25/81 = 1 – 41/81 = 0.493

### (β)

ID

GINI(id1) = 1 – (1/1)^2 = 1 – 1 = 0

GINI(id2) = 1 – (1/1)^2 = 1 – 1 = 0

…

GINI(id9) = 1 – (1/1)^2 = 1 – 1 = 0

GINI(ID_split) = 1*0 + 1*0 + … + 1*0 = 0

GAIN(ID) = 0,493 – 0 = 0,493

A1

GINI(T) = 1 - (3/4)^2 – (1/4)^2 = 1 – 9/16 – 1/16 = 1 – 10/16 = 0,375

GINI(F) = 1 – (1/5)^2 – (4/5)^2 = 1 – 1/25 – 16/25 = 1 – 17/25 = 0,32

GINI(a1_split) = (4/9) * 0,375 + (5/9) * 0,32 = 0,166 + 0,177 = 0,343

GAIN(a1) = 0,493 – 0,343 = 0,149

A2

GINI(T) = 1 – (2/5)^2 – (3/5)^2 = 1 – 4/25 – 9/25 = 1 – 13/25 = 0,48

GINI(F) = 1 – (2/4)^2 – (2/4)2 = 1 – 4/16 – 4/16 = 1 – 8/16 = 0,5

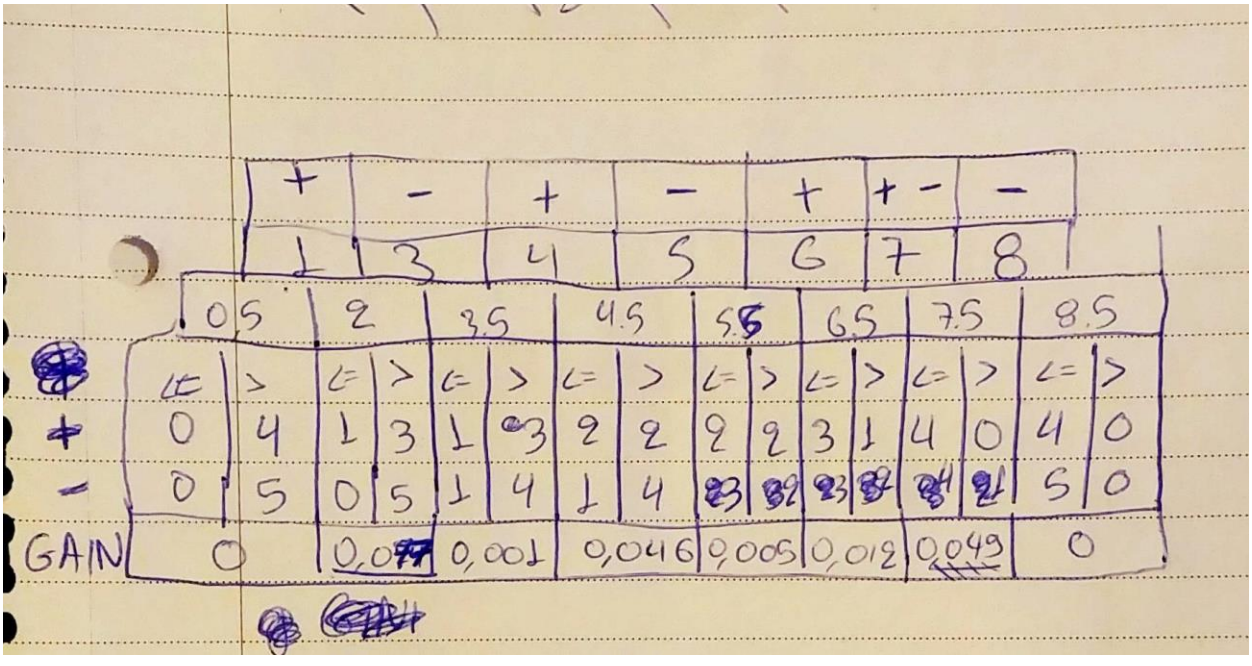
GINI(a2_split) = 5/9 * 0,48 + (4/9) * 0,5 = 0,266 + 0,222 = 0,488

GAIN(a2) = 0,493 – 0,488 = 0,005

Δεν επιλέγουμε το GAIN(ID) γιατί σε μελλοντικές καταχωρίσεις καμία καταχώριση δεν θα μπει σε κάποιον από τους κόμβους που ήδη υπάρχουν αλλά θα δημιουργείται κάθε φορά ένας νέος.

## (γ)

Ο πίνακας με τις διασπάσεις και τα GAIN είναι ο παρακάτω:




Ακολουθούν οι πράξεις που έγιναν στο χαρτί:

**0,5**

$GINI(\leq) = 1 - (0/0)^2 - (\frac{0}{0})^2 = 1$

$GINI(>) = 1 - (4/9)^2 - (5/9)^2 = 1 - \frac{16}{81} - \frac{25}{81} = 1 - \frac{41}{81} =$

$= 1 - 0,506 \simeq 0,493$

$GINI(0,5\_split) = 0 \cdot 1 + 9/9 \cdot 0,494 = 0,493$

$GAIN(0,5) = 0,493 - 0,493 = 0$

**2**

$GINI(\leq) = 1 - (1/1)^2 - (0/1)^2 = 1 - 1 - 0 = 0$

$GINI(>) = 1 - (3/8)^2 - (5/8)^2 = 1 - \frac{9}{64} - \frac{25}{64} = 1 - \frac{34}{64} =$

$= 1 - 0,531 = 0,469$

$GINI(2\_split) = 1/9 \cdot 0 + 8/9 \cdot 0,469 = 0,416$

$GAIN(2) = 0,493 - 0,416 = 0,077$

**3,5**

$GINI(\leq) = 1 - (1/2)^2 - (1/2)^2 = 1 - 0,5 = 0,5$

$GINI(>) = 1 - (3/7)^2 - (4/7)^2 = 1 - \frac{9}{49} - \frac{16}{49} = 1 - \frac{25}{49} =$

$= 1 - 0,510 = 0,49$

$GINI(3,5\_split) = 2/9 \cdot 0,5 + 7/9 \cdot 0,49 = 0,111 + 0,381 =$

$= 0,492$

$GAIN(3,5) = 0,493 - 0,492 = 0,001$

4,5

GINI(<=) = 1 - (2/3)^2 - (1/3)^2 = 1 - 4/9 - 1/9 = 1 - 5/9 = 0,445

GINI(>) = 1 - (2/6)^2 - (4/6)^2 = 1 - 4/36 - 16/36 = 1 - 20/36 =

= 1 - 0,555 = 0,445

GINI(4,5-split) = 3/9 · 0,445 + 6/9 · 0,445 = 0,151 + 0,296 = 0,447

GAIN(4,5) = 0,493 - 0,447 = 0,046

5,5

GINI(<=) = 1 - (2/5)^2 - (3/5)^2 = 1 - 4/25 - 9/25 = 1 - 13/25 = 1 - 0,52 = 0,48

GINI(>) = 1 - (2/4)^2 - (2/4)^2 = 1 - 4/16 - 4/16 = 1 - 8/16 = 0,5

GINI(5,5-split) = 5/9 · 0,48 + 4/9 · 0,5 = 0,266 + 0,222 = 0,488

GAIN(5,5) = 0,493 - 0,488 = 0,005

6,5

GINI(<=) = 1 - (3/6)^2 - (3/6)^2 = 1 - 9/36 - 9/36 = 1 - 18/36 = 0,5

GINI(>) = 1 - (1/3)^2 - (2/3)^2 = 1 - 1/9 - 4/9 = 1 - 5/9 = 0,445

GINI(6,5-split) = 6/9 · 0,5 + 3/9 · 0,445 = 0,333 + 0,148 = 0,481

GAIN(6,5) = 0,493 - 0,481 = 0,012

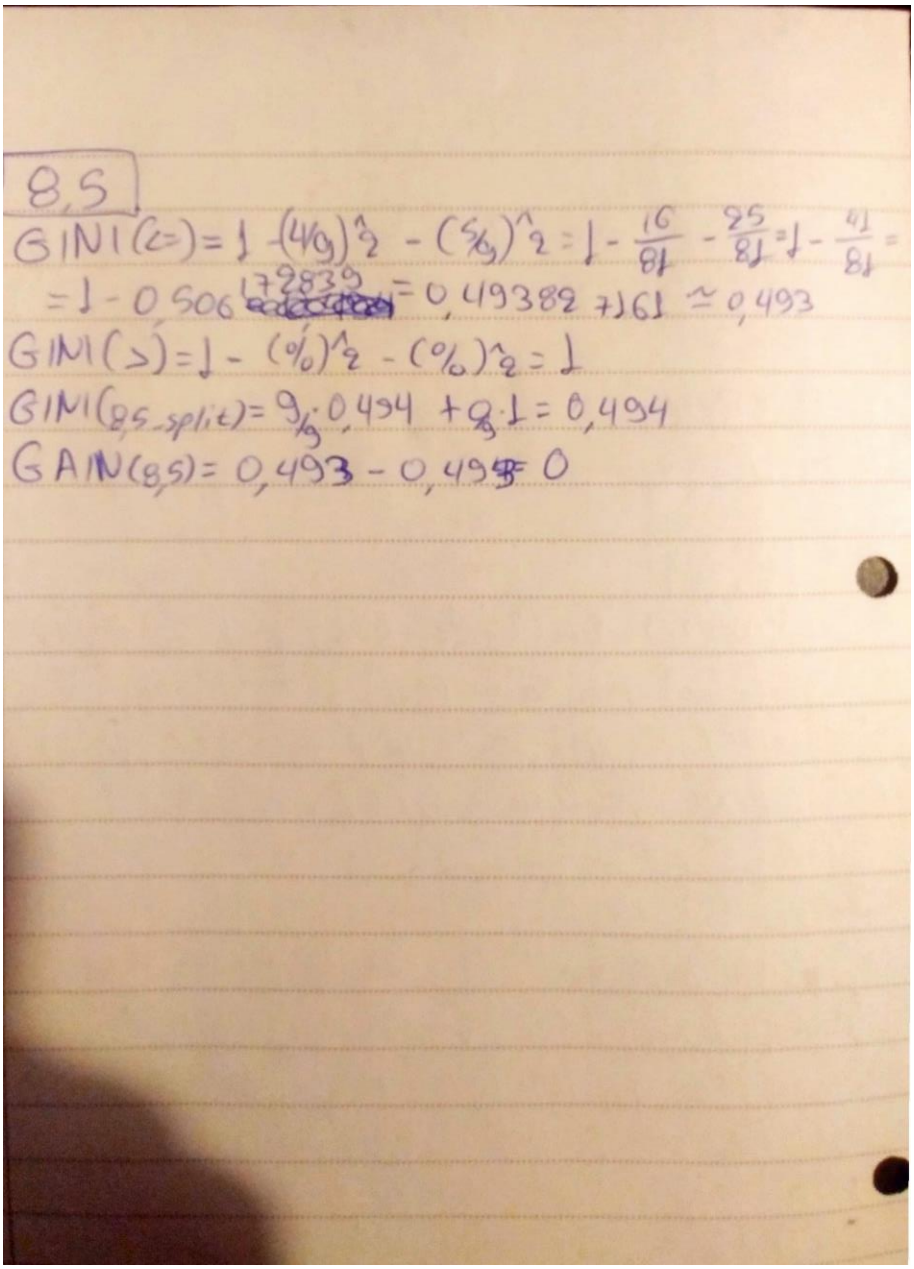
7,5

GINI(<=) = 1 - (4/8)^2 - (4/8)^2 = 1 - 16/64 - 16/64 = 1 - 32/64 = 1 - 0,5 = 0,5

GINI(>) = 1 - (0/1)^2 - (1/1)^2 = 1 - 0 - 1 = 0

GINI(7,5-split) = 8/9 · 0,5 + 1/9 · 0 = 0,444

GAIN(7,5) = 0,493 - 0,444 = 0,049




Η καλύτερη διάσπαση προκύπτει εκεί που υπάρχει το μεγαλύτερο GAIN, δηλαδή στην διάσπαση 2.

## (δ)

Η ρίζα του δέντρου θα είναι η στήλη a1 γιατί έχει το μεγαλύτερο GAIN από όλες τις στήλες (εκτός από την στήλη ID που δεν την μετράμε). Το δέντρο είναι το παρακάτω:

|--- T (3,1) {+}

A1 ----- |

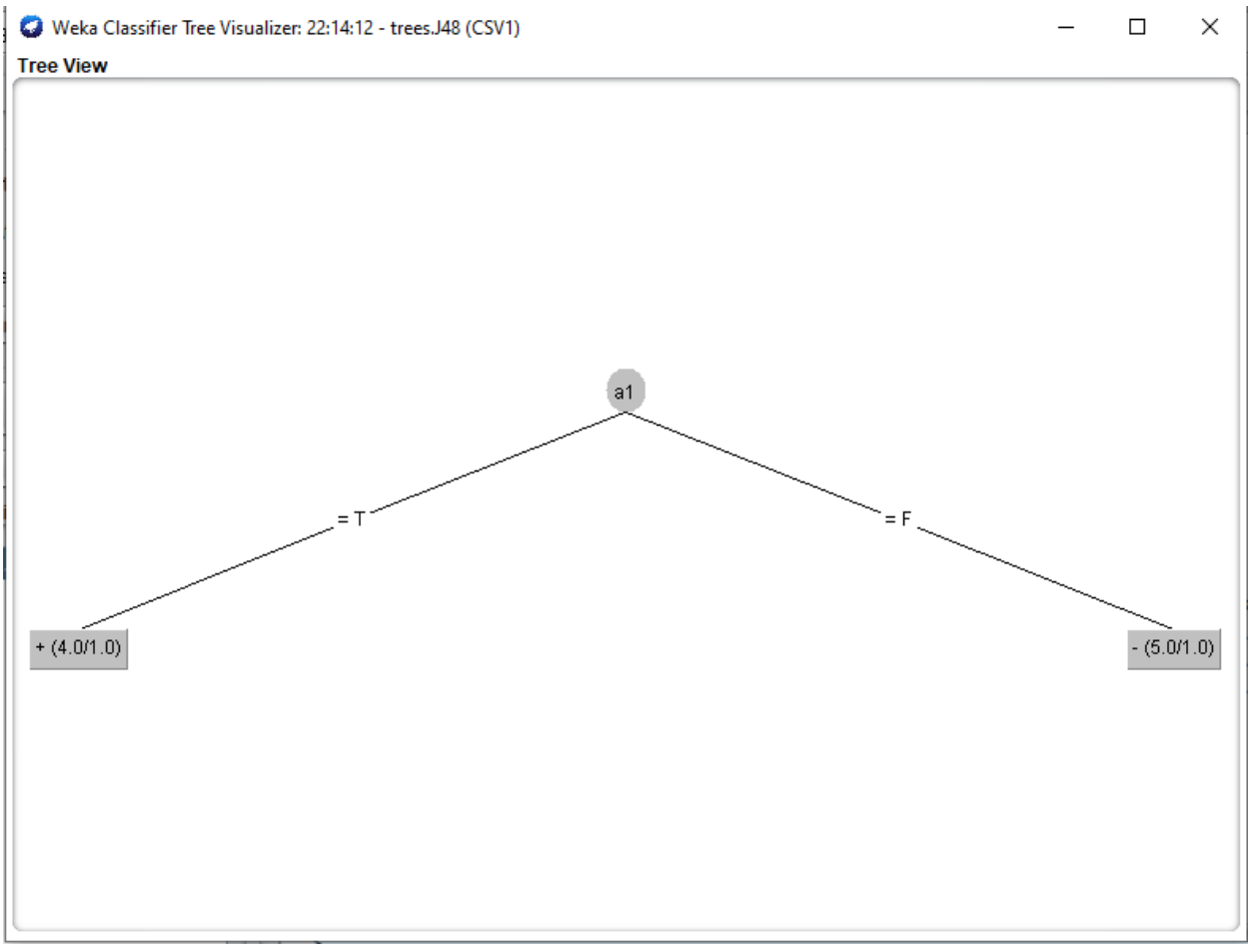
|--- F (1,4) {-}

Το ποσοστό των εγγραφών που κατηγοριοποιούνται σωστά είναι 7/9, δηλαδή 77,7 % και το υπόλοιπο 2/9 εγγραφές, δηλαδή 22,3% κατηγοριοποιούνται λάθος.

## (ε)

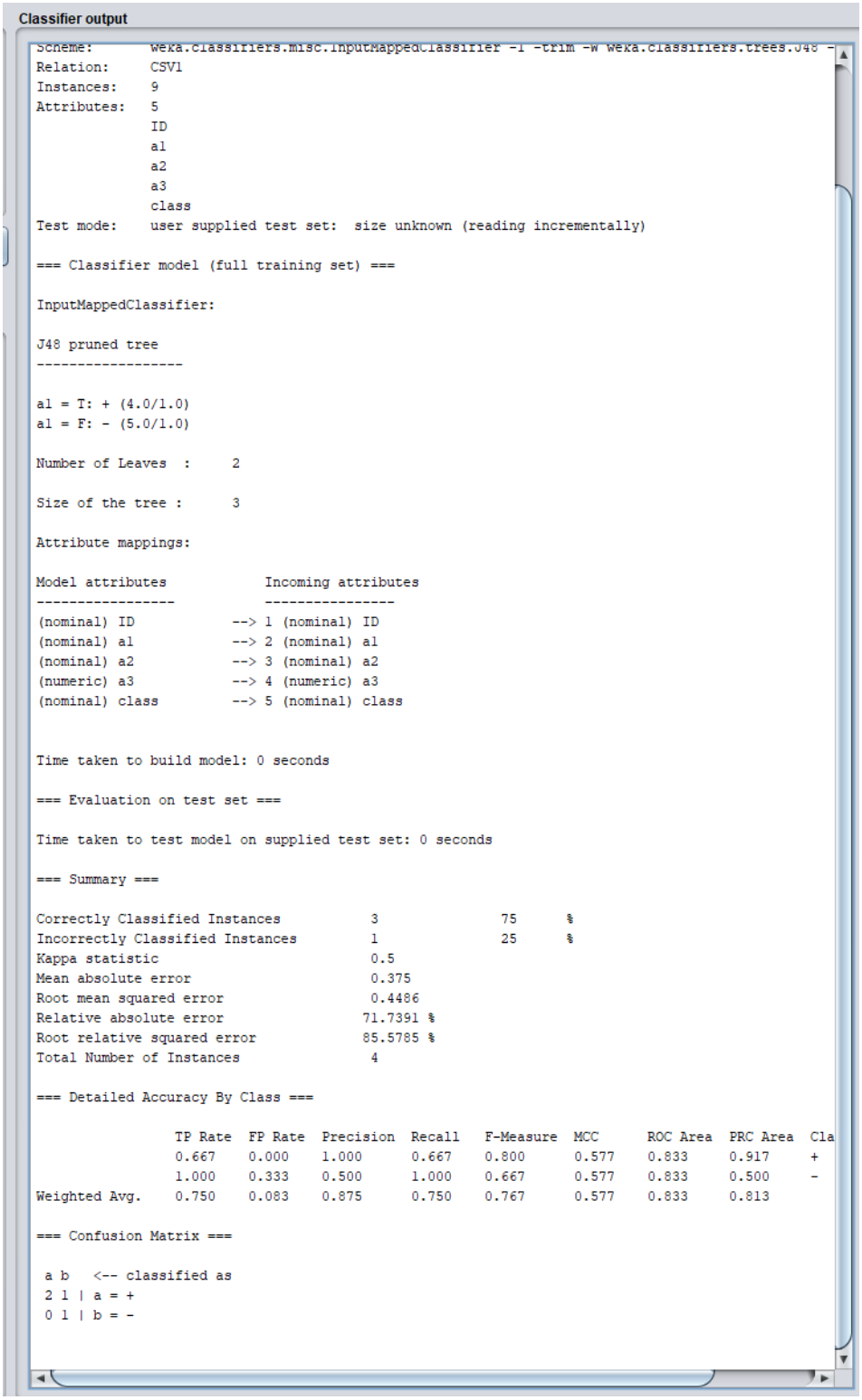
Το WEKA διάλεξε την μεταβλητή a1 όπως υπολογίστηκε και προηγουμένως με το χέρι.

Η εικόνα του δέντρου είναι η παρακάτω:



## (στ)

Με το WEKA πετυχαίνουμε 75% επιτυχία.



Classifier output

```
Scheme:       weka.classifiers.misc.InputMappedClassifier -I -trim -W weka.classifiers.trees.J48 --
Relation:     CSV1
Instances:    9
Attributes:   5
              ID
              a1
              a2
              a3
              class
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

InputMappedClassifier:

J48 pruned tree
------------------

a1 = T: + (4.0/1.0)
a1 = F: - (5.0/1.0)

Number of Leaves  :     2

Size of the tree :      3

Attribute mappings:

Model attributes            Incoming attributes
----------------            ----------------
(nominal) ID            --> 1 (nominal) ID
(nominal) a1            --> 2 (nominal) a1
(nominal) a2            --> 3 (nominal) a2
(numeric) a3            --> 4 (numeric) a3
(nominal) class         --> 5 (nominal) class


Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances           3               75      %
Incorrectly Classified Instances         1               25      %
Kappa statistic                          0.5
Mean absolute error                      0.375
Root mean squared error                  0.4486
Relative absolute error                 71.7391 %
Root relative squared error             85.5785 %
Total Number of Instances                4

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.667    0.000    1.000      0.667    0.800      0.577    0.833     0.917     +
                 1.000    0.333    0.500      1.000    0.667      0.577    0.833     0.500     -
Weighted Avg.    0.750    0.083    0.875      0.750    0.767      0.577    0.833     0.813

=== Confusion Matrix ===

 a b   <-- classified as
 2 1 | a = +
 0 1 | b = -
```

Με υπολογισμό στο χέρι λέμε πως πάλι είναι 75% γιατί είναι:

id10 (T = +) → Σωστό

id11 (T = +) → Σωστό

id12 (F = -) → Σωστό

id13(F = +) → Λάθος.

Οπότε καταλήγουμε ότι ¾ είναι σωστά και ¼ λάθος.