(a)

Gini(parent)=1-(4/9)^2-(5/9)^2=1-(16/81)-(25/81)=1-(41/81)=(81/81)-(41/81)
=40/81=0.494

a1:
Gini(T)=1-(3/4)^2-(1/4)^2=1-(9/16)-(1/16)=1-(10/16)=(16/16)-(10/16)=6/16=0.375
Gini(F)=1-(1/5)^2-(4/5)^2=1-(1/25)-(16/25)=1-(17/25)=(25/25)-(17/25)=8/25=0.32
Gini(a1_split)=(4/9 * 0.375) + (5/9 * 0.32)=(0.444 * 0.375)+(0.556 * 0.32)
=0.167+0.178=0.345

a2:
Gini(T)=1-(2/5)^2-(3/5)^2=1-(4/25)-(9/25)=1-(13/25)=(25/25)-(13/25)=12/25=0.48
Gini(F)=1-(2/4)^2-(2/4)^2=1-(4/16)-(4/16)=1-(8/16)=1-(1/2)=1/2=0.5
Gini(a2_split)=(5/9 * 0.48)+(4/9 * 0.5)=(0.556 * 0.48)+(0.444 *0.5)
=0.267+0.222=0.489

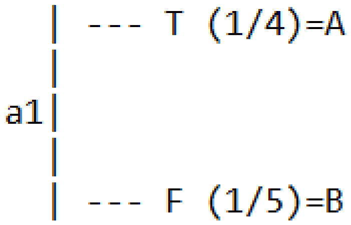
a3:

	a3													
	0.0		3.0		4.0		5.0			6.0		7.0		10.0
		1.5		3.5		4.5		5.5		6.5		8.5		
		<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	
A		1	3	1	3	2	2	2	2	3	1	4	0	
B		0	5	1	4	1	4	3	2	3	2	4	1	
Gini	-	0.417		0.492		0.444		0.489		0.481		0.445		-

Τα ακριανά σενάρια έχουν το πιο υψηλό Gini συνεπώς δεν υπάρχει λόγος να τα υπολογίσουμε.
1.5:
Gini(<=1.5)=1-(1/1)^2=0,
Gini(>1.5)=1-(3/8)^2-(5/8)^2=1-(34/64)=30/64=0.469,
Gini(1.5_split)=(8/9) * 0.469=0.417
3.5:
Gini(<=3.5)=1-(1/2)^2-(1/2)^2=0.5,
Gini(>3.5)=1-(3/7)^2-(4/7)^2=1-(25/49)=24/49=0.49,
Gini(3.5_split)=(2/9) * 0.5 + (7/9) * 0.49=0.492
4.5:
Gini(<=4.5)=1-(2/3)^2-(1/3)^2=1-(5/9)=4/9=0.444,
Gini(>4.5)=1-(2/6)^2-(4/6)^2=1-(20/36)=16/36=0.444,
Gini(4.5_split)=(3/9) * 0.444 + (6/9) * 0.444=0.444
5.5:
Gini(<=5.5)=1-(2/5)^2-(3/5)^2=1-(13/25)=12/25=0.48,
Gini(>5.5)=1-(2/4)^2-(2/4)^2=0.5,
Gini(5.5_split)=(5/9) * 0.48 + (4/9) * 0.5=0.489
6.5:
Gini(<=6.5)=1-(3/6)^2-(3/6)^2=0.5,
Gini(>6.5)=1-(1/3)^2-(2/3)^2=1-5/9=4/9=0.444,
Gini(6.5_split)=(6/9) * 0.5 + (3/9) * 0.444=0.481
8.5:
Gini(<=8.5)=1-(4/8)^2-(4/8)^2=0.5,
Gini(>8.5)=1-(1/1)^2=0,
Gini(8.5_split)=(8/9) * 0.5=0.445

(b)
GAIN(a1_split)=0.494-0.345=0.149
GAIN(a2_split)=0.494-0.489=0.005
GAIN(a3_split)=0.494-0.417=0.077

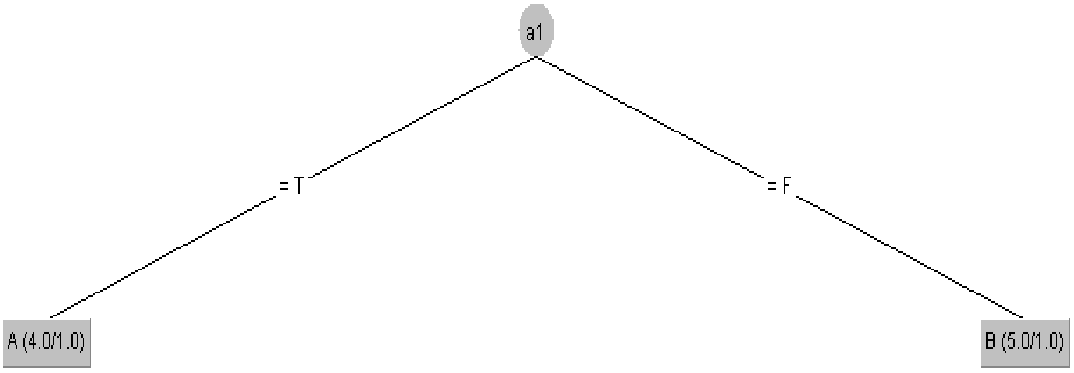
(c)
Το χαρακτηριστικό που θα έχει το μικρότερο Gini(split) θα έχει και το υψηλότερο Gain συνεπώς και θα αποτελεί τη ρίζα του δένδρου. Το χαρακτηριστικό αυτό είναι το a1.



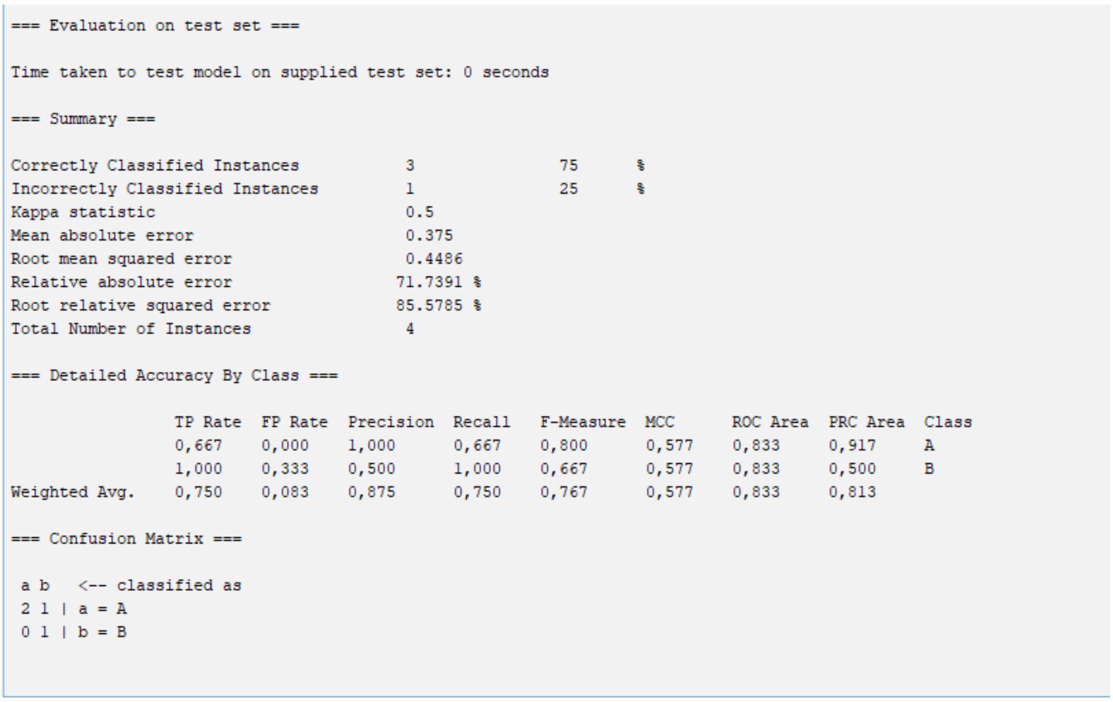
```
   | --- T (1/4)=A
   |
 a1|
   |
   | --- F (1/5)=B
```

Συνεπώς κατηγοριοποιούνται 2 εγγραφές λάθος, άρα πετυχαίνουμε 7/9=0.778=77.8% ακρίβεια.

(d) Ο αλγόριθμος J48 επιλέγει ως ρίζα του δένδρου το χαρακτηριστικό a1 και πετυχαίνει ακρίβεια 77.7778 % στο training dataset.

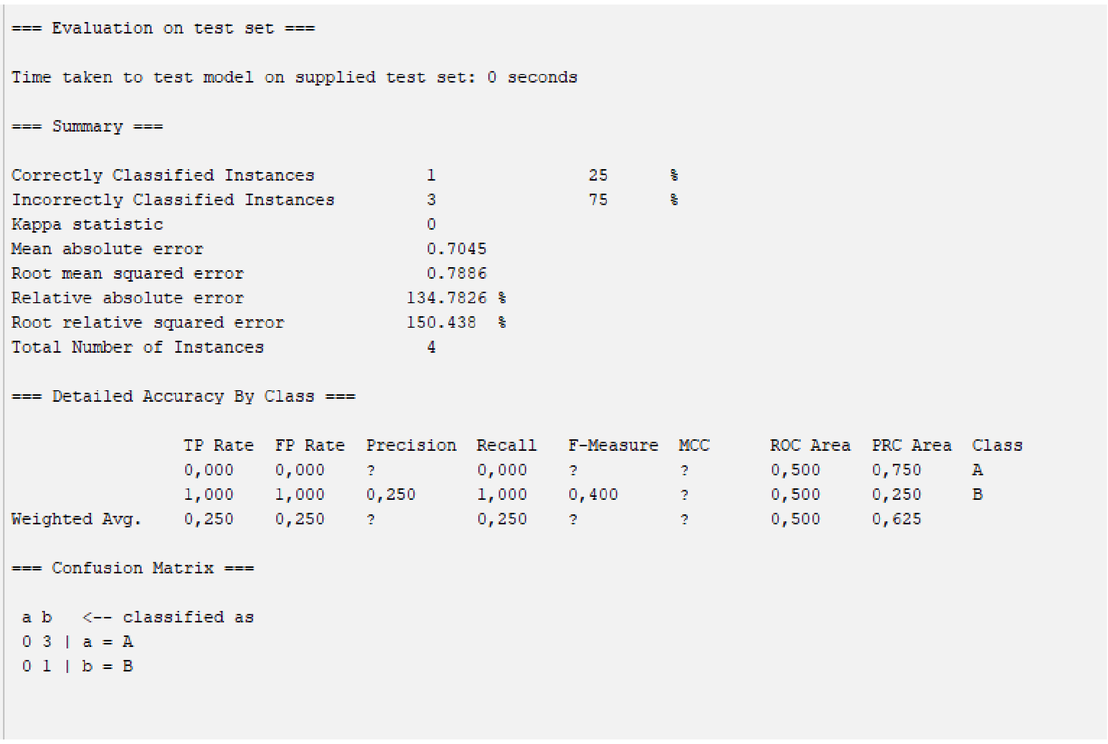


(e)
Ο αλγόριθμος J48 πετυχαίνει 75% ακρίβεια.



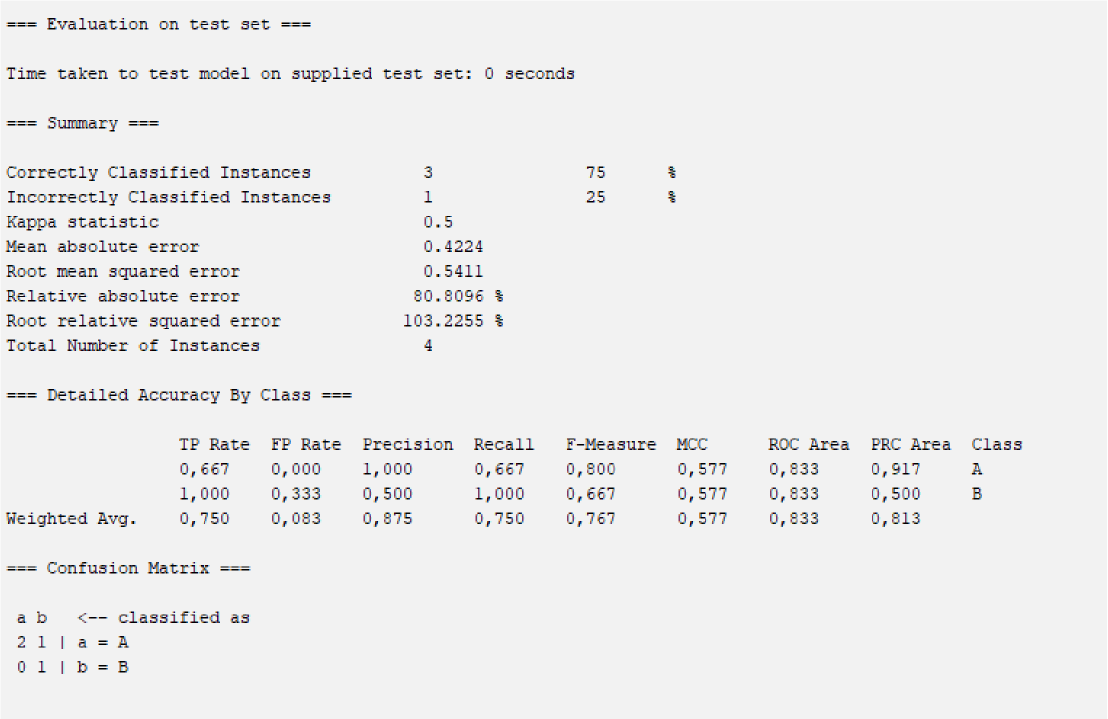
```
=== Evaluation on test set ===

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  Class
                 0,667    0,000    1,000      0,667    0,800      0,577    0,833     0,917     A
                 1,000    0,333    0,500      1,000    0,667      0,577    0,833     0,500     B
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 = A
 0 1 | b = B
```

(f)
i)kNN με k=1: Πετυχαίνει 25% ακρίβεια.



```
=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances           1               25      %
Incorrectly Classified Instances         3               75      %
Kappa statistic                          0
Mean absolute error                      0.7045
Root mean squared error                  0.7886
Relative absolute error                134.7826 %
Root relative squared error            150.438  %
Total Number of Instances                4

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,750     A
                 1,000    1,000    0,250      1,000    0,400      ?        0,500     0,250     B
Weighted Avg.    0,250    0,250    ?          0,250    ?          ?        0,500     0,625

=== Confusion Matrix ===

 a b   <-- classified as
 0 3 | a = A
 0 1 | b = B
```

ii)kNN με k=3: Πετυχαίνει 75% ακρίβεια.



```
=== 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                      0.4224
Root mean squared error                  0.5411
Relative absolute error                 80.8096 %
Root relative squared error            103.2255 %
Total Number of Instances                4

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,667    0,000    1,000      0,667    0,800      0,577    0,833     0,917     A
                 1,000    0,333    0,500      1,000    0,667      0,577    0,833     0,500     B
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 = A
 0 1 | b = B
```