Εργασία 1

(α) GINI = 1 - (4/9)^2 - (5/9)^2 = 1 - 16/81 - 25/81 = 1 - 41/81 = 40/81 = 0.493

(β)
ID
GINI(id1) = 1 - (1/1)^2 - (0/1)^2) = 0
ομοίως: GINI(id2) = GINI(id3) = ... = GINI(id9) = 0
GINI(ID_split) = 1/9 * 0 + ... + 1/9 * 0 = 0
GAIN(ID_split) = 40/81 - 0 = 40/81 = 0.493

a1
GINI(T) = 1 - (3/4)^2 - (1/4)^2 = 1 - 9/16 - 1/16 = 1 - 10/16 = 6/16 = 3/8
GINI(F) = 1 - (1/5)^2 - (4/5)^2 = 1 - 1/25 - 16/25 = 1 - 17/25 = 8/25
GINI(a1_split) = 4/9 * 3/8 + 5/9 * 8/25 = (4*3)/(9*8) + (5*8)/(9*25) = 1/6 + 8/45 = 15/90 + 16/90 = 31/90 = 0.344
GAIN(a1_split) = 0.493 - 0.344 = 0.149

a2
GINI(T) = 1 - (2/5)^2 - (3/5)^2 = 1 - 4/25 - 9/25 = 1 - 13/25 = 12/25
GINI(F) = 1 - (2/4)^2 - (2/4)^2 = 1 - 1/4 - 1/4 = 1 - 1/2 = 1/2
GINI(a2_split) = 5/9 * 12/25 + 4/9 * 1/2 = (5*12)/(9*25) + (4*1)/(9*2) = 4/15 + 2/9 = 12/45 + 10/45 = 22/45 = 0.488
GAIN(a2_split) = 0.493 - 0.488 = 0.005

Δεν επιλέγουμε το ID καθώς δεν έχει προβλεπτική ικανότητα (κάθε ID είναι μοναδικό, συνεπώς, τα ID των νέων instances δεν θα μπορούν να προβλέψουν την τιμή της κλάσης).

Άρα το a1 έχει το μεγαλύτερο GAIN.

(γ)

class	+		-		+		--		+		+-		-	
sorted	1		3		4		5		6		7		8	
split	0		2		3.5		4.5		5.5		6.5		7.5		8.5
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
-	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0
GINI	0.493		0.416		0.492		0.444		0.488		0.481		0.444		0.493	

0 <= 1 - 0 - 0 = 1
   > 1 - 4/9^2 - 5/9^2 = 40/81 = 0.493
   GINI = 0/9 * 1 + 9/9 * 0.493 = 0.493
2 <= 1 - 1/1^2 - 0/1^2 = 1 - 1 = 0
   > 1 - 3/8^2 - 5/8^2 = 1 - 9/64 - 25/64 = 30/64 = 15/32
   GINI = 8/9 * 15/32 = (8*15)/(9*32) = 5/12 = 0.416

```
3.5  <= 1 - 1/2^2 - 1/2^2 = 1 - 1/2 = 1/2
     >  1 - 3/7^2 - 4/7^2 = 1 - 25/49 = 24/49
     GINI = 2/9 * 1/2 + 7/9 * 24/49 = 1/9 + 24/63 = 7/63 + 24/63 = 31/63 = 0.492
4.5  <= 1 - 2/3^2 - 1/3^2 = 1 - 5/9 = 4/9
     >  1 - 2/6^2 - 4/6^2 = 1 - 5/9 = 4/9
     GINI = 3/9 * 4/9 + 6/9 * 4/9 = 36/81 = 12/27 = 0.444
5.5  <= 1 - 2/5^2 - 3/5^2 = 1 - 13/25 = 12/25
     >  1 - 2/4^2 - 2/4^2 = 1 - 1/2 = 1/2
     GINI = 5/9 * 12/25 + 4/9 * 1/2 = 12/45 + 2/9 = 22/45 = 0.488
6.5  <= 1 - 3/6^2 - 3/6^2 = 1 - 1/2 = 1/2
     >  1 - 1/3^2 - 2/3^2 = 1 - 5/9 = 4/9
     GINI = 6/9 * 1/2 + 3/9 * 4/9 = 3/9 + 12/81 = 39/81 = 0.481
7.5  <= 1 - 4/8^2 - 4/8^2 = 1 - 1/2 = 1/2
     >  1 - 0/1^2 - 1/1^2 = 1 - 1 = 0
     GINI = 8/9 * 1/2 + 1/9 * 0 = 4/9 = 0.444
8.5 Όμοια με το 0.
```
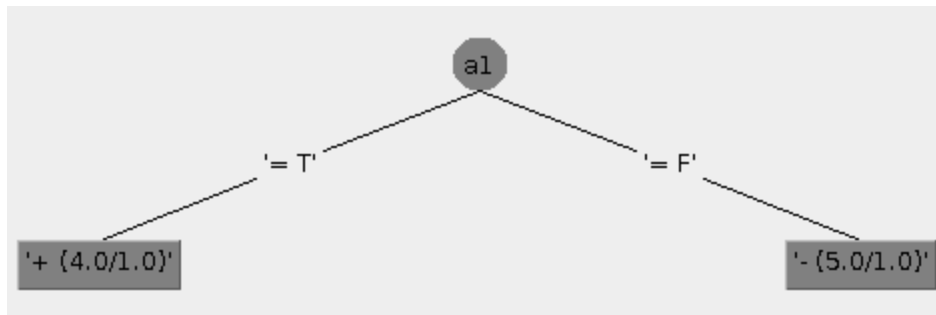
Best split 2: GAIN = 0.493 - 0.416 = 0.077

(δ) Η ρίζα θα είναι το a1 (μεγαλύτερο GAIN)
Ακρίβεια στο training set = 7/9 = 77.77% (χάνει ένα + και ένα -)

(ε)
Το WEKA διάλεξε την a1.

(στ) 75%