**ΕΡΓΑΣΙΑ 1**

**ΣΤΟ ΜΑΘΗΜΑ: ΣΥΣΤΗΜΑΤΑ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ ΑΠΟ ΒΔ**

Α) GINI(parent) = 1 – [(4/9)^2+(5/9)^2] = 0,49

Β) ID

GINI(id1) = 0

GINI(id2) = 0

:

:

GINI(id9) = 0

GINI(id\_split) = 0

a1

GINI(T) = 1 – (3/4)^2 – (1/4)^2 = 0,375

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

GINI(a1\_split) = (4/9\*0,357) + (5/9\*0,32) = 0,17 κοιταζω την στηλη α1 και λεω ότι 4/9 είναι Τ οποτε θα πολλαπλασιασω με το gini(T). Ομοια 5/9 είναι F αρα \* gini(F).

a2

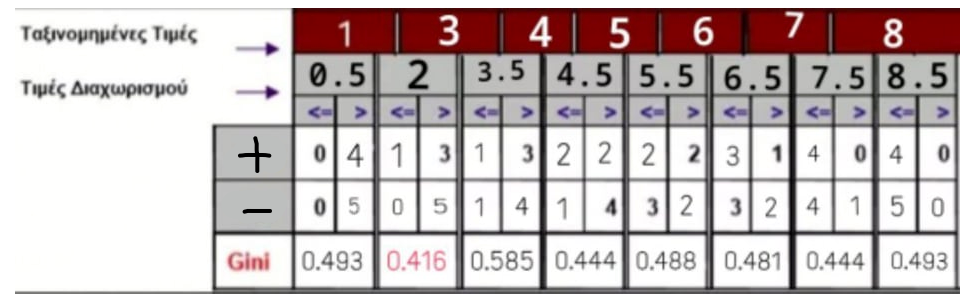
GINI(T) = 1 – (2/5)^2 – (3/5)^2 = 0,52

GINI(F) = 1 – (2/4)^2 – (2/4)^2 = 0,5

GINI(a2\_split) = (5/9\*0,52) + (4/9\*0,5) = 0,22

**ΣΧΟΛΙΟ**: Εύκολα προκύπτει ότι το ID έχει το μεγαλύτερο GAIN (GAIN(id\_split) = 0,49 – 0 = 0,49) όμως δεν θα το επιλέξουμε ως ρίζα γιατί προκύπτουν εννέα διασπάσεις και δεν μας βολεύουν σε καμία περίπτωση.

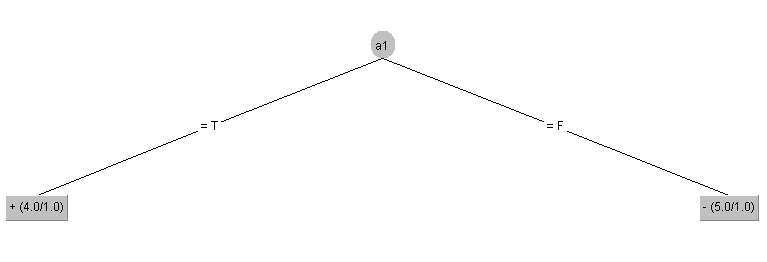
Γ) Για την διάσπαση του α3, φτιάχνουμε τον παρακάτω πίνακα γιατί είναι συνεχής μεταβλητή και πρέπει να δούμε σε πιο σημείο θα την διασπάσουμε. Επιλεγούμε το μικρότερο GINI, δηλαδή τον αριθμό 0.416, οπότε θα έχουμε διάσπαση στον αριθμό 2.



**ΣΗΜΕΙΩΣΗ**: Παρέλειψα τις πράξεις για εξοικονόμηση χρόνου μιας και είναι παρόμοιας λογικής με τα προαναφερθέντα ερωτήματα.

Δ) Η ρίζα του δέντρου μας θα είναι η **α1** με παιδιά **Τ** και **F,** γιατί έχει το μικρότερο GINI και άρα το μεγαλύτερο κέρδος GAIN(a1\_split) = 0,49 – 0,17 = 0,32. Δηλαδή, έχουμε 32% εγγράφων που κατηγοριοποιούνται σωστά.

Ε) Αφότου τρέξαμε το Weka Explorer πρόεκυψε το παρακάτω δέντρο με ρίζα το attribute a1 και παιδιά τα T και F (όπως ακριβώς και με την αρχική μας εκτίμηση).



ΣΤ) Για το συγκεκριμένο testing dataset, το παραπάνω δέντρο πετυχαίνει ακρίβεια 75% (3 σωστά και 1 λάθος).

**ΑΠΟ ΤΟΝ ΦΟΙΤΗΤΗ: Τσιαούση Σταύρο (dai17173)**