Εργασία 1 – Δέντρα Απόφασης (με το χέρι και με το WEKA)

Θεωρείστε το παρακάτω training dataset για ένα δυαδικό πρόβλημα κατηγοριοποίησης.

ID	a1	a2	a3	class
id1	T	T	1.0	+
id2	T	T	6.0	+
id3	T	F	5.0	−
id4	F	F	4.0	+
id5	F	T	7.0	−
id6	F	T	3.0	−
id7	F	F	8.0	−
id8	T	F	7.0	+
id9	F	T	5.0	−

(α) Υπολογίστε το GINI του συνόλου των εγγραφών.

(β) Υπολογίστε το GAIN των διασπάσεων ως προς ID, a1, και a2. Γιατί, ενώ το ID έχει το μεγαλύτερο GAIN, δεν θα πρέπει να το επιλέξετε ως ρίζα του δέντρου;

(γ) Υπολογίστε το GAIN για κάθε δυνατή διάσπαση της συνεχούς μεταβλητής a3. Σε ποιο σημείο προκύπτει η καλύτερη διάσπαση;

(δ) Ποια θα είναι τελικά η ρίζα του δέντρου; Τί ακρίβεια πετυχαίνει το δέντρο απόφασης πάνω στο training dataset; (με άλλα λόγια, ποιο είναι το ποσοστό των εγγραφών που κατηγοριοποιούνται σωστά;)

(ε) Δημιουργήστε ένα csv αρχείο με τα περιεχόμενα του παραπάνω πίνακα και τρέξτε τον κατηγοριοποιητή J48 στο WEKA με τις default ρυθμίσεις και επιλέγοντας “Use training set” στο “Test options”. Ποια μεταβλητή διάλεξε το WEKA ως ρίζα του δέντρου; Δώστε την εικόνα του δέντρου από το WEKA. **ΠΡΟΣΟΧΗ: Σιγουρευτείτε ότι στο αρχείο csv, οι τιμές της μεταβλητής class είναι μέσα σε εισαγωγικά. Το αρχείο μπορεί να είναι comma ή tab delimited. Μπορείτε να το δημιουργήσετε σε κάποια εφαρμογή τύπου Notepad (ή στο MS Excel ή το Libreoffice Calc και να το αποθηκεύσετε ως csv).**

(στ) Πόση ακρίβεια πετυχαίνει το δέντρο απόφασης στο παρακάτω testing dataset;

ID	a1	a2	a3	class
id10	T	F	1.0	+
id11	T	F	3.0	+
id12	F	T	2.0	−
id13	F	T	8.0	+

Θα καταθέσετε ένα αρχείο κειμένου (doc/odt) με τις απαντήσεις σας.