Δέντρα Απόφασης

Έχουμε το παρακάτω training dataset που αφορά τον τρόπο που μετακινούνται εργαζόμενοι στο χώρο εργασίας τους. Για κάθε εργαζόμενο καταγράφεται το φύλο (gender), το αν έχει αυτοκίνητο (car), το κόστος μετακίνησης που είναι διατεθειμένος/η να πληρώσει σε ευρώ/χλμ (travel_cost) και το εισόδημα (income). Θέλουμε να δούμε αν μπορούν να χρησιμοποιηθούν τα γνωρίσματα αυτά για την κατηγοριοποίηση των μετακινούμενων σε σχέση με το μέσο μεταφοράς που επιλέγουν. Το dataset που σας δίνεται έχει πάρει τιμές στο γνώρισμα κλάσης μεταφορικό μέσο (transportation_mode) μετά από συμπλήρωση ερωτηματολογίων.

Training dataset (αρχείο **transportation_train.arff**)

gender	**car**	**travel_cost**	**income**	**transportation_mode**
m	n	cheap	low	bus
m	y	cheap	medium	bus
f	y	cheap	medium	train
m	n	cheap	low	bus
m	y	cheap	medium	bus
m	n	cheap	medium	train
f	y	standard	medium	train
f	y	expensive	high	car
m	y	expensive	medium	car
f	y	expensive	high	car

(Α) Εκτελέστε με το χέρι τον αλγόριθμο του Hunt με δείκτη μη-καθαρότητας το Gini index:

1. Ποιο γνώρισμα θα αποτελέσει τη ρίζα του δέντρου απόφασης; Εξηγείστε αναλυτικά (μέσω υπολογισμού του Gini index για κάθε σενάριο) γιατί επιλέγετε το συγκεκριμένο γνώρισμα. Για τα κατηγορικά γνωρίσματα travel_cost και income δοκιμάστε μόνο την πολλαπλή διάσπαση (δηλαδή σε τρία παιδιά).
2. Πόσες εγγραφές του training dataset κατηγοριοποιούνται λάθος αν σταματήσουμε τον αλγόριθμο μετά την κατασκευή της ρίζας;

(Β) Επαληθεύστε το συμπέρασμά σας στο WEKA:

1. Ποιο γνώρισμα επιλέγει ο αλγόριθμος J48 ως ρίζα του δέντρου απόφασης;
2. Δώστε το πλήρες δέντρο που κατασκευάζει το WEKA με τις default τιμές στις παραμέτρους του αλγορίθμου.
3. Δώστε το πλήρες δέντρο που κατασκευάζει το WEKA αν αλλάξετε τη τιμή της παραμέτρου minNumObj σε 1.

(Γ) Έστω ότι έχουμε το παρακάτω test dataset (αρχείο **transportation_test.arff**):

gender	**car**	**travel_cost**	**income**	**transportation_mode**
f	y	cheap	high	train
f	n	standard	medium	bus
f	n	expensive	medium	car
f	n	expensive	high	car

Τι ακρίβεια πετυχαίνει το δέντρο με ένα κόμβο (μόνο τη ρίζα) που δημιουργήσατε με το χέρι και τι τα δέντρα απόφασης του WEKA των περιπτώσεων Β2 και Β3;