Εξόρυξη Γνώσης από Δεδομένα - Εργασία 1 (Classification)

(Α) Δέντρα Απόφασης με το χέρι

Έχουμε το παρακάτω training dataset που έχει να κάνει με χαρακτηριστικά μανιταριών (χρώμα, ύψος, ρίγες, υφή). Θέλουμε να δούμε αν μπορούν να χρησιμοποιηθούν τα χαρακτηριστικά αυτά για την κατηγοριοποίηση των μανιταριών ως δηλητηριωδών ή μη-δηλητηριωδών. Το dataset που σας δίνεται έχει πάρει τιμές στο χαρακτηριστικό Poisonous (class variable) από ειδικούς βοτανολόγους.

Colour	Height	Stripes	Texture	Poisonous
Purple	Tall	Yes	Rough	Yes
Purple	Tall	Yes	Smooth	Yes
Red	Short	Yes	Hairy	No
Blue	Short	No	Smooth	No
Blue	Short	Yes	Hairy	Yes
Red	Tall	No	Hairy	No
Blue	Tall	Yes	Smooth	Yes
Blue	Short	Yes	Smooth	Yes
Blue	Tall	No	Hairy	No
Blue	Short	Yes	Rough	Yes
Red	Short	No	Smooth	No
Purple	Short	No	Hairy	Yes
Red	Tall	Yes	Rough	No
Purple	Tall	Yes	Hairy	Yes
Purple	Tall	No	Rough	No
Purple	Tall	No	Smooth	No

(i) Ποιο χαρακτηριστικό θα αποτελέσει τη ρίζα του δέντρου απόφασης σύμφωνα με τον αλγόριθμο του Hunt; Εξηγείστε αναλυτικά (μέσω υπολογισμού του gini index για κάθε σενάριο) γιατί επιλέγετε το συγκεκριμένο χαρακτηριστικό. **Για να μην κάνετε πολλούς υπολογισμούς, για τις categorical variables δοκιμάστε μόνο multi-way splits (δηλαδή για τις Colour και Texture διάσπαση σε 3 παιδιά)**. Πόσες εγγραφές του training dataset κατηγοριοποιούνται λάθος αν σταματήσουμε το αλγόριθμο μετά την κατασκευή της ρίζας;

(ii) Επαληθεύστε το συμπέρασμά σας στο WEKA επιλέγοντας “Use training set” στο “Test options” (δίνεται το mushrooms.csv). Ποιο χαρακτηριστικό επιλέγει ο αλγόριθμος J48 ως ρίζα του δέντρου απόφασης; Δώστε το πλήρες δέντρο που κατασκευάζει το WEKA με τις default τιμές στις παραμέτρους του αλγορίθμου.

(iii) Έστω ότι έχουμε το παρακάτω test dataset:

Colour	Height	Stripes	Texture	Poisonous
Purple	Tall	Yes	Rough	Yes
Red	Tall	Yes	Smooth	No
Red	Short	No	Hairy	Yes
Blue	Short	No	Smooth	No

Τι ακρίβεια πετυχαίνει το δέντρο με ένα κόμβο (μόνο τη ρίζα) που δημιουργήσατε με το χέρι και τι το δέντρο απόφασης του WEKA; (η απάντηση προκύπτει εξετάζοντας μια-μια τις εγγραφές του test dataset με το μάτι πάνω στο κάθε δέντρο)

(Β) Μελέτη περίπτωσης με το WEKA

Το αρχείο δεδομένων **car.arff** περιλαμβάνει την αξιολόγηση 1728 αυτοκινήτων και την κατάταξή τους σε τέσσερεις κατηγορίες (στήλη class): μη αποδεκτό (unacc), αποδεκτό (acc), καλό (good) και πολύ καλό (vgood). Επίσης, περιλαμβάνει τα ακόλουθα χαρακτηριστικά των αυτοκινήτων:

- buying: τιμή αγοράς (χιλιάδες ευρώ)
- maint: έξοδα συντήρησης (χιλιάδες ευρώ)
- doors: πλήθος θυρών
- persons: πλήθος ατόμων
- lug_boot: μέγεθος αποθηκευτικού χώρου (λίτρα)
- safety: επίπεδο ασφάλειας (1-χαμηλό, 2-μέτριο, 3-υψηλό)

Χρησιμοποιώντας 10-fold cross validation, πειραματιστείτε με διαφορετικές τιμές για την παράμετρο minNumObj του κατηγοριοποιητή J48 (decision tree).

Επιλέξτε το καλύτερο μοντέλο και αναφέρετε την ακρίβεια και τον πίνακα σύγχυσης. Σχολιάστε ποιες κατηγορίες αυτοκινήτων δεν προβλέπονται ικανοποιητικά. Χρησιμοποιώντας το δέντρο απόφασης, να καταγράψετε έναν κανόνα που θεωρείτε ισχυρό, για κάθε μία από τις τέσσερις κατηγορίες αυτοκινήτων. Ένας κανόνας είναι η διαδρομή από τη ρίζα ως ένα φύλλο και περιγράφει τα χαρακτηριστικά ενός μεγάλου αριθμού στιγμιοτύπων που ανήκουν στην κατηγορία που πλειοψηφεί στο φύλλο.