Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Δέντρα Αποφάσεων

Τι είναι ένα δέντρο αποφάσεων (ΔΑ);

- Τα ΔΑ αποτελούν τα δημοφιλέστερα μοντέλα κατηγοριοποίησης.
- Τα ΔΑ αναπαριστούν ένα μοντέλο πρόβλεψης το οποίο χτίζεται μέσα από μια σειρά Boolean αποφάσεων του τύπου ΝΑΙ/ΟΧΙ, μεγαλύτερο/μικρότερο,….
- Ένα ΔΑ αποτελείται από ενδιάμεσους κόμβους και φύλλα. Τα φύλλα είναι οι κόμβοι στο τελευταίο επίπεδο και δεν έχουν απογόνους.

Δέντρο Αποφάσεων

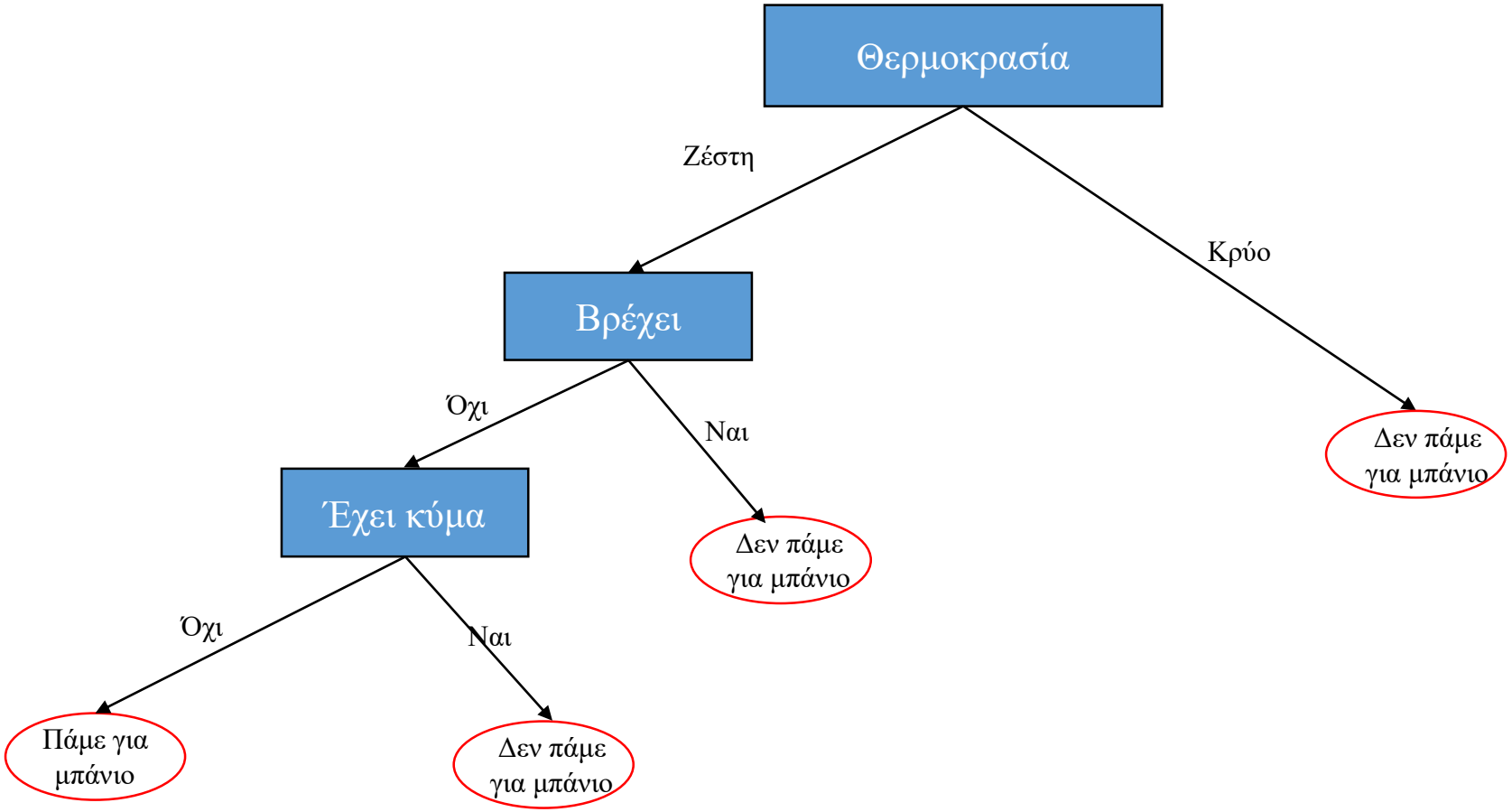
- Κάθε εγγραφή περιέχει ένα σύνολο από γνωρίσματα/χαρακτηριστικά (attributes)
- Ένα από τα γνωρίσματα είναι η κλάση/κατηγορία (class)

attributes

Α	Θερμοκρασία	Έχει κύμα	Βροχή	Μπάνιο
1	Ζέστη	Όχι	Όχι	Ναι
2	Ζέστη	Ναι	Ναι	Όχι
3	Ζέστη	Όχι	Όχι	Ναι
4	Κρύο	Όχι	Όχι	Όχι
5	Κρύο	Ναι	Ναι	Όχι
6	Κρύο	Όχι	Όχι	Όχι

class

Τι είναι ένα δέντρο αποφάσεων (ΔΑ);



Δέντρο Αποφάσεων

- Στο προηγούμενο ΔΑ, μέσα από μια σειρά ερωτήσεων-απαντήσεων χτίζουμε τα αντίστοιχα κλαδιά.
  - Έχει ζέστη;
  - Βρέχει;
  - Έχει κύμα η θάλασσα;

- Απαντώντας στις παραπάνω ερωτήσεις (ΝΑΙ/ΟΧΙ), οδηγούμαστε στο συμπέρασμα αν μπορούμε να κάνουμε μπάνιο ή όχι.

Δένδρα Απόφασης

- Τα βασικά στάδια ενός αλγορίθμου ΔΑ είναι τα παρακάτω:
  - Επιλογή του κατάλληλου χαρακτηριστικού προκειμένου να αποτελέσει την κορυφή του δέντρου.
  - Επανάληψη της διαδικασίας στο επόμενο επίπεδο για κάθε παιδί του κόμβου που προκύπτει.
  - Η διαδικασία τερματίζει όταν:
    - Όλες οι εγγραφές καταλήγουν στην ίδια παράμετρο.
    - Δεν υπάρχουν υποσύνολα που περιέχουν περισσότερες από μία κατηγορίες
    - Έχουν χρησιμοποιηθεί όλα τα χαρακτηριστικά

Δένδρα Απόφασης

- Οι περισσότεροι αλγόριθμοι δημιουργίας ΔΑ έχουν την ίδια λογική και διαφοροποιούνται μόνο στον τρόπο επιλογής των υποψήφιων κόμβων.
- Οι γνωστοί αλγόριθμοι είναι οι ID3, CART και C4.5.

Δένδρα Απόφασης - ID3 Heuristic

Ο αλγόριθμος κατασκευής δένδρου απόφασης ID3 έχει τα εξής βήματα:

1. Για κάθε μεταβλητή υπολογίζουμε το πληροφοριακό κέρδος.
2. Η μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος ορίζεται ως ρίζα του δένδρου.
3. Δημιουργούμε τόσα κλαδιά όσες είναι και οι διακριτές τιμές του κόμβου που επιλέχθηκε.
4. Χωρίζουμε το σύνολο δεδομένων σε τόσα υποσύνολα όσα και οι διακριτές τιμές της μεταβλητής που επιλέχθηκαν.
5. Επιλέγουμε μια τιμή από το υποσύνολο έτσι όπως αυτό διαμορφώθηκε στο βήμα 4. Αν ο κόμβος που επιλέγουμε αντιστοιχεί σε μόνο μια τιμή τότε επιλέγουμε τον επόμενο κόμβο. Αν όμως αντιστοιχεί σε περισσότερες τιμές τότε υπολογίζουμε το πληροφοριακό κέρδος των υπόλοιπων μεταβλητών για το συγκεκριμένο υποσύνολο και επιλέγεται ο κόμβος με το μεγαλύτερο πληροφοριακό κέρδος.
6. Επαναλαμβάνουμε από το βήμα 3, μέχρι να μην μπορούν να δημιουργηθούν νέα φύλλα.

ID3 Heuristic

Πληροφοριακό κέρδος G(S, A) = E(S) − I (S, A) όπου

$$I(S,A)=\sum_{j=1}^{K}\frac{|Sj|}{|S|}E(Sj)$$

S= όλα τα δείγματα, Sj=τα δείγματα με τιμή j για το χαρακτηριστικό A, |S|=το πλήθος τους, και E(Sj)= είναι η εντροπία για το υποσύνολο δειγμάτων με τιμή j για το χαρακτηριστικό A.

Η εντροπία E για ένα σύνολο δεδομένων (S) υπολογίζεται από τον παρακάτω τύπο

$$E(S) = -\sum_{i=1}^{k} P_i log_2(Pi)$$

όπου pi είναι η πιθανότητα της κλάσης i στο S.

Εντροπία

- Η εντροπία ελαχιστοποιείται όταν οι πιθανές καταστάσεις στις οποίες μπορεί να περιέλθουν τα αντικείμενά μας είναι ίδιες.
- Μεγιστοποιείται όταν υπάρχουν ίσες πιθανότητες για τις διαφορετικές καταστάσεις στις οποίες μπορεί να βρεθούν τα αντικείμενά μας.

C4.5

Ο αλγόριθμος C4.5 αποτελεί επέκταση του ID3 και έρχεται με σκοπό να διορθώσει τα παρακάτω μειονεκτήματα του ID3

- Προβλήματα όταν λείπουν πολλά δεδομένα
- Μεγάλα σε όγκο δένδρα με πολλά κλαδιά χωρίς όμως να εμφανίζουν κάποιο ενδιαφέρον
- Εφαρμόζεται μόνο για διακριτά δεδομένα

Overfitting

Ένα από τα σημαντικότερα προβλήματα που προκύπτουν κατά την δημιουργία ενός ΔΑ είναι ο μεγάλος αριθμός κλαδιών που έχει σαν αποτέλεσμα τον μεγάλο όγκο δεδομένων και ταυτόχρονα την αύξηση των απαιτήσεων για προσπέλαση των δεδομένων.

- Ένας τρόπος για να ξεπεράσουμε αυτό το πρόβλημα είναι το «κλάδεμα» εκείνων των κλαδιών τα οποία δεν δίνουν κάποια γνώση
- Υπάρχουν 2 τεχνικές κλαδέματος
  - Pre-Pruning: Να μην επιτρέπουμε σε συγκεκριμένα κλαδιά να εξελίσσονται
  - Post-Pruning: Το κλάδεμα να γίνεται όταν έχει ολοκληρωθεί η δημιουργία των κλαδιών.

Prepruning

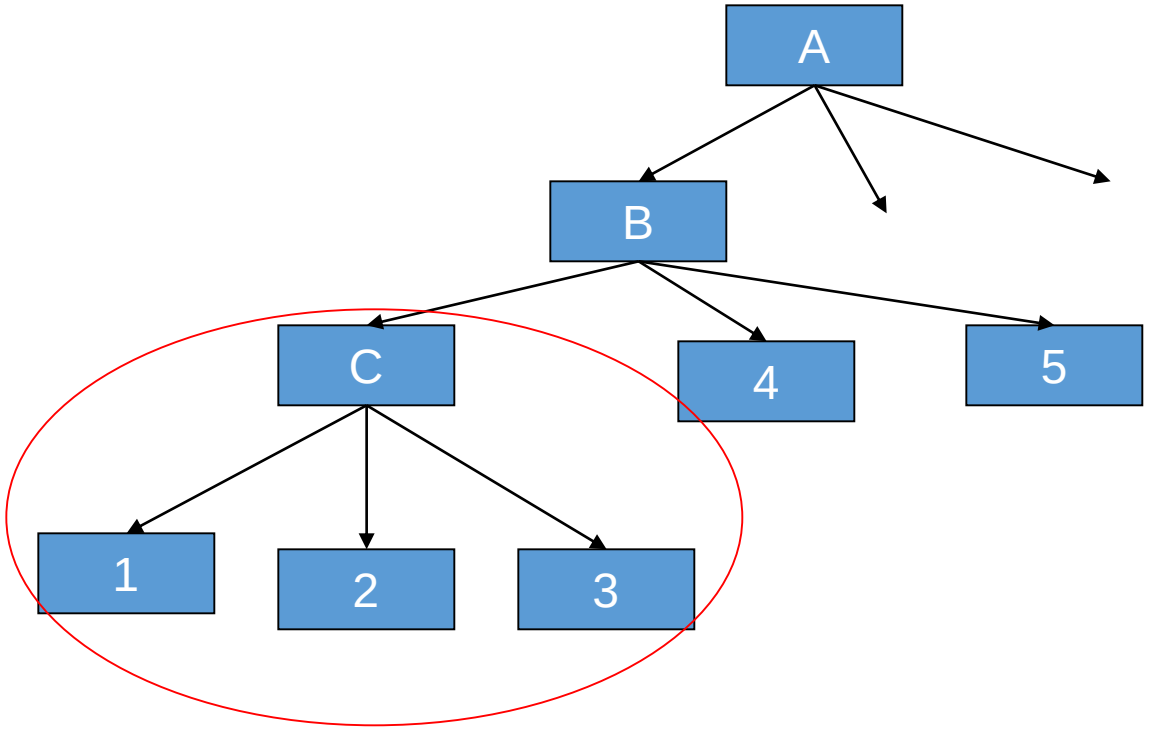
- Σε αυτή την μέθοδο επιλέγουμε κατά την διάρκεια του χτισίματος του δέντρου πότε θα πρέπει να σταματήσουμε να προσθέτουμε νέα αντικείμενα/κλαδιά.
- Αυτή η τεχνική ενδεχομένως να δημιουργήσει πρόβλημα γιατί υπάρχουν αντικείμενα που από μόνα τους δεν δίνουν κάποια πληροφορία αλλά σε συνδυασμό με κάποια άλλα είναι πολύ χρήσιμα.

Postpruning

- Σε αυτή την τεχνική πρώτα χτίζεται όλο το δένδρο και στην συνέχεια κλαδεύονται τα κλαδιά. Σε αυτή την περίπτωση υπάρχουν 2 τεχνικές.
  - Αντικατάσταση του υποδέντρου
  - Ανέβασμα υποδέντρου ένα επίπεδο

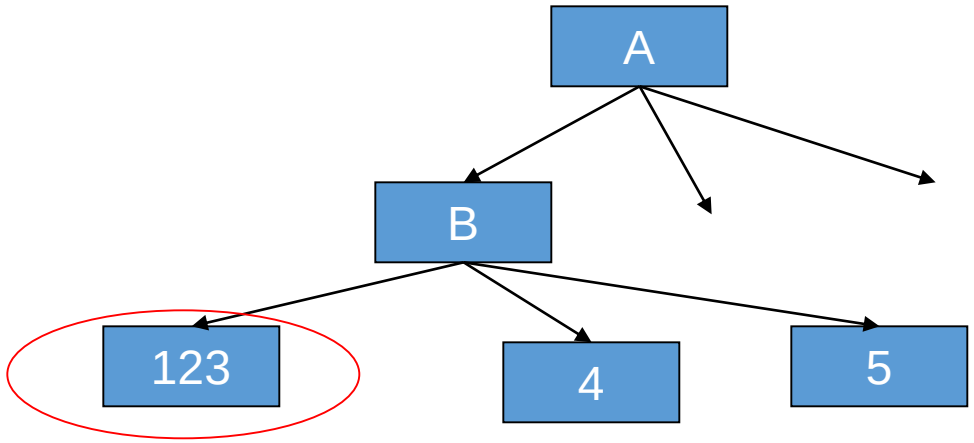
Postpruning - Αντικατάσταση υποδέντρου

- Όλο το υποδέντρο αντικαθίσταται από ένα φύλο.



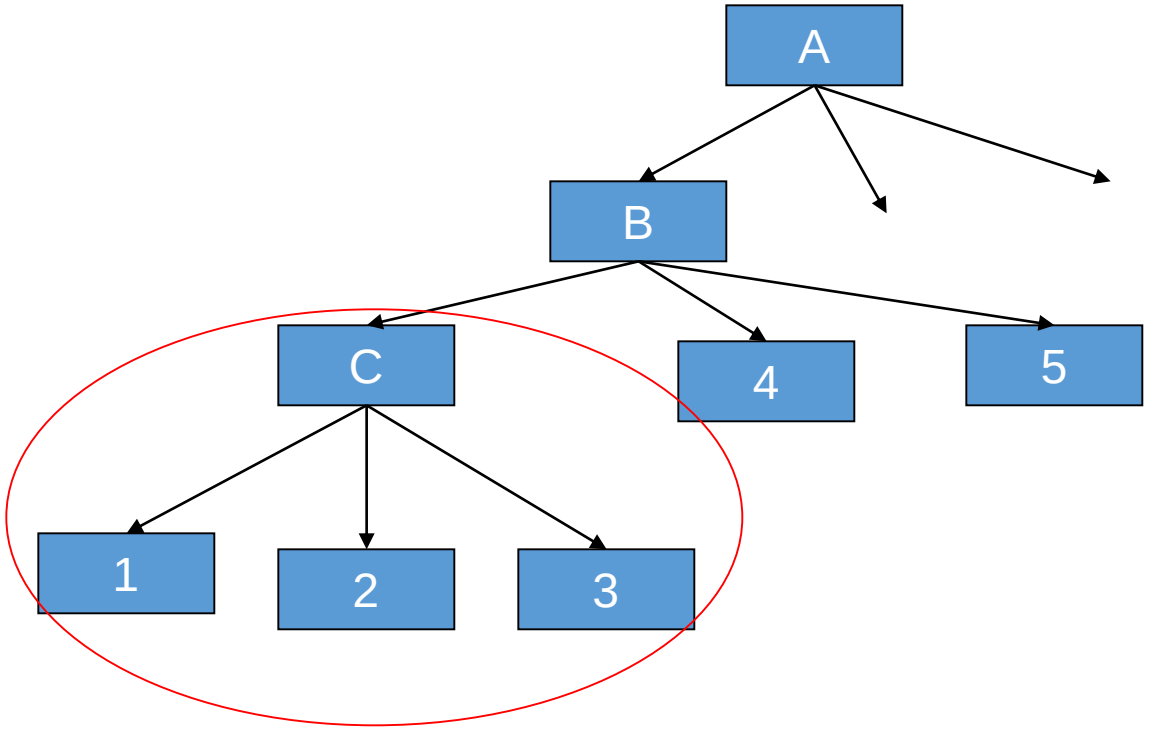
Postpruning - Αντικατάσταση υποδέντρου

- Ο κόμβος 123 αντικατέστησε το υποδέντρο



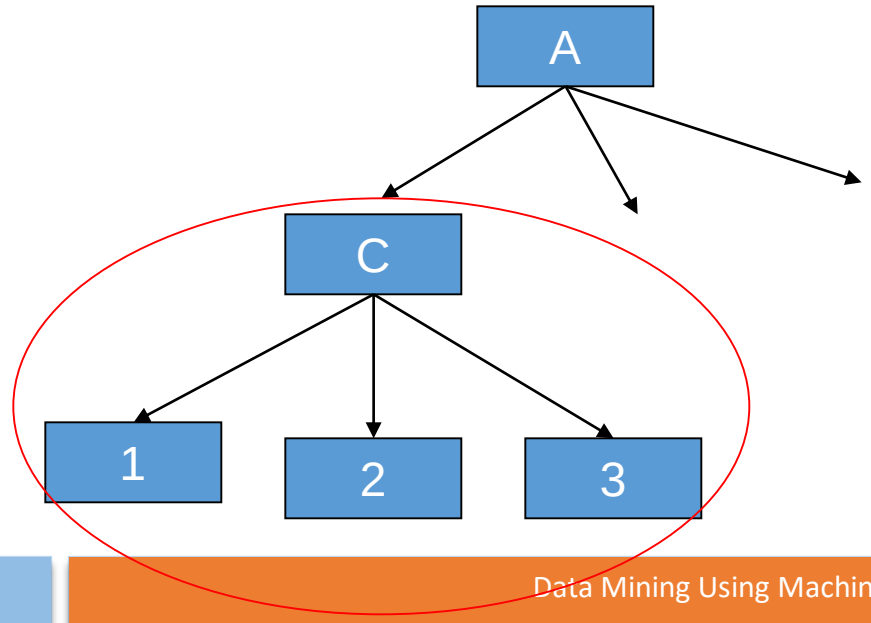
Postpruning - Ανέβασμα υποδέντρου ένα επίπεδο

- Όλο το υποδέντρο ανεβαίνει ένα επίπεδο.



Postpruning - Ανέβασμα υποδέντρου ένα επίπεδο

- Το υποδέντρο παίρνει τη θέση του κόμβου B



Προβλήματα με τον ID3

- Επίσης ένα άλλο πρόβλημα που αντιμετωπίζουμε σχετίζεται με το είδος των μεταβλητών. Για παράδειγμα, έχουμε μια στήλη η οποία περιέχει τις μετρήσεις της ταχύτητας με την οποία κινούνται τα αυτοκίνητα.

60 Km/h	90 Km/h	100 Km/h	120 Km/h	140 Km/h	180 Km/h

- Για αυτά τα δεδομένα είναι δύσκολο να παράγουμε κανόνες που να μπορούν να γενικευτούν καθώς θα είχαμε 6 κλαδιά με 6 φύλλα και σαν αποτέλεσμα η εντροπία της πηγής θα ήταν πολύ μικρή.

Προβλήματα με τον ID3

- Σε αυτή την περίπτωση χρησιμοποιούμε την τεχνική της κατηγοριοποίησης (discretization).
- Επιλέγουμε ένα σημείο τομής και θέτουμε αντίστοιχα την κατηγορία που θα ανήκουν όσες έχουν τιμή μικρότερη και μεγαλύτερη από την τιμή που έχουμε θέσει ως τομή.

Προβλήματα με τον ID3

- Οπότε αν θεωρήσουμε τις 2 τομές στα 100 και 140 αντίστοιχα

60 Km/h | 90 Km/h | 100 Km/h | 120 Km/h | 140 Km/h | 180 Km/h

ΠΡΙΝ	ΜΕΤΑ
60 Km/h	ΜΙΚΡΗ
90 Km/h	ΜΙΚΡΗ
100 Km/h	ΜΕΤΡΙΑ
120 Km/h	ΜΕΤΡΙΑ
140 Km/h	ΥΨΗΛΗ
180 Km/h	ΥΨΗΛΗ

- Οι τιμές διαμορφώνονται όπως στον παραπάνω πίνακα, όπου δυο ταχύτητες αντιστοιχούν σε μια τιμή π.χ. «ΜΙΚΡΗ» με αποτέλεσμα να δημιουργούμε ένα καλύτερο προγνωστικό μοντέλο.

ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ