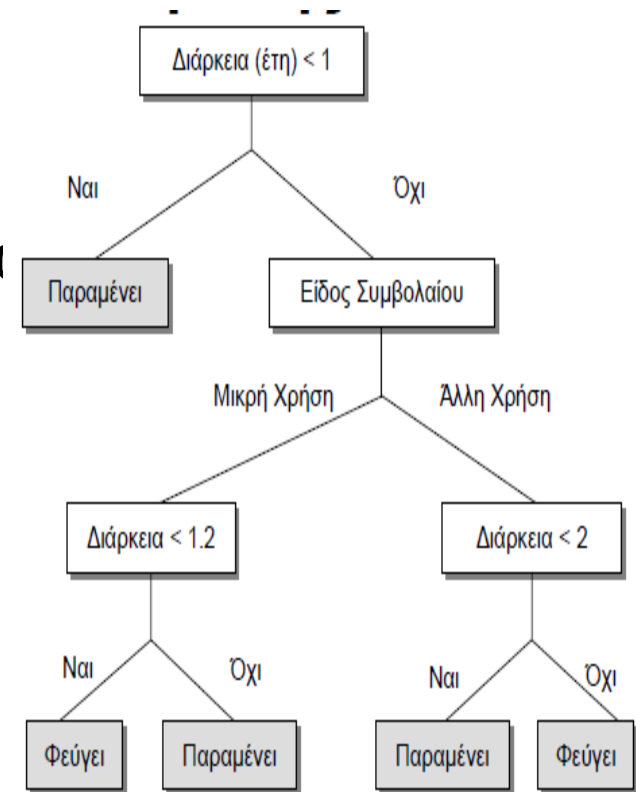


Τεχνολογίες Ευφυούς Διαχείρισης Ανθρωπιστικών Δεδομένων

Δέντρα Απόφασης

Δέντρα απόφασης

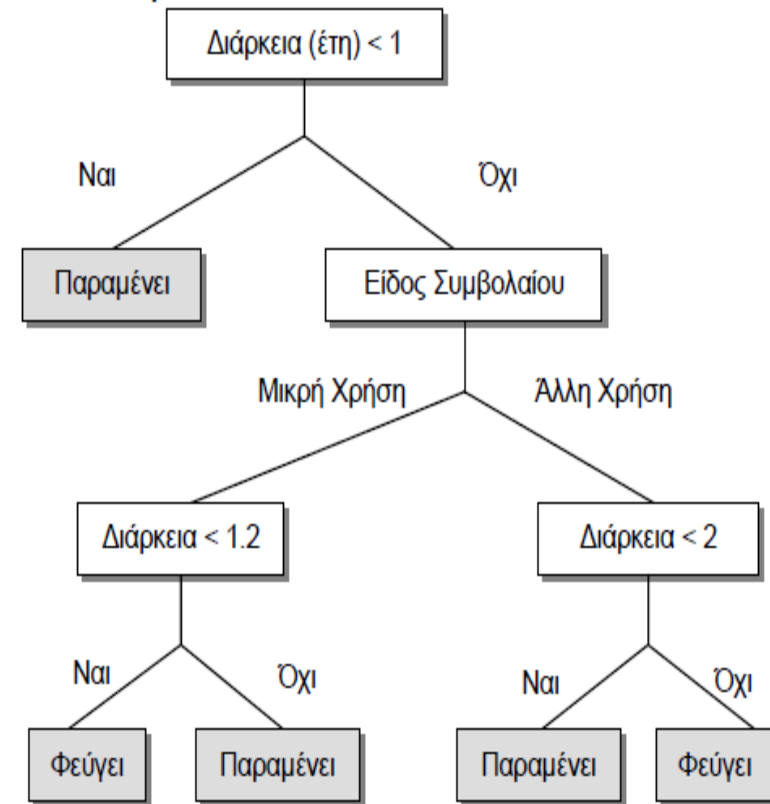
- Το μοντέλο που επάγεται είναι μια αφαιρετική δενδροειδής δομή που αναπαριστά τα δεδομένα.
- Έστω τα δεδομένα εταιρείας κινητής τηλεφωνίας που περιγράφουν περιπτώσεις συνδρομητών που παρέμειναν ή έφυγαν μετά τη λήξη του συμβολαίου τους, με βάση τη διάρκεια και το είδος αυτού.
- Μια αναπαράσταση σε δένδρο θα μπορούσε να έχει την μορφή του σχήματος δεξιά.
 - Κάθε κόμβος ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού των περιπτώσεων
 - Κάθε κλαδί που φεύγει από ένα κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού που σχετίζεται με τον κόμβο.
 - Στα κλαδιά φύλλα έχουμε το τι συνέβη.



Αναπαράσταση με Κανόνες

Classification Rules

- Εναλλακτικά, το δένδρο μπορεί να αναπαρασταθεί και ως σύνολο κανόνων if-then, που ονομάζονται **κανόνες ταξινόμησης (classification rules)**.
- π.χ. για τα δεδομένα της εταιρίας κινητής τηλεφωνίας προκύπτουν 5 κανόνες:
 - 1) if Διάρκεια < 1 then Παραμένει
 - 2) if Διάρκεια > 1 and Είδος_Συμβολαίου = Μικρή Χρήση and Διάρκεια < 1.2 then Φεύγει
 - 3) if Διάρκεια > 1 and Είδος_Συμβολαίου = Μικρή Χρήση and Διάρκεια > 1.2 then Παραμένει
 - 4) if Διάρκεια > 1 and Είδος_Συμβολαίου = Άλλη Χρήση and Διάρκεια < 2 then Παραμένει
 - 5) if Διάρκεια > 1 and Είδος_Συμβολαίου = Άλλη Χρήση and Διάρκεια > 2 then Φεύγει



Ο αλγόριθμος ID3

- Είναι ο πιο γνωστός αλγόριθμος μάθησης δένδρων ταξινόμησης.
- Είναι αναδρομικός
 - Βρες το χαρακτηριστικό οι τιμές του οποίου διαχωρίζουν βέλτιστα τα παραδείγματα εκπαίδευσης σε σχέση με την τιμή τους στην κλάση ταξινόμησης και τοποθέτησέ το στη ρίζα του δέντρου
 - Κάνε τον διαχωρισμό των παραδειγμάτων εκπαίδευσης βάσει των τιμών του χαρα/κου
 - Για κάθε τιμή του χαρα/κού δημιούργησε έναν καινούριο κόμβο στο δέντρο. Για κάθε κόμβο επανέλαβε την διαδικασία. Επέλεξε δηλ. το χαρακτηριστικό εκείνο που διαχωρίζει τα παραδείγματα του συγκεκριμένου κόμβου βέλτιστα.
 - Η διαδικασία τερματίζει όταν
 - Όλα τα παραδείγματα που ανήκουν σε αυτόν έχουν ίδια τιμή στην κλάση ταξινόμησης
 - Έχω ξεμείνει από χαρακτηριστικά προς έλεγχο
- Η επιλογή του πιο κατάλληλου χαρα/κου σε κάθε κόμβο πραγματοποιείται με κάποιο στατιστικό μέτρο

Εντροπία (Entropy)

- Η απάντηση σε μια ερώτηση που έχει n δυνατές απαντήσεις προσφέρει πληροφορία ίση με

$$I(P(v_1), \dots, P(v_n)) = \sum_{i=1}^n -P(v_i) \log_2 P(v_i)$$

- Όπου $p(v_i)$ η πιθανότητα εμφάνισης της απάντησης i .
- Πχ στην ρήψη του ζαριού

$$I\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \text{ δυαδικό ψηφίο}$$

- Εάν το ζάρι είναι πειραγμένο και έρχεται 99% κορώνα και 1% γράμματα
- $I(1/100, 99/100) = 0,08$ δυαδικά ψηφία

Επιλογή χαρακτηριστικών

- Έστω ότι έχω ένα πρόβλημα ταξινόμησης με δυο τιμές στην κλάση ταξινόμησης
- Αρχικά, υπολογίζεται από τα παραδείγματα εκπαίδευσης η πληροφορία που απαιτείται για την απάντηση στο ερώτημα αν ένα παράδειγμα είναι θετικό ή αρνητικό
- $$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$
- p : το σύνολο των θετικών παραδειγμάτων στα δεδομένα εκπαίδευσης
- n : το σύνολο των αρνητικών παραδειγμάτων στα δεδομένα εκπαίδευσης

Επιλογή χαρακτηριστικών

- Μετά την επιλογή κάποιου χαρακτηριστικού A , η πληροφορία που υπολείπεται για να απαντηθεί το ερώτημα αν ένα παράδειγμα είναι θετικό ή αρνητικό ισούται με

$$\text{Υπόλοιπο}(A) = \sum_{i=1}^v \frac{p_i + n_i}{p + n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

- όπου v ο αριθμός των τιμών του χαρακτηριστικού
- p_i ο αριθμός θετικών παραδειγμάτων που έχουν τιμή i στο χαρακτηριστικό A
- n_i ο αριθμός αρνητικών παραδειγμάτων που έχουν τιμή i στο χαρακτηριστικό A

Κέρδος Πληροφορίας (Information Gain)

- Το κέρδος σε πληροφορία επομένως που έχω επιλέγοντας το χαρ/κο A είναι

$$\text{Κέρδος}(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - \text{Υπόλοιπο}(A)$$

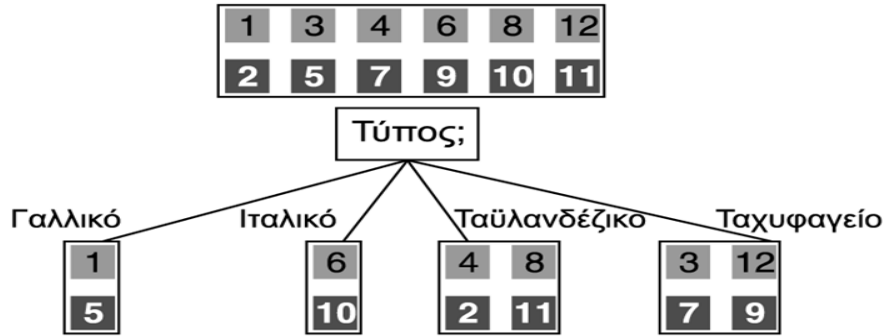
Το Παράδειγμα του Εστιατορίου

- Χαρακτηριστικά του προβλήματος:
 - Εναλλακτικό: Ναι, Όχι.
 - Μπαρ: Ναι, Όχι.
 - Π/Σ: Ναι, Όχι.
 - Πεινασμένος: Ναι, Όχι.
 - Πελάτες: Κανένας, Μερικοί, και Πλήρες.
 - Τιμή: \$, \$\$, \$\$\$.
 - Βρέχει: Ναι, Όχι.
 - Κράτηση: Ναι, Όχι.
 - Τύπος: Γαλλικό, Ιταλικό, Ταϊλανδέζικο, ή ταχυφαγείο.
 - Εκτίμηση Αναμονής: 0'-10', 10'-30', 30'-60', >60'.

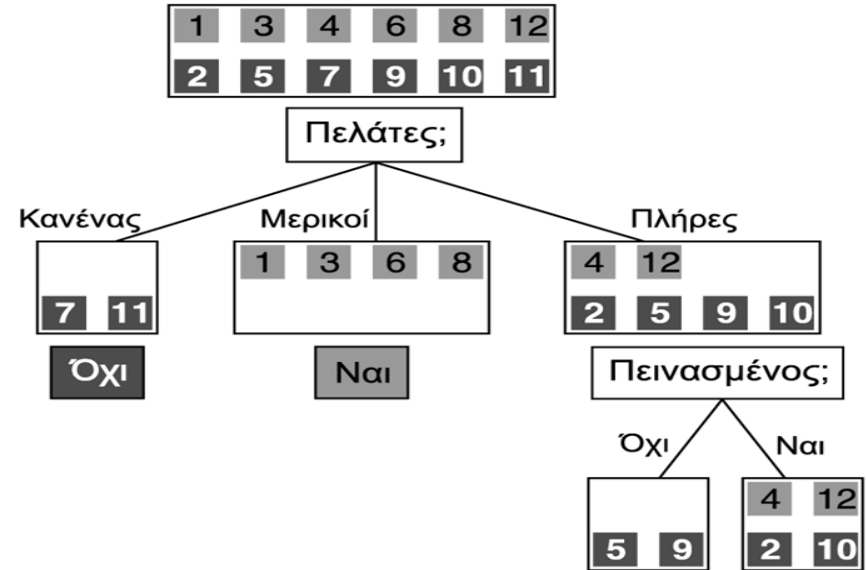
Παραδείγματα εκπαίδευσης

#	Εναλ	Μπαρ	Π/Σ	Πεινασμ	Πελατες	Τιμή	Βρέχει	Κράτηση	Τύπος	Εκτιμ	ΘαΠεριμένει
X ₁	Ναι	Όχι	Όχι	Ναι	Μερικοί	\$\$\$	Όχι	Ναι	Γαλλικό	0-10	Ναι
X ₂	Ναι	Όχι	Όχι	Ναι	Πλήρες	\$	Όχι	Όχι	Ταϋλ	30-60	Όχι
X ₃	Όχι	Ναι	Όχι	Όχι	Μερικοί	\$	Όχι	Όχι	Ταχυφ.	0-10	Ναι
X ₄	Ναι	Όχι	Ναι	Ναι	Πλήρες	\$	Ναι	Όχι	Ταϋλ	10-30	Ναι
X ₅	Ναι	Όχι	Ναι	Όχι	Πλήρες	\$\$\$	Όχι	Ναι	Γαλλικό	>60	Όχι
X ₆	Όχι	Ναι	Όχι	Ναι	Μερικοί	\$\$	Ναι	Ναι	Ιταλικό	0-10	Ναι
X ₇	Όχι	Ναι	Όχι	Όχι	Κανένας	\$	Ναι	Όχι	Ταχυφ.	0-10	Όχι
X ₈	Όχι	Όχι	Όχι	Ναι	Μερικοί	\$\$	Ναι	Ναι	Ταϋλ	0-10	Ναι
X ₉	Όχι	Ναι	Ναι	Όχι	Πλήρες	\$	Ναι	Όχι	Ταχυφ.	>60	Όχι
X ₁₀	Ναι	Ναι	Ναι	Ναι	Πλήρες	\$\$\$	Όχι	Ναι	Ιταλικό	10-30	Όχι
X ₁₁	Όχι	Όχι	Όχι	Όχι	Κανένας	\$	Όχι	Όχι	Ταϋλ	0-10	Όχι
X ₁₂	Ναι	Ναι	Ναι	Ναι	Πλήρες	\$	Όχι	Όχι	Ταχυφ.	30-60	Ναι

Επαγωγή δένδρου αποφάσεων



(α)



(β)

$$\text{Κέρδος(Πελάτες)} = 1 - \left[\frac{2}{12} I(0,1) + \frac{4}{12} I(1,0) + \frac{6}{12} I\left(\frac{2}{6}, \frac{4}{6}\right) \right] \approx 0,541 \text{ δυαδικά ψηφία}$$

$$\text{Κέρδος(Τύπος)} = 1 - \left[\frac{2}{12} I\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{2}{12} I\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{4}{12} I\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{4}{12} I\left(\frac{2}{4}, \frac{2}{4}\right) \right] = 0$$

Γαλλικό Ιταλικό Ταυλ Ταχυφ

$$I(0,1) = -0 \log 0 - 1 \log 1$$

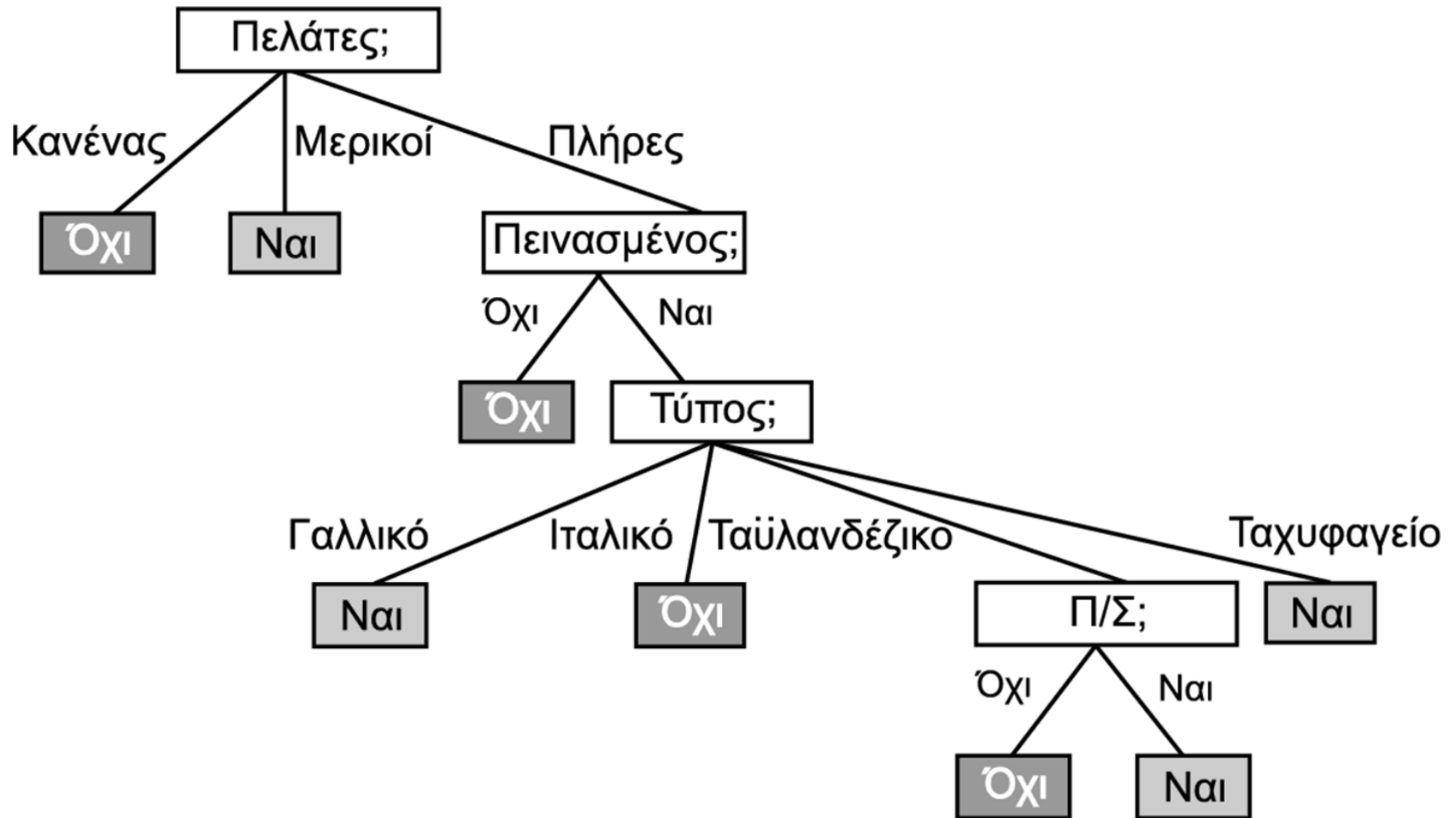
$$I(1,0) = -1 \log 1 - 0 \log 0$$

$$I\left(\frac{2}{6}, \frac{4}{6}\right) = -\frac{2}{6} \log \frac{2}{6} - \frac{4}{6} \log \frac{4}{6}$$

$$I\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2}$$

$$I\left(\frac{2}{4}, \frac{2}{4}\right) = -\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4}$$

Τελικό δένδρο



Το πρόβλημα δανειοδότησης

Πελάτης	Τρέχουσες Οφειλές	Εισόδημα	Παντρεμένος(η)	Χαρακτηρισμός
1	Υψηλές	Υψηλό	Ναι	Καλός (p)
2	Χαμηλές	Υψηλό	Όχι	Κακός (n)
3	Χαμηλές	Υψηλό	Ναι	Καλός (p)
4	Υψηλές	Χαμηλό	Ναι	Κακός (n)
5	Χαμηλές	Χαμηλό	Ναι	Κακός (n)

$$p = 2 \quad n = 3$$

$$I_{\text{αρχ}} = -p^+ \log_2 p^+ - p^- \log_2 p^- = -2/5 * \log_2(2/5) - 3/5 * \log_2(3/5) = 0.97$$

Αν $A = \text{Εισόδημα}$:

Για $i = \text{υψηλό}$ έχουμε $p_i = 2/3$, $n_i = 1/3$

Για $i = \text{χαμηλό}$ έχουμε $p_i = 0$, $n_i = 2/2 = 1$

$$\begin{aligned} \text{Υπόλοιπο}(A) &= 3/5 * (-2/3 * \log_2 2/3 - 1/3 * \log_2 1/3) + \\ &\quad 2/5 * (-0 * \log_2 0 - 1 * \log_2 1) = 3/5 * 0.92 + 2/5 * 0 \end{aligned}$$

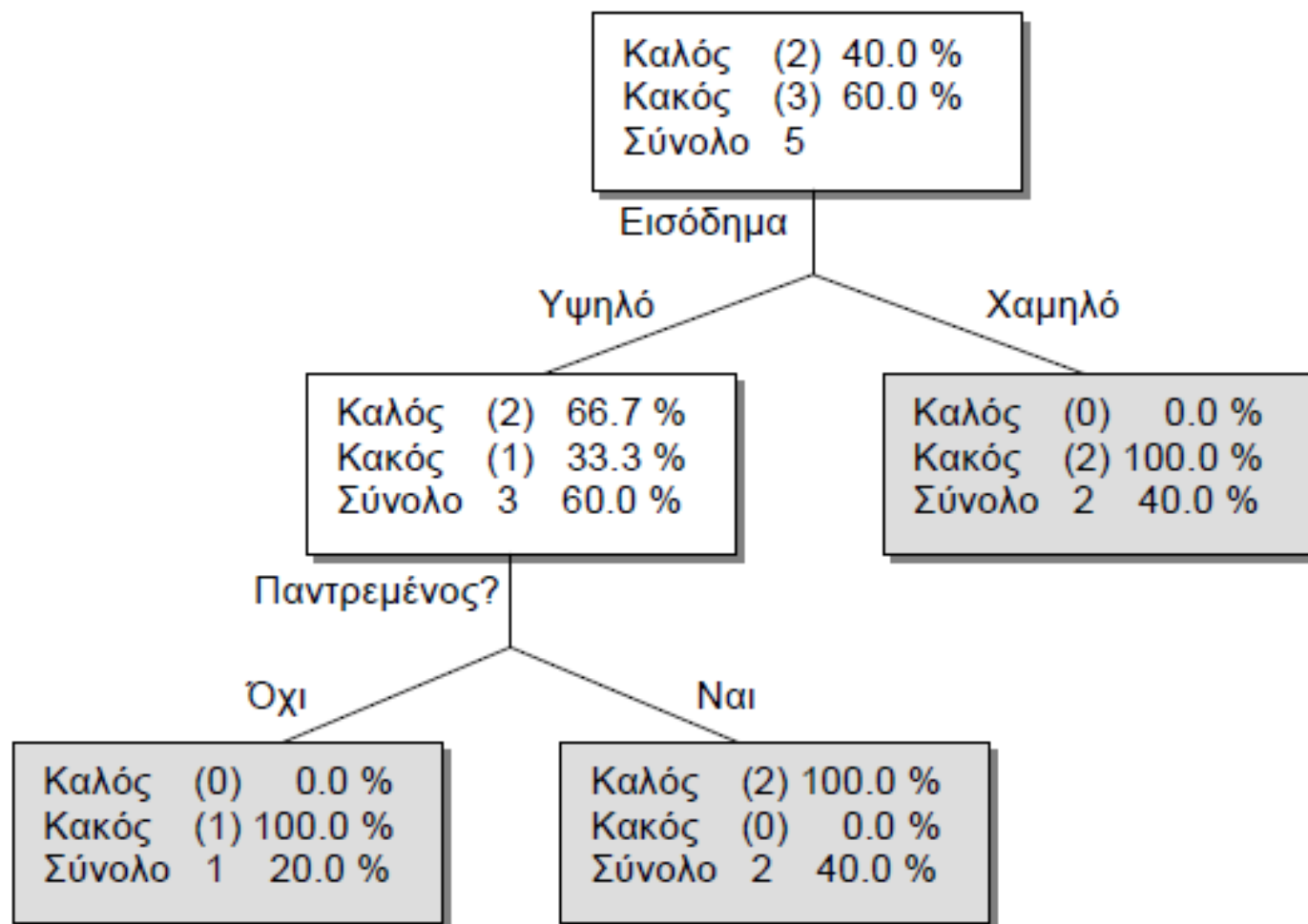
$$\text{Κέρδος}(A) = 0.97 - \text{Υπόλοιπο}(A) = 0.42$$

Αντίστοιχα, αν $A = \text{Χρέος}$, $\text{Κέρδος}(A) = 0.02$

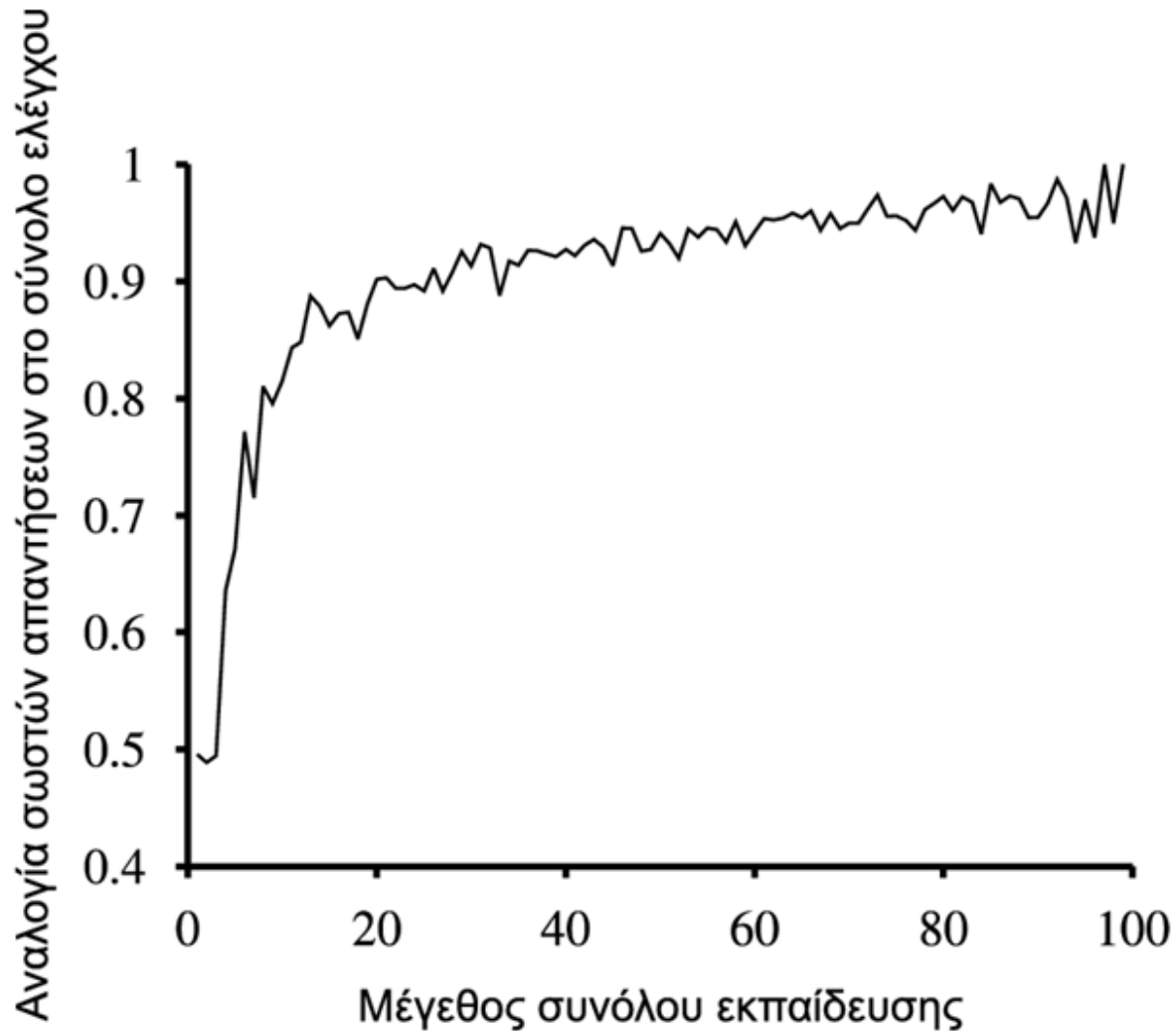
Αν $A = \text{Παντρεμένος}$, $\text{Κέρδος}(A) = 0.17$

Άρα ο κόμβος ρίζα είναι το Εισόδημα.

Τελικό Δένδρο

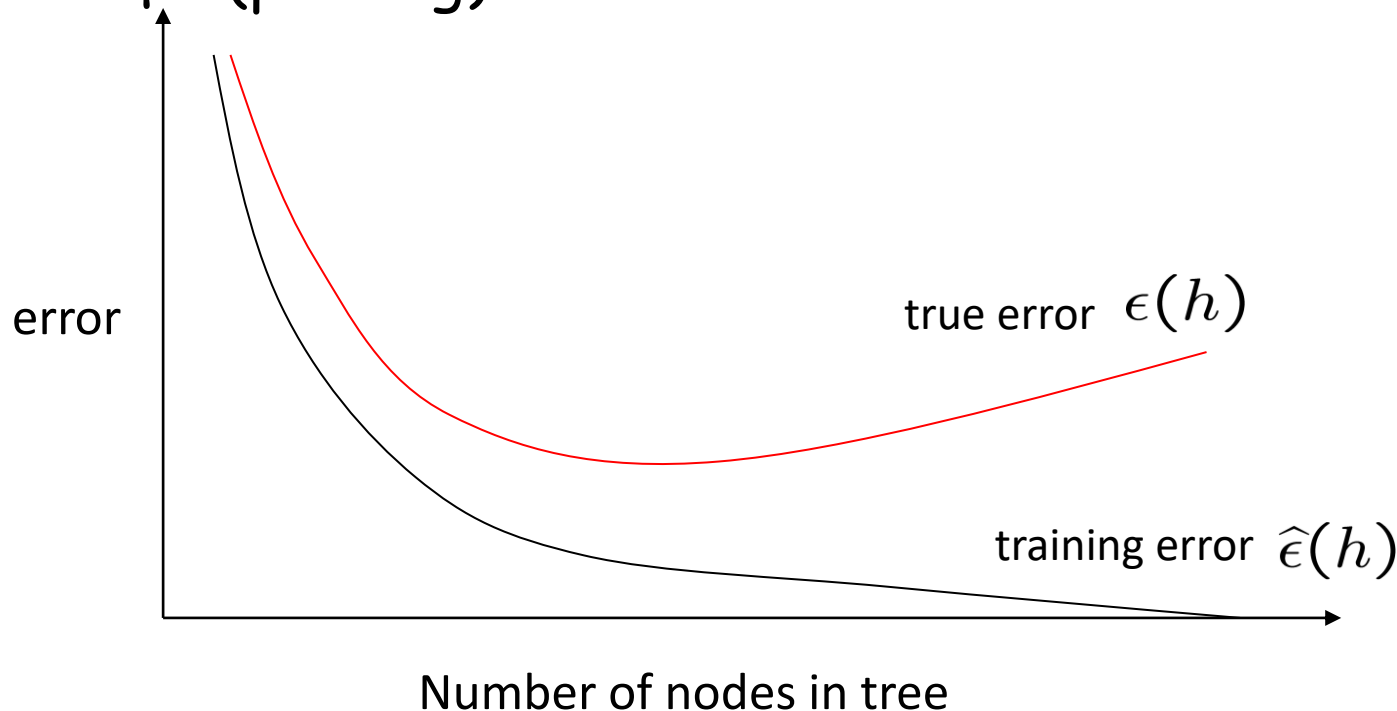


Αξιολόγηση ταξινόμησης



Το πρόβλημα της υπερπροσαρμογής (Overfitting)

- Το επαγόμενο δένδρο έχει την τάση να αποδίδει καλύτερα στην ταξινόμηση των παραδειγμάτων εκπαίδευσης, παρά των παραδειγμάτων ελέγχου.
- Χαμηλή ικανότητα γενίκευσης
- Κλάδεμα (pruning)



Reduced error pruning

Table 1.2 **The weather data.**

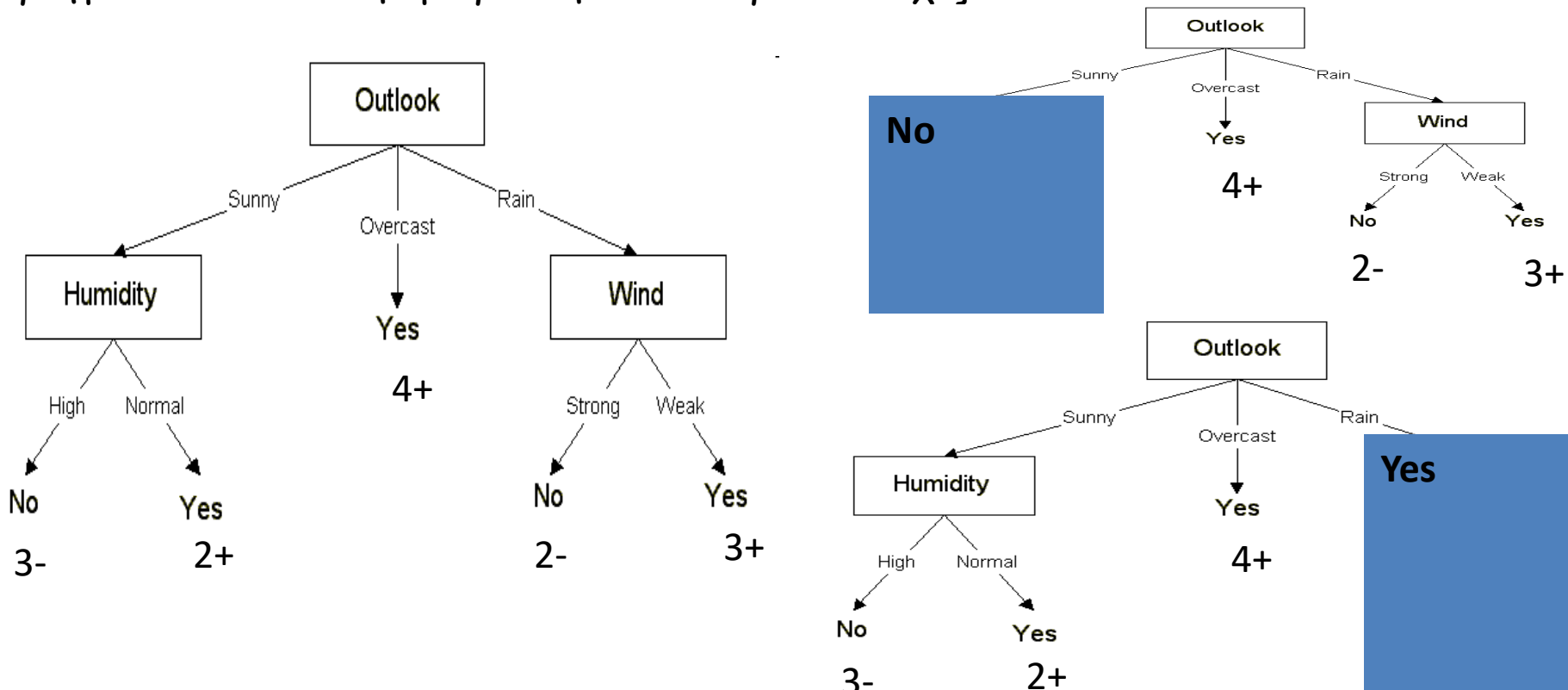
Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Reduced error pruning

Ξεχωρίζεται ένα σετ επικύρωσης (validation set) πριν την εκπαίδευση
Πραγματοποιείται εκπαίδευση επομένως με λιγότερα δεδομένα
Υπολογίζεται το σφάλμα ταξινόμησης του δέντρου που προκύπτει στα δεδομένα επικύρωσης

Από κάτω προς τα πάνω, ένας-ένας κόμβος αφαιρείται και αντικαθίσταται με φύλλο
Υπολογίζεται ξανά το σφάλμα ταξινόμησης του μικρότερου δέντρου στα δεδομένα επικύρωσης

Αν είναι μικρότερο από το σφάλμα στο μεγαλύτερο δέντρο, η αντικατάσταση πραγματοποιείται & η σμίκρυνση του δέντρου συνεχίζεται.



Παράδειγμα

- Έστω ένα σύστημα που ταξινομεί μια ελιά αυτόματα αν είναι μελάνωμα ή όχι.
- Τα χαρακτηριστικά είναι το χρώμα, το σχήμα, η υφή και το μέγεθος της ελιάς
- Έστω τα παρακάτω παραδείγματα εκπαίδευσης

	ΧΡΩΜΑ	ΣΧΗΜΑ	ΥΦΗ	ΜΕΓΕΘΟΣ	ΜΕΛΑΝΩΜΑ
1	ΜΑΥΡΟ	ΣΤΡΟΓΓΥΛΟ	ΛΕΙΑ	ΜΕΓΑΛΗ	ΝΑΙ
2	ΜΑΥΡΟ	ΑΚΑΝΟΝΙΣΤΟ	ΌΧΙ ΛΕΙΑ	ΜΙΚΡΗ	ΝΑΙ
3	ΚΑΦΕ	ΑΚΑΝΟΝΙΣΤΟ	ΛΕΙΑ	ΜΙΚΡΗ	ΌΧΙ
4	ΚΑΦΕ	ΣΤΡΟΓΓΥΛΟ	ΌΧΙ ΛΕΙΑ	ΜΕΣΑΙΑ	ΟΧΙ

- Σχεδιάστε το δέντρο απόφασης που ο id3 θα επάγει από τα παραπάνω παραδείγματα εκπαίδευσης.

Μ-2Ν 0Ο	Σ-1Ν 1Ο	Λ-1Ν 1Ο	ΜΕΓ-1Ν 0Ο
Χρώμα Κ- 0Ν 2Ο	Σχήμα Α-1Ν 1Ο	Υφή Ο-1Ν 1Ο	Μέγεθος ΜΙΚ-1Ν 1Ο ΜΕΣ-0Ν 1Ο

- Έστω πως έρχεται προς εξέταση μια ελιά μικρή, λεία, στρογγυλή, καφέ.
- Πώς θα ταξινομήσει την ελιά αυτή ο id3;

