ΚΑΤΑΣΚΕΥΗ ΔΕΝΔΡΩΝ ΑΠΟΦΑΣΗΣ

Ο αλγόριθμος ID3 (http://repfiles.kallipos.gr/html_books/93/04a-main.html)

Ο **αλγόριθμος ID3 (Iterative Dichotomizer 3)** αναπτύχθηκε αρχικά στο Πανεπιστήμιο του Σίδνεϋ από τον J. Ross Quinlan και βασίζεται στον αλγόριθμο **Concept Learning System (CLS)** που αποτελεί τυπικό αλγόριθμο δημιουργίας ΔΑ.

Αυτό που διαθέτει επιπλέον ο ID3 είναι η ευρετική αναζήτηση χαρακτηριστικού για διαχωρισμό. Δηλαδή, σε κάθε κόμβο του δένδρου αναζητά μεταξύ των χαρακτηριστικών του συνόλου δειγμάτων εκπαίδευσης αυτό το χαρακτηριστικό το οποίο διαχωρίζει καλύτερα τα δεδομένα δείγματα.

Εάν το χαρακτηριστικό διαχωρίζει πλήρως το σύνολο εκπαίδευσης, τότε ο ID3 σταματά.

Αλλιώς, λειτουργεί αναδρομικά στα n (όπου n = αριθμός των πιθανών τιμών ενός χαρακτηριστικού) διαχωρισμένα υποσύνολα, για να εντοπίσει το «καλύτερό» τους χαρακτηριστικό.

Ο αλγόριθμος εφαρμόζει αναδρομικά μια άπληστη αναζήτηση, δηλαδή επιλέγει το καλύτερο χαρακτηριστικό και δεν ανατρέχει σε προηγούμενα χαρακτηριστικά που έχει χρησιμοποιήσει στη δεδομένη πορεία του για να τα επανεξετάσει.

Τα χαρακτηριστικά του αλγόριθμου ID3 είναι τα ακόλουθα:

- Μεγαλώνει επαναληπτικά ένα μικρό σύνολο εκπαίδευσης από δεδομένα που υπάρχουν σε μία περιορισμένη βάση δειγμάτων.
- Αφαιρεί χαρακτηριστικά που περιέχονται στην κατηγορία προς ταξινόμηση με σκοπό τη συμπίεση της βάσης.
- Χρησιμοποιεί στατιστικές μεθόδους (Shannon’s Information Statistic, και τα μεγέθη εντροπία και κέρδος πληροφορίας), για να επιλέξει χαρακτηριστικά για διαχωρισμό.

Ο ID3, όπως και κάθε άλλος επαγωγικός αλγόριθμος, μπορεί να κατηγοριοποιήσει λανθασμένα ένα νέο στιγμιότυπο για πολλούς διαφορετικούς λόγους.

Ο πιο απλός είναι να μην έχει γίνει καλός σχεδιασμός και η συγκέντρωση των δειγμάτων να μην έχει βασιστεί σε σωστά χαρακτηριστικά.

Το πρόβλημα δεν αφορά τον αλγόριθμο, αλλά οδηγεί σε κακής ποιότητας συμπεράσματα στα φύλλα του δένδρου που παράγει.

Άλλος λόγος είναι ότι μη επαρκές πλήθος των δειγμάτων θα οδηγήσει σε πρόωρες συγκλίσεις και υπεραπλουστευμένα συμπεράσματα.

Τέλος, βασικότερο πρόβλημα για το οποίο ευθύνεται ο αλγόριθμος είναι η κακή επιλογή μεθόδου για τον ευρετικό ή συνήθως στατιστικό εντοπισμό του «καλού» χαρακτηριστικού.

Επιλογή χαρακτηριστικού

Προκειμένου να αποφασιστεί η καλύτερη επιλογή χαρακτηριστικού, ο αλγόριθμος ID3 βασίζεται στις έννοιες εντροπία και κέρδος πληροφορίας.

Εντροπία

Η **εντροπία πληροφορίας** (information entropy) χαρακτηρίζει το βαθμό αβεβαιότητας ενός συνόλου δεδομένων S:

$$Entropy\ (S) \ = \ \sum_{1=n}^{n} -p_i\ log p_i$$

όπου p1, p2,…pi οι πιθανότητες του κάθε ενδεχομένου που περιλαμβάνεται στο σύνολο.

Σημείωση: Ο log στην παραπάνω σχέση είναι με βάση το 2. Για να αλλάξουμε τη βάση από το b στο c, μπορούμε να χρησιμοποιήσουμε τον κανόνα αλλαγής λογάριθμου βάσης. Ο λογάριθμος βάσης b του x είναι ίσος με τον λογάριθμο βάσης c του x διαιρούμενος με τον λογάριθμο βάσης c του b:

$\log_b(x) = \log_c(x) / \log_c(b)$

Ο αλγόριθμος ID3 χρησιμοποιεί την εντροπία ως μέτρο ομοιογένειας των τιμών του αρχικού συνόλου δεδομένων προς κάποιο χαρακτηριστικό, καθώς και τα υποσύνολα που προκύπτουν σε κάθε βήμα μετά την ταξινόμηση.

Εάν όλα τα δείγματα του συνόλου εκπαίδευσης είναι ομοιογενή ως προς μια κατηγορία, τότε η εντροπία του ισούται με μηδέν. Εάν οι κατηγορίες είναι διαφορετικές και ***τα δείγματα που ανήκουν σε καθεμία έχουν το ίδιο πλήθος, τότε η εντροπία είναι 1***.

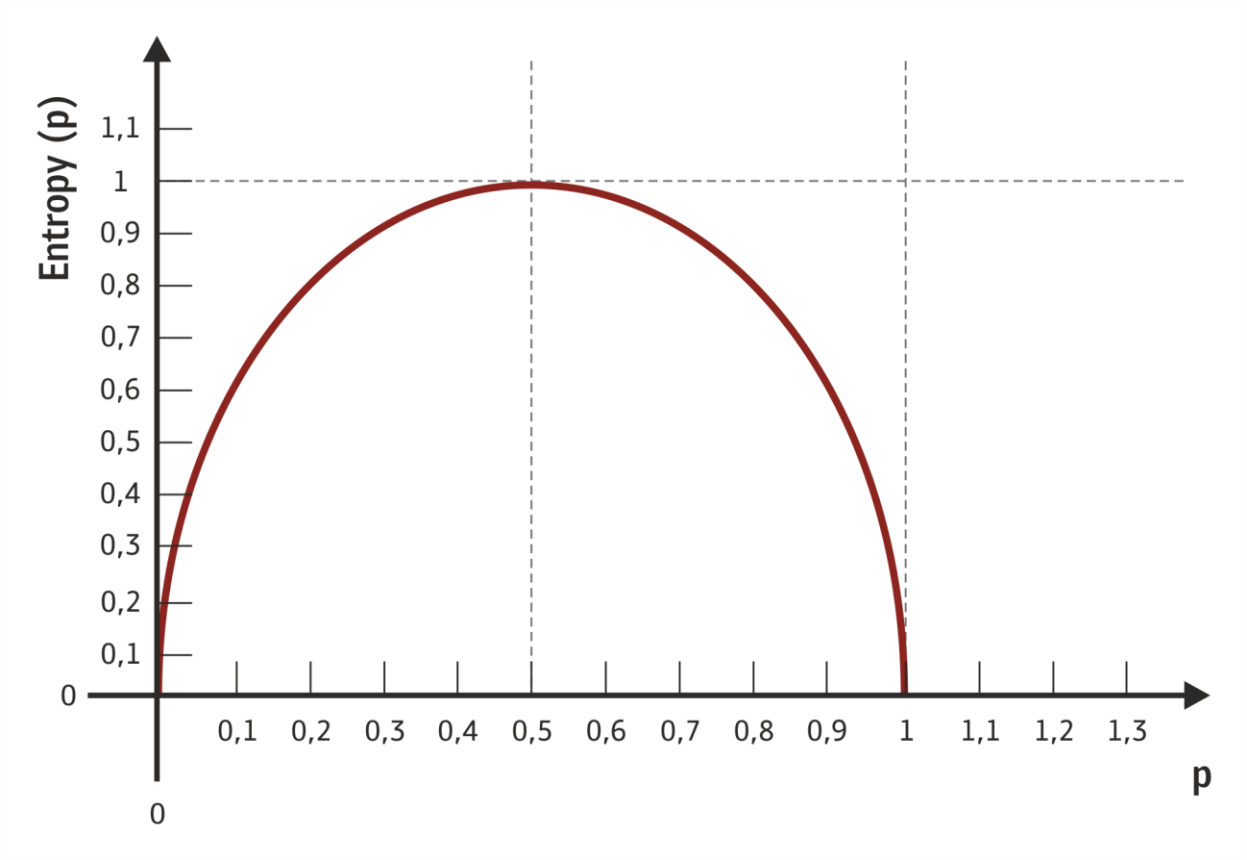
Στη δυαδική περίπτωση, όπου οι κατηγορίες είναι δύο, η εντροπία δίνεται από τη συνάρτηση:

$$Entropy\,(p) \ = \ -plogp \ - \ (1-p)\,log\,(1-p)$$

Παράδειγμα, έστω δοχείο με N μπάλες με πιθανότητα ενδεχομένου η μπάλα να είναι λευκή N(p) και να είναι μαύρη N(1-p). Τότε ισχύει:

- Αν οι μπάλες είναι όλες μαύρες ή όλες άσπρες, δηλαδή p=1 ή (1-p)=0,τότε Εντροπία=0

- Αν οι μπάλες είναι μισές μαύρες και μισές άσπρες, δηλαδή p=(1-p)= ½ (πιθανότητα επιλογής 50%) τότε Εντροπία=1 (μέγιστη)



Σχήμα 4.3 Εντροπία σε μια δυαδική περίπτωση κατηγοριών

Εάν σε ένα σύνολο S 14 μπαλών υπήρχαν 5 άσπρες μπάλες και 9 μαύρες, τότε η εντροπία θα υπολογιζόταν ως εξής:

$$Entropy(S) = \sum_{1=n}^{2} -p_i \, log p_i = -\frac{5}{14} log \frac{5}{14} - \frac{9}{14} log \frac{9}{14}$$
$$= -0{,}36 log 0{,}36 - 0{,}64 log 0{,}64 = 0{,}94$$

Μπάλες	
Μαύρες	Άσπρες
9	5

Entropy (Μπάλες) = Entropy (5,9)
= Entropy (0.36,0.64)
= -(0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

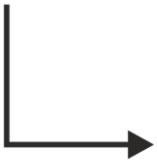
Σχήμα 4.4 Εύρεση εντροπίας με την χρήση πίνακα συχνοτήτων ενός χαρακτηριστικού

Έστω τώρα ότι μας ενδιαφέρει η εντροπία ενός χαρακτηριστικού X με n διαφορετικές τιμές c σε σχέση με την κατηγορία S. Τότε η εντροπία δίνεται με τη συνάρτηση:

$$Entropy(S, X) = \sum_{1}^{n} P(c) * Entropy(c)$$

Στο σχήμα 4.5 δίνεται ένα παράδειγμα υπολογισμού της εντροπίας για το πρόβλημα της διαφήμισης αντηλιακού που παρουσιάστηκε παραπάνω, με βάση το σύνολο εκπαίδευσης που δόθηκε στον πίνακα 4.1:

		Κάπκε		
		NAI	OXI	
Μαλλιά	μαύρα	3	2	5
	κόκκινα	4	0	4
	ξανθά	2	3	5
				14



Entropy (Κάπκε, Μαλλιά) = P (μαύρα)

* Entropy (3,2)

+ P (κόκκινα)

* Entropy (4,0)

+ P (ξανθά)

* Entropy (2,3)

$= \frac{5}{14}$ * 0,971 + *0,0

$+ \frac{5}{14}$ * 0,971 = 0,693

Σχήμα 4.5 Εύρεση εντροπίας με τη χρήση πίνακα συχνοτήτων δύο χαρακτηριστικών

$$\boldsymbol{Entropy}(\boldsymbol{K}\acute{\alpha}\boldsymbol{\eta\kappa\varepsilon}, \boldsymbol{M\alpha\lambda\lambda\iota}\acute{\alpha})$$
$$= P(\mu\alpha\acute{\upsilon}\rho\alpha) * Entropy(3,2) + P(\kappa\acute{o}\kappa\kappa\iota\nu\alpha) * Entropy(4,0) + P(\xi\alpha\nu\theta\acute{\alpha}) * Entropy(2,3) = \frac{5}{14} * 0.971 + \frac{4}{14} * 0.0 + \frac{5}{14} * 0.971 = 0.693$$

Κέρδος Πληροφορίας

Εάν A είναι χαρακτηριστικό σε ένα S σύνολο δειγμάτων από το σύνολο μάθησης, τότε το **κέρδος πληροφορίας** (information gain) χαρακτηρίζει το πόση πληροφορία «φέρει» το χαρακτηριστικό A:

$$G(S, A) = E(S) - \sum_{i=1}^{m} f_S(A_i) * E(S_{A_i})$$

όπου

- **E(...)** η συνάρτηση εντροπίας
- **m** το πλήθος των τιμών Ai που παίρνει το A στο S
- **fs(Ai)** το ποσοστό των δειγμάτων στο S που παίρνουν την τιμή Ai
- **S**Ai το υποσύνολο του S όπου η τιμή του A είναι Ai

Ή πιο απλά:

$$Gain(S, A) = Entropy(S) - Entropy(S, A)$$

Η γενική ιδέα του ID3, για να διαχωρίσει ένα μη ομοιογενές σύνολο δειγμάτων σε έναν κόμβο του δένδρου απόφασης που ανήκει σε ένα συγκεκριμένο κλαδί, είναι:

- Για όλα τα αχρησιμοποίητα στο συγκεκριμένο κλαδί χαρακτηριστικά υπολόγισε την εντροπία σε σχέση με τα δείγματα.
- Διάλεξε το χαρακτηριστικό που παρουσιάζει το μέγιστο κέρδος πληροφορίας.
- Κατασκεύασε κόμβο για το χαρακτηριστικό αυτό.

<u>Παράδειγμα ID3</u>

Αν χρησιμοποιήσουμε τον αλγόριθμο ID3 για το πρόβλημα της διαφήμισης αντηλιακού που παρουσιάσαμε προηγουμένως, τα χαρακτηριστικά (FS) που θα χρησιμοποιήσει ο αλγόριθμος θα είναι:

- Μαλλιά = {ξανθά, καστανά, κόκκινα}
- Μάτια = {καφέ, μπλε, πράσινα}
- Δέρμα = {ανοιχτό, σκούρο}
- Αντηλιακό = {ναι, όχι}

Οι κατηγορίες ταξινόμησης (C) είναι:

- Κάηκε = {ΝΑΙ, ΟΧΙ}

Το σύνολο δειγμάτων εκπαίδευσης είναι συγκεντρωμένο στον επόμενο πίνακα 4.2.

Δείγμα	Χαρακτηριστικά				Κατηγορία Ταξινόμησης
	Μαλλιά	Μάτια	Δέρμα	Αντηλιακό	Κάηκε
Τ1	ξανθά	καφέ	σκούρο	όχι	ΟΧΙ
Τ2	ξανθά	καφέ	σκούρο	ναι	ΟΧΙ
Τ3	κόκκινα	καφέ	σκούρο	όχι	ΝΑΙ
Τ4	μαύρα	μπλε	σκούρο	όχι	ΝΑΙ
Τ5	μαύρα	πράσινα	ανοιχτό	όχι	ΝΑΙ
Τ6	μαύρα	πράσινα	ανοιχτό	ναι	ΟΧΙ
Τ7	κόκκινα	πράσινα	ανοιχτό	ναι	ΝΑΙ
Τ8	ξανθά	μπλε	σκούρο	όχι	ΟΧΙ
Τ9	ξανθά	πράσινα	ανοιχτό	όχι	ΝΑΙ
Τ10	μαύρα	μπλε	ανοιχτό	όχι	ΝΑΙ
Τ11	ξανθά	μπλε	ανοιχτό	ναι	ΝΑΙ
Τ12	κόκκινα	μπλε	σκούρο	ναι	ΝΑΙ
Τ13	κόκκινα	καφέ	ανοιχτό	όχι	ΝΑΙ
Τ14	μαύρα	μπλε	σκούρο	ναι	ΟΧΙ

Πίνακας 4.2 Σύνολο εκπαίδευσης για το πρόβλημα της διαφήμισης αντηλιακού

Παράδειγμα δημιουργίας δένδρου απόφασης με χρήση εντροπίας και κέρδους πληροφορίας

Βήμα 1: Υπολογισμός εντροπίας της κατηγορίας του στόχου

$$\begin{aligned} \boldsymbol{Entropy}(\boldsymbol{Κάηκε}) &= Entropy(5,9) \\ &= Entropy(0.36, 0.64) \\ &= -(0.36 log 0.36) - (0.64 log 0.64) \\ &= 0.94 \end{aligned}$$

Βήμα 2: Το σύνολο εκπαίδευσης διαχωρίζεται ως προς τα διαφορετικά χαρακτηριστικά.

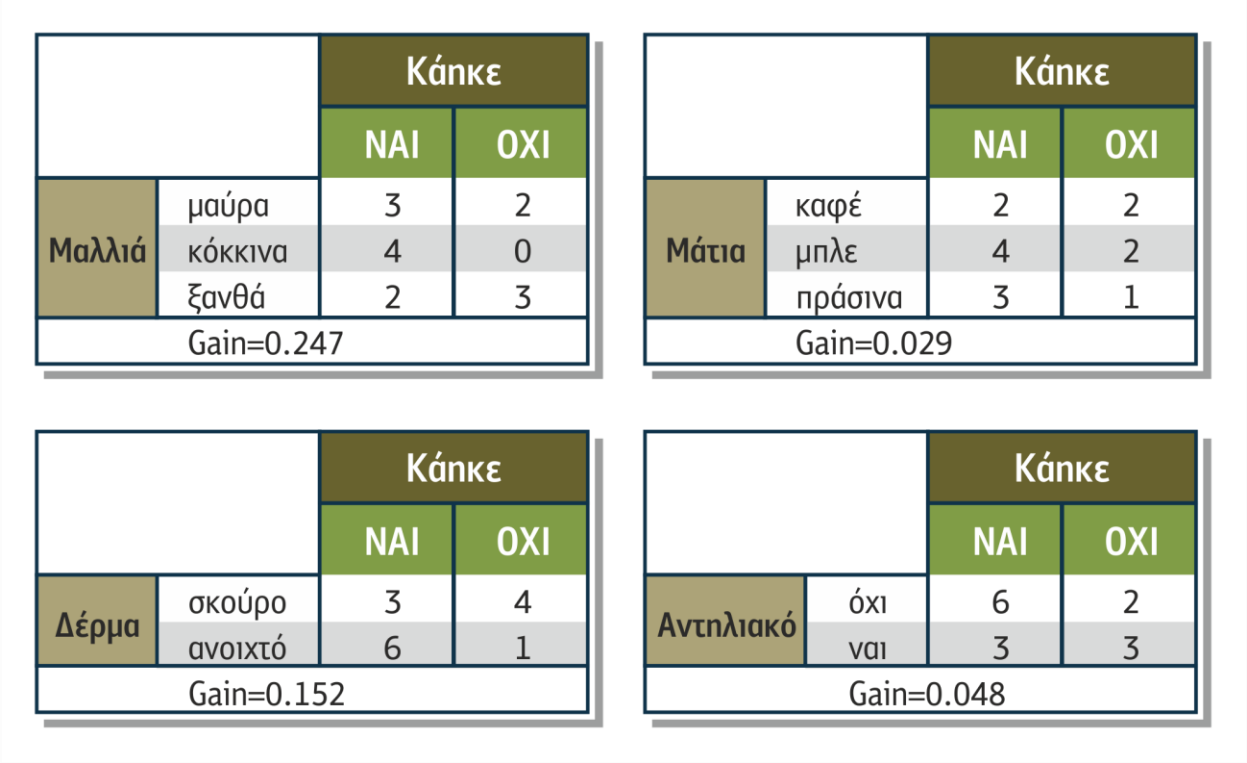
Για κάθε διαφορετικό χαρακτηριστικό γίνεται ο ακόλουθος υπολογισμός: Υπολογίζεται η εντροπία για κάθε διαφορετική τιμή του χαρακτηριστικού και στη συνέχεια, αφού πολλαπλασιαστεί με το ποσοστό των δειγμάτων που διαθέτουν την αντίστοιχη τιμή, προστίθεται για να υπολογιστεί η συνολική εντροπία του διαχωρισμένου τμήματος.

$$\begin{aligned}
&\boldsymbol{Entropy}(\boldsymbol{K}\acute{\alpha}\boldsymbol{\eta\kappa\varepsilon}, \boldsymbol{M\alpha\lambda\lambda\iota}\acute{\alpha}) \\
&\quad = P(\mu\alpha\acute{\upsilon}\rho\alpha) * Entropy(3,2) \\
&\quad + \ P(\kappa\acute{o}\kappa\kappa\iota\nu\alpha) * Entropy(4,0) \\
&\quad + P(\xi\alpha\nu\theta\acute{\alpha}) * Entropy(2,3) \\
&\quad = \frac{5}{14} * 0.971 + \frac{4}{14} * 0.0 + \frac{5}{14} * 0.971 \\
&\quad = 0.693
\end{aligned}$$

Η εντροπία που προκύπτει ως αποτέλεσμα αφαιρείται από την εντροπία πριν από το διαχωρισμό. Το αποτέλεσμα είναι το κέρδος πληροφορίας (Gain).

$$\begin{aligned}
&\boldsymbol{Gain}(\boldsymbol{K}\acute{\alpha}\boldsymbol{\eta\kappa\varepsilon}, \boldsymbol{M\alpha\lambda\lambda\iota}\acute{\alpha}) \\
&\quad = Entropy(K\acute{\alpha}\eta\kappa\varepsilon) - Entropy(K\acute{\alpha}\eta\kappa\varepsilon, M\alpha\lambda\lambda\iota\acute{\alpha}) \\
&\quad = 0.940 - 0.693 = 0.247
\end{aligned}$$

Το κέρδος πληροφορίας πρέπει να υπολογιστεί για κάθε διαφορετικό χαρακτηριστικό που μπορεί να εφαρμοστεί για να διαχωρίσει ένα συγκεκριμένο υποσύνολο μη πλήρως διαχωρισμένων δειγμάτων που είναι συγκεντρωμένα σε έναν κόμβο του δένδρου (βλέπε Σχήμα 4.6).



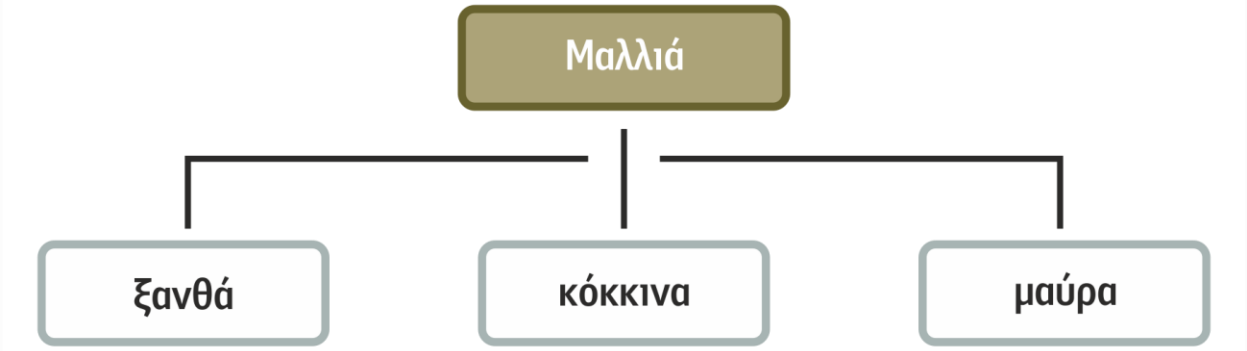
Σχήμα 4.6 Υπολογισμός κέρδους πληροφορίας για επιλογή χαρακτηριστικού

Βήμα 3: Επιλογή του χαρακτηριστικού με το μεγαλύτερο κέρδος.

Στον πρώτο κύκλο διαχωρισμού επιλέγεται το χαρακτηριστικό Μαλλιά για να διαχωρίσει το δείγμα, δεδομένου ότι έχει το μεγαλύτερο κέρδος πληροφορίας.

Παρατηρούμε ότι τα δείγματα που έχουν κόκκινα μαλλιά είναι ομοιογενή ως προς το αποτέλεσμα, δηλαδή έχουν όλα καεί.

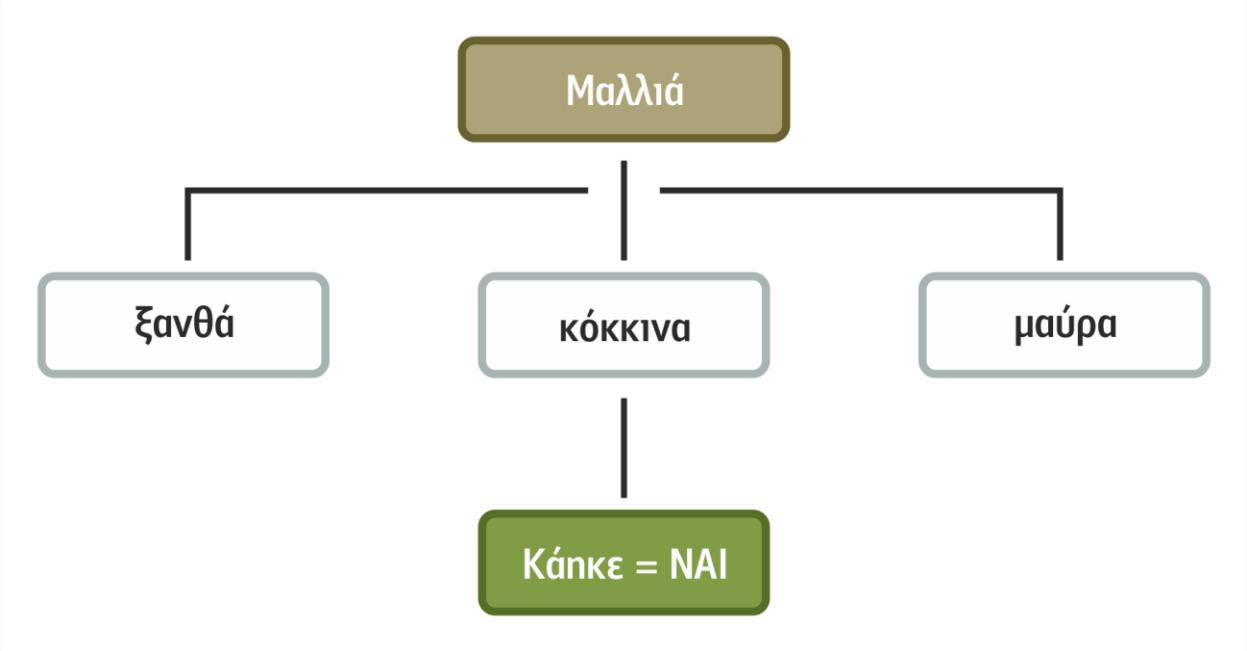
		Κάπκε	
		ΝΑΙ	ΟΧΙ
Μαλλιά	μαύρα	3	2
	κόκκινα	4	0
	ξανθά	2	3
Gain=0.247			



Σχήμα 4.7 Επιλογή του χαρακτηριστικού Μαλλιά για να διαχωρίσει το δείγμα

Βήμα 4Α: Ολοκλήρωση διαχωρισμού

Ένα κλαδί με εντοπία 0, δηλαδή όταν όλα τα δείγματα του κόμβου είναι ομοιογενή, θεωρείται φύλλο. Ένα κλαδί που ξεκινά από τη ρίζα του δένδρου και καταλήγει σε φύλλο, θεωρείται ολοκληρωμένο και μπορεί να εκφραστεί ως κανόνας.



Σχήμα 4.8 Ημιτελές δένδρο απόφασης με ένα πλήρως διαχωρισμένο κλαδί.

Βήμα 4Β: Συνέχιση διαχωρισμού

Ένα κλαδί με εντροπία μεγαλύτερη από το 0, δηλαδή όταν τα δείγματα του κόμβου δεν είναι ομοιογενή, χρειάζεται περαιτέρω διαχωρισμό. Το ημιτελές δένδρο απόφασης στο σχήμα 4.8 περιέχει ένα ολοκληρωμένο κλαδί και δύο που χρήζουν περαιτέρω διαχωρισμού.

Για να διαχωρίσομε τα ανομοιογενή δείγματα του τελευταίου κόμβου του αριστερού κλαδιού του δένδρου και επειδή χρησιμοποιήσαμε το χαρακτηριστικό Μαλλιά ήδη, πρέπει να αποφασίσουμε μόνο για το ποιο από τα υπόλοιπα τρία χαρακτηριστικά είναι καταλληλότερο να εφαρμοστεί: Μάτια, Δέρμα ή Αντηλιακό. Μελετώντας ξεχωριστά το κέρδος πληροφορίας για κάθε ένα από αυτά, έχουμε τα παρακάτω αποτελέσματα:

Μαλλιά ξανθά = {T1, T2, T8, T9, T11} = 5 δείγματα από τον πίνακα 1 με Μαλλιά = Ξανθά

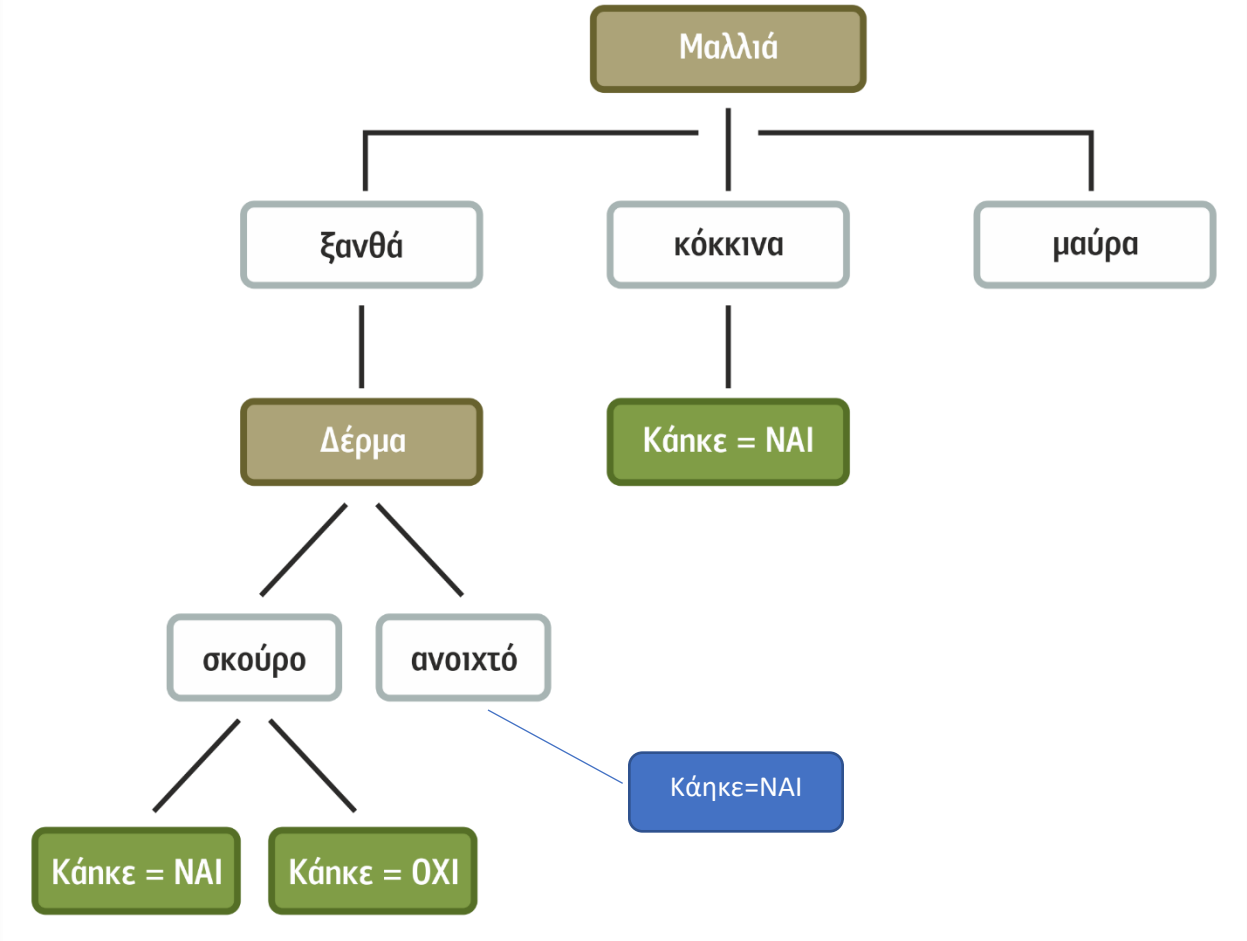
Κέρδος πληροφορίας (Μαλλιά ξανθά, Δέρμα) = 0.970

Κέρδος πληροφορίας (Μαλλιά ξανθά, Μάτια) = 0.570

Κέρδος πληροφορίας (Μαλλιά ξανθά, Αντηλιακό) = 0.019

Παρατηρούμε ότι το χαρακτηριστικό Δέρμα έχει το υψηλότερο κέρδος πληροφορίας. Ως εκ τούτου, εφαρμόζεται ως χαρακτηριστικό διαχωρισμού των δειγμάτων του συγκεκριμένου κόμβου (Μαλλιά Ξανθά), όπως φαίνεται στο παρακάτω σχήμα 4.9.

Δείγμα	Μαλλιά	Μάτια	Δέρμα	Αντηλιακό	Κάηκε
T1	ξανθά	καφέ	σκούρο	όχι	ΟΧΙ
T2	ξανθά	καφέ	σκούρο	ναι	ΟΧΙ
T8	ξανθά	μπλε	σκούρο	όχι	ΟΧΙ
T9	ξανθά	πράσινα	ανοιχτό	όχι	ΝΑΙ
T11	ξανθά	μπλε	ανοιχτό	ναι	ΝΑΙ

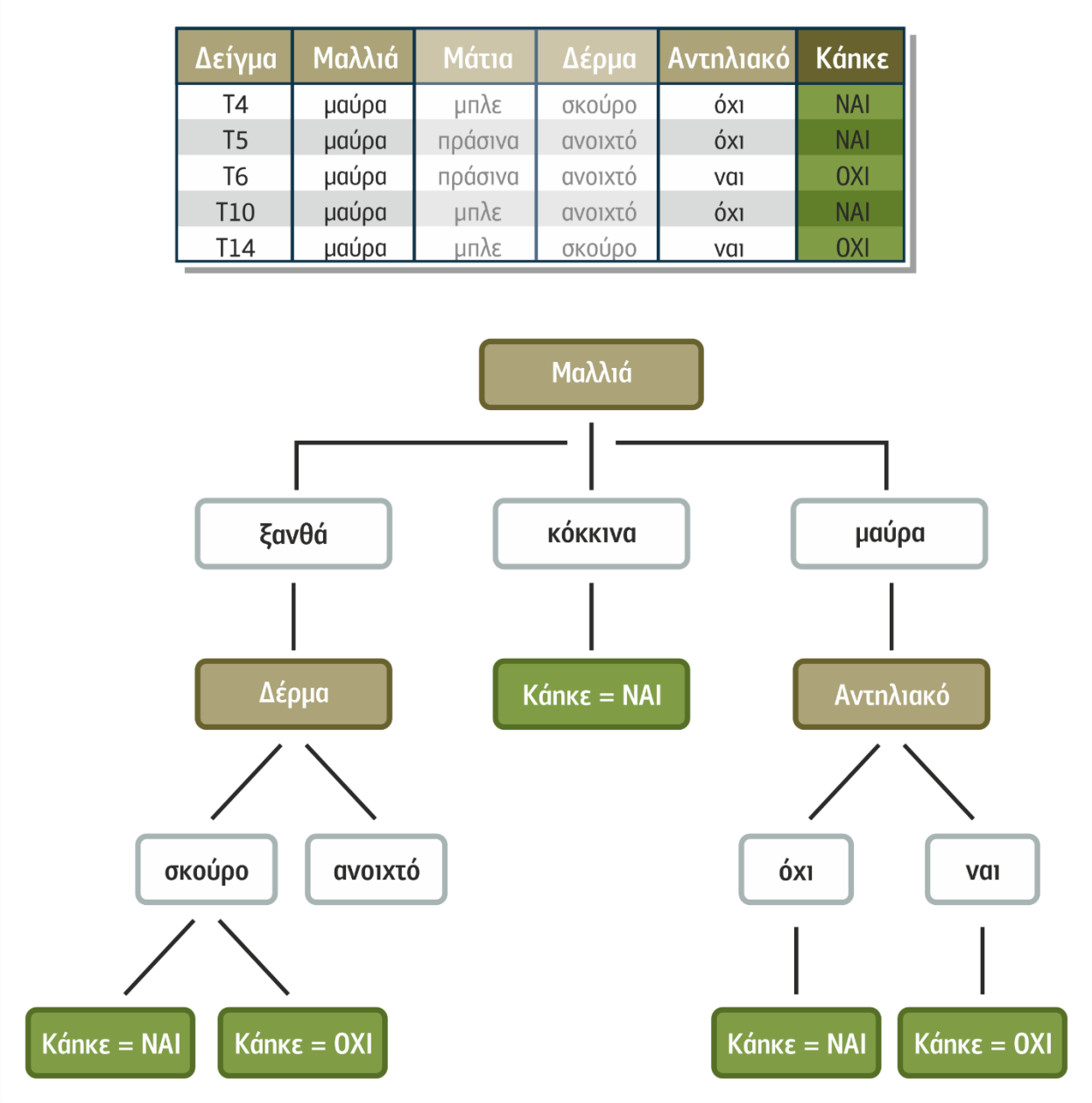


Σχήμα 4.9. Ημιτελές δένδρο απόφασης με δύο πλήρως διαχωρισμένα κλαδιά.

Βήμα 5: Ολοκλήρωση διαχωρισμού

Ο αλγόριθμος ID3 τρέχει αναδρομικά διαχωρίζοντας σε κάθε κύκλο έναν κόμβο του δένδρου που δεν είναι φύλλο, έως ότου όλα τα δείγματα κατηγοριοποιηθούν.

Δείγμα	Μαλλιά	Μάτια	Δέρμα	Αντηλιακό	Κάηκε
T4	μαύρα	μπλε	σκούρο	όχι	ΝΑΙ
T5	μαύρα	πράσινα	ανοιχτό	όχι	ΝΑΙ
T6	μαύρα	πράσινα	ανοιχτό	ναι	ΟΧΙ
T10	μαύρα	μπλε	ανοιχτό	όχι	ΝΑΙ
T14	μαύρα	μπλε	σκούρο	ναι	ΟΧΙ



Σχήμα 4.10 Δένδρο απόφασης με όλα τα δείγματα κατηγοριοποιημένα

Δείτε κινούμενη εικόνα 4.1 - Βηματική ανάπτυξη Δένδρου Απόφασης

Αν δεν υπάρχουν άλλα χαρακτηριστικά για να διαχωρίσουν έναν κόμβο που δεν είναι φύλλο, αυτό σημαίνει αποτυχία διαχωρισμού ελλείψει επαρκών χαρακτηριστικών και η όλη μελέτη για τη λύση του προβλήματος πρέπει να επαναληφθεί από την αρχή.

Βήμα 6: Έλεγχος αποτελεσμάτων

Ο αλγόριθμος μπαίνει σε φάση πιστοποίησης (recall) ελέγχοντας την εγκυρότητα των αποτελεσμάτων του με τη βοήθεια των δειγμάτων του συνόλου ελέγχου, δηλαδή δειγμάτων που δεν είχαν περιληφθεί στο σύνολο εκπαίδευσης.

Σε περίπτωση που υπάρχουν εξαιρέσεις, τα εξαιρούμενα δείγματα προστίθενται στο σύνολο εκπαίδευσης και η εκπαίδευση επαναλαμβάνεται από την αρχή.

Αν δεν υπάρχουν εξαιρέσεις, τα αποτελέσματα θεωρούνται έγκυρα και δίνονται προς χρήση.

Άλλα παραδείγματα εφαρμογής ID3

Η εντροπία και το κέρδος πληροφορίας δεν είναι ο μόνος τρόπος επιλογής χαρακτηριστικών.

Σε πολλά προβλήματα ως ευρετικοί μηχανισμοί χρησιμοποιούνται άλλες απλούστερες μέθοδοι, όπως για παράδειγμα ο δοκιμαστικός διαχωρισμός των δειγμάτων του συνόλου εκπαίδευσης σε υποσύνολα βάσει κάθε διαφορετικού χαρακτηριστικού και επιλογή εκείνου που (ιδανικά) θα αποδώσει υποσύνολα με δείγματα «όλα θετικά» ή «όλα αρνητικά» ή θα προσεγγίσει μια τέτοια κατάσταση.

Παράδειγμα όπου εφαρμόζεται μια τέτοια τακτική είναι το πρόβλημα της αναμονής σε εστιατόριο (Russel & Norvig, 2003).

Το συγκεκριμένο πρόβλημα αφορά τη λήψη απόφασης ενός πελάτη να περιμένει ή όχι για να φάει σε ένα εστιατόριο.

Το σύνολο των χαρακτηριστικών (FS) περιλαμβάνει τα ακόλουθα χαρακτηριστικά:

Εναλλακτικό:	υπάρχει μια εναλλακτική λύση εστιατορίου κοντά (Ναι/Όχι);
Μπαρ	υπάρχει κοντά ένα άνετο μπαρ για να περιμένει κανείς εκεί(Ναι/Όχι)
Π/Σ	η ημέρα είναι Παρασκευή ή Σάββατο(Ναι/Όχι)
Πεινασμένος	ο πελάτης είναι πεινασμένος (Ναι/Όχι)
Πελάτες	ποιο το πλήθος των πελατών στο εστιατόριο (Κανείς, Μερικοί, Πλήρες)
Τιμή	εύρος τιμών ($, $$, $$$)
Βρέχει	έξω βρέχει (Ναι/Όχι)

Κράτηση	έχει γίνει κράτηση (Ναι/Όχι)
Τύπος	είδος του εστιατορίου (Γαλλικό-fre, Ιταλικό-ital, Ταϋλανδέζικο-thai, Ταχυφαγείο-fast)
Αναμονή	εκτιμώμενος χρόνος αναμονής σε λεπτά (0-10, 10-30, 30-60,> 60)

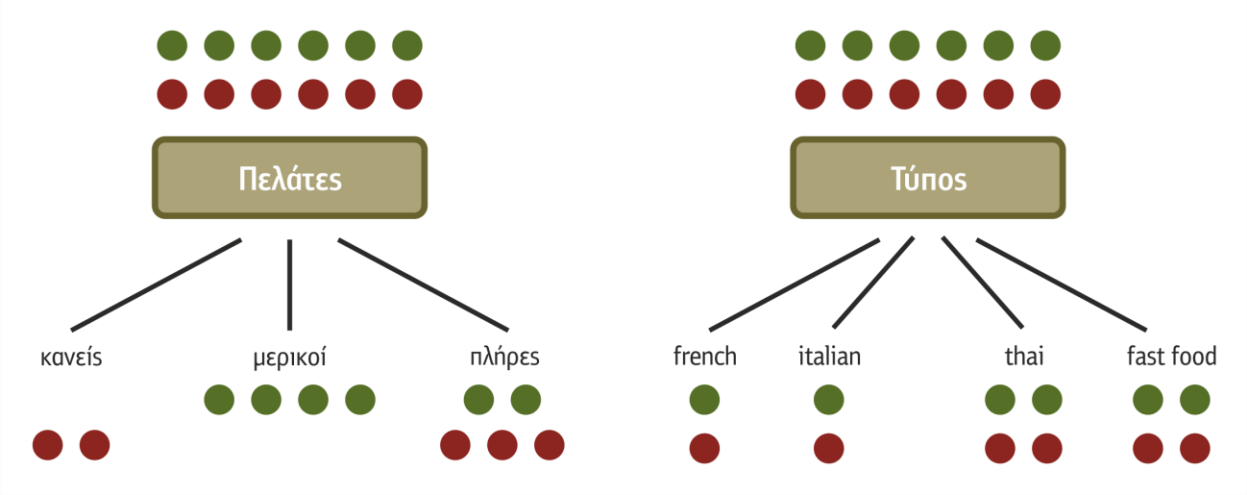
Το σύνολο εκπαίδευσης είναι καταχωρημένο στον ακόλουθο πίνακα 4.3.

Δείγμα	Χαρακτηριστικά										Αναμονή?
	Εναλλακτικό	Μπαρ	Π/Σ	Πεινασμένος	Πελάτες	Τιμή	Βρέχει	Κράτηση	Τύπος	Εκτίμηση Αναμονής	
X_1	ναι	όχι	όχι	ναι	μερικοί	$$$	όχι	ναι	fre	0-10	ΝΑΙ
X_2	ναι	όχι	όχι	ναι	πλήρες	$	όχι	όχι	thai	30-60	ΟΧΙ
X_3	όχι	ναι	όχι	όχι	μερικοί	$	όχι	όχι	fast	0-10	ΝΑΙ
X_4	ναι	όχι	ναι	ναι	πλήρες	$	όχι	όχι	thai	10-30	ΝΑΙ
X_5	ναι	όχι	ναι	όχι	πλήρες	$$$	όχι	ναι	fre	> 60	ΟΧΙ
X_6	όχι	ναι	όχι	ναι	μερικοί	$$	ναι	ναι	ital	0-10	ΝΑΙ
X_7	όχι	ναι	όχι	όχι	κανείς	$	ναι	όχι	fast	0-10	ΟΧΙ
X_8	όχι	όχι	όχι	ναι	μερικοί	$$	ναι	ναι	thai	0-10	ΝΑΙ
X_9	όχι	ναι	ναι	όχι	πλήρες	$	ναι	όχι	fast	> 60	ΟΧΙ
X_{10}	ναι	ναι	ναι	ναι	πλήρες	$$$	όχι	ναι	ital	10-30	ΟΧΙ
X_{11}	όχι	όχι	όχι	όχι	κανείς	$	όχι	όχι	thai	0-10	ΟΧΙ
X_{12}	ναι	ναι	ναι	ναι	πλήρες	$	όχι	όχι	fast	30-60	ΝΑΙ

Πίνακας 4.3 Σύνολο εκπαίδευσης για το πρόβλημα της αναμονής στο εστιατόριο

Για την επιλογή του 1ου χαρακτηριστικού που θα διαχωρίσει το αρχικό σύνολο δειγμάτων του προβλήματος, έστω ότι συγκρίνονται τα χαρακτηριστικά Πελάτες και Τύπος.

Από το αποτέλεσμα παρατηρούμε ότι το Πελάτες διαχωρίζει το σύνολο δημιουργώντας δυο φύλλα και ένα μη διαχωρισμένο κόμβο, ενώ το Ναι/Όχι δε δημιουργεί κανένα φύλλο (βλέπε Σχήμα 4.11).



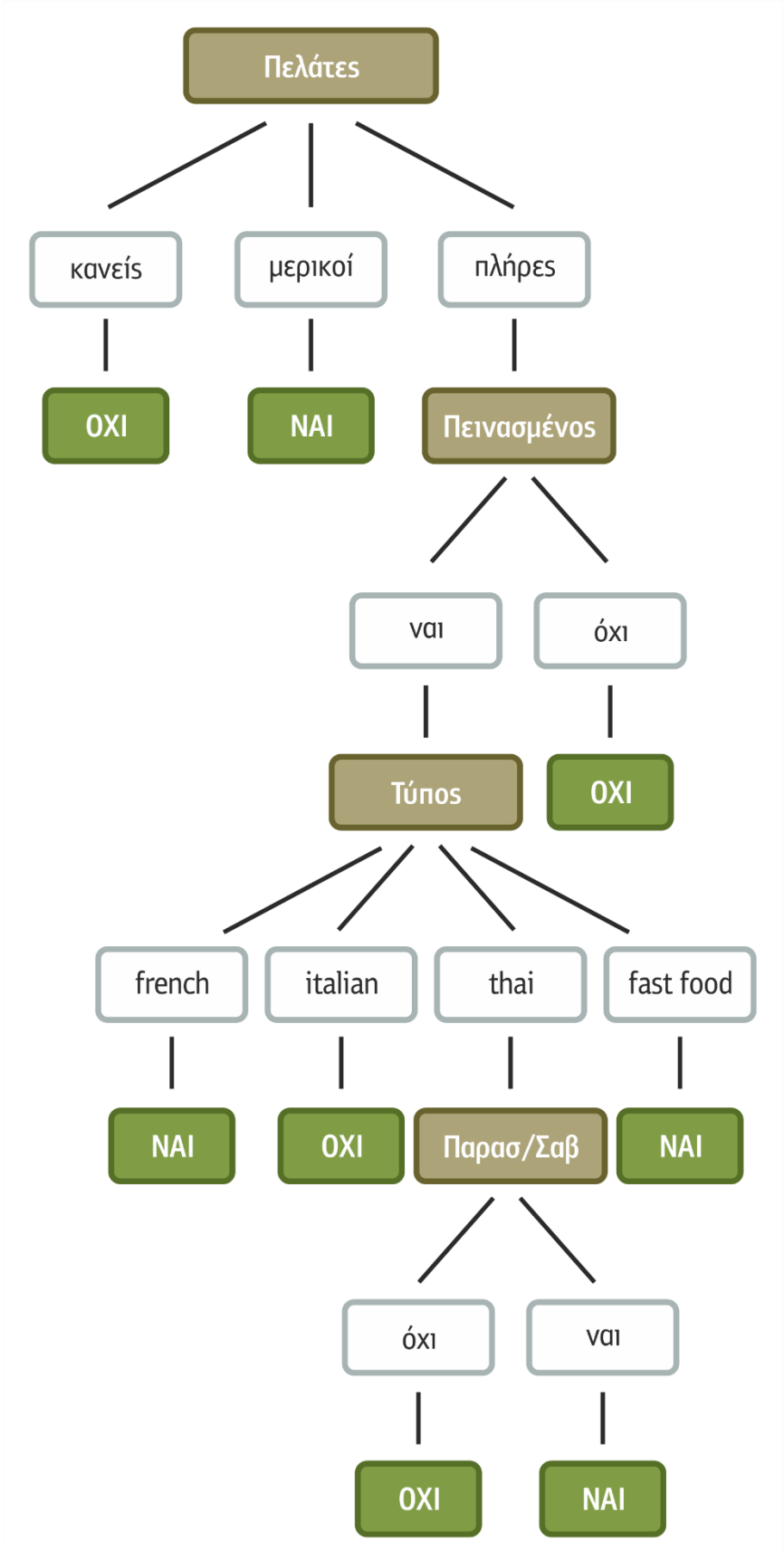
Σχήμα 4.11 Σύγκριση χαρακτηριστικών για το πρόβλημα αναμονής σε εστιατόριο

Αν ως ευρετικό κριτήριο θεωρεί ότι το καλύτερο χαρακτηριστικό είναι αυτό με το οποίο ο διαχωρισμός οδηγεί στα περισσότερα φύλλα, τότε ο αλγόριθμος θα επιλέξει το Πελάτες έναντι του Τύπος.

Προφανώς, η σύγκριση πρέπει να γίνει μεταξύ όλων των χαρακτηριστικών που μπορούν να εφαρμοστούν.

Έστω ότι το χαρακτηριστικό πελάτες είναι καλύτερο όλων και εφαρμόζεται στην κορυφή του δένδρου και στη συνέχεια το παραπάνω ευρετικό κριτήριο εφαρμόζεται σε κάθε κόμβο που πρέπει να διαχωριστεί.

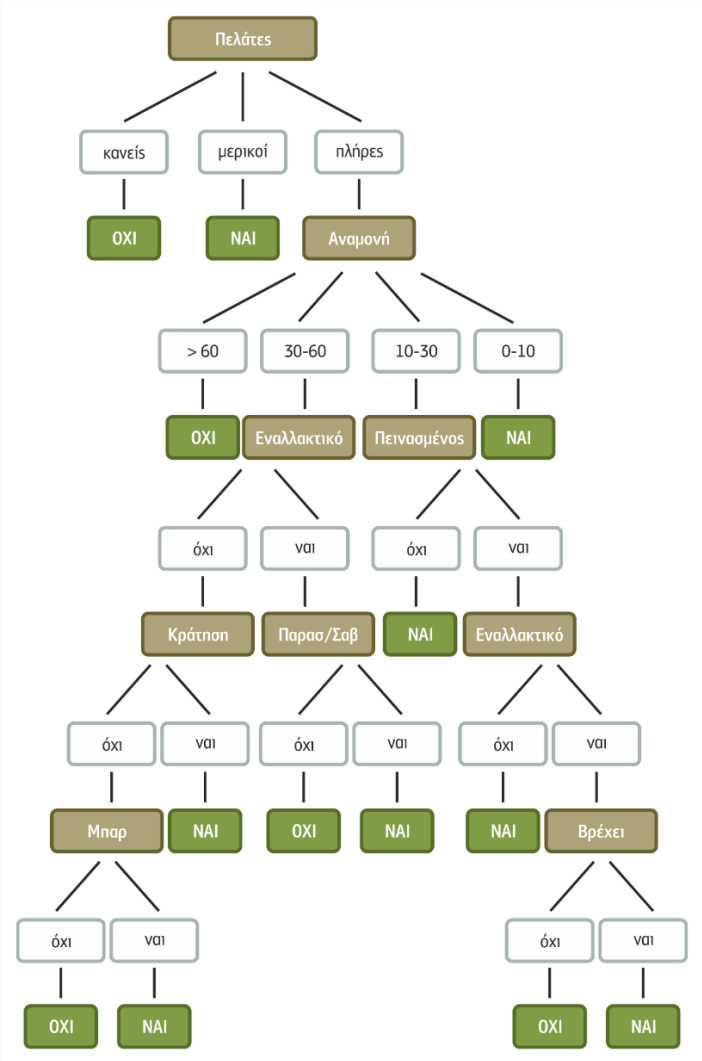
Το δένδρο που θα προκύψει θα είναι αυτό του σχήματος 4.12.


Πελάτες
κανείς
μερικοί
πλήρες
ΟΧΙ
ΝΑΙ
Πεινασμένος
ναι
όχι
Τύπος
ΟΧΙ
french
italian
thai
fast food
ΝΑΙ
ΟΧΙ
Παρασ/Σαβ
ΝΑΙ
όχι
ναι
ΟΧΙ
ΝΑΙ

Σχήμα 4.12 Δένδρο απόφασης για το πρόβλημα αναμονής σε εστιατόριο

Ανάλογα με το ευρετικό κριτήριο που έχει, ο ID3 θα δημιουργήσει και διαφορετικό ΔΑ. Αν, για παράδειγμα, το κριτήριο είναι να εφαρμοστούν όσο γίνεται περισσότερα χαρακτηριστικά, για να αποφευχθεί η πρόωρη σύγκλιση και να είναι πιο πλούσια η γνώση που θα προκύψει, πρέπει να επιλεγούν τα χαρακτηριστικά με διαφορετική σειρά.

Στο σχήμα 4.13 παρουσιάζεται ένα πιθανό ΔΑ για το πρόβλημα αναμονής σε εστιατόριο που λαμβάνει υπόψη του το συγκεκριμένο κριτήριο.



Σχήμα 4.13 Δένδρο απόφασης για το πρόβλημα αναμονής σε εστιατόριο που προσπαθεί να εξαντλήσει τα χαρακτηριστικά