ΔΕΝΔΡΑ ΑΠΟΦΑΣΕΩΝ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

ΕΙΣΑΓΩΓΗ

- Μια κατηγορία αλγορίθμων που χρησιμοποιούνται για την επίλυση προβλημάτων κατηγοριοποίησης είναι αυτή των Δένδρων Απόφασης (Decision Trees). Το μοντέλο κατηγοριοποίησης αυτής της κατηγορίας αλγορίθμων είναι μια δενδρική δομή. Μόλις χτιστεί η δενδρική δομή, εφαρμόζεται σε κάθε πλειάδα της Βάσης Δεδομένων και καταλήγει για κάθε μια από αυτές σε μια κατηγοριοποίηση.
- Η διαδικασία κατηγοριοποίησης χωρίζεται σε δύο φάσεις: (α) η κατασκευή του δένδρου και (β) η εφαρμογή του στη Βάση Δεδομένων..

ΕΙΣΑΓΩΓΗ

- Η τεχνικές δένδρων αποφάσεων βασίζονται στη διαίρεση του χώρου αναζήτησης σε ορθογ ώνιες περιοχές (χρήση της τεχνικής του «διαίρει και βασίλευε»). Κάθε πλειάδα της Βάσης Δεδομένων τοποθετείται με βάση την περιοχή όπου ανήκει

ΟΡΙΣΜΟΣ

Ορισμός : Έστω μια Βάση Δεδομένων D = $\{t_1, t_2,\ldots, t_n\}$, όπου $t_i = t_{i1}, t_{i2},\ldots, t_{ih}$ και έστω ότι το σχήμα της Βάσης Δεδομένων περιέχει τα εξής χαρακτηριστικά (πεδία) $\{A_1, A_2, \ldots, A_h\}$.

Επίσης, έστω ότι έχουμε ένα σύνολο κατηγοριών C = $\{C_1, C_2, \ldots, C_m\}$. Ένα δένδρο απόφασης ή δένδρο κατηγοριοποίησης είναι μια δενδρική δομή που συσχετίζεται με το D και έχει τις εξής ιδιότητες:

1. Κάθε εσωτερικός κόμβος παίρνει το όνομα του από ένα γνώρισμα, A_i
2. Κάθε τόξο παίρνει το όνομα του από ένα κατηγόρημα το οποίο μπορεί να εφαρμοστεί στο γνώρισμα που συνδέεται με τον πατέρα–κόμβο
3. Κάθε φύλλο έχει ως όνομα μια κατηγορία C_j.

ΟΡΙΣΜΟΣ

- **Ορισμός** : Ένα δένδρο απόφασης, είναι ένα δένδρο όπου η ρίζα και κάθε εσωτερικός κόμβος έχει χαρακτηριστεί με μια ερώτηση. Τα τόξα που προέρχονται από κάθε κόμβο αντιπροσωπεύουν κάθε πιθανή απάντηση στη σχετική ερώτηση. Κάθε φύλλο αντιπροσωπεύει μια πρόβλεψη της λύσης στο πρόβλημα που εξετάζεται. Στα προβλήματα κατηγοριοποίηση, η πρόβλεψη είναι η κατηγορία της πλειάδας που εξετάζεται.

ΟΡΙΣΜΟΣ

- Ένα δένδρο απόφασης κατασκευάζεται συνήθως σε δύο φάσεις:

Στην πρώτη φάση, τη φάση της ανάπτυξης, κατασκευάζεται ένα μεγάλο δένδρο. Το δένδρο αυτό απεικονίζει τις πλειάδες τις Βάσης Δεδομένων με μεγάλη ακρίβεια. Για παράδειγμα, το δένδρο μπορεί να περιέχει φύλλα για μεμονωμένες πλειάδες της Βάσης Δεδομένων.

Στη δεύτερη φάση, η οποία ονομάζεται φάση κλαδέματος, προσδιορίζεται το τελικό μέγεθος του δένδρου. Οι κανόνες που μπορούν να παραχθούν από το δένδρο πριν τη φάση του κλαδέματος είναι αρκετά εξειδικευμένοι.

ΟΡΙΣΜΟΣ

- Ο πίνακας περιέχει τα δεδομένα που θα χρησιμοποιήσουμε παρακάτω στο παράδειγμα ώστε να γίνει πιο κατανοητό.
- Το παράδειγμα αυτό υποθέτει ότι το πρόβλημα μας είναι να κατηγοριοποιήσουμε ενήλικές σαν short, medium ή tall. Ο πίνακας περιέχει μια στήλη για τα ύψη (σε μέτρα). Οι τελευταίες δύο στήλες του πίνακα παρουσιάζουν δύο κατηγοριοποιήσεις που θα μπορούσαν να γίνουν (κατηγοριοποίηση 1 και κατηγοριοποίηση 2). Η κατηγοριοποίηση 1 χρησιμοποιεί την απλή διαίρεση που φαίνεται παρακάτω:

ΟΡΙΣΜΟΣ

Κατασκευή του δέντρου

1. ξεκίνα με *έναν* κόμβο που περιέχει *όλες* τις εγγραφές

2. *διάσπαση* του κόμβου (μοίρασμα των εγγραφών) με βάση μια συνθήκη-διαχωρισμού σε κάποιο από τα γνωρίσματα

3. Αναδρομική κλήση του 2 σε κάθε κόμβο (top-down, recursive, divide-and-conquer προσέγγιση)

4. Αφού κατασκευαστεί το δέντρο, κάποιες βελτιστοποιήσεις (tree pruning)

Το βασικό θέμα είναι

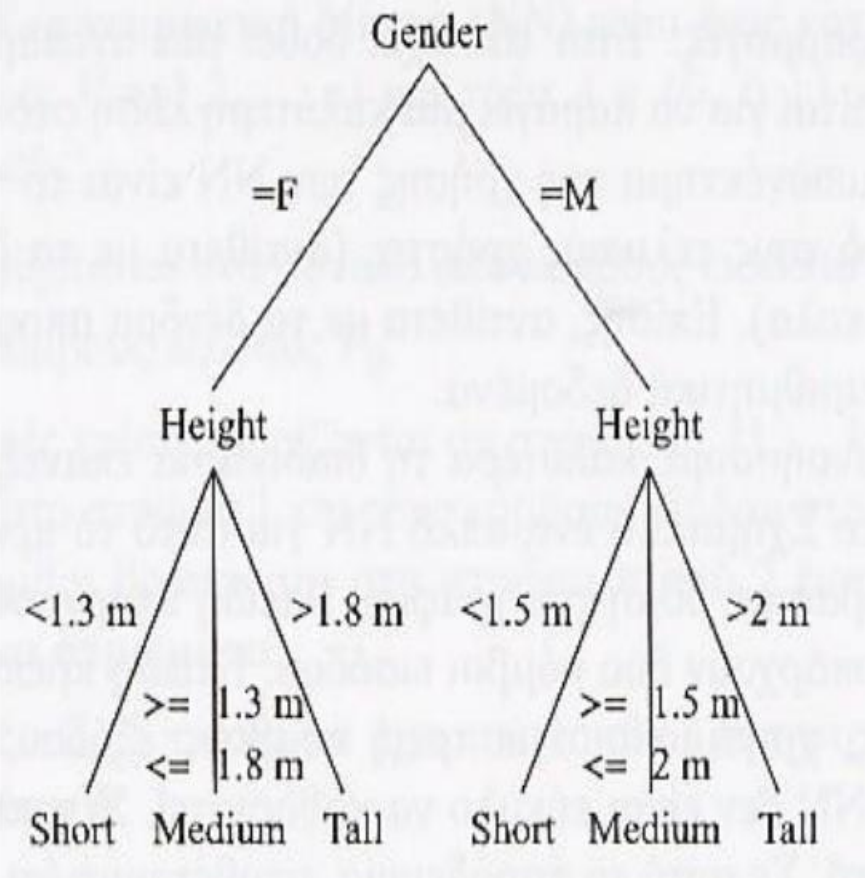
Ποιο γνώρισμα-συνθήκη διαχωρισμού να χρησιμοποιήσουμε για τη διάσπαση των εγγραφών κάθε κόμβου;

ΟΡΙΣΜΟΣ

$2\mu \leq \text{ύψος} \rightarrow \psi\eta\lambda\acute{o}\varsigma,$
$1{,}7\mu \leq \text{ύψος} < 2\mu \rightarrow \mu\acute{\varepsilon}\tau\rho\iota\text{ος},$
$\text{Ύψος} \leq 1{,}7 \rightarrow \kappa\text{οντός}$

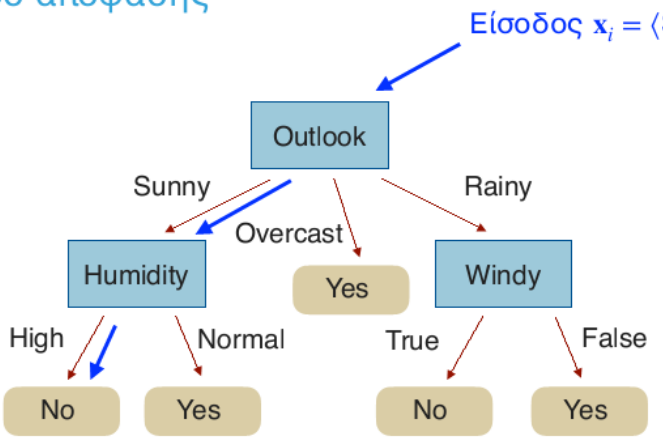
Όνομα	Φύλο	Ύψος (μ)	Κατηγοριοποίηση 1	Κατηγοριοποίηση 2
Kristina	Θ	1,6	Κοντός	Μέτριος
Jim	Α	2	Ψηλός	Μέτριος
Maggie	Θ	1,9	Μέτριος	Ψηλός
Martha	Θ	1,88	Μέτριος	Ψηλός
Stephanie	Θ	1,7	Κοντός	Μέτριος
Bob	Α	1,85	Μέτριος	Μέτριος
Kathy	Θ	1,6	Κοντός	Μέτριος
Dave	Α	1,7	Κοντός	Μέτριος
Worth	Α	2,2	Ψηλός	Ψηλός
Steven	Α	2,1	Ψηλός	Ψηλός
Debbie	Θ	1,8	Μέτριος	Μέτριος
Todd	Α	1,95	Μέτριος	Μέτριος
Kim	Θ	1,9	Μέτριος	Ψηλός
Amy	Θ	1,8	Μέτριος	Μέτριος
Wynette	Θ	1,75	Μέτριος	Μέτριος

ΟΡΙΣΜΟΣ


Gender
=F
=M
Height
Height
<1.3 m
>1.8 m
>= 1.3 m
<= 1.8 m
<1.5 m
>2 m
>= 1.5 m
<= 2 m
Short
Medium
Tall
Short
Medium
Tall

ΠΑΡΑΔΕΙΓΜΑ

Δέντρο απόφασης



έξοδος $y_i = \langle \text{No} \rangle$

Outlook	Temperature	Humidity	Windy	PlayTennis
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	TRUE	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

Σε κάθε εσωτερικό κόμβο ελέγχεται η τιμή του χαρακτηριστικού x_i

Σε κάθε διακλάδωση επιλέγεται μία τιμή του χαρακτηριστικού x_i

Σε κάθε φύλλο αποδίδεται μία ετικέτα y στο στοιχείο

ΠΛΕΟΝΕΚΤΗΜΑΤΑ

- Τα δένδρα απόφασης, ως τεχνική κατηγοριοποίησης, έχουν αρκετά πλεονεκτήματα. Ένα από τα πιο βασικά πλεονεκτήματα είναι το ότι μπορούν να χρησιμοποιηθούν εύκολα και αποτελεσματικά. Επίσης, ένα δένδρο απόφασης μπορεί να εξάγει κανόνες οι οποίοι μπορούν εύκολα να κατανοηθούν και να ερμηνευτούν από το χρήστη. Ένα ακόμη βασικό πλεονέκτημα των δένδρων απόφασης είναι το ότι μπορούν να χρησιμοποιηθούν με επιτυχία σε μεγάλες Βάσεις Δεδομένων και αυτό επειδή το μέγεθος της Βάσης Δεδομένων είναι ανεξάρτητο από το μέγεθος του δένδρου. Κάθε πλειάδα προς κατηγοριοποίηση πρέπει να περάσει από το δένδρο.

ΜΕΙΟΝΕΚΤΗΜΑΤΑ

- Υπάρχουν και αρκετά μειονεκτήματα όταν εφαρμόζουμε τα δένδρα απόφασης για να επιλύσουμε προβλήματα κατηγοριοποίησης. Ένα από τα βασικά μειονεκτήματα τους είναι ότι δεν μπορούν να χειριστούν συνεχή δεδομένα. Επίσης, τα δένδρα απόφασης προϋποθέτουν ότι ο χώρος του πεδίου διαιρείται σε ορθογώνιες περιοχές Άλλου είδους σχήματα δε μπορούν να χειριστούν από αυτή την τεχνική. Τα ελλιπή δεδομένα είναι ένα ακόμη πρόβλημα για τα δένδρα απόφασης και αυτό γιατί δε μπορούν να βρεθούν οι σωστές διακλαδώσεις για να ακολουθηθούν

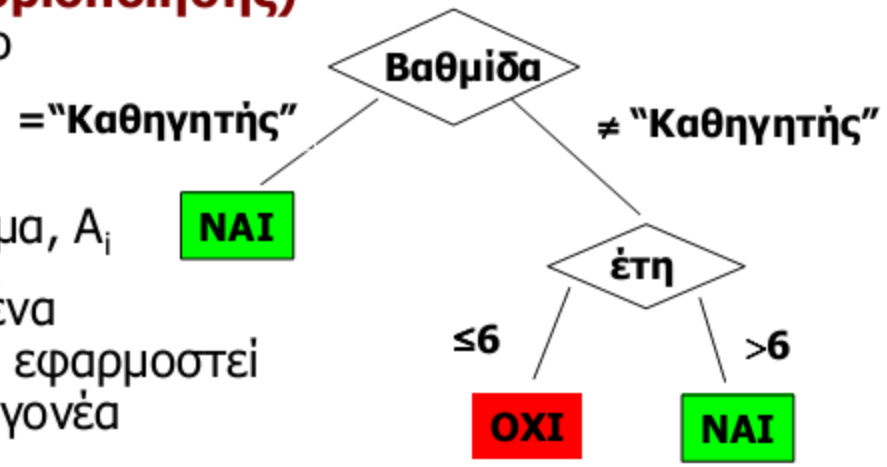
ΓΕΝΙΚΗ ΜΟΡΦΗ

Δοθέντων:

- μιας βάσης δεδομένων D = $\{t_1, \ldots, t_n\}$ όπου $t_i=<t_{i1}, \ldots, t_{ih}>$
- του σχήματος της ΒΔ $\{A_1, A_2, \ldots, A_h\}$
- ενός συνόλου κλάσεων $C=\{C_1, \ldots., C_m\}$

Δένδρο απόφασης (ή κατηγοριοποίησης) είναι ένα δένδρο συσχετισμένο με τη D έτσι ώστε

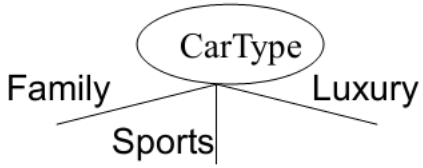
- Κάθε εσωτερικός κόμβος έχει ως ετικέτα ένα γνώρισμα, A_i
- Κάθε τόξο έχει ως ετικέτα ένα κατηγόρημα που μπορεί να εφαρμοστεί στο γνώρισμα του κόμβου-γονέα
- Κάθε φύλλο (τερματικός κόμβος) έχει ως ετικέτα μια κλάση, C_j



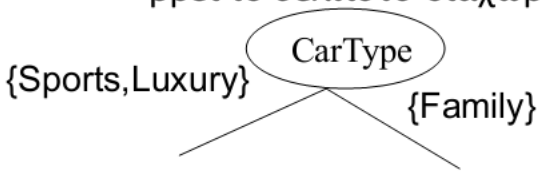
ΔΙΑΧΩΡΙΣΜΟΙ

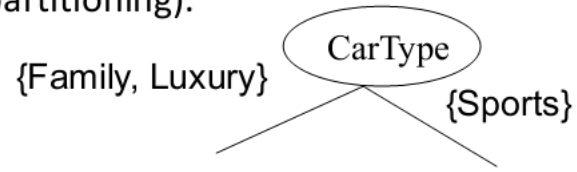
1. Διαχωρισμός σε πεδίο με διακριτές τιμές

- Πολλαπλός διαχωρισμός: Χρησιμοποίησε τόσες διασπάσεις όσες οι διαφορετικές τιμές

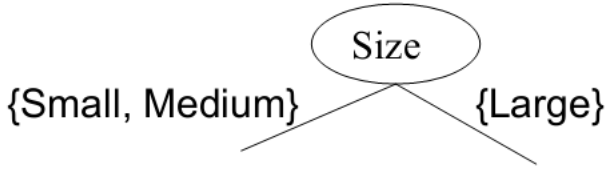


- Δυαδικός Διαχωρισμός: Χωρίζει τις τιμές σε δύο υποσύνολα. Πρέπει να βρει το *βέλτιστο* διαχωρισμό (partitioning).





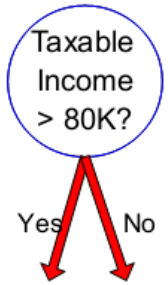
- Αν έχω κ τιμές, υπάρχουν $2^{κ-1}$-1 τρόποι διαχωρισμού
- Όταν έχω <u>διάταξη</u>, πρέπει ο διαχωρισμός να τη διατηρεί



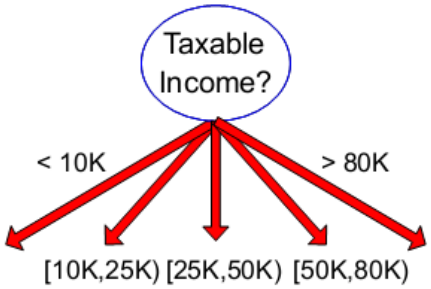
ΔΙΑΧΩΡΙΣΜΟΣ

1. Διαχωρισμός σε πεδίο με συνεχείς τιμές

- Διακριτοποίηση ώστε να προκύψει ένα διατεταγμένο κατηγορικό γνώρισμα
  - Ταξινόμηση των τιμών και χωρισμός τους σε περιοχές καθορίζοντας n – 1 σημεία διαχωρισμού, απεικόνιση όλων των τιμών μιας περιοχής στην ίδια κατηγορική τιμή
    - Στατικό – μια φορά στην αρχή
    - Δυναμικό – εύρεση των περιοχών πχ έτσι ώστε οι περιοχές να έχουν το ίδιο διάστημα ή τις ίδιες συχνότητες εμφάνισης ή με χρήση συσταδοποίησης
- Δυαδική Απόφαση: (A < v) or (A >= v)
  - εξετάζει όλους τους δυνατούς διαχωρισμούς (τιμές του v) και επιλέγει τον καλύτερο – υπολογιστικά βαρύ



Δυαδικός διαχωρισμός



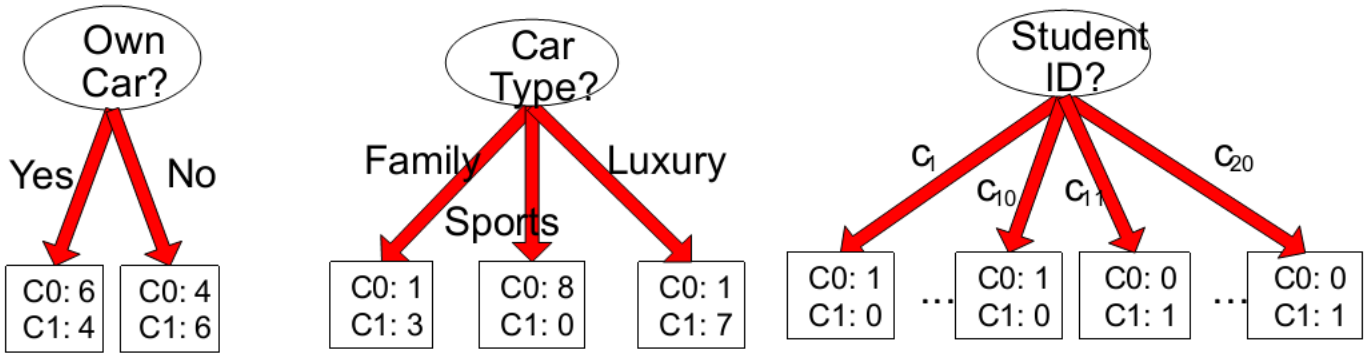
Πολλαπλός διαχωρισμός

- Καλύτερο=Ισοπληθείς ομάδες
  - Διάταξε τις τιμές σε αύξουσα διάταξη
  - Βρες τις ενδιάμεσες γειτονικές τιμές a_i και a_{i+1} (median όχι average)
  - Υπολόγισε το σημείο διαχωρισμού στη μέση των διαχωριστικών τιμών $(a_i+a_{i+1})/2$

ΔΙΑΧΩΡΙΣΜΟΣ

2. Βέλτιστος Διαχωρισμός

- Πριν το διαχωρισμό: 10 εγγραφές από κάθε κλάση (0,1)



- Ποια από τις 3 διασπάσεις να προτιμήσουμε; (Δηλαδή, ποια συνθήκη ελέγχου είναι καλύτερη;)
- => ορισμός κριτηρίου βέλτιστου διαχωρισμού

ΚΑΘΑΡΟΤΗΤΑ ΚΟΜΒΟΥ

- Για κάθε κόμβο n, μετράμε την καθαρότητα του, I(n)
  - Έστω μια διάσπαση ενός κόμβου (parent) με N εγγραφές σε k παιδιά u_i
  - Έστω N(ui) ο αριθμός εγγραφών κάθε παιδιού ($\Sigma N(u_i) = N$)
- Για να χαρακτηρίσουμε μια διάσπαση, κοιτάμε το κέρδος, δηλαδή τη διαφορά μεταξύ της καθαρότητας του γονέα (πριν τη διάσπαση) και των παιδιών του (μετά τη διάσπαση)

$$\Delta = I(parent) - \sum_{i=1}^{k} \frac{N(u_i)}{N} I(u_i)$$

Βάρος (εξαρτάται από τον αριθμό εγγραφών $N(u_i)$ του κάθε παιδιού)

- "Καλύτερη" διάσπαση = μεγαλύτερο Δ

ΚΑΤΑΛΛΗΛΟΤΗΤΑ

- Ποιο είναι το καλύτερο;
  - Αυτό που θα οδηγήσει στο μικρότερο δένδρο
  - Ένας ευριστικός κανόνας (heuristic): επιλέγουμε το γνώρισμα που παράγει τους πιο "αγνούς" κόμβους. Για το σκοπό αυτό, χρησιμοποιείται μια **συνάρτηση καταλληλότητας** (fitness function).
- Στρατηγική: επιλέγουμε το γνώρισμα που μεγιστοποιεί τη συνάρτηση καταλληλότητας
  - Χαρακτηριστικές συναρτήσεις καταλληλότητας:
    - Κέρδος πληροφορίας – Gain (ID3)
    - Λόγος κέρδους πληροφορίας – GainRatio (C4.5)
    - gini index (SPRINT)

ΠΛΗΡΟΦΟΡΙΑ

- Είναι η πληροφορία που απαιτείται για την κατηγοριοποίηση ενός δείγματος
- Αν s τα δείγματα και m οι κατηγορίες

$$I(s_1, \dots, s_m) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

- s_i είναι ο αριθμός των δειγμάτων στην κατηγορία C_i και p_i η πιθανότητα ένα δείγμα να ανήκει στην κατηγορία C_i ($p_i=s_i/s$)

ΠΛΗΡΟΦΟΡΙΑ

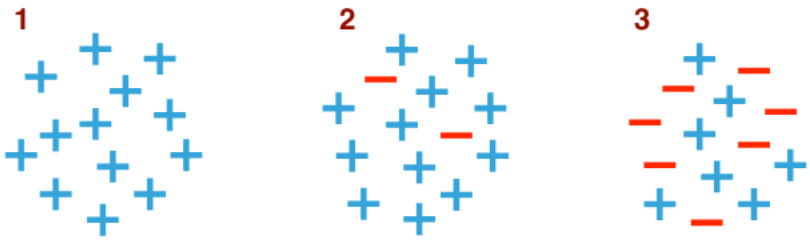
Μέθοδος

- **Τυχαία**: Επίλεξε ένα χαρακτηριστικό χωρίς κάποιο συγκεκριμένο κριτήριο
- **Λιγότερες τιμές**: Επίλεξε το χαρακτηριστικό με τη μικρότερη πληθικότητα του πεδίου τιμών
- **Περισσότερες τιμές**: Επίλεξε το χαρακτηριστικό με τη μεγαλύτερη πληθικότητα του πεδίου τιμών
- **Μεγαλύτερο όφελος**: Επίλεξε το χαρακτηριστικό με το μεγαλύτερο κέρδος πληροφορίας (information gain)

Εντροπία

$$E = \sum_i - p_i \log_2 p_i$$

- Μέτρο της καθαρότητας ενός συνόλου παραδειγμάτων
  - μικρότερη εντροπία, μεγαλύτερη καθαρότητα
- Μείωση εντροπίας, κέρδος πληροφορίας



- Ποια από τις κατανομές έχει μεγαλύτερο κέρδος πληροφορίας;

$$E_1 = -1 \ \log_2 1 - 0 \ \log_2 0 = 0$$

$$E_2 = -0.133 \ \log_2 0.133 - 0.87 \ \log_2 0.87 \simeq 0{,}565$$

$$E_3 = -0.5 \ \log_2 0.5 - 0.5 \ \log_2 0.5 = 1$$

Επιλέγουμε το χαρακτηριστικό που γνωρίζοντας την τιμή του πετυχαίνουμε τη μεγαλύτερη μείωση της εντροπίας

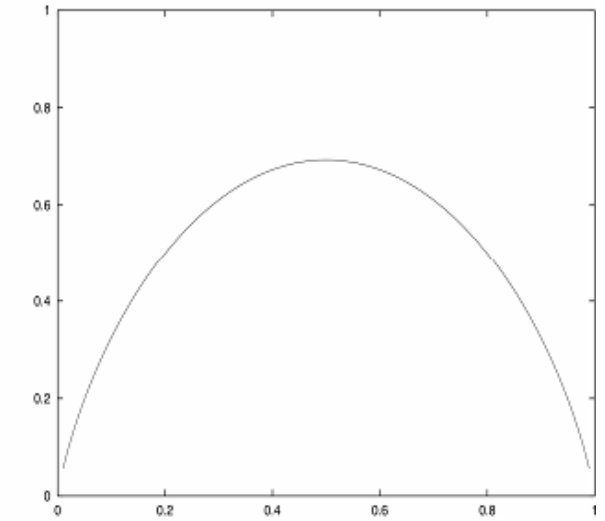
ΕΝΤΡΟΠΙΑ ΚΟΜΒΟΥ

- Έστω πιθανότητες p_1, p_2, .., p_s των οποίων το άθροισμα είναι 1.
  Η **Εντροπία** ορίζεται ως εξής:

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} \left(p_i \log\left(1/p_i \right) \right)$$

η βάση του λογάριθμου δεν προσδιορίζεται (συνήθως, 10 ή 2)

- Η εντροπία είναι ποσοτικοποίηση της τυχαιότητας (έκπληξης, αβεβαιότητας)
- Ο στόχος της κατηγοριοποίησης
  - καθόλου έκπληξη
  - εντροπία = 0



H(p,1-p)

ΕΝΤΡΟΠΙΑ ΔΙΑΣΠΑΣΗΣ

- Έστω ότι από τα s δείγματα που έχω συνολικά θεωρώ ένα υποσύνολο S_j με τα δείγματα που έχουν τιμή a_j για το γνώρισμα A
- Η εντροπία του A ορίζεται ως εξής:

$$E(A) = \sum_{j=1}^{v} \frac{\sum_{i=1}^{m} s_{ij}}{s} I(s_{1j}, \dots, s_{mj})$$

- όπου $$I(s_{1j}, \dots, s_{mj}) = -\sum_{i=1}^{m} \frac{s_{ij}}{|S_j|} log_2 \left(\frac{s_{ij}}{|S_j|}\right)$$

- Το κέρδος πληροφορίας (information gain) από την επιλογή του A είναι:

$$Gain(A) = E(S) - \sum_{k \text{ οι διαχωρισμοί του } S} E(S_k)$$

ΠΑΡΑΔΕΙΓΜΑ

- Έχω συνολικά 100 δείγματα πιστωτών (καλοί και κακοί)
- 40 από αυτά είναι άντρες και 60 από αυτά είναι γυναίκες
- Θέλω να υπολογίσω την εντροπία του γνωρίσματος ‘φύλο’ ως προς την τελική απόφαση που διαχωρίζει τα δείγματα σε καλούς και κακούς πιστωτές

	i=1:Άντρες	i=2:Γυναίκες
j=1:Καλοί	$s_{1,1}=5$	$s_{2,1}=50$
j=2:Κακοί	$s_{1,2}=35$	$S_{2,2}=10$

$$I(s_{11},s_{21}) = -\left(\frac{5}{40}log_2\left(\frac{5}{40}\right)+\frac{50}{60}log_2\left(\frac{50}{60}\right)\right) = 0.594$$

$$I(s_{12},s_{22}) = -\left(\frac{35}{40}log_2\left(\frac{35}{40}\right)+\frac{10}{60}log_2\left(\frac{10}{60}\right)\right) = 0.599$$

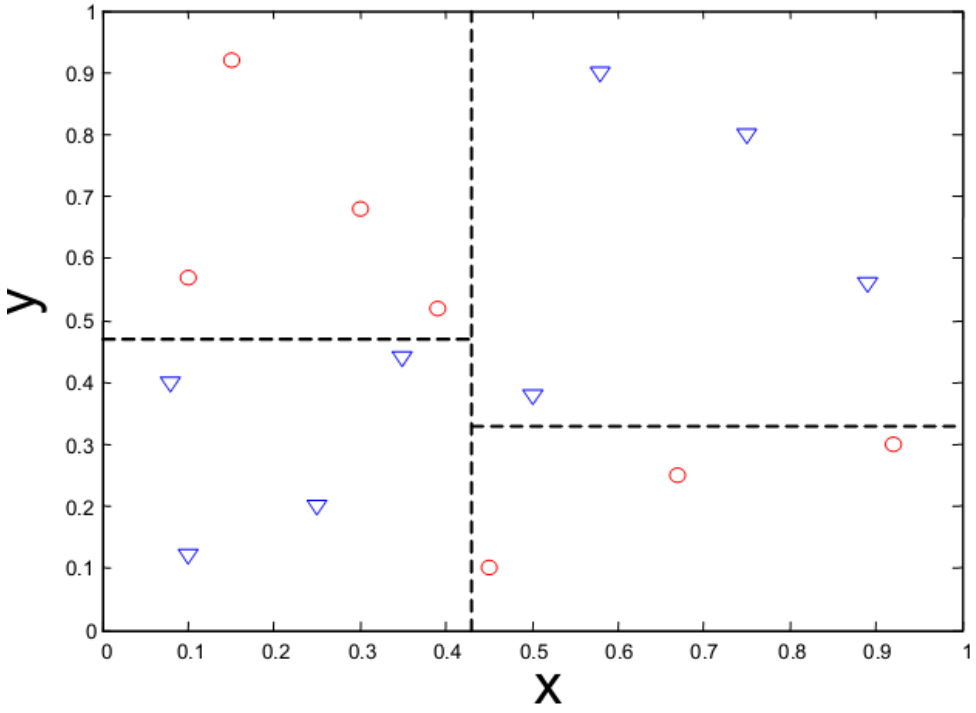
$$E(A) = \frac{50+5}{100}0.594 + \frac{35+10}{100}0.599 = 0.597$$

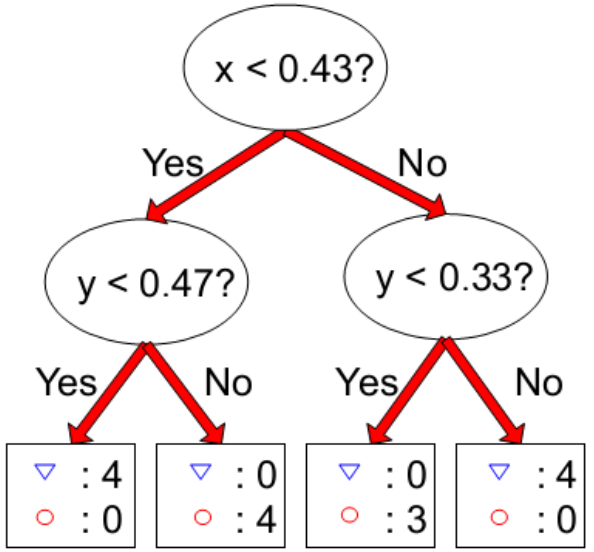
$$E(A) = \sum_{j=1}^{v}\frac{\sum_{i=1}^{m}s_{ij}}{s}I(s_{1j},\dots,s_{mj})$$

$$I(s_{1j},\dots,s_{mj}) = -\sum_{i=1}^{m}\frac{s_{ij}}{|S_j|}log_2\left(\frac{s_{ij}}{|S_j|}\right)$$

ΔΙΑΧΩΡΙΣΜΟΣ

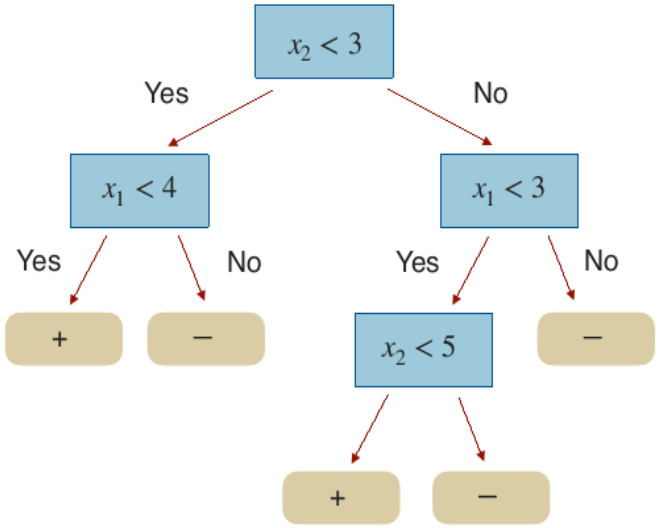
Όταν η συνθήκη ελέγχου περιλαμβάνει μόνο ένα γνώρισμα τη φορά τότε το Decision boundary είναι παράλληλη στους άξονες (τα decision boundaries είναι *ορθογώνια παραλληλόγραμμα*)



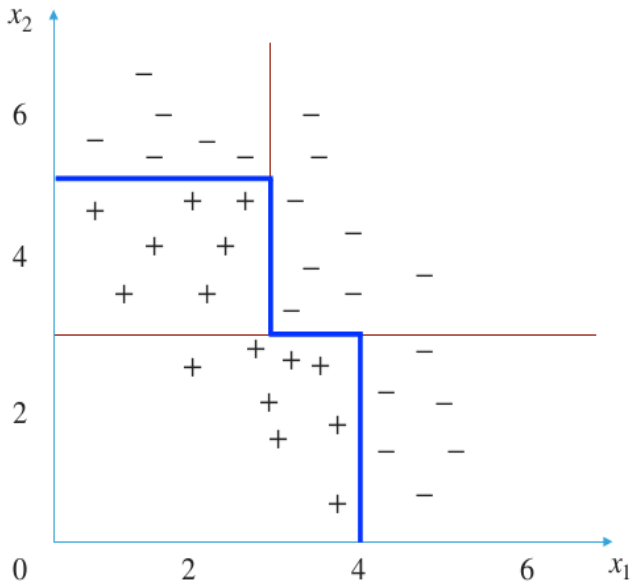


ΚΑΜΠΥΛΕΣ ΔΙΑΧΩΡΙΣΜΟΥ

Δέντρο απόφασης



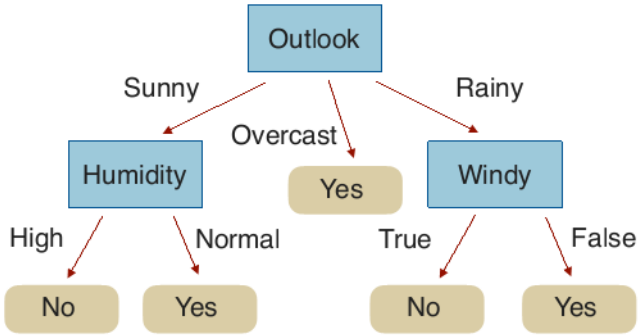
Καμπύλη διαχωρισμού



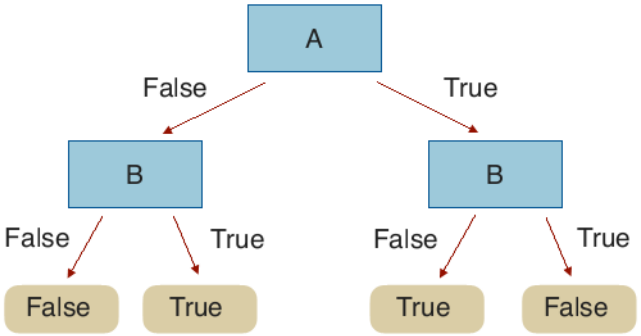
- Η καμπύλη διαχωρισμού χωρίζει το χώρο χαρακτηριστικών σε (υπερ-)κύβους παράλληλους των αξόνων
- Κάθε (υπερ-)κυβική επιφάνεια αντιστοιχίζεται σε μία ετικέτα - ή (στη γενική περίπτωση) σε μία κατανομή πιθανοτήτων πάνω στις ετικέτες

ΕΚΦΡΑΣΤΙΚΟΤΗΤΑ

Ικανότητα αναπαράστασης

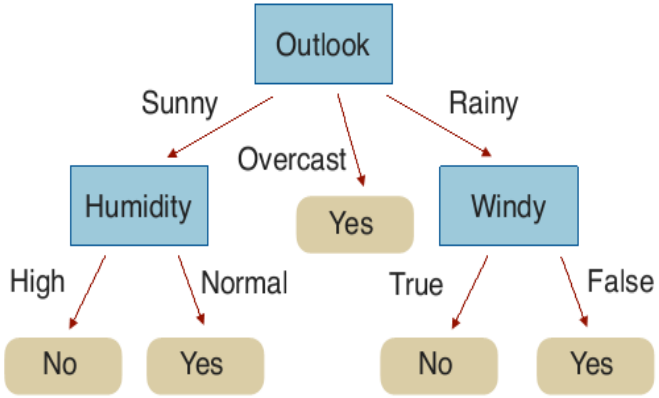


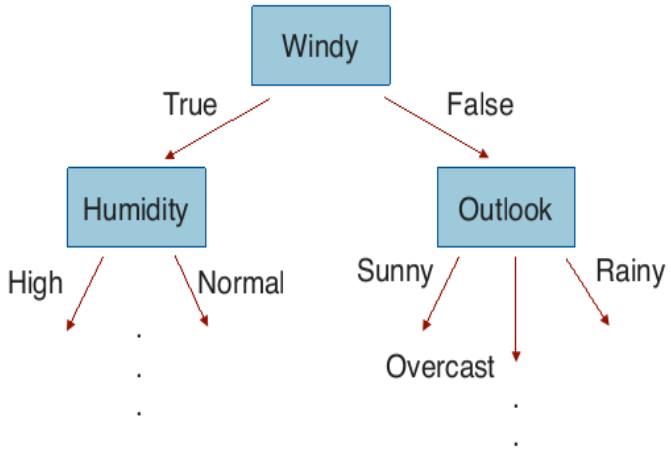
Δένδρο απόφασης για XOR



Outlook	Temperature	Humidity	Windy	PlayTennis
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	TRUE	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

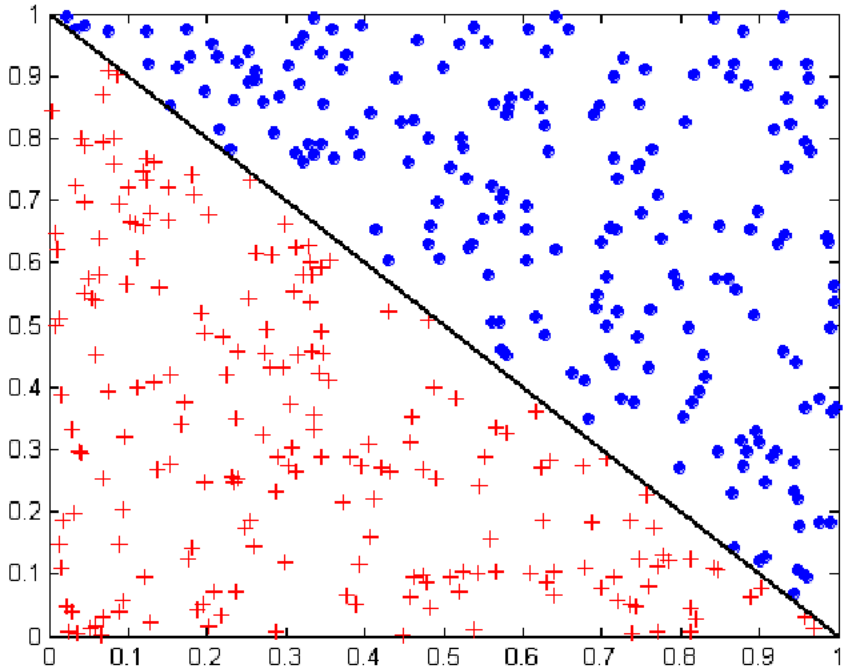
ΕΚΦΡΑΣΤΙΚΟΤΗΤΑ



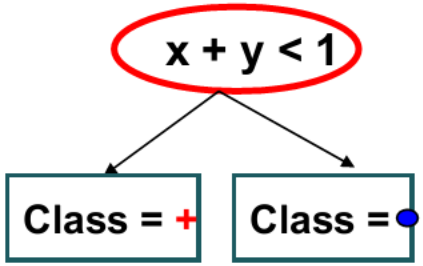


Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

ΠΛΑΓΙΟ ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ



Oblique (πλάγιο) Δέντρο Απόφασης



- Οι συνθήκες ελέγχου μπορούν να περιλαμβάνουν περισσότερα από ένα γνωρίσματα
- Μεγαλύτερη εκφραστικότητα
- Η εύρεση βέλτιστων συνθηκών ελέγχου είναι υπολογιστικά ακριβή

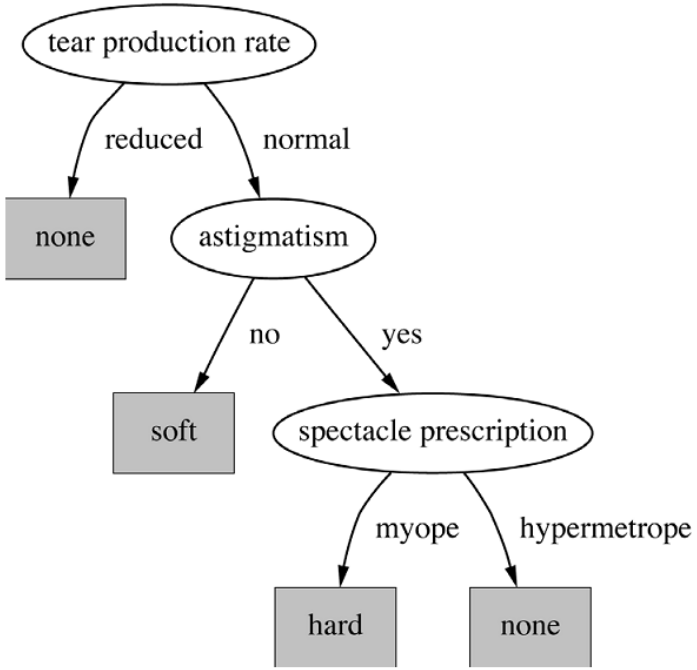
ΠΑΡΑΔΕΙΓΜΑ

Επιλογή φακών επαφής

Ηλικία	Πάθηση	Αστιγματισμός	Δάκρυα	Είδος φακών
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

ΠΑΡΑΔΕΙΓΜΑ

```
If tear production rate = reduced then recommendation = none

If age = young and astigmatic = no
    and tear production rate = normal then recommendation = soft

If age = pre-presbyopic and astigmatic = no
    and tear production rate = normal then recommendation = soft

If age = presbyopic and spectacle prescription = myope
    and astigmatic = no  then recommendation = none

If spectacle prescription = hypermetrope and astigmatic = no
    and tear production rate = normal then recommendation = soft

If spectacle prescription = myope and astigmatic = yes
    and tear production rate = normal then recommendation = hard

If age young and astigmatic = yes
    and tear production rate = normal then recommendation = hard

If age = pre-presbyopic
    and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none

If age = presbyopic and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
```



Αλγόριθμος ID3

Είσοδος: Δείγματα εκπαίδευσης τα οποία παρουσιάζονται με διακριτές τιμές γνωρισμάτων.

Έξοδος: Δέντρο Απόφασης

Διαδικασία:

Βήμα 1ο: Το δέντρο ξεκινάει με έναν μόνο κόμβο που αντιπροσωπεύει ολόκληρο το σύνολο των δεδομένων εκπαίδευσης.

Βήμα 2ο: Αν τα δείγματα είναι όλα της ίδιας κατηγορίας, τότε ο κόμβος γίνεται φύλλο και προστίθεται η ετικέτα της κατηγορίας.

Βήμα 3ο: Ο αλγόριθμος χρησιμοποιεί ένα μέτρο εντροπίας, γνωστό σαν κέρδος πληροφορίας, για την επιλογή των γνωρισμάτων που διαχωρίζουν καλύτερα τα δείγματα στις διαφορετικές κατηγορίες. Στην συνέχεια το κέρδος πληροφορίας υπολογίζεται για κάθε γνώρισμα. Το γνώρισμα με το μέγιστο κέρδος πληροφορίας επιλέγεται σαν γνώρισμα ελέγχου.

Αλγόριθμος ID3

Επιλογή γνωρίσματος.

Έστω S το σύνολο των s δειγμάτων δεδομένων. Υποθέτοντας ένα σύνολο m κατηγοριών C_i (για i = 1,2,..., η), η αναμενόμενη πληροφορία που απαιτείται για την κατηγοριοποίηση του ενός δείγματος δίνεται από την εξίσωση:

$$I(s_1, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i)$$

όπου S_i είναι ο αριθμός των δειγμάτων στην κατηγορία C_i και P_i είναι η πιθανότητα να χρησιμοποιηθεί για το διαχωρισμό του S σε ν υποσύνολα $\{S_1, S_2,..., S_v)$, όπου S_j περιέχει εκείνα τα δείγματα του S που έχουν την τιμή α για το γνώρισμα Α.

Αλγόριθμος ID3

Βήμα 4ο: Ένας κόμβος δημιουργείται και χαρακτηρίζεται γνώρισμα ελέγχου (test attribute),όσο δημιουργούνται κλαδιά για κάθε τιμή του. Στην συνέχεια το δείγμα δεδομένων διαχωρίζεται αναλόγως

Βήμα 5ο: Ο αλγόριθμος εφαρμόζεται συνεχώς για τη

μορφοποίηση ενός δέντρου απόφασης με βάση τα δείγματα σε κάθε προκαθορισμένη κατηγορία.

Αλγόριθμος ID3

Ο συνεχής διαχωρισμός σταματάει μόνο όταν κάποια από τις παρακάτω συνθήκες ικανοποιείται:

1. Όλα τα δείγματα του δοσμένου κόμβου ανήκουν στην ίδια κατηγορία, ή
2. Δεν υπάρχουν άλλα γνωρίσματα με βάση τα οποία τα δείγματα θα μπορούσαν να διαχωριστούν περαιτέρω, ή
3. Δεν υπάρχουν μη κατηγοριοποιημένα δείγματα για το κλαδί του γνωρίσματος ελέγχου.

Αλγόριθμος ID3

Name	Gender	Height	Output1
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium
Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} \left(p_i \log\left(1/p_i \right) \right)$$

- Αρχική κατάσταση εντροπίας:

H(D) = 4/15 log(15/4) + 8/15 log(15/8) + 3/15 log(15/3) = 0.4384

Short Medium Tall

Αλγόριθμος ID3

- Κέρδος αν γίνει διάσπαση στο gender:
  - Gender='F':
    3/9 log(9/3) + 6/9 log(9/6)=0.2764

    9 F 6 M
  - Gender='M':
    1/6 log(6/1) + 2/6 log(6/2)
    + 3/6 log(6/3) = 0.4392

    1 M - Shot 2 M - Medium

    3 M - Tall
  - Weighted sum:
    (9/15)(0.2764) +
    (6/15)(0.4392) = 0.3415

    9 F 6 M
- Gain: 0.4384–0.3415 = **0.0969**

Name	Gender	Height	Output1
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium
Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

Αλγόριθμος ID3

Ο αλγόριθμος κατασκευής δένδρου απόφασης ID3 έχει τα εξής βήματα:

1. Υπολόγισε το πληροφοριακό κέρδος κάθε μεταβλητής.
2. Θέσε ως ρίζα του δένδρου τη μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος.
3. Δημιούργησε τόσα κλαδιά όσες και οι διακριτές τιμές της μεταβλητής.
4. Χώρισε το σύνολο δεδομένων σε τόσα υποσύνολα όσα και οι διακριτές τιμές της μεταβλητής που επιλέχθηκε.
5. Επέλεξε μια τιμή-υποσύνολο, που δεν έχει ήδη επιλεχθεί. Αν στην τρέχουσα τιμή – υποσύνολο αντιστοιχεί μόνο μια τιμή κλάσης, πήγαινε στο βήμα 6, αλλιώς στο βήμα 7.
6. Βάλε την τιμή κλάσης ως φύλλο και προχώρησε στην επόμενη τιμή μεταβλητής-υποσύνολο και πήγαινε στο βήμα 5.
7. Υπολόγισε το πληροφοριακό κέρδος των υπόλοιπων μεταβλητών για το συγκεκριμένο υποσύνολο.
8. Επέλεξε τη μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος και πρόσθεσε έναν νέο κόμβο στον κλάδο που αντιστοιχεί στην τρέχουσα τιμή-υποσύνολο.
9. Επανέλαβε από το βήμα 3, μέχρι να μην μπορούν να δημιουργηθούν νέα φύλλα.

ΠΑΡΑΔΕΙΓΜΑ

Α/Α	Θέα	Θερμοκρασία	Υγρασία	Αέρας	Κλάση
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω
7	Βροχή	Κανονική	Κανονική	Ασθενής	Έξω
8	Συννεφιά	Υψηλή	Κανονική	Ασθενής	Έξω

Για τη μεταβλητή κλάσης έχουμε 3 φορές την τιμή Μέσα και 5 φορές την τιμή Έξω.

$$I(S,A)=\sum_i \frac{|S_j|}{|S|}E(S_j),$$

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

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.

pi είναι η πιθανότητα της κλάσης i στο S.

$$G(S,A)=E(S)-I(S,A)$$

$$E(S)=-\frac{3}{8}\log_2\left(\frac{3}{8}\right)-\frac{5}{8}\log_2\left(\frac{5}{8}\right)=0.53+0.42=0.95$$

ΠΑΡΑΔΕΙΓΜΑ

Στη συνέχεια υπολογίζουμε το πληροφοριακό κέρδος για κάθε μεταβλητή. Ξεκινάμε με τη μεταβλητή Θέα. Έχουμε συνολικά 8 δείγματα και η μεταβλητή Θέα παίρνει 2 φορές την τιμή Ήλιοφάνεια, και από 3 φορές τις τιμές Συννεφιά και Βροχή. Για τα 2 δείγματα με τιμή Θέα=Ήλιοφάνεια και τα 2 έχουν τιμή κλάσης Μέσα. Για τα 3 δείγματα με τιμή Θέα=Συννεφιά και τα 3 έχουν τιμή κλάσης Έξω. Για τα 3 δείγματα με τιμή Θέα=Βροχή, 1 έχει τιμή κλάσης Μέσα και 2 έχουν τιμή κλάσης Έξω.

$$G(S,\Theta\acute{\varepsilon}\alpha) = E(S) - I(S,\Theta\acute{\varepsilon}\alpha) = E(S) - \frac{2}{8}E(S_{\text{Ηλιοφάνεια}}) - \frac{3}{8}E(S_{\Sigma\upsilon\nu\nu\varepsilon\phi\iota\acute{\alpha}}) - \frac{3}{8}E(S_{\text{Βροχή}})$$

$$E(S_{\text{Ηλιοφάνεια}}) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) = 0$$

$$E(S_{\Sigma\upsilon\nu\nu\varepsilon\phi\iota\acute{\alpha}}) = -\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right) = 0$$

$$E(S_{\text{Βροχή}}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.53 + 0.39 = 0.92$$

$$G(S,\Theta\acute{\varepsilon}\alpha) = 0.95 - \frac{2}{8}\cdot 0 - \frac{3}{8}\cdot 0 - \frac{3}{8}\cdot 0.92 = 0.345$$

ΠΑΡΑΔΕΙΓΜΑ

Στη συνέχεια υπολογίζουμε το πληροφοριακό κέρδος για τη μεταβλητή Θερμοκρασία. Έχουμε συνολικά 8 δείγματα και η μεταβλητή Θερμοκρασία παίρνει 4 φορές την τιμή Υψηλή, 2 φορές την τιμή Κανονική και 2 φορές την τιμή Χαμηλή. Για τα 4 δείγματα με τιμή Θερμοκρασία=Υψηλή, 2 έχουν τιμή κλάσης Μέσα και 2 τιμή κλάσης Έξω.Και τα 2 δείγματα με τιμή Θερμοκρασία=Κανονική έχουν τιμή κλάσης Έξω. Για τα 2 δείγματα με τιμή Θερμοκρασία=Χαμηλή, 1 έχει τιμή κλάσης Μέσα και 1 έχει τιμή κλάσης Έξω

$$G(S,\Theta\varepsilon\rho\mu οκρασία) = E(S) - I(S,\Theta\varepsilon\rho\mu οκρασία) =$$

$$= E(S) - \frac{4}{8}E(S_{\Upsilon\psi\eta\lambda ή}) - \frac{2}{8}E(S_{\text{Κανονική}}) - \frac{2}{8}E(S_{\text{Χαμηλή}})$$

$$E(S_{\Upsilon\psi\eta\lambda ή}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

$$E(S_{\text{Κανονική}}) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$$

$$E(S_{\text{Χαμηλή}}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

$$G(S,\Theta\varepsilon\rho\mu οκρασία) = 0.95 - \frac{4}{8}\cdot 1 - \frac{2}{8}\cdot 0 - \frac{2}{8}\cdot 1 = 0.2$$

ΠΑΡΑΔΕΙΓΜΑ

Συνεχίζουμε με τη μεταβλητή Υγρασία. Έχουμε συνολικά 8 δείγματα και η μεταβλητή Υγρασία παίρνει 4 φορές την τιμή Υψηλή και 4 φορές την τιμή Κανονική. Για τα 4 δείγματα με τιμή Υγρασία=Υψηλή, 2 έχουν τιμή κλάσης Μέσα και 2 έχουν τιμή κλάσης Έξω.Για τα 2 δείγματα με τιμή Υγρασία=Κανονική, 1 έχει τιμή κλάσης Μέσα και 3 έχουν τιμή κλάσης Έξω.

$$G(S,\text{Υγρασία}) = E(S) - I(S,\text{Υγρασία}) = E(S) - \frac{4}{8}E(S_{\text{Υψηλή}}) - \frac{4}{8}E(S_{\text{Κανονική}})$$

$$E(S_{\text{Υψηλή}}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

$$E(S_{\text{Κανονική}}) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.81$$

$$G(S,\text{Υγρασία}) = 0.95 - \frac{4}{8}\cdot 1 - \frac{4}{8}\cdot 0.81 = 0.045$$

ΠΑΡΑΔΕΙΓΜΑ

Τέλος, έχουμε τη μεταβλητή Αέρας. Έχουμε συνολικά 8 δείγματα και η μεταβλητή Αέρας παίρνει 6 φορές την τιμή Ασθενής και 2 φορές την τιμή Δυνατός. Για τα 6 δείγματα με τιμή Αέρας=Ασθενής, 1 έχει τιμή κλάσης Μέσα και 5 έχουν τιμή κλάσης Έξω.Για τα 2 δείγματα με τιμή Αέρας=Δυνατός, 1 έχει τιμή κλάσης Μέσα και 1 έχει τιμή κλάσης Έξω.

$$G(S,\text{Αέρας}) = E(S) - I(S,\text{Αέρας}) = E(S) - \frac{6}{8}E(S_{\text{Ασθενής}}) - \frac{2}{8}E(S_{\text{Δυνατός}})$$

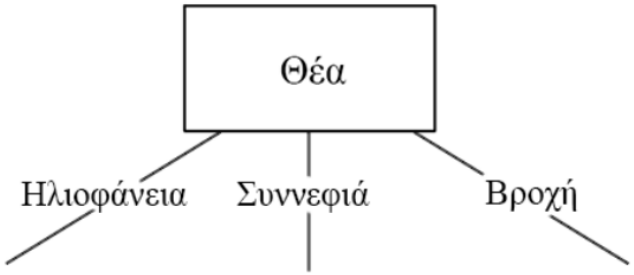
$$E(S_{\text{Ασθενής}}) = -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{5}{6}\log_2\left(\frac{5}{6}\right) = 0.65$$

$$E(S_{\text{Δυνατόςή}}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

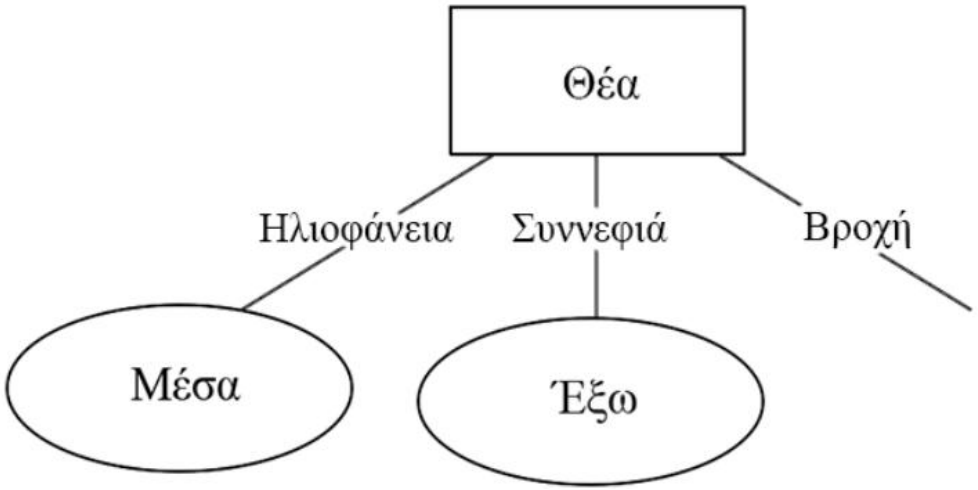
$$G(S,\text{Υγρασία}) = 0.95 - \frac{4}{8}\cdot 0.65 - \frac{4}{8}\cdot 1 = 0.125$$

ΠΑΡΑΔΕΙΓΜΑ

η μεταβλητή Θέα έχει το υψηλότερο πληροφοριακό κέρδος



Για τις τιμές Ήλιοφάνεια και Συννεφιά παρατηρούμε ότι όλα τα δείγματα ανήκουν στην ίδια κλάση, Μέσα και Έξω, αντίστοιχα.



ΠΑΡΑΔΕΙΓΜΑ

Α/Α	Θέα	Θερμοκρασία	Υγρασία	Αέρας	Κλάση
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
7	Βροχή	Κανονική	Κανονική	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα

Για τη μεταβλητή Θερμοκρασία(Αέρας) έχουμε 2 με τιμή Κανονική(Ασθενής) και 1 με τιμή Χαμηλή(Δυνατός). Για την τιμή Θερμοκρασία=Κανονική(Αέρας=Ασθενής) έχουμε 2 φορές την τιμή κλάσης Έξω και 0 φορές την τιμή κλάσης Μέσα, ενώ για την τιμή Θερμοκρασία=Χαμηλή (Αέρας=Δυνατός) έχουμε 1 φορά την τιμή κλάσης Μέσα και 0 φορές τιμή κλάσης Έξω.

$$G(S_{\text{Βροχή}}, \Theta\varepsilon\rho\mu οκρασία) = G(S_{\text{Βροχή}}, \text{Αέρας})$$

$$G(S_{\text{Βροχή}}, \Theta\varepsilon\rho\mu οκρασία) = E(S_{\text{Βροχή}}) - I(S_{\text{Βροχή}}, \Theta\varepsilon\rho\mu οκρασία) =$$

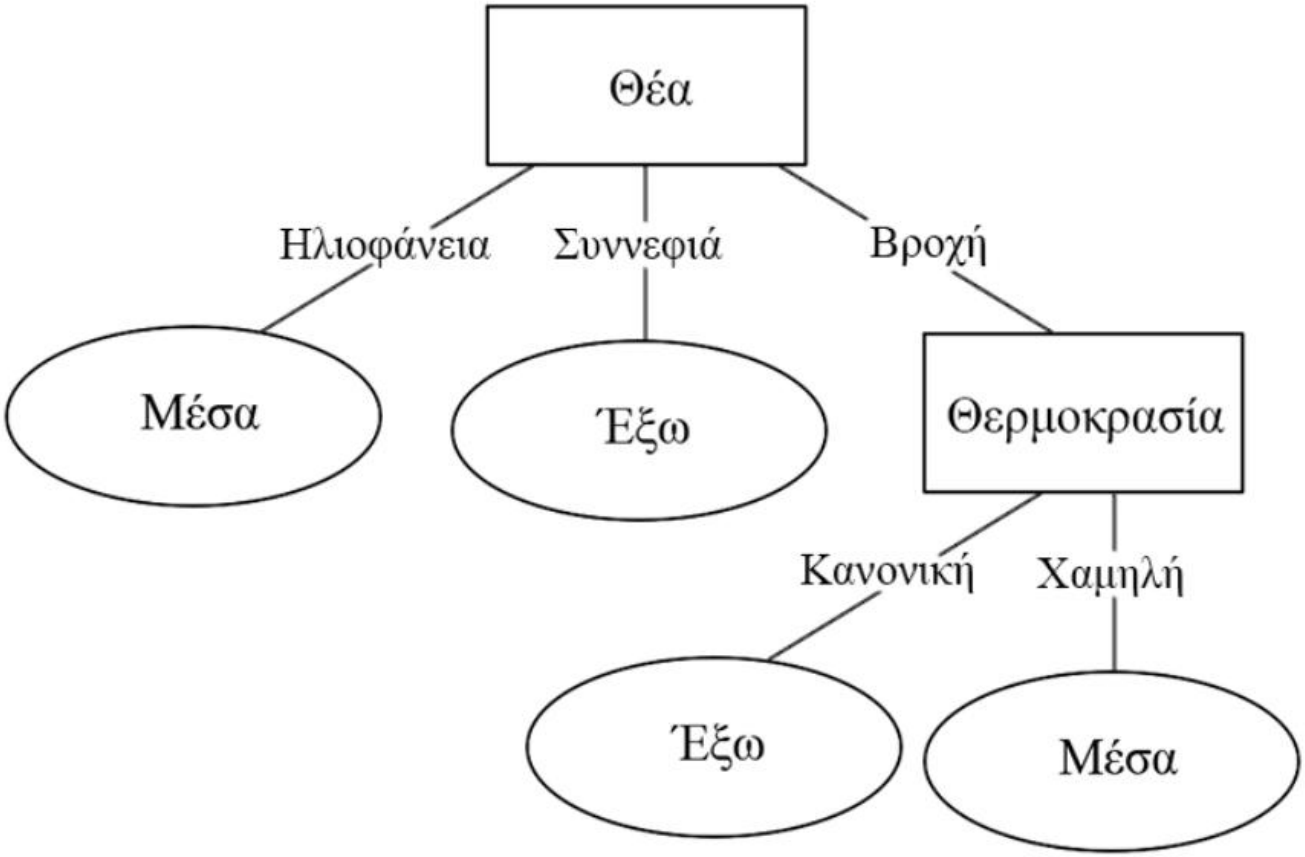
$$E(S_{\text{Βροχή}}) - \frac{2}{3}E(S_{\text{Κανονική}}) - \frac{1}{3}E(S_{\text{Χαμηλή}})$$

$$E(S_{\text{Κανονική}}) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$$

$$E(S_{\text{Χαμηλή}}) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0$$

$$G(S_{\text{Βροχή}}, \Theta\varepsilon\rho\mu οκρασία) = 0.92 - \frac{2}{3}\cdot 0 - \frac{1}{3}\cdot 0 = 0.92$$

ΠΑΡΑΔΕΙΓΜΑ

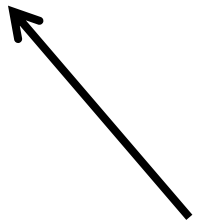

Θέα
Ηλιοφάνεια
Συννεφιά
Βροχή
Μέσα
Έξω
Θερμοκρασία
Κανονική
Χαμηλή
Έξω
Μέσα

ΠΑΡΑΔΕΙΓΜΑ

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

$$E(S) = -p_p * log_2(p_p) - p_n * log_2(p_n)$$

$$E(S) = -(9/14)*\log_2(9/14) - (5/14)*\log_2(5/14) = 0{,}94.$$



Στο σύνολο δεδομένων υπάρχουν εννέα θετικές και πέντε αρνητικές παρατηρήσεις

ΠΑΡΑΔΕΙΓΜΑ

$$E(S,A) = \sum_{j=1}^{u} \frac{s_j}{s} * E(S_j)$$

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

Το υποσύνολο S_1 περιέχει τρεις θετικές και μια αρνητική παρατήρηση

$$E(S_1) = -(3/4)*\log_2(3/4) - (1/4)*\log_2(1/4) = 0{,}811.$$

Το υποσύνολο S_2 περιέχει δύο θετικές και δύο αρνητικές παρατηρήσεις

$$E(S_2) = -(2/4)*\log_2(2/4) - (2/4)*\log_2(2/4) = 1.$$

Το υποσύνολο S_3 περιέχει τέσσερεις θετικές και δύο αρνητικές παρατηρήσεις

$$E(S_3) = -(4/6)*\log_2(4/6) - (4/6)*\log_2(4/6) = 0{,}918.$$

Εάν επιλεγεί το Εισόδημα ως μεταβλητή διαχωρισμού, τότε το σύνολο δεδομένων θα διαχωριστεί σε τρία υποσύνολα, όπου στο πρώτο υποσύνολο S_1 θα περιλαμβάνονται οι υποψήφιοι με χαμηλό εισόδημα, στο δεύτερο υποσύνολο S_2 οι υποψήφιοι με υψηλό εισόδημα και στο τρίτο υποσύνολο S_3 οι υποψήφιοι με μεσαίο εισόδημα.

ΠΑΡΑΔΕΙΓΜΑ

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

Το υποσύνολο S_1 περιέχει τέσσερεις παρατηρήσεις, το υποσύνολο S_2 περιέχει τέσσερεις παρατηρήσεις, το υποσύνολο S_3 περιέχει έξι παρατηρήσεις, και το αρχικό σύνολο περιέχει δέκα τέσσερεις παρατηρήσεις.

$$E(S, \text{Εισόδημα}) = (4/14)*E(S_1) + (4/14)*E(S_2) + (6/14)*E(S_3) = 0{,}911.$$

Κέρδος Πληροφορίας

$$IG(S, \text{Εισόδημα}) = E(S) - E(S, \text{Εισόδημα}) = 0{,}94 - 0{,}911 = 0{,}029.$$

ΑΛΓΟΡΙΘΜΟΣ CART – GINI INDEX

- Ένας άλλος τρόπος κατασκευής δένδρων απόφασης γίνεται με τη χρήση του Gini Index για την επιλογή των κόμβων. Το Gini Index μετράει την ανισότητα μεταξύ τιμών μιας κατανομής συχνοτήτων. Οι τιμές του κυμαίνονται από 0 έως 1, με το 0 να δηλώνει πλήρη ισότητα και το 1 να δηλώνει πλήρη ανισότητα. Για ένα σύνολο δεδομένων S με m δείγματα και k κλάσεις το gini(S) υπολογίζεται με τον τύπο

$$gini(S) = 1 - \sum_{j=1}^{k} p_j^2$$

όπου p_j είναι η πιθανοτητα εμφανισης της κλάσης j στο σύνολο δεδομένων S. Αν το S διαχωριστεί σε S_1 και S_2, τότε

$$gini(S) = \frac{n_1}{n} gini(S_1) + \frac{n_2}{n} gini(S_2)$$

όπου n_1 και n_2 είναι το σύνολο των δειγμάτων στο S_1 και S_2 αντίστοιχα.

ΠΑΡΑΔΕΙΓΜΑ

Α/Α	**Θέα**	Θερμοκρασία	**Υγρασία**	**Αέρας**	**Κλάση**
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω

μεταβλητή Θέα

Α/Α	**Θέα**	Θερμοκρασία	**Υγρασία**	**Αέρας**	**Κλάση**
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω

ΠΑΡΑΔΕΙΓΜΑ

Α/Α	**Θέα**	Θερμοκρασία	**Υγρασία**	**Αέρας**	**Κλάση**
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω

$$gini(\text{Ηλιοφάνεια}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(1^2 + 0\right) = 1 - 1 = 0 \qquad (\text{Μέσα})$$

$$gini(\text{Συννεφιά}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(0 + 1^2\right) = 1 - 1 = 0 \qquad (\text{Έξω})$$

$$gini(\text{Βροχή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) = 1 - \frac{1}{2} = 0.5 \quad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Θέα}) = \frac{2}{6} gini(\text{Ηλιοφάνεια}) + \frac{2}{6} gini(\text{Συννεφιά}) + \frac{2}{6} gini(\text{Βροχή})$$

$$= \frac{2}{6} \cdot 0 + \frac{2}{6} \cdot 0 + \frac{2}{6} \cdot 0.5 = \frac{1}{6} = 0.16$$

ΠΑΡΑΔΕΙΓΜΑ

Α/Α	**Θέα**	Θερμοκρασία	**Υγρασία**	**Αέρας**	**Κλάση**
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω

$$gini(\text{Υψηλή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right) = 1 - \frac{5}{9} = \frac{4}{9} = 0.55 \quad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Κανονική}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(0 + 1^2\right) = 1 - 1 = 0 \quad (\text{Έξω})$$

$$gini(\text{Χαμηλή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) = 1 - \frac{1}{2} = 0.5 \quad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Θερμοκρασία}) = \frac{3}{6} gini(\text{Υψηλή}) + \frac{1}{6} gini(\text{Κανονική}) + \frac{2}{6} gini(\text{Χαμηλή}) =$$

$$= \frac{2}{6} \cdot 0.55 + \frac{2}{6} \cdot 0 + \frac{2}{6} \cdot 0.5 = \frac{1}{6} = 0.35$$

ΠΑΡΑΔΕΙΓΜΑ

Α/Α	**Θέα**	Θερμοκρασία	**Υγρασία**	**Αέρας**	**Κλάση**
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω

$$gini(\text{Υψηλή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right) = 1 - \frac{1}{2} = 0.5 \qquad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Κανονική}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) = 1 - \frac{1}{2} = 0.5 \qquad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Υγρασία}) = \frac{4}{6} gini(\text{Υψηλή}) + \frac{2}{6} gini(\text{Κανονική}) = \frac{4}{6} \cdot 0.5 + \frac{2}{6} \cdot 0.5 = \frac{3}{6} = 0.5$$

ΠΑΡΑΔΕΙΓΜΑ

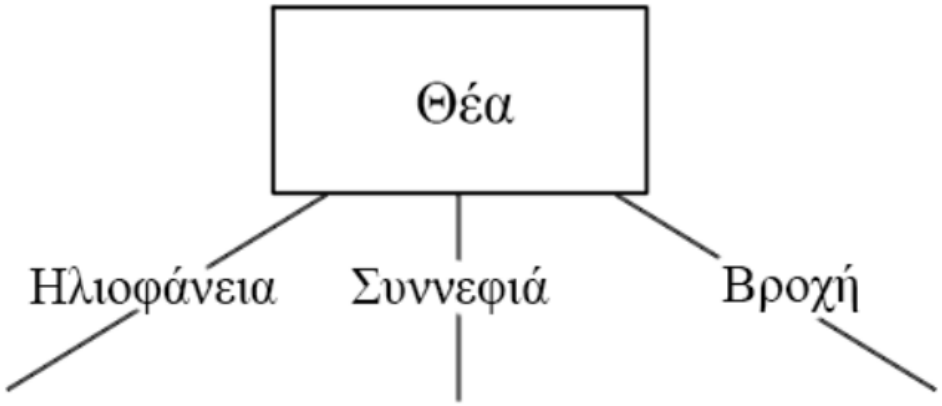
Α/Α	Θέα	Θερμοκρασία	Υγρασία	Αέρας	Κλάση
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω

$$gini(\text{Ασθενής}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2 \right) = 1 - \frac{10}{16} = \frac{6}{16} = 0.375 \qquad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Δυνατός}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(1^2 + 0\right) = 1 - 1 = 0 \qquad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Αέρας}) = \frac{4}{6} gini(\text{Ασθενής}) + \frac{2}{6} gini(\text{Δυνατός}) = \frac{4}{6} \cdot 0.375 + \frac{2}{6} \cdot 0 = 0.25$$

ΠΑΡΑΔΕΙΓΜΑ

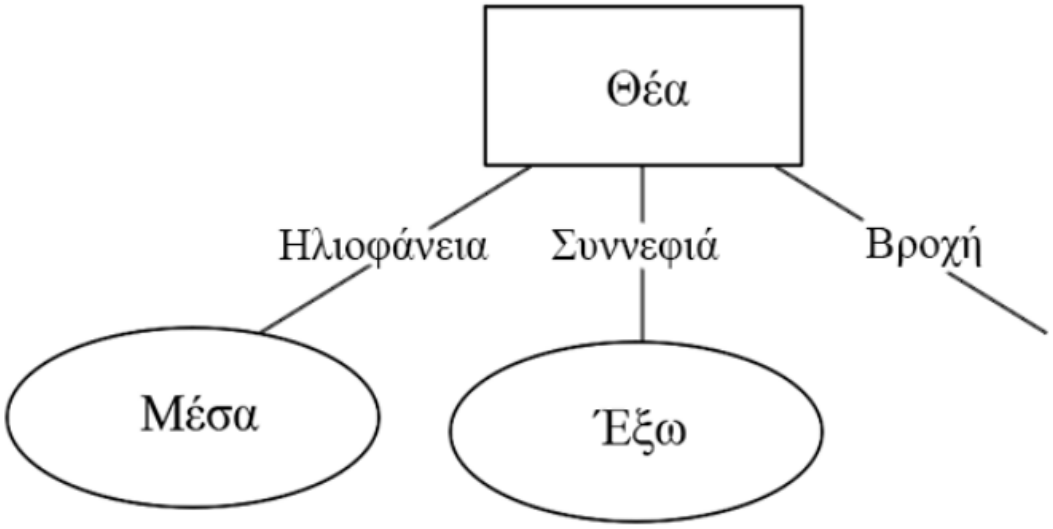


Επιλέγουμε ως αρχικό κόμβο το χαρακτηριστικό με μικρότερη τιμή Gini, δηλαδή τη μεταβλητή Θέα

ΠΑΡΑΔΕΙΓΜΑ

Α/Α	**Θέα**	**Θερμοκρασία**	**Υγρασία**	**Αέρας**	**Κλάση**
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Μέσα
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Δυνατός	Μέσα
3	Συννεφιά	Υψηλή	Υψηλή	Ασθενής	Έξω
6	Συννεφιά	Χαμηλή	Κανονική	Ασθενής	Έξω
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα

Για τις τιμές Ήλιοφάνεια και Συννεφιά, παρατηρούμε ότι όλα τα δείγματα ανήκουν στην ίδια κλάση, Μέσα και Έξω, αντίστοιχα



ΠΑΡΑΔΕΙΓΜΑ

Για την τιμή Βροχή θα πρέπει να εξετάσουμε περαιτέρω τον διαχωρισμό. Αρκεί να εξετάσουμε μόνο τα δείγματα, για τα οποία η μεταβλητή Θέα έχει τιμή Βροχή

Α/Α	**Θέα**	Θερμοκρασία	**Υγρασία**	**Αέρας**	**Κλάση**
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα

$$gini(\text{Βροχή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) = 1 - \frac{1}{2} = 0.5 \quad (\text{Μέσα}, \text{Έξω})$$

$$gini(\text{Θέα}) = \frac{2}{2} gini(\text{Βροχή}) = 1 \cdot 0.5 = 0.5$$

ΠΑΡΑΔΕΙΓΜΑ

Παρατηρούμε ότι για τις μεταβλητές Θερμοκρασία, Υγρασία και Αέρας έχουμε παρόμοιο διαχωρισμό, δηλαδή αντιστοιχία διαφορετικής τιμής μεταβλητής και τιμής κλάσης. Συνεπώς, ο υπολογισμός γίνεται με τον ίδιο τρόπο και οι τιμές που θα προκύψουν θα είναι ίσες. Αρκεί, λοιπόν, να υπολογίσουμε για μια από αυτές τις μεταβλητές το Gini Index. Έστω για τη μεταβλητή Θερμοκρασία.

A/A	**Θέα**	Θερμοκρασία	**Υγρασία**	**Αέρας**	**Κλάση**
4	Βροχή	Κανονική	Υψηλή	Ασθενής	Έξω
5	Βροχή	Χαμηλή	Κανονική	Δυνατός	Μέσα

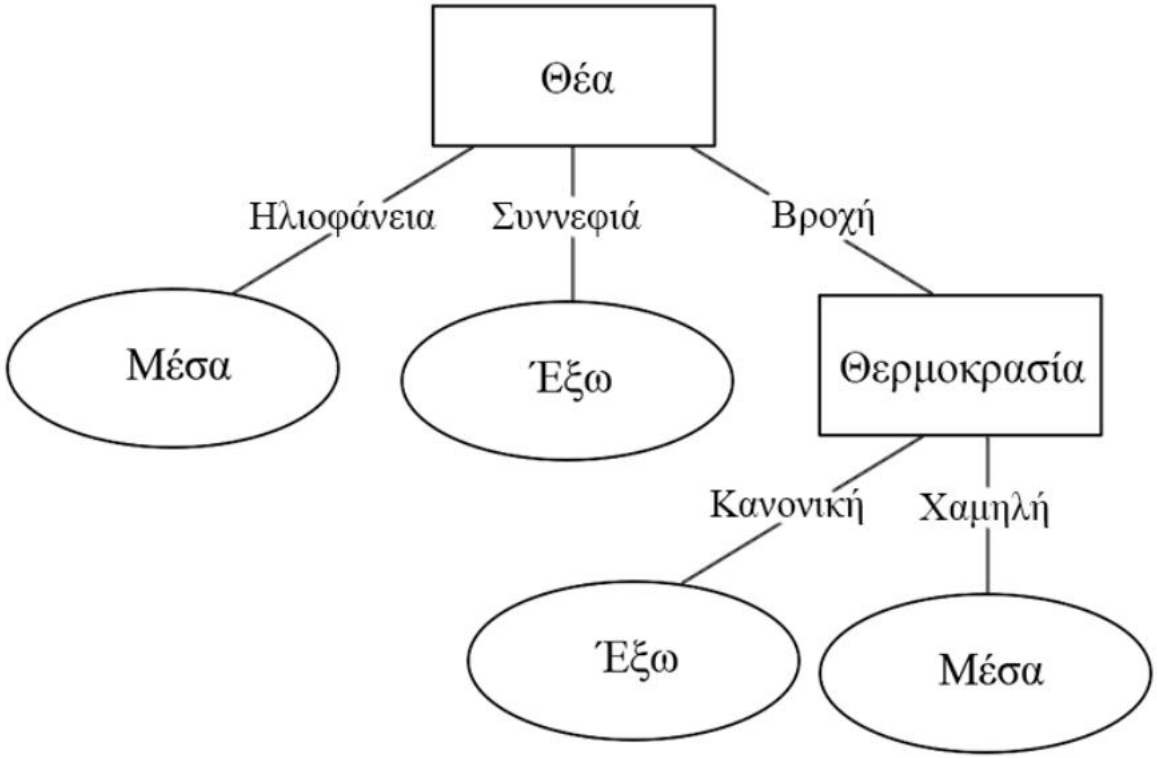
$$gini(\text{Κανονική}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(0 + 1^2\right) = 1 - 1 = 0 \qquad (\text{Έξω})$$

$$gini(\text{Χαμηλή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{Έξω}} \right) = 1 - \left(1^2 + 0\right) = 1 - 1 = 0 \qquad (\text{Μέσα})$$

$$gini(\text{Θερμοκρασία}) = \frac{1}{2} gini(\text{Κανονική}) + \frac{1}{2} gini(\text{Χαμηλή}) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$$

$$gini(\text{Θερμοκρασία}) = gini(\text{Υγρασία}) = gini(\text{Αέρας}) = 0$$

ΠΑΡΑΔΕΙΓΜΑ


Θέα
Ηλιοφάνεια
Συννεφιά
Βροχή
Μέσα
Έξω
Θερμοκρασία
Κανονική
Χαμηλή
Έξω
Μέσα

ΑΛΓΟΡΙΘΜΟΣ C4.5

Ο αλγόριθμος ID3 μεροληπτεί υπέρ των γνωρισμάτων με μεγάλο αριθμό διαιρέσεων.

Ο αλγόριθμος C4.5 αποτελεί βελτιωμένη εκδοχή του ID3

Βελτιωμένη συνάρτηση καταλληλότητας

$$GainRatio(D,S) = \frac{Gain(D,S)}{H\left(\frac{|D_1|}{|D|},\ldots,\frac{|D_s|}{|D|}\right)}$$

Δημιουργεί δυαδικό δένδρο

Χρησιμοποιεί εντροπία

Μαθηματικός τύπος για την επιλογή του σημείου διάσπασης, s, για τον κόμβο t

$$\Phi(s/t) = 2P_L P_R \sum_{j=1}^{m} | P(C_j \mid t_L) - P(C_j \mid t_R) |$$

Οι πιθανότητες P_L,P_R αντιστοιχούν στην πιθανότητα μια εγγραφή να βρεθεί στην αριστερή ή τη δεξιά πλευρά, αντίστοιχα, του δένδρου.

ΠΑΡΑΔΕΙΓΜΑ

Name	Gender	Height	Output1
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium
Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

Gender='M', height=1.6, height=1.7, height=1.8, height=1.9, height=2.0

Φ(Gender='M') = 2·(6/15)·(9/15)·(2/15 + 4/15 + 3/15)=0.224

M F

M – Tall - 3
M – Short - 1
M- Medium - 2

F – Tall - 0
F – Medium - 6
M- Short - 3

Tall – 3/15
Medium – 4/15
Short – 2/15

$$\Phi(s/t) = 2P_L P_R \sum_{j=1}^{m} | P(C_j \mid t_L) - P(C_j \mid t_R) |$$

M F

ΠΑΡΑΔΕΙΓΜΑ

Name	Gender	Height	Output1
Kristina	F	1.6m	Short
Jim	M	2m	Tall
Maggie	F	1.9m	Medium
Martha	F	1.88m	Medium
Stephanie	F	1.7m	Short
Bob	M	1.85m	Medium
Kathy	F	1.6m	Short
Dave	M	1.7m	Short
Worth	M	2.2m	Tall
Steven	M	2.1m	Tall
Debbie	F	1.8m	Medium
Todd	M	1.95m	Medium
Kim	F	1.9m	Medium
Amy	F	1.8m	Medium
Wynette	F	1.75m	Medium

Male

$\Phi(\text{height}=1.6) = 0$

$\Phi(\text{height}=1.7) = 2\cdot(2/15)\cdot(13/15)\cdot(0 + 8/15 + 3/15) = 0.169$

Short Medium Tall

$\Phi(\text{height}=1.8) = 2\cdot(5/15)\cdot(10/15)\cdot(4/15 + 6/15 + 3/15) = 0.385$

$\Phi(\text{height}=1.9) = 2\cdot(9/15)\cdot(6/15)\cdot(4/15 + 2/15 + 3/15) = 0.256$

$\Phi(\text{height}=2.0) = 2\cdot(12/15)\cdot(3/15)\cdot(4/15 + 8/15 + 3/15) = 0.32$

Αποφασίζεται διάσπαση στο height=1.8

Προβληματικές

- 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

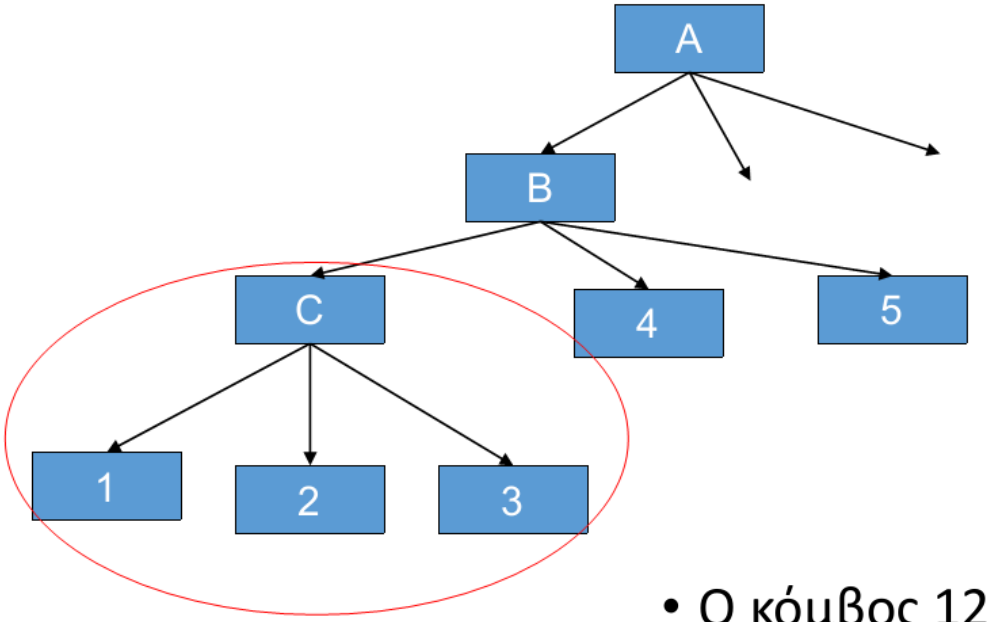
Προβληματικές

- **Postpruning**

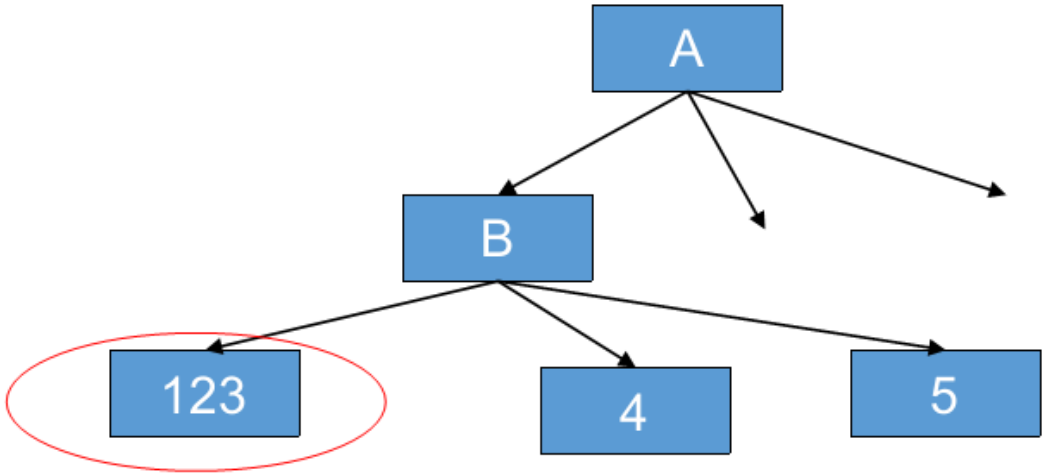
Σε αυτή την τεχνική πρώτα χτίζεται όλο το δένδρο και στην συνέχεια κλαδεύονται τα κλαδιά. Σε αυτή την περίπτωση υπάρχουν 2 τεχνικές.

- Αντικατάσταση του υποδέντρου
- Ανέβασμα υποδέντρου ένα επίπεδο

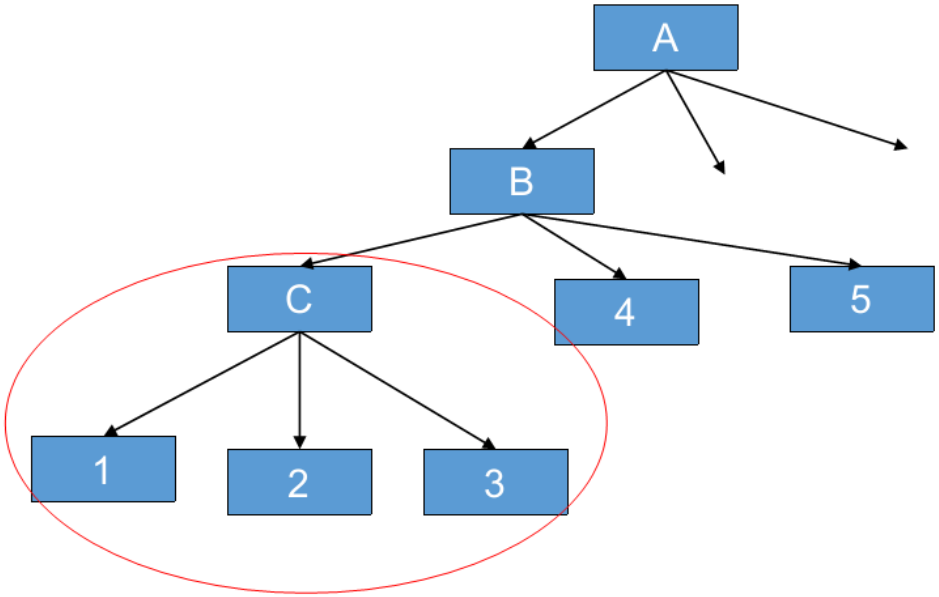
• Όλο το υποδέντρο αντικαθίσταται από ένα φύλο.



• Ο κόμβος 123 αντικατέστησε το υποδέντρο



- Όλο το υποδέντρο ανεβαίνει ένα επίπεδο.



- Το υποδέντρο παίρνει τη θέση του κόμβου B