Κατηγοριοποίηση και Πρόβλεψη

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Συνοπτική Στατιστική και Οπτικοποίηση

Κατηγοριοποίηση και Πρόβλεψη

Κατηγοριοποίηση και Πρόβλεψη

Ο βασικός στόχος αυτού του μαθήματος είναι η εισαγωγή στις έννοιες της κατηγοριοποίησης και της πρόβλεψης.

Η κατηγοριοποίηση έχει ως σκοπό τη δημιουργία ενός μοντέλου κατηγοριοποίησης με τη χρήση ενός συνόλου εκπαίδευσης και ενός αλγόριθμου μάθησης, μέσω του οποίου μπορεί να γίνει η ανάθεση τιμών στο γνώρισμα της κατηγορίας σε μη κατηγοριοποιημένες εγγραφές.

Υπάρχουν διαφόρων ειδών μοντέλα κατηγοριοποίησης, όπως κανόνες, λίστες, δέντρα απόφασης, σύνολο υποδειγμάτων ή παραδειγμάτων δεδομένων, νευρωνικά δίκτυα, μέθοδοι ομάδων κ.λπ.

Κατηγοριοποίηση και Πρόβλεψη

Θα ασχοληθούμε με την επαγωγή μοντέλων δέντρων απόφασης και θα δούμε τις τεχνικές διάσπασης, οι οποίες χρησιμοποιούνται για την ανάπτυξη των δέντρων αυτών.

Στη συνέχεια, θα μελετηθεί η έννοια της πρόβλεψης και θα εξεταστεί η γραμμική παλινδρόμηση, ένα από τα πιο απλά μοντέλα πρόβλεψης για αριθμητικά δεδομένα.

Τέλος, θα ασχοληθούμε με θέματα, που αφορούν τη γενίκευση των μοντέλων, όπως είναι η υπερπροσαρμογή ενός μοντέλου στα δεδομένα.

Η κατηγοριοποίηση αποτελεί μια από τις βασικές εργασίες στο στάδιο της Εξόρυξης Δεδομένων.

Βασίζεται στην εξέταση των χαρακτηριστικών ενός αντικειμένου, το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων.

Η βασική ιδέα είναι η εξής: έχοντας ένα σύνολο από κατηγορίες (κλάσεις) και ένα σύνολο δεδομένων με δείγματα, για τα οποία ξέρουμε σε ποια κλάση ανήκουν, στόχος της κατηγοριοποίησης είναι η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να κατηγοριοποιήσει αυτόματα σε αυτές τις κατηγορίες νέα, άγνωστα, μη-κατηγοριοποιημένα δείγματα.

Ένα από τα δημοφιλέστερα μοντέλα κατηγοριοποίησης είναι τα δένδρα απόφασης.

Τα δέντρα απόφασης είναι μια απλή μορφή αναπαράστασης κανόνων και είναι ευρέως διαδεδομένα, επειδή είναι εύκολα κατανοητά από τον άνθρωπο.

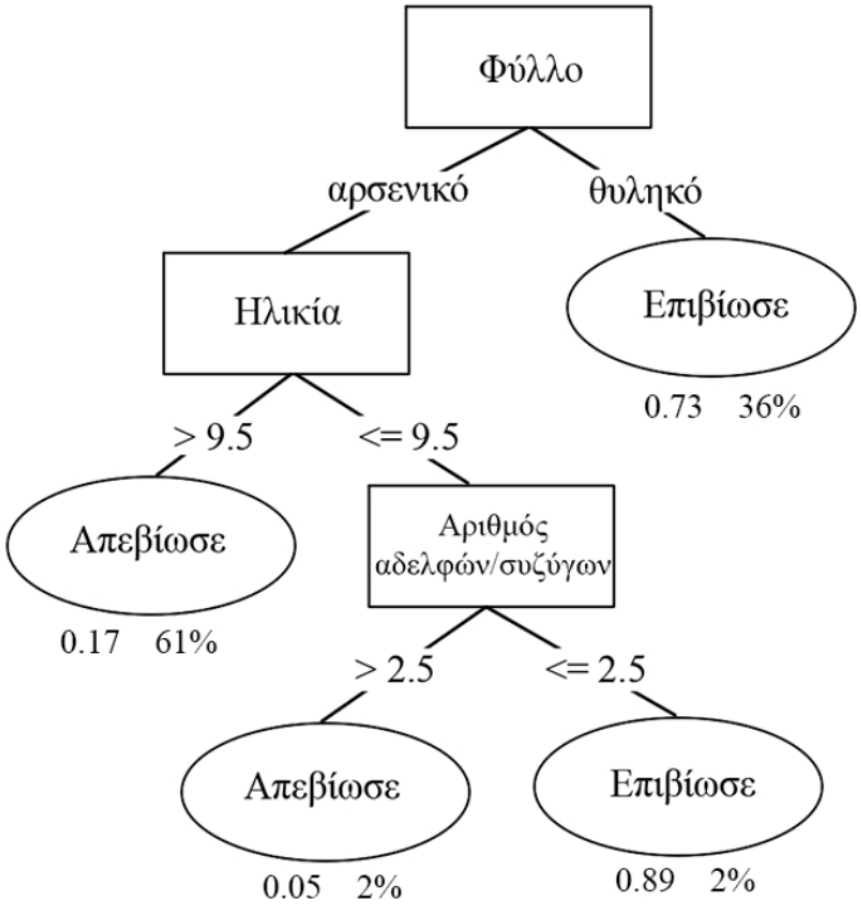
Τα δένδρα απόφασης είναι το απλούστερο μοντέλο κατηγοριοποίησης.

Γενικά, ένα δένδρο αποτελείται από εσωτερικούς κόμβους και φύλλα.

Εσωτερικούς κόμβους λέμε τους κόμβους, οι οποίοι έχουν παιδιά, ενώ φύλλα λέμε τους κόμβους του κατώτερου επιπέδου, τα οποία δεν έχουν απογόνους.

Ο τρόπος αναπαράστασης γίνεται ως εξής:

- κάθε εσωτερικός κόμβος του δένδρου ονοματίζεται με το όνομα ενός χαρακτηριστικού,
- κάθε κλαδί/σύνδεση δυο κόμβων ονοματίζεται με μια συνθήκη ή τιμή για το χαρακτηριστικό του γονικού κόμβου,
- κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης.



Σχήμα: Παράδειγμα δένδρου απόφασης.

Στην προηγούμενη εικόνα είδαμε ένα παράδειγμα δένδρου απόφασης.

Πρόκειται για ένα δένδρο απόφασης, το οποίο δημιουργήθηκε με βάση το σύνολο δεδομένων των επιβατών του Τιτανικού.

Κάτω από τα φύλλα εμφανίζεται η πιθανότητα επιβίωσης και το ποσοστό δειγμάτων, που καταλήγουν στο συγκεκριμένο φύλλο.

Είναι αναμενόμενο οι περισσότεροι άντρες να απεβίωσαν, αφού για τις σωστικές λέμβους δόθηκε προτεραιότητα στα παιδιά και στις γυναίκες.

Στο παραπάνω παράδειγμα χρησιμοποιήθηκαν οι μεταβλητές φύλο, ηλικία και αριθμός συνεπιβατών αδελφών/συζύγων, για να προσδιοριστεί η τιμή της κλάσης.

Εφόσον έχουμε πεπερασμένο αριθμό τιμών (Επιβίωσε, Απεβίωσε), αναφερόμαστε σε ένα δένδρο απόφασης, το οποίο κάνει κατηγοριοποίηση.

Κατασκευή Δένδρου Απόφασης – Αλγόριθμος ID3

Κατασκευή Δένδρου Απόφασης – Αλγόριθμος ID3

Ένας από τους δημοφιλέστερους αλγόριθμους κατασκευής δένδρων απόφασης είναι ο αλγόριθμος ID3.

Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί τις έννοιες της εντροπίας και του πληροφοριακού κέρδους για την επιλογή των κόμβων του δένδρου απόφασης.

Υπενθυμίζουμε ότι το κέρδος πληροφορίας υπολογίζεται από τον τύπο:

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

όπου:

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

όπου με S_j συμβολίζουμε τα δείγματα με τιμή j για το χαρακτηριστικό , με $|S_j|$ το πλήθος τους, με S συμβολίζουμε όλα τα δείγματα και με $|S|$ το πλήθος τους, ενώ με $E(S_j)$ συμβολίζουμε την εντροπία για το υποσύνολο δειγμάτων του συνόλου δεδομένων με τιμή j για το χαρακτηριστικό A.

Η εντροπία για ένα δεδομένο σύνολο υπολογίζεται με βάση την κατανομή της κλάσης των δειγμάτων στο σύνολο.

Αν έχουμε k κλάσεις, η εντροπία για το σύνολο δεδομένων S είναι:

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

όπου p_i είναι η πιθανότητα της κλάσης i στο S.

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.

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

Σχήμα: Σύνολο δεδομένων για το παράδειγμα κατασκευής δένδρου απόφασης με τον ID3.

Ας δούμε ένα παράδειγμα κατασκευής δένδρου απόφασης με τον ID3, για το σύνολο δεδομένων που παρουσιάζει ο προηγούμενος πίνακας.

Αρχικά υπολογίζουμε την εντροπία E(S).

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

Συνεπώς:

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

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 Ηλιοφάνεια, και από 3 φορές τις τιμές Συννεφιά και Βροχή.

Για τα 2 δείγματα με τιμή Θέα=Ηλιοφάνεια 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.

Συνεπώς, έχουμε:

$$G(S, \text{Θέα}) = E(S) - I(S, \text{θέα}) =$$

$$E(S) - \frac{2}{8}E(S, \text{Ηλιοφάνεια}) - \frac{3}{8}E(S, \text{Συνεφιά}) - \frac{3}{8}E(S, \text{Βροχή})$$

Όπου

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

$$E(S, \text{Συνεφιά}) = -\frac{0}{3} log_2 \left(\frac{0}{3}\right) - \frac{3}{3} log_2 \left(\frac{3}{3}\right) = 0$$

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

Επομένως, τελικά:

$$E(S, \text{Θέα}) = 0.95 - \frac{2}{8} \cdot 0 - \frac{3}{8} \cdot 0 - \frac{3}{8} \cdot 0.92 = 0.345$$

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. Συνεπώς, έχουμε:

$$G(S, \text{Θερμογρασία}) = E(S) - I(S,) =$$

$$E(S) - \frac{4}{8} - E(S_{\text{Υψηλή}}) - \frac{2}{8} - E(S_{\text{Κανονική}}) - \frac{2}{8} - E(S_{\text{Χαμηλή}})$$

Όπου

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

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

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

Επομένως, τελικά:

$$G(S, \text{Θερμοκρασία}) = 0.95 - \frac{4}{8} \cdot 1 - \frac{2}{8} \cdot 0 - \frac{2}{8} \cdot 1 = 0.2$$

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.

Συνεπώς, έχουμε:

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

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

Όπου

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

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

Επομένως, τελικά:

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

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. Συνεπώς„ έχουμε:

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

Όπου

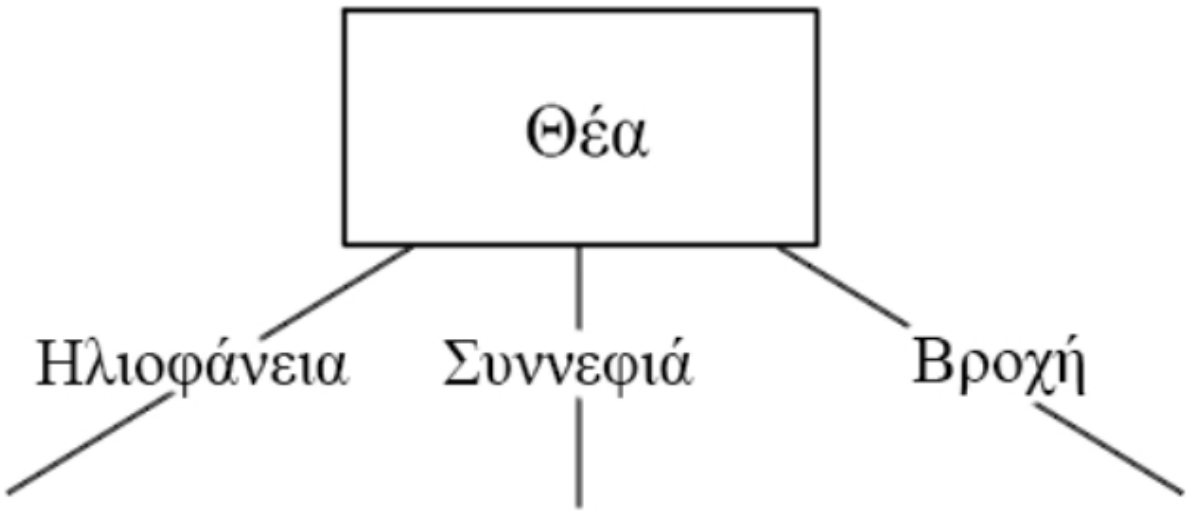
$$E(S, \text{Ασθενής}) = -\frac{1}{6} log_2 \left(\frac{1}{6}\right) - \frac{5}{6} log_2 \left(\frac{5}{6}\right) = 0.65$$

$$E(S, \text{Δυνατός}) = -\frac{2}{2} log_2 \left(\frac{2}{2}\right) - \frac{0}{2} log_2 \left(\frac{0}{2}\right) = 0$$

Επομένως, τελικά:

$$G(S, \text{Αέρας}) = 0.95 - \frac{6}{8} \cdot 0.65 - \frac{2}{8} \cdot 0 = 0.3$$

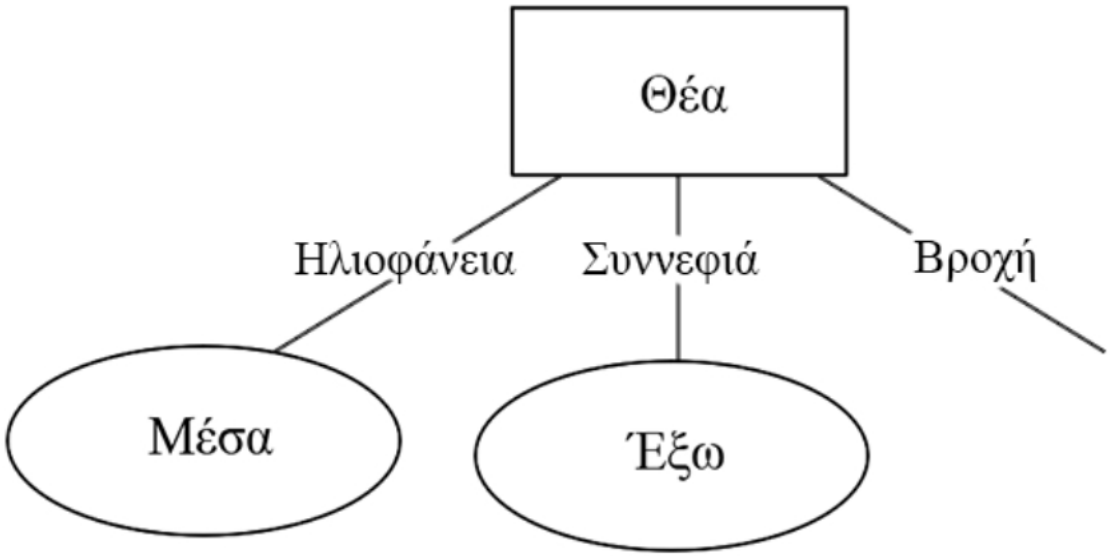
Από τα παραπάνω, η μεταβλητή Θέα έχει το υψηλότερο πληροφοριακό κέρδος. Επομένως, την επιλέγουμε για ρίζα του δένδρου



Σχήμα: ”Αρχικός κόμβος παραδείγματος κατασκευής δένδρου απόφασης με χρήση του αλγορίθμου ID3.”

Στην συνέχεια, πρέπει να εξετάσουμε πώς θα συνεχίσει το κάθε κλαδί του δένδρου.

Για τις τιμές Ηλιοφάνεια καιΣυννεφιά παρατηρούμε ότι όλα τα δείγματα ανήκουν στην ίδια κλάση, Μέσα και Έξω, αντίστοιχα. Συνεπώς, οδηγούμαστε σε φύλλα



Σχήμα: ”Παράδειγμα κατασκευής δένδρου απόφασης με χρήση του αλγορίθμου ID3 (συνέχεια).”

Τώρα μένει να εξετάσουμε τα δείγματα με τιμή Θέα=Βροχή

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

Σχήμα: ”Διαχωρισμός βάσει της μεταβλητής Θέα.”

Αρχικά υπολογίζουμε το πληροφοριακό κέρδος των υπόλοιπων μεταβλητών. 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. Συνεπώς, έχουμε:

$$G(S_{\text{Βροχή, Θερμοκρασία}}) = G(S_{\text{Βροχή, Αέρας}})$$

$$G(S_{\text{Βροχή, Θερμοκρασία}}) = E(S_{\text{Βροχή}}) - I(S_{\text{Βροχή,Θερμοκρασία}}) =$$

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

Όπου

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

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

Επομένως, τελικά:

$$G(S, \text{Βροχή, θερμοκρασία}) = 0.92 - \frac{2}{3} \cdot 0 - \frac{1}{3} \cdot 0 = 0.92$$

Τέλος, για τη μεταβλητή Υγρασία έχουμε 2 δείγματα με τιμή Κανονική και 1 με τιμή Υψηλή. Για το 1 δείγμα με τιμή Υγρασία=Υψηλή έχουμε 1 φορά τιμή κλάσης Έξω και 0 φορές τιμή κλάσης Μέσα. Για τα 2 δείγματα με τιμή Υγρασία=Κανονική έχουμε 1 φορά τιμή κλάσης Μέσα και 1 φορά τιμή κλάσης Έξω.

$$G(S, \text{Βροχή, Υγρασία}) = E(S_{\text{Βροχή}}) - I(S_{\text{Βροχή, Υγρασία}}) =$$

$$E(S, \text{Βροχή}) = -\frac{2}{3}E(S_{\text{Κανονική}}) - \frac{1}{3}E(S_{\text{Υψηλή}})$$

Όπου

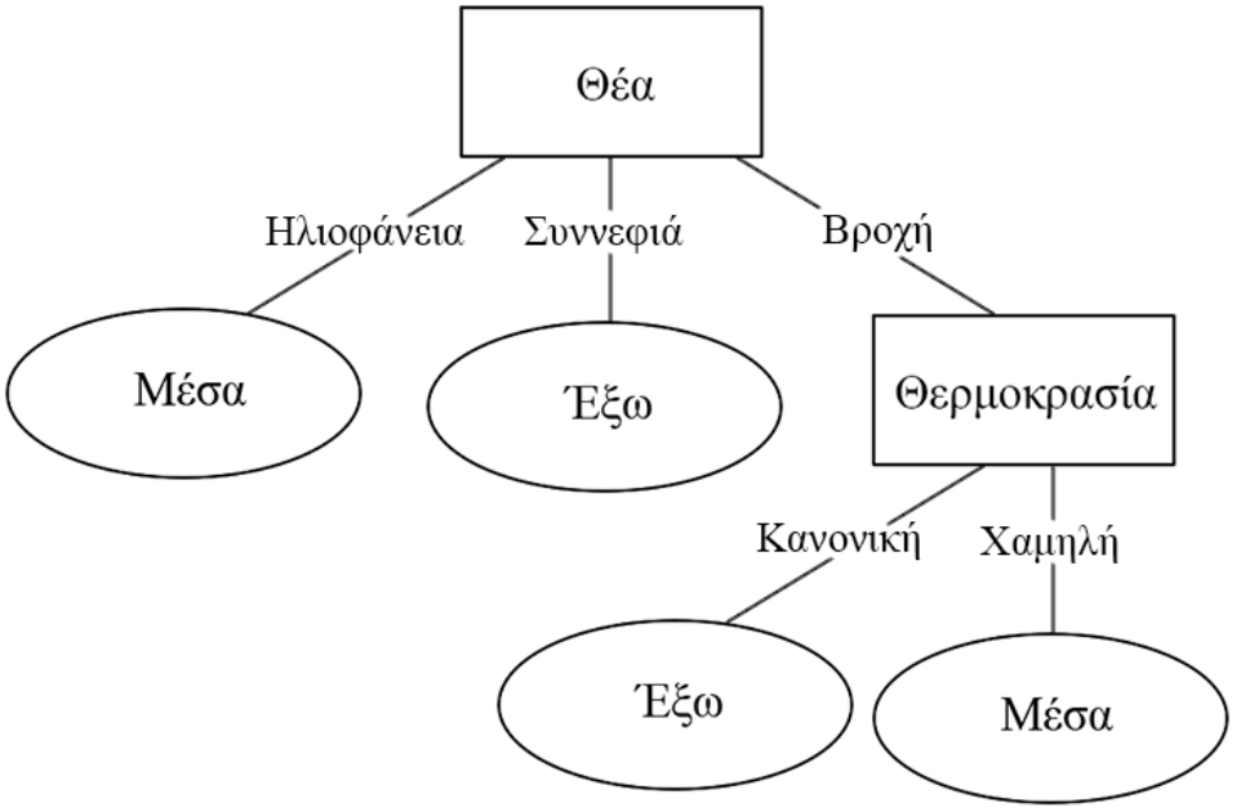
$$E(S, \text{Κανονική}) = -\frac{1}{2} log_2 \left(\frac{1}{2}\right) - \frac{1}{2} log_2 \left(\frac{1}{2}\right) = 1$$

$$E(S, \text{Υψηλή}) = -\frac{1}{1} log_2 \left(\frac{1}{1}\right) - \frac{0}{1} log_2 \left(\frac{0}{1}\right) = 0$$

Επομένως, τελικά:

$$G(S, \text{Βροχή, Υγρασία}) = 0.92 - \frac{2}{3} \cdot 1 - \frac{1}{3} \cdot 0 = 0.25$$

Επιλέγουμε τη μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος, δηλαδή ή τη μεταβλητή Θερμοκρασία ή τη μεταβλητή Αέρας, αφού έχουν ίσο πληροφοριακό κέρδος. Στην Εικόνα φαίνεται το τελικό δένδρο απόφασης με χρήση του αλγορίθμου ID3.



Κατασκευή Δένδρου Απόφασης – Gini Index

Κατασκευή Δένδρου Απόφασης – 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.

Κατασκευή Δένδρου Απόφασης – Gini Index

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 πιθανότητα εμφάνισης 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.

Το πλεονέκτημα της μεθόδου αυτήςείναι ότι για τον υπολογισμό απαιτείται μόνο ο διαχωρισμός των κλάσεων σε κάθε υποσύνολο.

Το καλύτερο χαρακτηριστικό είναι εκείνο με τη μικρότερη τιμή Gini.

Κατασκευή Δένδρου Απόφασης – Gini Index

Ας δούμε ένα παράδειγμα χρήσης του Gini Index για κατασκευή δένδρου απόφασης.

Έστω το σύνολο δεδομένων που φαίνεται παρακάτω

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

Σχήμα: Σύνολο δεδομένων παραδείγματος κατασκευής δένδρου απόφασης με Gini Index.

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

Σχήμα: Διαχωρισμός βάσει της μεταβλητής Θέα.

Κατασκευή Δένδρου Απόφασης – Gini Index

Ξεκινάμε από τη μεταβλητή Θέα. Αρχικά κάνουμε τον διαχωρισμό με βάση τις τιμές της μεταβλητής (βλέπε προηγούμενο Πίνακα), οπότε και έχουμε:

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

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

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

Κατασκευή Δένδρου Απόφασης – Gini Index

Συνεπώς, για τη μεταβλητή Θέα καταλήγουμε:

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

$$gini(\text{Θέα}) = \frac{2}{6}0 + \frac{2}{6}0 + \frac{2}{6}0.5 = 0.16$$

Κατασκευή Δένδρου Απόφασης – Gini Index

A/A	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

Σχήμα: Διαχωρισμός βάσει της μεταβλητής Θερμοκρασία.

Συνεχίζουμε με τη μεταβλητή Θερμοκρασία. Κάνουμε τον διαχωρισμό με βάση τις τιμές της μεταβλητής (βλέπε πίνακα), οπότε και έχουμε:

$$gini(\text{Υψηλή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{έξω}}\right) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) =$$

$$1 - \frac{5}{9} = \frac{4}{9} = 0.55 \quad (\text{Μέσα, Έξω})$$

Κατασκευή Δένδρου Απόφασης – Gini Index

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

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

$$1 - \frac{1}{2} = 0.5 \ (\text{Μέσα, Έξω})$$

Συνεπώς, για τη μεταβλητή θερμοκρασία καταλήγουμε:

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

$$\frac{3}{6} 0,55 + \frac{1}{6} 0 + \frac{2}{6} 0,5 = 0.35$$

Κατασκευή Δένδρου Απόφασης – Gini Index

A/A	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

Σχήμα: Διαχωρισμός βάσει της μεταβλητής Θερμοκρασία.

Συνεχίζουμε με τη μεταβλητή Θερμοκρασία. Κάνουμε τον διαχωρισμό με βάση τις τιμές της μεταβλητής (βλέπε πίνακα), οπότε και έχουμε:

$$gini(\text{Υψηλή}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{έξω}}\right) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) =$$

$$1 - \frac{5}{9} = \frac{4}{9} = 0.55 \quad (\text{Μέσα, Έξω})$$

Κατασκευή Δένδρου Απόφασης – Gini Index

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

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

$$1 - \frac{1}{2} = 0.5 \quad (\text{Μέσα, Έξω})$$

Κατασκευή Δένδρου Απόφασης – Gini Index

Συνεπώς, για τη μεταβλητή Υγρασία καταλήγουμε:

$$gini(\text{Υγρασία}) = \frac{4}{6} gini(\text{Υψηλή}) + \frac{2}{6} gini(\text{Κανονική}) = \frac{4}{6} 0,5 + \frac{2}{6} 0.5 = 0.5$$

Κατασκευή Δένδρου Απόφασης – Gini Index

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

Σχήμα: Διαχωρισμός βάσει της μεταβλητής Θερμοκρασία.

Τέλος, για τη μεταβλητή Αέρας ο διαχωρισμός φαίνεται στον παραπάνω πίνακα. Έχουμε:

$$gini(\text{Ασθενής}) = 1 - \left(p^2_{\text{Μέσα}} + p^2_{\text{έξω}}\right) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) =$$

$$1 - \frac{10}{16} = \frac{6}{16} = 0.375 \quad (\text{Μέσα, Έξω})$$

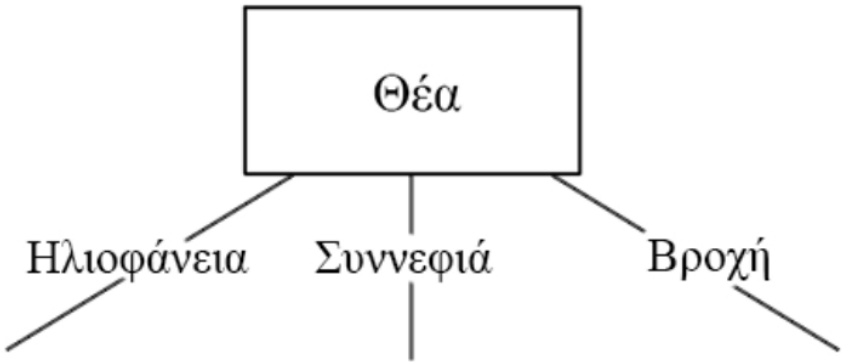
Κατασκευή Δένδρου Απόφασης – Gini Index

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

Συνεπώς, για τη μεταβλητή Αέρας καταλήγουμε:

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

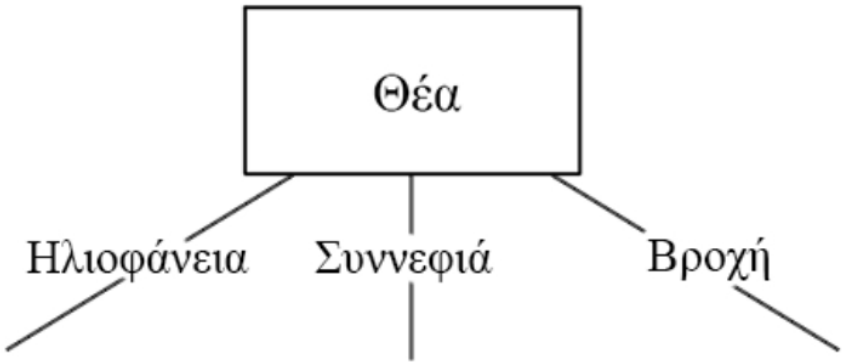
Κατασκευή Δένδρου Απόφασης – Gini Index



Σχήμα: Αρχικός κόμβος παραδείγματος κατασκευής δένδρου απόφασης με χρήση Gini Index.

Επιλέγουμε ως αρχικό κόμβο το χαρακτηριστικό με μικρότερη τιμή Gini, δηλαδή τη μεταβλητή Θέα.
Στη συνέχεια πρέπει να εξεταστούν οι τιμές Ηλιοφάνεια, Συννεφιά και Βροχή ξεχωριστά.
Συνεπώς, ο αρχικός πίνακας πρέπει να διασπαστεί σε 3 τμήματα, όπως φαίνεται στον επόμενο πίνακα.

Κατασκευή Δένδρου Απόφασης – Gini Index



Σχήμα: Αρχικός κόμβος παραδείγματος κατασκευής δένδρου απόφασης με χρήση Gini Index.

Επιλέγουμε ως αρχικό κόμβο το χαρακτηριστικό με μικρότερη τιμή Gini, δηλαδή τη μεταβλητή Θέα.
Στη συνέχεια πρέπει να εξεταστούν οι τιμές Ηλιοφάνεια, Συννεφιά και Βροχή ξεχωριστά.
Συνεπώς, ο αρχικός πίνακας πρέπει να διασπαστεί σε 3 τμήματα, όπως φαίνεται στον επόμενο πίνακα.

Παράδειγμα Κατηγοριοποίησης με το πακέτο kernlab

(α) Χρησιμοποιήστε τη συνάρτηση rpart() με τις default τιμές για τις παραμέτρους που αυτή δέχεται, για να κτίσετε ένα δέντρο κατηγοριοποίησης για την πρόβλεψη του κατηγορικού γνωρίσματος Type με τη χρήση όλων των άλλων γνωρισμάτων του συνόλου δεδομένων (Προσοχή! Εξαιρουμένου φυσικά και του ίδιου του Type).

Για την εκπαίδευση και τον έλεγχο του μοντέλου θα πρέπει να δημιουργήσετε δύο τυχαία δείγματα μεγέθους ίσου με το 70% και το 30% αντίστοιχα του μεγέθους του αρχικού συνόλου δεδομένων.

Γράψτε κώδικα, που να κάνει αυτό τον διαχωρισμό.

Αν χρησιμοποιήσετε έτοιμες συναρτήσεις, χρησιμοποιήστε seed = 70.

Αν επιλέξετε να γράψετε δικό σας κώδικα, επιλέξτε ως σύνολο ελέγχου τις γραμμές, των οποίων το υπόλοιπο της διαίρεσης με το 3 θα ισούται με 0 (π.χ. γραμμή 3, 6, 9, κ.ο.κ).

Παράδειγμα Κατηγοριοποίησης με το πακέτο rpart

Σε κάθε περίπτωση, φροντίστε, ώστε ο διαχωρισμός να είναι ο ίδιος σε κάθε εκτέλεση του κώδικα και τα δύο σύνολα να μην περιέχουν κοινές εγγραφές.

β) Χρησιμοποιήστε το μοντέλο που δημιουργήσατε για το ερώτημα (α), για να κατηγοριοποιήσετε τις 10 πρώτες γραμμές του αρχικού συνόλου δεδομένων.

γ) Δημιουργήστε μία γραφική αναπαράσταση του μοντέλου (του δέντρου απόφασης) που δημιουργήσατε στο ερώτημα (α), κάνοντας χρήση των συναρτήσεων plot(), text() και par().

Αναφέρετε 2 μειονεκτήματα των δένδρων απόφασης.

Τι είναι η υπερπροσαρμογή (overfitting) και πώς μπορούμε να την αποφύγουμε;

Παράδειγμα Κατηγοριοποίησης με το πακέτο rpart

α) Ο κώδικας για την επίλυση του ερωτήματος παρατίθεται παρακάτω:

```
rm(list=ls())
library(rpart)
library(kernlab)
data(spam)

dim(spam)
[1] 4601   58

str(spam)
'data.frame':   4601 obs. of  58 variables:
$ make            : num  0 0.21 0.06 0 0 0 0 0 0.15 0.06 ...
$ address         : num  0.64 0.28 0 0 0 0 0 0 0 0.12 ...
$ order           : num  0 0 0.64 0.31 0.31 0 0 0 0.92 0.06 ...
$ capitalTotal    : num  278 1028 2259 191 191 ...
$ type            : Factor w/ 2 levels "nonspam","spam": 2 2 2 2 ...
```

Παράδειγμα Κατηγοριοποίησης με το πακέτο rpart

```
vars=c(1:57,58)
sum(spam$type=="nonspam")
[1] 2788
sum(spam$type=="spam")
[1] 1813
No.sub=c(1:nrow(spam))[spam$type=="nonspam"]
Yes.sub=c(1:nrow(spam))[spam$type=="spam"]
set.seed(70)
```

Παράδειγμα Κατηγοριοποίησης με το πακέτο rpart

Ηλικία	Φύλο	Επάγγελμα	Αγορά Βιβλίου
20-30	Άνδρας	Φοιτητής	Όχι
30-40	Γυναίκα	Δάσκαλος	Ναι
40-50	Γυναίκα	Γιατρός	Ναι
20-30	Γυναίκα	Δικηγόρος	Ναι
30-40	Άνδρας	Δάσκαλος	Όχι
40-50	Άνδρας	Γιατρός	Ναι

Πίνακας: Δείγμα δεδομένων

$$H_{\text{total}} = -\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6}\right) = 1$$

Παράδειγμα Κατηγοριοποίησης με το πακέτο rpart

Η εντροπία ορίζεται ως:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot log_2(P(x_i))$$

όπου X είναι η τυχαία μεταβλητή, $P(x_i)$ είναι η πιθανότητα της εμφάνισης της τιμής x_i, και log_2 είναι ο λογάριθμος βάσης 2.
Η αρχική εντροπία (H_{total}) του συνόλου δεδομένων υπολογίζεται ως:

$$H_{\text{total}} = -\sum_{i=1}^{K} p_i \cdot \log_2(p_i)$$

To calculate the information gain for each attribute (Has_Fur and Lays_Eggs) with respect to the target variable (Class), we can use the formula for information gain:

$$\text{Information Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \cdot \text{Entropy}(S_v)$$

Where:

- S is the dataset - A is the attribute for which we are calculating the information gain - S_v is the subset of S for which attribute A has value v - $\text{Entropy}(S)$ is the entropy of the dataset S - $\text{Entropy}(S_v)$ is the entropy of the subset S_v

First, let's calculate the entropy of the dataset: