
**Εθνικό Μετσόβιο Πολυτεχνείο**

Διατμηματικό Μεταπτυχιακό Πρόγραμμα Εφαρμοσμένων Μαθηματικών Επιστημών

# Μέθοδοι ταξινόμησης στην εξόρυξη πληροφορίας από βάσεις δεδομένων

# Classification methods in Data Mining

Διπλωματική εργασία

της

Νεφέλης Ροδοπούλου

Επιβλέπων: Ιωάννης Κολέτσος

Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2015

# Περιεχόμενα

# Ευρετήριο πινάκων

# Ευρετήριο σχημάτων

# Ευχαριστίες

*Στον Θανάση.*

# Περίληψη

Τις τελευταίες δεκαετίες έχει σημειωθεί σημαντική πρόοδος στην εξόρυξη δεδομένων και την μηχανική εκμάθηση. Ο συνδυασμός της στατιστικής, της μηχανικής εκμάθησης, της θεωρίας πληροφορίας και της επιστήμης των υπολογιστών έχει δημιουργήσει μία ολοκληρωμένη επιστήμη, με ισχυρό μαθηματικό υπόβαθρο και πολύ χρήσιμα εργαλεία. Αυτό καθιστά δυνατή την επεξεργασία του τεράστιου όγκου δεδομένων που παράγεται καθημερινά, με απώτερο σκοπό την ανακάλυψη κρυμμένων σχέσεων και την εξαγωγή σημαντικών πληροφοριών. Οι πληροφορίες αυτές μάς βοηθούν στη λήψη αποφάσεων και στην πρόβλεψη μελλοντικών συμπεριφορών.

Στην εργασία αυτή μελετώνται διάφορες μέθοδοι ταξινόμησης των δεδομένων σε κατηγορίες με βάση τις ιδιαιτερότητες και τα χαρακτηριστικά τους. Μέσω της ταξινόμησης ανακαλύπτονται συσχετίσεις και προκύπτουν κανόνες που διευκολύνουν τον χρήστη των αποτελεσμάτων να εξάγει αξιόπιστα συμπεράσματα. Τα προβλήματα ταξινόμησης που χρησιμοποιούμε ως παραδείγματα καλύπτουν ένα μεγάλο εύρος καταστάσεων και αυτό φανερώνει το γεγονός ότι η ταξινόμηση των δεδομένων μπορεί να φανεί χρήσιμη, ή και απαραίτητη, σε πολλούς τομείς όπως στη βιοστατιστική, την οικονομία, τις επιχειρήσεις κ.ά.

Αρχικά αναλύεται το γενικό πρόβλημα της ταξινόμησης και τα βήματα που πρέπει να ακολουθηθούν για την αντιμετώπισή του. Στη συνέχεια, περιγράφονται τρόποι επίλυσης τέτοιων προβλημάτων με τη χρήση μεθόδων της κλασσικής στατιστικής. Τέλος, ακολουθεί η ανάλυση τεχνικών που εφαρμόζονται με τη βοήθεια στατιστικών προγραμμάτων.

# Κεφάλαιο 1. Εισαγωγή

## 1.1 Η αρχή και η εξέλιξη

Η ανάγκη για την εξαγωγή προτύπων από δεδομένα υπάρχει εδώ και αιώνες. Οι πρώτες μέθοδοι για τον προσδιορισμό προτύπων περιγράφονται από τη θεωρία Bayes και την ανάλυση παλινδρόμησης. Καθώς οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα, η χειρωνακτική ανάλυση δεδομένων αντικαταστάθηκε από την αυτόματη επεξεργασία δεδομένων. Σε αυτό συνέβαλαν ανακαλύψεις της επιστήμης των υπολογιστών όπως τα νευρωνικά δίκτυα, οι γενετικοί αλγόριθμοι και τα δέντρα απόφασης. Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής των μεθόδων αυτών, με σκοπό την αποκάλυψη άγνωστων προτύπων από μεγάλα σύνολα δεδομένων.

Η εξόρυξη δεδομένων συνδέει τα πεδία της Στατιστικής και της Μηχανικής Εκμάθησης (machine learning). Η Μηχανική Εκμάθηση είναι μία περιοχή της τεχνητής νοημοσύνης η οποία μελετά αλγορίθμους και μεθόδους που επιτρέπουν στους υπολογιστές να «μαθαίνουν», και επικαλύπτεται σημαντικά με την Στατιστική αφού και τα δύο πεδία ασχολούνται με την ανάλυση δεδομένων.

Υπάρχουν διάφορες τεχνικές που χρησιμοποιεί η κλασσική στατιστική για την εξερεύνηση δεδομένων και την κατασκευή μοντέλων, όπως για παράδειγμα η λογιστική παλινδρόμηση, η διακριτή ανάλυση και η ανάλυση σε κύριες συνιστώσες. Όμως, η εξόρυξη δεδομένων υπερτερεί, και αυτό γιατί διαθέτει άφθονα δεδομένα και άπειρες υπολογιστικές δυνατότητες, σε αντίθεση με την κλασσική στατιστική όπου οι υπολογισμοί είναι δύσκολοι και τα δεδομένα είναι σπάνια.

Λόγω των περιορισμένων δεδομένων, η κλασσική στατιστική χρησιμοποιεί το ίδιο δείγμα για να κάνει μία εκτίμηση, αλλά και για να προσδιορίσει το πόσο αξιόπιστη είναι η εκτίμηση αυτή. Έτσι, λοιπόν, τα συμπεράσματα που προκύπτουν από τα διαστήματα εμπιστοσύνης και τον έλεγχο υποθέσεων, μπορεί να θεωρηθούν παραπλανητικά. Αντιθέτως, με την εξόρυξη δεδομένων μπορούμε να προσαρμόσουμε ένα μοντέλο βασιζόμενοι σε ένα δείγμα, και στη συνέχεια να αξιολογήσουμε την επίδοσή του με ένα άλλο δείγμα.

Η έννοια της στατιστικής συμπερασματολογίας, δηλαδή το να προσδιορίσουμε αν μία συγκεκριμένη συμπεριφορά ή ένα ενδιαφέρον αποτέλεσμα συνέβη τυχαία, δεν υπάρχει στην εξόρυξη δεδομένων˙ ο όγκος των δεδομένων είναι τόσο μεγάλος, που καθιστά σχεδόν αδύνατο το να τεθούν αυστηρά όρια γύρω από το ερώτημα που μας απασχολεί, όπως θα απαιτούσε η συμπερασματολογία.

Σαν αποτέλεσμα, η γενική προσέγγιση με την εξόρυξη δεδομένων, είναι ευάλωτη στον κίνδυνο της υπερπροσαρμογής (overfitting). Η υπερπροσαρμογή συμβαίνει όταν υπάρχει σφάλμα στην μοντελοποίηση, το οποίο οφείλεται στο γεγονός ότι το μοντέλο μας προσαρμόστηκε υπερβολικά κοντά στα δεδομένα, κάποια από τα οποία μπορεί να ήταν και ανακριβή(random errors) ή να είχαν τυχαίες ιδιαιτερότητες (noise). Έτσι, τα στοιχεία αυτά «μολύνουν» το μοντέλο με ουσιώδη σφάλματα και μειώνουν την προβλεπτική του ισχύ.

Ο σημαντικότερος παράγοντας που να προωθεί την ανάπτυξη της εξόρυξης δεδομένων, είναι η ανάπτυξη του όγκου των ίδιων των δεδομένων. Στη δεκαετία του ' 50, οι μεγαλύτερες εταιρείες διέθεταν δεδομένα τα οποία καταλάμβαναν, σε ηλεκτρονική μορφή, μερικές εκατοντάδες megabytes. Το 2002 παράχθηκαν περίπου 5 exabytes πληροφορίας που είναι ο διπλάσιος όγκος σε σχέση με το 1999 (1 exabyte= $10^6$ terabytes και 1 terabyte= $10^6$ megabytes). Εκτιμάται ότι πλέον κάθε μέρα παράγονται περίπου 2,5 exabytes ($2{,}5 \cdot 10^{18}$ bytes) δεδομένων και μάλιστα το 90% της παγκόσμιας παραγωγής συνέβη μόνο τα τελευταία δύο χρόνια (2013-2014).

Η ανάπτυξη του όγκου των δεδομένων δεν οφείλεται μόνο στην επέκταση της οικονομίας και της βάσης γνώσεων, αλλά και στην μείωση του κόστους καθώς και την αύξηση της διαθεσιμότητας μηχανισμών οι οποίοι «συλλαμβάνουν» πληροφορίες αυτόματα. Παραδείγματα τέτοιων μηχανισμών είναι οι σαρώσιμοι bar codes, οι συσκευές Point of Sale (POS), τα «ίχνη» από τα mouse clicks και οι συσκευές global positioning satellite (GPS).

Η εξάπλωση του διαδικτύου έχει δημιουργήσει μια νέα γενιά πληροφορίας. Πολλές δραστηριότητες όπως το λιανικό εμπόριο, η εξερεύνηση μιας βιβλιοθήκης και ο κατάλογος αγορών, μπορούν πλέον να γίνουν μέσω του διαδικτύου, και μάλιστα να παρέχουν πληροφορίες οι οποίες μπορούν να μετρηθούν μέχρι την παραμικρή λεπτομέρεια.

Το marketing επικεντρώνεται πλέον, όχι μόνο στα προϊόντα και τις υπηρεσίες που προσφέρονται, αλλά και στον ίδιο τον πελάτη και τις δικές του ξεχωριστές ανάγκες. Αυτό έχει σαν αποτέλεσμα να απαιτούνται λεπτομερείς πληροφορίες σχετικά με τους πελάτες.

Οι περισσότερες από τις τεχνικές διερεύνησης και ανάλυσης που χρησιμοποιούνται στην εξόρυξη δεδομένων θα ήταν σχεδόν αδύνατες χωρίς τη σημερινή υπολογιστική ισχύ. Η συνεχής μείωση του κόστους αποθήκευσης και ανάκτησης δεδομένων κατέστησε δυνατή τη δημιουργία τεχνολογιών, που να επιτρέπουν την αποθήκευση και τη διάθεση τεράστιου όγκου δεδομένων. Εν ολίγοις, η ταχεία και συνεχής βελτίωση της υπολογιστικής ικανότητας αποτελεί βασικό καταλύτη για την ανάπτυξη της εξόρυξης δεδομένων.

## 1.2 Ορισμοί και δραστηριότητες

Η εξόρυξη δεδομένων (data mining) λέγεται αλλιώς και εξόρυξη γνώσης από βάσεις δεδομένων (knowledge discovery in databases). Συνήθως ορίζεται ως η διαδικασία ανακάλυψης χρήσιμων πληροφοριών, προτύπων(patterns) και τάσεων από πηγές δεδομένων, όπως για παράδειγμα σύνολα δεδομένων (datasets), κείμενα, το διαδίκτυο κ.α. Στη συνέχεια θα αναφέρουμε μερικούς από τους ορισμούς που έχουν δοθεί για την εξόρυξη δεδομένων:

Το 2001 οι Hand, Mannila και Smyth στο βιβλίο τους "Principles of Data Mining" έδωσαν τον ορισμό:

> "*Εξόρυξη δεδομένων είναι η ανάλυση συνήθως μεγάλων βάσεων δεδομένων με σκοπό την ανακάλυψη μη αναμενόμενων/κρυμμένων σχέσεων και την*

*σύνοψη των δεδομένων ώστε να είναι κατανοητά και χρήσιμα στον χρήστη.*"[1]

Ένας πιο σύντομος ορισμός δόθηκε από τους Han και Kamber:

"*Εξόρυξη δεδομένων είναι η εξαγωγή σημαντικών πληροφοριών από μεγάλο όγκο δεδομένων.*"[2]

Άλλοι συναφείς ορισμοί είναι:

"*Εξόρυξη δεδομένων ορίζεται ως η αυτόματη ή ημι-αυτόματη διαδικασία ανακάλυψης προτύπων στα δεδομένα.*"[3]

"*Εξόρυξη δεδομένων είναι η εξερεύνηση και ανάλυση μεγάλου όγκου δεδομένων με σκοπό την ανακάλυψη σημαντικών προτύπων και κανόνων.*"[4]

Υπάρχουν δύο τύποι της εξόρυξης δεδομένων:

1) ***Κατευθυνόμενη εξόρυξη δεδομένων (directed data mining)***

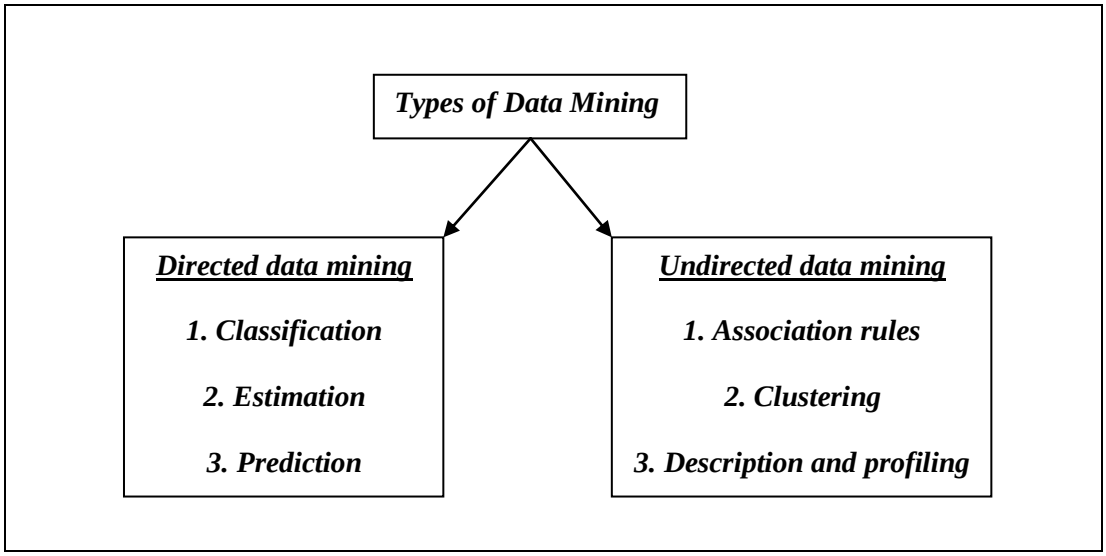
Η κατευθυνόμενη εξόρυξη δεδομένων είναι μία προσέγγιση από πάνω προς τα κάτω, δηλαδή υπάρχει μία μεταβλητή-στόχος και με βάση τα δεδομένα θέλουμε να προβλέψουμε την τιμή της. Με άλλα λόγια, έχουμε από την αρχή μία ιδέα γι' αυτό που ψάχνουμε και το τι θέλουμε να προβλέψουμε. Στόχος μας είναι να κατασκευάσουμε ένα μοντέλο από τα δεδομένα και να το εφαρμόσουμε για να προβλέψουμε μελλοντικές συμπεριφορές.

2) ***Μη κατευθυνόμενη εξόρυξη δεδομένων (undirected data mining)***

Η μη κατευθυνόμενη εξόρυξη δεδομένων είναι μία προσέγγιση από κάτω προς τα πάνω, δηλαδή δεν υπάρχει μεταβλητή-στόχος και θέλουμε να ανακαλύψουμε κάποια δομή στα δεδομένα. Τα ίδια τα δεδομένα θα καθορίσουν τις σχέσεις που μπορεί να υπάρχουν μεταξύ τους και αν βρεθεί κάποια δομή ο χρήστης θα αποφασίσει αν είναι χρήσιμη ή όχι. Στόχος μας είναι η ανακάλυψη προτύπων στα δεδομένα, δηλαδή η ανακάλυψη κάποιων χαρακτηριστικών που μπορεί να σχετίζονται μεταξύ τους ή να συμβαίνουν μαζί συχνά.

Στη συνέχεια θα αναλύσουμε τις δραστηριότητες που χρησιμοποιούνται σε κάθε τύπο.

[1] D. Hand, H. Mannila, P. Smyth "Principles of Data Mining" (2001)
[2] J. Han, M. Kamber "Data Mining: Concepts and Techniques" (2000)
[3] I. Witten, E. Frank, M. Hall "Data Mining: Practical Machine Learning Tools and Techniques" (2005)
[4] M. Berry, G. Linoff "Data Mining Techniques for Marketing, Sales and Customer Support (1997)




### 1.2.1 Κατευθυνόμενη εξόρυξη δεδομένων

Οι δραστηριότητες που μπορούν να γίνουν στην κατευθυνόμενη εξόρυξη δεδομένων είναι η κατηγοριοποίηση, η εκτίμηση και η πρόβλεψη.

- ***Ταξινόμηση (Classification)***

  Η ταξινόμηση περιλαμβάνει την εξέταση των χαρακτηριστικών του αντικειμένου που μας ενδιαφέρει και την ανάθεσή του σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που θέλουμε να ταξινομήσουμε βρίσκονται συνήθως καταχωρημένα σε έναν πίνακα βάσης δεδομένων και η ταξινόμησή τους γίνεται με την προσθήκη μιας νέας στήλης με κάποιον κωδικό κλάσης.

  Για να ολοκληρωθεί η δραστηριότητα αυτή απαιτείται ένας σαφής ορισμός των κλάσεων, καθώς και ένα σύνολο εκπαίδευσης (training set) που να περιλαμβάνει παραδείγματα δεδομένων για τα οποία γνωρίζουμε ήδη σε ποια κλάση ανήκουν. Ο στόχος είναι να κατασκευάσουμε ένα μοντέλο βασιζόμενοι στο σύνολο εκπαίδευσης και στη συνέχεια να το εφαρμόσουμε ώστε να κατατάξουμε τα αταξινόμητα δεδομένα.

- ***Εκτίμηση (Estimation)***

  Η ταξινόμηση ασχολείται με διακριτές μεταβλητές όπως για παράδειγμα μία δίτιμη απόκριση ναι ή όχι, ενώ η εκτίμηση αναφέρεται σε συνεχείς μεταβλητές. Με βάση τα δεδομένα μπορούμε να εκτιμήσουμε την τιμή της συνεχούς μεταβλητής που μας ενδιαφέρει όπως για παράδειγμα το εισόδημα, το ύψος, το υπόλοιπο της πιστωτικής κάρτας κ.α.

  Στην πράξη χρησιμοποιούμε αρκετά συχνά τη μέθοδο της εκτίμησης για να μπορέσουμε να ταξινομήσουμε τα δεδομένα. Ας υποθέσουμε ότι μία εταιρεία πιστωτικών καρτών θέλει να πουλήσει διαφημιστικό χώρο πάνω στους φακέλους που χρησιμοποιεί για την αποστολή των λογαριασμών σε μία εταιρεία που κατασκευάζει αθλητικά παπούτσια. Θα πρέπει, λοιπόν, η πρώτη εταιρεία να κατασκευάσει ένα μοντέλο το οποίο θα κατατάσσει

όλους τους κατόχους πιστωτικών καρτών σε δύο κατηγορίες: αυτούς που αθλούνται και αυτούς που δεν αθλούνται. Μια άλλη προσέγγιση θα ήταν η κατασκευή ενός μοντέλου που να αποδίδει σε κάθε κάτοχο μία τιμή στην μεταβλητή «τάση για άθληση». Αυτή η μεταβλητή θα μπορούσε να παίρνει τιμές από 0 έως 1 και να δείχνει την εκτίμηση της πιθανότητας ένας κάτοχος κάρτας να έχει σαν χόμπι κάποιο άθλημα. Στη συνέχεια η κατηγοριοποίηση θα δώσει ένα κατώτατο όριο στη βαθμολογία της «τάσης για άθληση». Έτσι, τα άτομα με βαθμολογία μεγαλύτερη ή ίση από αυτό το όριο θα χαρακτηριστούν ως «αθλητές», ενώ όσοι έχουν βαθμολογία κατώτερη του ορίου θα θεωρηθούν «μη αθλητές». Με αυτόν τον τρόπο, η εταιρεία μπορεί να επιλέξει σε ποιους λογαριασμούς θα μπει η διαφήμιση.

- ***Πρόβλεψη (Prediction)***

  Η πρόβλεψη είναι παρόμοια με την ταξινόμηση ή την εκτίμηση, με τη διαφορά ότι τα δεδομένα αναθέτονται σε κατηγορίες με βάση κάποια προβλεπόμενη μελλοντική συμπεριφορά ή την εκτιμώμενη μελλοντική τιμή. Στην τεχνική της πρόβλεψης, ο μόνος τρόπος για να ελέγξουμε την ακρίβεια της ταξινόμησης είναι να περιμένουμε και να τη συγκρίνουμε με τα αποτελέσματα εκ των υστέρων. Ο βασικός λόγος για τον οποίο ξεχωρίζουμε την πρόβλεψη από την ταξινόμηση και την εκτίμηση είναι ότι στην προγνωστική μοντελοποίηση υπάρχουν πρόσθετα θέματα που αφορούν τη χρονική σχέση ανάμεσα στις μεταβλητές εισόδου (ή τους προγνωστικούς παράγοντες) και την μεταβλητή-στόχο.

  Οι τεχνικές που χρησιμοποιούνται στην ταξινόμηση και την εκτίμηση μπορούν να εφαρμοστούν και στην πρόβλεψη με τη χρήση ιστορικών δεδομένων (historical data) για τα οποία γνωρίζουμε ήδη την τιμή της μεταβλητής που θέλουμε να προβλέψουμε και με τα οποία μπορούμε να «εκπαιδεύσουμε» το μοντέλο. Έτσι, με βάση τα ιστορικά δεδομένα κατασκευάζουμε ένα μοντέλο το οποίο εξηγεί την τρέχουσα παρατηρούμενη συμπεριφορά. Στη συνέχεια το μοντέλο εφαρμόζεται στα νέα δεδομένα και τελικά προκύπτει η πρόβλεψη της μελλοντικής συμπεριφοράς.

### 1.2.2 Μη κατευθυνόμενη εξόρυξη δεδομένων

Τρεις δραστηριότητες της μη κατευθυνόμενης εξόρυξης δεδομένων είναι οι κανόνες συσχέτισης, η συσταδοποίηση και η περιγραφή.

- ***Κανόνες συσχέτισης (Association rules)***

  Οι κανόνες συσχέτισης στοχεύουν στην ανακάλυψη κρυμμένων συσχετίσεων ανάμεσα στα χαρακτηριστικά μιας βάσης δεδομένων. Ένα τυπικό παράδειγμα είναι ο προσδιορισμός των συνόλων των προϊόντων που συχνά αγοράζονται μαζί σε ένα σουπερμάρκετ και αποτελεί τη βάση της ανάλυσης καλαθιού αγοράς (market basket analysis). Οι αλυσίδες καταστημάτων λιανικής πώλησης χρησιμοποιούν τους κανόνες συσχέτισης για να βρουν ενδιαφέροντες συσχετισμούς στις αγορές των πελατών και να καθορίσουν την θέση των προϊόντων στα ράφια ή σε έναν κατάλογο, ώστε αυτά που αγοράζονται μαζί να βλέπονται και μαζί.

Με τους κανόνες συσχέτισης μπορούν επιπλέον να εντοπιστούν τα προϊόντα που είναι κατάλληλα για σταυροειδείς πωλήσεις (cross-selling) και να σχεδιαστούν ελκυστικά πακέτα προσφορών.

- ***Συσταδοποίηση (Clustering)***

  Με τη δραστηριότητα της συσταδοποίησης χωρίζουμε έναν ετερογενή πληθυσμό σε ομογενείς υποομάδες ή συστάδες. Η διαφορά της συσταδοποίησης και της ταξινόμησης είναι ότι στη συσταδοποίηση δε βασιζόμαστε σε προκαθορισμένες ομάδες, αλλά ομαδοποιούμε τα δεδομένα με βάση το πόσο μοιάζουν μεταξύ τους.

  Ο ίδιος ο χρήστης είναι αυτός που θα αποφασίσει ποια ερμηνεία θα δώσει στις συστάδες που προκύπτουν. Για παράδειγμα, οι ομάδες συμπτωμάτων μπορεί να υποδεικνύουν διαφορετικές ασθένειες και οι ομάδες που χωρίζονται με βάση τα χαρακτηριστικά των πελατών μπορεί να υποδεικνύουν διαφορετικά τμήματα της αγοράς.

- ***Περιγραφή και αξιολόγηση (Description and profiling)***

  Μερικές φορές ο σκοπός της εξόρυξης δεδομένων είναι απλά η περιγραφή μιας περίπλοκης βάσης δεδομένων ώστε να μπορέσουμε να κατανοήσουμε τους ανθρώπους, τα προϊόντα ή τις διαδικασίες που οδήγησαν στη δημιουργία των δεδομένων εξαρχής. Σε πολλές περιπτώσεις, μια καλή περιγραφή των δεδομένων μπορεί να οδηγήσει και στην ερμηνεία τους ή έστω να μας δώσει μία κατεύθυνση στο πού πρέπει να κοιτάξουμε πρώτα για να αναζητήσουμε μια ερμηνεία.

Στο επόμενο κεφάλαιο θα ασχοληθούμε με το πρόβλημα της ταξινόμησης.

## 1.3 Εφαρμογές της εξόρυξης δεδομένων

Η εξόρυξη δεδομένων χρησιμοποιείται σε πολλά πεδία και έχει πολλές εφαρμογές. Ο στρατός χρησιμοποιεί την εξόρυξη δεδομένων για να ελέγξει τον ρόλο που μπορεί να παίζουν διάφοροι παράγοντες στην ακρίβεια της ρίψης μίας βόμβας. Οι υπηρεσίες πληροφοριών τη χρησιμοποιούν για να καθορίσουν ποιες από τις επικοινωνίες που υποκλέπτονται θα μπορούσαν να κρύβουν σημαντικές πληροφορίες. Επίσης, οι ειδικοί σε θέματα ασφάλειας μπορούν μέσω της εξόρυξης δεδομένων να καθορίσουν αν κάποια από τα πακέτα δεδομένων του διαδικτύου αποτελούν απειλή, ενώ οι ιατρικοί ερευνητές μπορούν να προβλέψουν την πιθανότητα υποτροπής κάποιας ασθένειας.

Γενικά, οι μέθοδοι και τα εργαλεία της εξόρυξης δεδομένων μπορούν να εφαρμοστούν σε πολλούς τομείς. Συγκεκριμένα στον τομέα των επιχειρήσεων, μερικά από τα ερωτήματα στα οποία μπορούμε να απαντήσουμε με τη βοήθεια της εξόρυξης δεδομένων είναι τα ακόλουθα:

1. Από μία μεγάλη λίστα με υποψήφιους πελάτες, ποια άτομα είναι πιθανότερο να ανταποκριθούν; Μπορούμε να χρησιμοποιήσουμε τεχνικές ταξινόμησης όπως η λογιστική παλινδρόμηση, τα δέντρα ταξινόμησης (classification trees) κ.α., για να ξεχωρίσουμε ποιοι από τους υποψήφιους πελάτες έχουν παρόμοια δημογραφικά και άλλα δεδομένα με τα δεδομένα των καλύτερων από τους ήδη υπάρχοντες πελάτες.

Έτσι, μπορούμε να προβλέψουμε ακόμη και πόσα χρήματα διατίθεται να ξοδέψει ένα άτομο.

2. Ποιοι από τους πελάτες θα ήταν ικανοί να διαπράξουν μία απάτη (ή ίσως και να την έχουν ήδη διαπράξει); Με τις τεχνικές ταξινόμησης μπορούμε να αναγνωρίσουμε παραδείγματος χάριν αν κάποιες ιατρικές αποζημιώσεις μπορεί να είναι αποτέλεσμα απάτης, και να δώσουμε ιδιαίτερη προσοχή σε αυτές.

3. Ποια από τα άτομα που κάνουν αίτηση για χορήγηση δανείου είναι πιο πιθανό να αθετήσουν τη συμφωνία; Η λογιστική παλινδρόμηση μας επιτρέπει να θέσουμε μία τιμή στην «πιθανότητα αθέτησης συμφωνίας».

4. Ποιοι πελάτες είναι πιο πιθανό να διακόψουν μία συνδρομητική υπηρεσία (τηλεφωνική γραμμή, περιοδικό κ.α.); Και πάλι μέσω της λογιστικής παλινδρόμησης μπορούμε να ορίσουμε την τιμή της «πιθανότητας διακοπής». Με τον τρόπο αυτό, μπορούμε να επιλέξουμε σε ποια άτομα θα προσφέρουμε έκπτωση ή θα κάνουμε κάποια άλλη δελεαστική προσφορά.

# Κεφάλαιο 2. Το πρόβλημα της ταξινόμησης

## 2.1 Περιγραφή

Οι βάσεις δεδομένων κρύβουν πολλές πληροφορίες που μπορούν να φανούν χρήσιμες στη λήψη αποφάσεων. Η ταξινόμηση είναι μία μορφή ανάλυσης δεδομένων με την οποία εξάγουμε μοντέλα για την περιγραφή σημαντικών κατηγοριών. Για παράδειγμα, ένα άτομο που κάνει αίτηση για χορήγηση δανείου μπορεί να αποπληρώσει εγκαίρως, να αποπληρώσει καθυστερημένα ή να κηρύξει πτώχευση. Μία συναλλαγή μέσω πιστωτικής κάρτας μπορεί να είναι πραγματική ή δόλια. Ένα πακέτο δεδομένων στο διαδίκτυο μπορεί να είναι ασφαλές ή «κακόβουλο», ένα άτομο που πάσχει από μία ασθένεια μπορεί να θεραπεύτηκε, να νοσεί ακόμα ή να απεβίωσε. Με άλλα λόγια, το πρόβλημα της ταξινόμησης είναι το γενικό πρόβλημα της ανάθεσης μιας παρατήρησης σε μία ή περισσότερες προκαθορισμένες κατηγορίες.

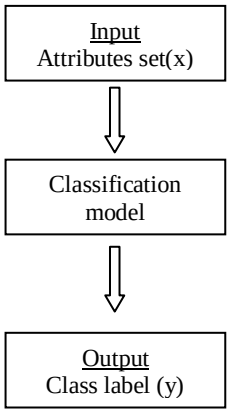
## 2.2 Ορισμός

Στη μέθοδο της ταξινόμησης τα δεδομένα εισόδου είναι ένα σύνολο παρατηρήσεων που χαρακτηρίζονται από μία πλειάδα (x, y), όπου x είναι ένα σύνολο χαρακτηριστικών και y είναι ένα ιδιαίτερο χαρακτηριστικό που ορίζεται ως η ετικέτα κατηγορίας. Στον επόμενο πίνακα παρουσιάζεται ένα δείγμα από ένα σύνολο δεδομένων που χρησιμοποιήθηκε για την ταξινόμηση σπονδυλωτών ζώων σε μία από τις ακόλουθες κατηγορίες: θηλαστικό, πτηνό, ψάρι, ερπετό ή αμφίβιο. Το σύνολο των χαρακτηριστικών αποτελείται από ιδιότητες όπως η θερμοκρασία του σώματος, το δέρμα, η μέθοδος αναπαραγωγής, η ικανότητα να πετάει και η ικανότητα να ζει στο νερό. Στο συγκεκριμένο παράδειγμα παρατηρούμε ότι τα διάφορα χαρακτηριστικά παίρνουν μόνο διακριτές τιμές, παρόλο που σε άλλες περιπτώσεις μπορούν να πάρουν και συνεχείς τιμές. Αντιθέτως, η ετικέτα κατηγορίας y παίρνει μόνο διακριτές τιμές, ενώ αν είναι συνεχής μεταβλητή αντιμετωπίζουμε το πρόβλημα με τη μέθοδο της παλινδρόμησης.

| Όνομα | Θερμοκρασία σώματος | Δέρμα | Γεννάει μικρά | Πλάσμα του νερού | Πλάσμα του αέρα | Έχει πόδια | Χειμερία νάρκη | Ετικέτα κατηγορίας |
|---|---|---|---|---|---|---|---|---|
| **άνθρωπος** | θερμόαιμο | τρίχωμα | ναι | όχι | όχι | ναι | όχι | θηλαστικό |
| **κροταλίας** | ψυχρόαιμο | λέπια | όχι | όχι | όχι | όχι | ναι | ερπετό |
| **σολομός** | ψυχρόαιμο | λέπια | όχι | ναι | όχι | όχι | όχι | ψάρι |
| **φάλαινα** | θερμόαιμο | τρίχωμα | ναι | ναι | όχι | όχι | όχι | θηλαστικό |
| **βάτραχος** | ψυχρόαιμο | τίποτα | όχι | ήμισυ | όχι | ναι | ναι | αμφίβιο |
| **νυχτερίδα** | θερμόαιμο | τρίχωμα | ναι | όχι | ναι | ναι | ναι | θηλαστικό |
| **περιστέρι** | θερμόαιμο | φτερά | όχι | όχι | ναι | ναι | όχι | πτηνό |
| **γάτα** | θερμόαιμο | γούνα | ναι | όχι | όχι | ναι | όχι | θηλαστικό |
| **λεοπάρδαλη** | θερμόαιμο | γούνα | ναι | όχι | όχι | ναι | όχι | θηλαστικό |
| **χελώνα** | ψυχρόαιμο | λέπια | όχι | ήμισυ | όχι | ναι | όχι | ερπετό |
| **πιγκουίνος** | θερμόαιμο | φτερά | όχι | ήμισυ | όχι | ναι | όχι | πτηνό |
| **σκαντζόχοιρος** | θερμόαιμο | αγκάθια | ναι | όχι | όχι | ναι | ναι | θηλαστικό |
| **χέλι** | ψυχρόαιμο | λέπια | όχι | ναι | όχι | όχι | όχι | ψάρι |
| **σαλαμάνδρα** | ψυχρόαιμο | τίποτα | όχι | ήμισυ | όχι | ναι | ναι | αμφίβιο |

**Πίνακας 2.1. Ταξινόμηση σπονδυλωτών ζώων**

*"Ταξινόμηση είναι η διαδικασία εκμάθησης μιας συνάρτησης στόχου f (target function) που απεικονίζει κάθε σύνολο χαρακτηριστικών x σε μία από τις προκαθορισμένες κατηγορίες y."*[5]




Η συνάρτηση στόχος λέγεται αλλιώς και μοντέλο ταξινόμησης. Ένα μοντέλο ταξινόμησης μπορεί να χρησιμοποιηθεί για έναν από τους παρακάτω σκοπούς:

[5] P.-N. Tan, M. Steinbach, V. Kumar "Introduction to Data Mining" Addison-Wesley (2005)

- Περιγραφικό μοντέλο (descriptive modeling)

  Το περιγραφικό μοντέλο παρουσιάζει τα βασικά χαρακτηριστικά των δεδομένων ή αλλιώς συνοψίζει τις πληροφορίες που παίρνουμε από τα δεδομένα. Είναι ένα επεξηγηματικό εργαλείο το οποίο μας βοηθά να διαχωρίσουμε τα δεδομένα που ανήκουν σε διαφορετικές κατηγορίες. Για παράδειγμα, επεξηγεί ποιες είναι οι ιδιότητες που κάνουν ένα ζώο να χαρακτηριστεί ως θηλαστικό.

| Όνομα | Θερμοκρασία σώματος | Δέρμα | Γεννάει μικρά | Πλάσμα του νερού | Πλάσμα του αέρα | Έχει πόδια | Χειμερία νάρκη | Ετικέτα κατηγορίας |
|---|---|---|---|---|---|---|---|---|
| *γάτα* | *θερμόαιμο* | γούνα | *ναι* | *όχι* | *όχι* | *ναι* | *όχι* | *θηλαστικό* |

- Μοντέλο πρόβλεψης (predictive modeling)

  Με το μοντέλο πρόβλεψης μπορούμε να προβλέψουμε σε ποια κατηγορία θα ανήκει μια άγνωστη παρατήρηση. Για παράδειγμα, δοσμένων των χαρακτηριστικών κάποιου ζώου μπορούμε να προβλέψουμε αν είναι θηλαστικό, πτηνό, ερπετό ή αμφίβιο.

| Όνομα | Θερμοκρασία σώματος | Δέρμα | Γεννάει μικρά | Πλάσμα του νερού | Πλάσμα του αέρα | Έχει πόδια | Χειμερία νάρκη | Ετικέτα κατηγορίας |
|---|---|---|---|---|---|---|---|---|
| **σαύρα Χίλα** | *ψυχρόαιμο* | *λέπια* | *όχι* | *όχι* | *όχι* | *ναι* | *ναι* | *?* |

## 2.3 Βήματα της ταξινόμησης

Η ταξινόμηση είναι μία διεργασία που ολοκληρώνεται σε τρία βήματα. Αρχικά χωρίζουμε το σύνολο των δεδομένων σε ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο ελέγχου (test set), συνήθως $^2/_3$ και $^1/_3$ αντίστοιχα. Θεωρούμε ότι γνωρίζουμε ήδη σε ποιες από τις κατηγορίες ανήκει η κάθε παρατήρηση του συνόλου εκπαίδευσης και ελέγχου. Στο πρώτο βήμα κατασκευάζεται ένα μοντέλο που περιγράφει ένα προκαθορισμένο σύνολο κατηγοριών. Αυτή είναι η φάση της «εκπαίδευσης», όπου ο αλγόριθμος ταξινόμησης εξάγει κάποιους κανόνες (ή «μαθαίνει») από το σύνολο εκπαίδευσης και με βάση αυτούς τους κανόνες κατασκευάζει το μοντέλο.

Στο δεύτερο βήμα αξιολογούμε τους κανόνες που έχουν προκύψει από το προηγούμενο βήμα. Σε αυτή τη φάση εφαρμόζουμε το μοντέλο στο σύνολο ελέγχου και ελέγχουμε την αξιοπιστία (accuracy) του ή αλλιώς την προβλεπτική του ισχύ. Είναι προφανές ότι αν εφαρμόζαμε το μοντέλο στο σύνολο εκπαίδευσης θα οδηγούμασταν σε υπερπροσαρμογή των δεδομένων. Για να αποφύγουμε, λοιπόν, τον κίνδυνο της υπερπροσαρμογής, χρησιμοποιούμε το σύνολο ελέγχου που είναι ανεξάρτητο του συνόλου εκπαίδευσης, δηλαδή οι παρατηρήσεις στο σύνολο ελέγχου δε λαμβάνουν μέρος στην κατασκευή του μοντέλου.

Η αξιοπιστία μετριέται από το ποσοστό των παρατηρήσεων του συνόλου ελέγχου που ταξινομήθηκαν σωστά από το μοντέλο. Γνωρίζοντας ήδη την ετικέτα κατηγορίας που αντιστοιχεί σε κάθε παρατήρηση, τη συγκρίνουμε με την ετικέτα που προβλέφθηκε από το

μοντέλο. Αν κρίνουμε ότι είναι επαρκώς αξιόπιστο, μπορούμε να το χρησιμοποιήσουμε για το τελικό βήμα της ταξινόμησης.

Ενδεικτικά, αναφέρουμε ένα παράδειγμα για τον τρόπο υπολογισμού της αξιοπιστίας ενός μοντέλου:

Κατασκευάζουμε έναν πίνακα με τις μετρήσεις του συνόλου ελέγχου, που είναι γνωστός ως πίνακας σύγχυσης (confusion matrix). Κάθε κελί $n_{ij}$ υποδηλώνει τον αριθμό των παρατηρήσεων που στην πραγματικότητα ανήκουν στην κατηγορία i και προβλέφθηκε η ανάθεσή τους στην κατηγορία j. Δηλαδή, το κελί $n_{01}$ είναι ο αριθμός των παρατηρήσεων από την "κατηγορία 0" που λανθασμένα προβλέφθηκαν ως "κατηγορία 1".

| | | *Προβλεπόμενη κατηγορία* | |
|---|---|---|---|
| | | ***Κατηγορία 0*** | ***Κατηγορία 1*** |
| *Πραγματική κατηγορία* | ***Κατηγορία 0*** | $\boldsymbol{n_{00}}$ | $\boldsymbol{n_{01}}$ |
| | ***Κατηγορία 1*** | $\boldsymbol{n_{10}}$ | $\boldsymbol{n_{11}}$ |

**Πίνακας 2.2. Πίνακας σύγχυσης**

Με βάση τα κελιά του πίνακα 2.2, ο συνολικός αριθμός των σωστών προβλέψεων είναι $(n_{00}+n_{11})$ και ο συνολικός αριθμός των λανθασμένων προβλέψεων είναι $(n_{01}+n_{10})$. Η αξιοπιστία θα μετριέται από τη σχέση:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{Αριθμός\ σωστών\ προβλέψεων}}{\boldsymbol{Συνολικός\ αριθμός\ προβλέψεων}} = \frac{\boldsymbol{n_{00}+n_{11}}}{\boldsymbol{n_{00}+n_{01}+n_{10}+n_{11}}} \ (\mathbf{2.1})$$

Ισοδύναμα, μπορούμε να μετρήσουμε την αξιοπιστία από το ποσοστό σφάλματος (error rate) που δίνεται από την ακόλουθη σχέση:

$$\boldsymbol{Error\ rate} = \frac{\boldsymbol{Αριθμός\ λανθασμένων\ προβλέψεων}}{\boldsymbol{Συνολικός\ αριθμός\ προβλέψεων}} = \frac{\boldsymbol{n_{01}+n_{10}}}{\boldsymbol{n_{00}+n_{01}+n_{10}+n_{11}}} \ (\mathbf{2.2})$$

Οι αλγόριθμοι ταξινόμησης αναζητούν μοντέλα με τη μεγαλύτερη δυνατή τιμή αξιοπιστίας ή αντίστοιχα με το μικρότερο ποσοστό σφάλματος.

Στο τελευταίο βήμα της ταξινόμησης, εφαρμόζουμε το μοντέλο σε νέες παρατηρήσεις για τις οποίες δεν γνωρίζουμε σε ποια κατηγορία ανήκουν.

Στόχος της ταξινόμησης είναι η κατασκευή ενός μοντέλου που αφενός θα μας βοηθά να εξηγήσουμε και να κατανοήσουμε καλύτερα τα δεδομένα, και αφετέρου θα μας δίνει τη δυνατότητα να προβλέψουμε την κατηγορία νέων δεδομένων.

## 2.4 Προετοιμασία των δεδομένων

Σε αυτήν την παράγραφο θα αναφέρουμε μερικά βήματα προεπεξεργασίας των δεδομένων ώστε να βελτιώσουμε την αξιοπιστία, την αποδοτικότητα και την επεκτασιμότητα της ταξινόμησης.

- ✓ ***Καθαρισμός δεδομένων (data cleaning)*** Είναι η προεπεξεργασία των δεδομένων με σκοπό να απομακρυνθούν ή να ελαττωθούν οι παράγοντες «θορύβου» (noise factors) εφαρμόζοντας τεχνικές εξομάλυνσης (smoothing techniques) και να αντικατασταθούν οι αγνοούμενες ή ελλιπείς τιμές (missing values) χρησιμοποιώντας στη θέση τους την τιμή που εμφανίζεται συχνότερα για το συγκεκριμένο χαρακτηριστικό. Με τον τρόπο αυτό αποφεύγουμε το ενδεχόμενο σύγχυσης κατά τη διάρκεια «μάθησης» του αλγορίθμου ταξινόμησης.

- ✓ ***Ανάλυση συσχετίσεων (relevance analysis)*** Πολλά από τα χαρακτηριστικά των δεδομένων μπορεί να περισσεύουν/πλεονάζουν. Με την ανάλυση συσχέτισης εντοπίζουμε εκείνα τα χαρακτηριστικά που μπορεί να έχουν κάποια στατιστική σχέση μεταξύ τους. Για παράδειγμα, αν το χαρακτηριστικό $x_1$ παρουσιάζει έντονη συσχέτιση με το $x_2$ θα μπορούσαμε να αφαιρέσουμε το ένα εκ των δύο από την ανάλυση δεδομένων. Επιπλέον, ένα σύνολο δεδομένων μπορεί να περιέχει κάποια χαρακτηριστικά που να μη συνδέονται με τη μεταβλητή-στόχο. Σε αυτήν την περίπτωση, βρίσκουμε ένα υποσύνολο χαρακτηριστικών που μας ενδιαφέρουν, του οποίου η κατανομή πιθανότητας να είναι όσο το δυνατόν κοντύτερα στην κατανομή του αρχικού συνόλου (attribute subset selection). Έτσι, αποφεύγουμε να συμπεριλάβουμε στην ανάλυση δεδομένων τα χαρακτηριστικά που δε συμβάλλουν στην δραστηριότητα της ταξινόμησης και που θα μπορούσαν να καθυστερήσουν ή και να αποπροσανατολίσουν τη διαδικασία της «μάθησης».

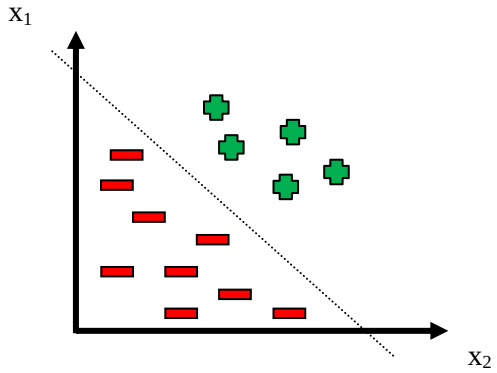
- ✓ ***Μετασχηματισμός και μείωση των δεδομένων (data transformation and reduction)*** Για τον μετασχηματισμό των δεδομένων χρησιμοποιούμε τη μέθοδο της κανονικοποίησης, με την οποία κατατάσσουμε τις τιμές ενός χαρακτηριστικού σε μία συγκεκριμένη κλίμακα όπως -1,0 έως 1,0 ή 0,0 έως 1,0 , ώστε να εμπίπτουν μέσα σε ένα μικρό καθορισμένο εύρος. Έτσι, αποφεύγεται η μεγάλη μεταβλητότητα μεταξύ των τιμών όταν έχουμε δεδομένα που συνήθως παίρνουν μεγάλες τιμές, όπως για παράδειγμα το εισόδημα.

  Μία μέθοδος μείωσης των δεδομένων είναι η γενίκευση σε άλλες έννοιες, όπως_π.χ. περιγραφικές. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη σε δεδομένα που παίρνουν συνεχείς τιμές. Για παράδειγμα, οι αριθμητικές τιμές ενός χαρακτηριστικού όπως το εισόδημα, μπορούν να γενικευτούν σε διακριτά εύρη όπως “χαμηλό”, “μεσαίο” και “υψηλό”. Επειδή λοιπόν η γενίκευση «συμπιέζει» το αρχικό σύνολο δεδομένων, προκύπτουν λιγότερα δεδομένα εισόδου και εξόδου.

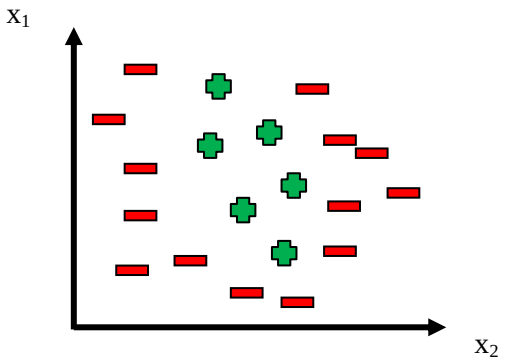
## 2.5 Γραμμικοί και μη γραμμικοί ταξινομητές

Για την αντιμετώπιση των προβλημάτων ταξινόμησης μπορούμε να χρησιμοποιήσουμε γραμμικούς ή μη γραμμικούς ταξινομητές. Οι γραμμικοί ταξινομητές χρησιμοποιούνται όταν τα δεδομένα παρουσιάζουν γραμμικές σχέσεις μεταξύ τους και σε αντίθεση με τους μη γραμμικούς ταξινομητές, αδυνατούν να μοντελοποιήσουν μη γραμμικά όρια μεταξύ των κατηγοριών.

Δύο ή περισσότερες κατηγορίες είναι γραμμικά διαχωρίσιμες (linearly separable), αν μπορούμε να τις διαχωρίσουμε απόλυτα με γραμμικά όρια απόφασης.

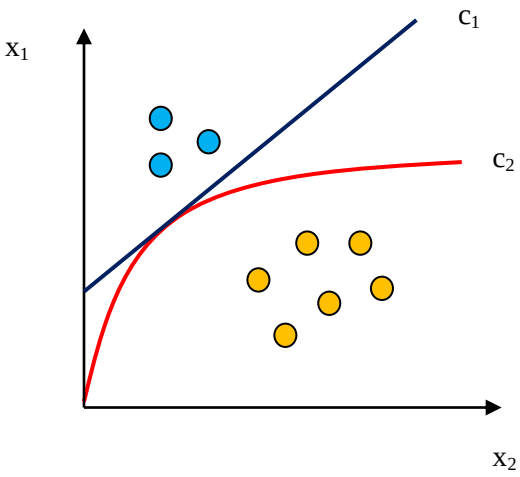



**Σχήμα 2.1. Γραμμικά διαχωρίσιμες κατηγορίες**



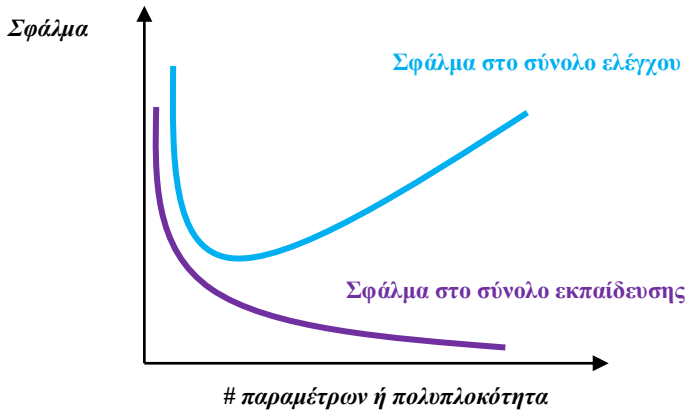

**Σχήμα 2.2. Μη γραμμικά διαχωρίσιμες κατηγορίες**

Στο ακόλουθο γράφημα δίνεται ένα παράδειγμα με δύο κατηγορίες αντικειμένων οι οποίες μπορούν να διαχωριστούν βάσει δύο χαρακτηριστικών $x_1$ και $x_2$. Η μία κατηγορία τείνει να έχει υψηλές τιμές στο $x_1$ και η άλλη στο $x_2$. Στο γράφημα φαίνονται επίσης δύο ταξινομητές $c_1$ και $c_2$ οι οποίοι διαχωρίζουν επιτυχώς τις δύο κατηγορίες. Ο $c_1$ είναι ένας γραμμικός ταξινομητής αφού αποτελεί γραμμικό συνδυασμό των χαρακτηριστικών $x_1$ και $x_2$, ενώ ο $c_2$ είναι μη γραμμικός καθώς η επίδραση του $x_1$ αρχίζει να μειώνεται όσο μεγαλώνει το $x_2$.



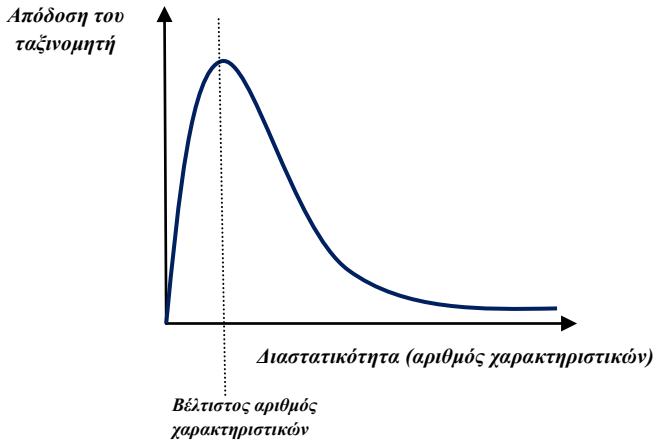

**Σχήμα 2.3. Γραμμικός και μη γραμμικός ταξινομητής**

Οι ταξινομητές που μοντελοποιούν με μη γραμμικά όρια απόφασης (π.χ. νευρωνικά δίκτυα, k-NN, δέντρα απόφασης), δεν είναι καλοί στη «γενίκευση» και είναι επιρρεπείς στην υπερπροσαρμογή. Με άλλα λόγια, τα πολύπλοκα μοντέλα προσαρμόζονται καλύτερα στα δεδομένα του συνόλου εκπαίδευσης, αλλά μετά από κάποιο σημείο τείνουν να «γενικεύουν» χειρότερα στα νέα δεδομένα. Συνεπώς, ο αριθμός της διάστασης πρέπει να διατηρείται σχετικά χαμηλός όταν χρησιμοποιούνται τέτοιοι ταξινομητές. Αντίθετα, όταν έχουμε ταξινομητές που γενικεύουν εύκολα (π.χ. αφελής Bayes, γραμμικός ταξινομητής), τότε ο αριθμός των χαρακτηριστικών μπορεί να είναι μεγαλύτερος αφού ο ίδιος ο ταξινομητής είναι λιγότερο «περιγραφικός»/εκφραστικός.




**Σχήμα 2.4. Σφάλμα συναρτήσει του πλήθους των παραμέτρων**

Τέλος, καθώς αυξάνεται η διαστατικότητα, η απόδοση του ταξινομητή μεγαλώνει μέχρι να φτάσουμε σε έναν βέλτιστο αριθμό χαρακτηριστικών. Αν η διάσταση συνεχίσει να αυξάνεται, χωρίς όμως να αυξηθεί και το πλήθος των δεδομένων, ο ταξινομητής αρχίζει πλέον να χάνει την αποδοτικότητά του. Αυτό το φαινόμενο είναι γνωστό ως "η κατάρα της διαστατικότητας" (curse of dimensionality).




**Σχήμα 2.5. Απόδοση του ταξινομητή συναρτήσει της διαστατικότητας**

# Κεφάλαιο 3. Στατιστικές μέθοδοι ταξινόμησης

## 3.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται τρεις απλές στατιστικές μέθοδοι ταξινόμησης στις οποίες δεν κάνουμε καμία υπόθεση για τη δομή των δεδομένων και βγάζουμε συμπεράσματα μόνο από τα ίδια τα δεδομένα κι όχι από κάποιο μοντέλο.

### 3.1.1 Παράδειγμα 1: Πρόβλεψη χρηματοοικονομικής απάτης

Μία εταιρεία ελεγκτών διαθέτει στο πελατολόγιό της πολλές μεγάλες επιχειρήσεις. Κάθε μία από τις επιχειρήσεις στέλνει την ετήσια οικονομική έκθεσή της στην εταιρεία, η οποία στη συνέχεια την ελέγχει. Επειδή η ελεγκτική εταιρεία θέλει να αποφύγει οποιαδήποτε νομική κατηγορία εναντίον της, θα πρέπει να μπορεί να εντοπίζει αν κάποια από τις οικονομικές εκθέσεις είναι δόλια. Στο παράδειγμα αυτό, οι επιχειρήσεις-πελάτες είναι τα καταγεγραμμένα δεδομένα και η μεταβλητή απόκρισης που μας ενδιαφέρει είναι Y={δόλιος, ειλικρινής}. Με βάση τη μεταβλητή Y μπορούμε να κατατάξουμε τις επιχειρήσεις σε δύο κατηγορίες: $C_1$=δόλιος και $C_2$=ειλικρινής. Η μοναδική επιπλέον πληροφορία που διαθέτει η ελεγκτική εταιρεία σε σχέση με τους πελάτες είναι αν εκδόθηκε κάποια νομική κατηγορία εναντίον της ή όχι και δεδομένου ότι θέλει να βελτιώσει την εκτίμηση της πρόβλεψης απάτης εισάγεται η κατηγορική μεταβλητή $X$=”νομική κατηγορία” που παίρνει την τιμή 1 αν εκδόθηκε κάποια νομική κατηγορία εναντίον της και την τιμή 0 αν όχι.

Η εταιρεία ελεγκτών έχει στη διάθεσή της τα δεδομένα 1500 επιχειρήσεων τα οποία έχει ερευνήσει στο παρελθόν, δηλαδή γνωρίζει ήδη για κάθε επιχείρηση-πελάτη αν τελικά ήταν δόλια ή ειλικρινής και αν υπήρξε νομική κατηγορία ή όχι. Αφού χώρισε τα δεδομένα σε ένα σύνολο εκπαίδευσης (1000 πελάτες) και ένα σύνολο αξιολόγησης (500 πελάτες), προέκυψαν τα παρακάτω αποτελέσματα από το σύνολο εκπαίδευσης:

| | **Νομική κατηγορία (X=1)** | **Όχι νομική κατηγορία (X=0)** | **Σύνολο** |
|---|---|---|---|
| **$C_1$= δόλιος** | 50 | 50 | 100 |
| **$C_2$= ειλικρινής** | 180 | 720 | 900 |
| **Σύνολο** | 230 | 770 | 1000 |

Πώς μπορεί να χρησιμοποιηθεί αυτή η πληροφορία για να ταξινομηθεί ένας πελάτης ως δόλιος ή ειλικρινής;

## 3.2 Ο αφελής κανόνας

Ένας πολύ απλός κανόνας για να ταξινομούμε τα δεδομένα σε μία από τις k κατηγορίες, αγνοώντας όλες τις πληροφορίες που μπορεί να έχουμε από τους προγνωστικούς παράγοντες, είναι να κατατάξουμε τα δεδομένα στην κατηγορία με τις περισσότερες παρατηρήσεις. Σύμφωνα με αυτόν τον κανόνα, στο παράδειγμα με την ελεγκτική εταιρεία θα κατατάσσαμε όλους τους πελάτες στην κατηγορία $C_2$=ειλικρινής, δεδομένου ότι το 90% των επιχειρήσεων που εξετάστηκαν στο σύνολο εκπαίδευσης ήταν ειλικρινείς. Ο αφελής κανόνας χρησιμοποιείται κυρίως για την αξιολόγηση της επίδοσης πιο περίπλοκων ταξινομητών.

## 3.3 Ο αφελής ταξινομητής Bayes

Ο αφελής ταξινομητής Bayes είναι μία πιο εξελιγμένη μέθοδος από τον αφελή κανόνα. Η βασική ιδέα είναι ότι ενσωματώνουμε τις πληροφορίες που έχουμε από τους προγνωστικούς παράγοντες στον αφελή κανόνα, για να προκύψει μία πιο ακριβής ταξινόμηση των δεδομένων. Με άλλα λόγια, για να εκτιμήσουμε την πιθανότητα μια παρατήρηση να ανήκει σε μία συγκεκριμένη κατηγορία βασιζόμαστε αρχικά στην επικράτηση ή επιπολασμό (prevalence) της κάθε κατηγορίας, δηλαδή στο ποσοστό των παρατηρήσεων που ανήκουν στην κάθε κατηγορία. Επιπλέον, όμως, βασιζόμαστε και στις πρόσθετες πληροφορίες που έχουμε για αυτήν την παρατήρηση. Στο παράδειγμα με την ελεγκτική εταιρεία, θα βασιστούμε στο ποσοστό των επιχειρήσεων που ανήκουν στις δύο κατηγορίες ($C_1$=δόλιος και $C_2$=ειλικρινής) δεδομένης της πληροφορίας $X$, δηλαδή αν υπήρξε ή όχι νομική κατηγορία εναντίον της εταιρείας.

Ο αφελής ταξινομητής Bayes χρησιμοποιείται μόνο όταν έχουμε κατηγορικές μεταβλητές και στην περίπτωση που οι μεταβλητές μας είναι συνεχείς θα πρέπει να τις κωδικοποιήσουμε και να τις μετασχηματίσουμε σε κατηγορικές για να μπορέσουμε να τον εφαρμόσουμε.

Η μέθοδος αυτή είναι πολύ χρήσιμη όταν έχουμε μεγάλα σύνολα δεδομένων. Για παράδειγμα, οι εταιρείες μηχανών αναζήτησης όπως η Google χρησιμοποιούν τον αφελή ταξινομητή Bayes για να διορθώσουν τα ορθογραφικά λάθη που κάνουν οι χρήστες. Όταν γράφουμε μία φράση που περιέχει κάποιο ορθογραφικό λάθος, το Google μάς προτείνει μία διόρθωση για αυτή τη φράση. Η πρόταση αυτή βασίζεται αφενός στη συχνότητα των λέξεων με παρόμοια ορθογραφία που πληκτρολογήθηκαν από εκατομμύρια άλλους χρήστες, και αφετέρου στις υπόλοιπες λέξεις που υπάρχουν μέσα στη φράση.

### 3.3.1 Το θεώρημα Bayes

Η μέθοδος του αφελή ταξινομητή Bayes βασίζεται σε δεσμευμένες πιθανότητες και συγκεκριμένα σε ένα θεμελιώδες θεώρημα της Θεωρίας Πιθανοτήτων, το θεώρημα του Bayes. Με το θεώρημα αυτό μπορούμε να υπολογίσουμε την πιθανότητα να έχει συμβεί πρώτα το ενδεχόμενο Α δεδομένου ότι μεταγενέστερα συνέβη το ενδεχόμενο Β και συμβολίζεται με $P(A|B)$. Για παράδειγμα, ποια είναι η πιθανότητα μία επιχείρηση να παρέδωσε μία δόλια οικονομική έκθεση στην ελεγκτική εταιρεία, δεδομένου ότι τελικά

υπήρξε νομική κατηγορία εναντίον της; Είναι προφανές ότι η παράνομη πράξη εκ μέρους της επιχείρησης προηγήθηκε της νομικής κατηγορίας εναντίον της εταιρείας.

Στο παράδειγμα με την ελεγκτική εταιρεία μάς ενδιαφέρει να υπολογίσουμε την πιθανότητα P(δόλια οικονομική έκθεση | νομική κατηγορία) ή αλλιώς $P(Y = C_1|X = 1)$. Το γεγονός ότι αυτή η πιθανότητα είναι δεσμευμένη, σημαίνει ότι διαθέτουμε πρόσθετες πληροφορίες κι έτσι τα αποτελέσματα είναι πιο αξιόπιστα.

Στην τεχνική της ταξινόμησης το θεώρημα του Bayes μάς δίνει έναν τύπο για να υπολογίσουμε την πιθανότητα μίας παρατήρησης να ανήκει σε μία κατηγορία, δεδομένων των χαρακτηριστικών της παρατήρησης. Έστω ότι έχουμε k κατηγορίες $C_1, C_2, \dots, C_k$ και γνωρίζουμε ότι το ποσοστό των παρατηρήσεων που ανήκουν σε κάθε κατηγορία είναι $P(C_1), P(C_2), \dots, P(C_k)$ αντίστοιχα. Θέλουμε να ταξινομήσουμε μία νέα παρατήρηση με ένα σύνολο προγνωστικών παραγόντων (ή χαρακτηριστικών) $X_1, X_2, \dots, X_p$ βάσει των παραγόντων αυτών. Αν γνωρίζουμε την πιθανότητα να συμβεί το ενδεχόμενο $X_1$ *και* το $X_2$ *και… και* το $X_p$ ($P(X_1, X_2, \dots, X_p)$), τότε από το θεώρημα Bayes μπορούμε να βρούμε την πιθανότητα η νέα παρατήρηση να ανήκει σε μία από τις κατηγορίες $C_i$ (i=1, 2, …,k) και θα είναι:

$$\boldsymbol{P(C_i|X_1, \dots, X_p) = \frac{P(X_1, \dots, X_p|C_i)P(C_i)}{P(X_1, \dots, X_p|C_1)P(C_1) + \cdots + P(X_1, \dots, X_p|C_k)P(C_k)}} \qquad \textbf{(3.1)}$$

Αυτή είναι η εκ των υστέρων πιθανότητα να ανήκει μία παρατήρηση στην κατηγορία $C_k$ και περιλαμβάνει τις πληροφορίες από τους προγνωστικούς παράγοντες. Αντιθέτως, η εκ των προτέρων πιθανότητα $P(C_i)$ να ανήκει σε μία από τις κατηγορίες δεν περιέχει τις πρόσθετες πληροφορίες.

Για να ταξινομήσουμε μία παρατήρηση υπολογίζουμε την πιθανότητα να ανήκει σε μία από τις κατηγορίες $P(C_i|X_1, \dots, X_p)$ για κάθε i κατηγορία. Στη συνέχεια βάζουμε την παρατήρηση στην κατηγορία που έχει τη μεγαλύτερη πιθανότητα. Στην πράξη, χρειάζεται να υπολογίσουμε μόνο τον αριθμητή της σχέσης (3.1) αφού ο παρονομαστής είναι κοινός για όλες τις κατηγορίες. Οι πληροφορίες που πρέπει να έχουμε για τον υπολογισμό του αριθμητή είναι:

1. Το ποσοστό της κάθε κατηγορίας για όλο τον πληθυσμό $P(C_1), P(C_2), \dots, P(C_k)$

2. Την πιθανότητα να συμβούν τα ενδεχόμενα $X_1, X_2, \dots, X_p$ μέσα σε κάθε κατηγορία

Αν υποθέσουμε ότι το σύνολο των δεδομένων που διαθέτουμε είναι ένα αντιπροσωπευτικό δείγμα του πληθυσμού, μπορούμε να εκτιμήσουμε τα ποσοστά του πληθυσμού που θα ανήκουν σε κάθε κατηγορία και τον παράγοντα πρόβλεψης (τον παράγοντα που θέλουμε να προβλέψουμε) από το σύνολο εκπαίδευσης. Για να εκτιμήσουμε την πιθανότητα $P(X_1, \dots, X_p|C_i)$, θα βρούμε ποια από τα χαρακτηριστικά $X_1, \dots, X_p$ εμφανίζονται στην κατηγορία $C_i$ και θα τα διαιρέσουμε με τον συνολικό αριθμό των παρατηρήσεων αυτής της κατηγορίας.

Στο παράδειγμα με την ελεγκτική εταιρεία μπορούμε να βρούμε τις εκτιμημένες πιθανότητες για το σύνολο εκπαίδευσης $\hat{P}(C_1) = \frac{100}{1000} = 0{,}1$ και $\hat{P}(C_2) = \frac{900}{1000} = 0{,}9$ βάσει του πίνακα 1.1. Η πρόσθετη πληροφορία που διαθέτουμε σχετικά με τον αν εκδόθηκε νομική κατηγορία ή όχι εναντίον της εταιρείας μάς δίνει την πιθανότητα να εμφανίζεται ή όχι το χαρακτηριστικό ($X = 1$ ή $X = 0$ αντίστοιχα) μέσα σε κάθε κατηγορία:

$$\hat{\mathrm{P}}(\text{νομική κατηγορία}|\text{δόλιος}) = \hat{\mathrm{P}}(\mathrm{X} = 1|\mathrm{C}_1) = \frac{50}{100} = 0{,}5$$

$$\hat{\mathrm{P}}(\text{νομική κατηγορία}|\text{ειλικρινής}) = \hat{\mathrm{P}}(\mathrm{X} = 1|\mathrm{C}_2) = \frac{180}{900} = 0{,}2$$

$$\hat{\mathrm{P}}(\text{όχι νομική κατηγορία}|\text{δόλιος}) = \hat{\mathrm{P}}(\mathrm{X} = 0|\mathrm{C}_1) = \frac{50}{100} = 0{,}5$$

$$\hat{\mathrm{P}}(\text{όχι νομική κατηγορία}|\text{ειλικρινής}) = \hat{\mathrm{P}}(\mathrm{X} = 0|\mathrm{C}_2) = \frac{720}{900} = 0{,}8$$

Ας υποθέσουμε τώρα ότι υπήρξε νομική κατηγορία εναντίον της εταιρείας λόγω της οικονομικής έκθεσης μιας συγκεκριμένης επιχείρησης-πελάτη. Για να κατατάξουμε τον πελάτη ως δόλιο ή ειλικρινή θα υπολογίσουμε την πιθανότητα να ανήκει σε μία από τις δύο κατηγορίες:

$$\hat{P}(\textit{δόλιος}|\textit{νομική κατηγορία}) \propto \hat{P}(\textit{νομική κατηγορία}|\textit{δόλιος})\hat{P}(\textit{δόλιος})$$

$$= 0{,}5 \cdot 0{,}1 = 0{,}05$$

$$\hat{P}(\textit{ειλικρινής}|\textit{νομική κατηγορία}) \propto \hat{P}(\textit{νομική κατηγορία}|\textit{ειλικρινής})\hat{P}(\textit{ειλικρινής})$$

$$= 0{,}2 \cdot 0{,}9 = 0{,}018$$

Αυτό σημαίνει ότι ο πελάτης είναι πιο πιθανό να είναι δόλιος και όπως παρατηρούμε έρχεται σε αντίθεση με τον αφελή κανόνα ο οποίος αγνοεί την πληροφορία για τις νομικές κατηγορίες και τους κατατάσσει όλους ως ειλικρινείς.

### 3.3.2 Από τον Bayes στον αφελή Bayes

Η δυσκολία με την εφαρμογή του παραπάνω τύπου είναι ότι αν ο αριθμός των προγνωστικών παραγόντων p είναι αρκετά μεγάλος, έστω 20, και οι κατηγορίες είναι δύο, τότε ακόμα και αν οι προγνωστικοί παράγοντες είναι δυαδικοί θα χρειαστούμε έναν πολύ μεγάλο αριθμό παρατηρήσεων για να πάρουμε μία καλή εκτίμηση της πιθανότητας $P(X_1, \dots, X_p|C_i)$. Επιπλέον, ο τύπος αυτός απαιτεί να εμφανίζονται όλοι οι προγνωστικοί παράγοντες μέσα σε κάθε κατηγορία, κάτι που μπορεί να μη συμβαίνει στην πραγματικότητα˙ είναι πιθανό ακόμη και κάποιοι από τους παράγοντες να μην υπάρχουν στο σύνολο των δεδομένων.

Μία λύση σε αυτό το πρόβλημα είναι η υπόθεση της ανεξαρτησίας των προγνωστικών παραγόντων. Αν υποθέσουμε ότι όλοι οι παράγοντες είναι ανεξάρτητοι μεταξύ τους, τότε απλουστεύεται σημαντικά η παραπάνω έκφραση και γίνεται πιο εύχρηστη στην πράξη. Έτσι,

η ανεξαρτησία των προγνωστικών παραγόντων μέσα σε κάθε κατηγορία μάς δίνει την ακόλουθη σχέση:

$$\boldsymbol{P(X_1, X_2, \dots, X_p|C_i) = P(X_1|C_i)P(X_2|C_i) \dots P(X_p|C_i)} \qquad (\mathbf{3.2})$$

Οι όροι $P(X_1|C_i), P(X_2|C_i), \dots, P(X_p|C_i)$ εκτιμώνται με βάση τις μετρήσεις από το σύνολο εκπαίδευσης. Η εκτίμηση της πιθανότητας $P(X_j|C_i)$ θα είναι ίση με τον αριθμό των παρατηρήσεων της κατηγορίας $C_i$ που έχουν το χαρακτηριστικό $X_j$ διαιρούμενο με τον συνολικό αριθμό των παρατηρήσεων της κατηγορίας $C_i$.

Ο αριθμός των παρατηρήσεων που θα χρειαζόμασταν θα ήταν πολύ μεγαλύτερος αν δεν είχαμε κάνει την υπόθεση της ανεξαρτησίας των προγνωστικών παραγόντων. Αυτή είναι μία πολύ απλοϊκή υπόθεση δεδομένου ότι οι παράγοντες συνήθως συσχετίζονται. Παραδόξως η προσέγγιση με τον αφελή ταξινομητή του Bayes «δουλεύει» πολύ καλά στην πράξη όταν έχουμε πολλές μεταβλητές οι οποίες είναι δυαδικές ή κατηγορικές με μερικά διακριτά επίπεδα.

Στο παράδειγμα που ακολουθεί φαίνονται οι διαφορές ανάμεσα στους υπολογισμούς με βάση το θεώρημα Bayes και με βάση τον αφελή ταξινομητή Bayes. Έστω ότι διαθέτουμε τα δεδομένα από δέκα επιχειρήσεις. Για την κάθε επιχείρηση γνωρίζουμε αν υπήρξε νομική κατηγορία εναντίον της ή όχι, αν είναι μικρή ή μεγάλη επιχείρηση και αν έχει χαρακτηριστεί ως δόλια ή ειλικρινής με βάση την οικονομική έκθεση που παρέδωσε.

| **Νομική κατηγορία?** | **Μέγεθος επιχείρησης** | **Κατάσταση** |
|---|---|---|
| **ναι** | μικρή | ειλικρινής |
| **όχι** | μικρή | ειλικρινής |
| **όχι** | μεγάλη | ειλικρινής |
| **όχι** | μεγάλη | ειλικρινής |
| **όχι** | μικρή | ειλικρινής |
| **όχι** | μικρή | ειλικρινής |
| **ναι** | μικρή | δόλια |
| **ναι** | μεγάλη | δόλια |
| **όχι** | μεγάλη | δόλια |
| **ναι** | μεγάλη | δόλια |

Πρώτα θα υπολογίσουμε τις δεσμευμένες πιθανότητες απάτης, δεδομένων όλων των δυνατών συνδυασμών {ναι, μικρή}, {ναι, μεγάλη}, {όχι, μικρή} και {όχι, μεγάλη}. Για τον συνδυασμό {ναι, μικρή}, βρίσκουμε το ποσοστό των ζευγών {ναι, μικρή} τα οποία χαρακτηρίστηκαν ως "δόλια" και θα είναι $P(\delta\acute{o}\lambda\iota\alpha|\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}) = \frac{1}{4}$. Το συνολικό ποσοστό των "δόλιων" επιχειρήσεων είναι $P(\delta\acute{o}\lambda\iota\alpha) = \frac{4}{10}$, ενώ το συνολικό ποσοστό των επιχειρήσεων με χαρακτηρισμό {ναι, μικρή} είναι $P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}) = \frac{2}{10}$. Εφαρμόζοντας τον τύπο (3.1) υπολογίζουμε την πιθανότητα:

$$P(\delta\acute{o}\lambda\iota\alpha|\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}) = \frac{P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}|\delta\acute{o}\lambda\iota\alpha)P(\delta\acute{o}\lambda\iota\alpha)}{P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\})} = \frac{{}^{1}/_{4} \cdot {}^{4}/_{10}}{{}^{2}/_{10}} = 0{,}5$$

Ομοίως υπολογίζονται και οι πιθανότητες για τους υπόλοιπους συνδυασμούς:

$$P(\delta\acute{o}\lambda\iota\alpha|\{\nu\alpha\iota, \mu\epsilon\gamma\acute{\alpha}\lambda\eta\}) = \frac{{}^{2}/_{4} \cdot {}^{4}/_{10}}{{}^{2}/_{10}} = 1$$

$$P(\delta\acute{o}\lambda\iota\alpha|\{\acute{o}\chi\iota, \mu\iota\kappa\rho\acute{\eta}\}) = \frac{0 \cdot {}^{4}/_{10}}{{}^{3}/_{10}} = 0$$

$$P(\delta\acute{o}\lambda\iota\alpha|\{\acute{o}\chi\iota, \mu\epsilon\gamma\acute{\alpha}\lambda\eta\}) = \frac{{}^{1}/_{4} \cdot {}^{4}/_{10}}{{}^{3}/_{10}} = 0{,}33$$

Τώρα θα υπολογίσουμε τις ίδιες πιθανότητες με βάση τον αφελή ταξινομητή Bayes. Για τον συνδυασμό {ναι, μικρή}, θα ισχύει:

$$P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}|\delta\acute{o}\lambda\iota\alpha) = P(\nu\alpha\iota|\delta\acute{o}\lambda\iota\alpha)P(\mu\iota\kappa\rho\acute{\eta}|\delta\acute{o}\lambda\iota\alpha) = \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$$

$$P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}|\epsilon\iota\lambda\iota\kappa\rho\iota\nu\acute{\eta}\varsigma) = P(\nu\alpha\iota|\epsilon\iota\lambda\iota\kappa\rho\iota\nu\acute{\eta}\varsigma)P(\mu\iota\kappa\rho\acute{\eta}|\epsilon\iota\lambda\iota\kappa\rho\iota\nu\acute{\eta}\varsigma) = \frac{1}{6} \cdot \frac{4}{6} = \frac{1}{9}$$

Επομένως, η δεσμευμένη πιθανότητα απάτης από τον τύπο (3.2) θα είναι:

$$P_{NB}(\delta\acute{o}\lambda\iota\alpha|\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}) = \frac{P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}|\delta\acute{o}\lambda\iota\alpha)P(\delta\acute{o}\lambda\iota\alpha)}{P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}|\delta\acute{o}\lambda\iota\alpha)P(\delta\acute{o}\lambda\iota\alpha) + P(\{\nu\alpha\iota, \mu\iota\kappa\rho\acute{\eta}\}|\epsilon\iota\lambda\iota\kappa\rho\iota\nu\acute{\eta}\varsigma)P(\epsilon\iota\lambda\iota\kappa\rho\iota\nu\acute{\eta}\varsigma)} =$$
$$= \frac{{}^{3}/_{16} \cdot {}^{4}/_{10}}{{}^{3}/_{16} \cdot {}^{4}/_{10} + {}^{1}/_{9} \cdot {}^{6}/_{10}} = 0{,}53$$

Ομοίως οι υπόλοιπες πιθανότητες είναι:

$$P_{NB}(\delta\acute{o}\lambda\iota\alpha|\{\nu\alpha\iota, \mu\epsilon\gamma\acute{\alpha}\lambda\eta\}) = \frac{{}^{3}/_{4} \cdot {}^{3}/_{4} \cdot {}^{4}/_{10}}{{}^{3}/_{4} \cdot {}^{3}/_{4} \cdot {}^{4}/_{10} + {}^{1}/_{6} \cdot {}^{2}/_{6} \cdot {}^{6}/_{10}} = 0{,}87$$

$$P_{NB}(\delta\acute{o}\lambda\iota\alpha|\{\acute{o}\chi\iota, \mu\iota\kappa\rho\acute{\eta}\}) = \frac{{}^{1}/_{4} \cdot {}^{1}/_{4} \cdot {}^{4}/_{10}}{{}^{1}/_{4} \cdot {}^{1}/_{4} \cdot {}^{4}/_{10} + {}^{5}/_{6} \cdot {}^{4}/_{6} \cdot {}^{6}/_{10}} = 0{,}07$$

$$P_{NB}(\delta\acute{o}\lambda\iota\alpha|\{\acute{o}\chi\iota, \mu\epsilon\gamma\acute{\alpha}\lambda\eta\}) = \frac{{}^{1}/_{4} \cdot {}^{3}/_{4} \cdot {}^{4}/_{10}}{{}^{1}/_{4} \cdot {}^{3}/_{4} \cdot {}^{4}/_{10} + {}^{5}/_{6} \cdot {}^{2}/_{6} \cdot {}^{6}/_{10}} = 0{,}31$$

Παρατηρούμε ότι οι τιμές των πιθανοτήτων που προκύπτουν με τις δύο μεθόδους είναι πολύ κοντά. Παρά το γεγονός ότι δεν είναι ακριβώς ίσες, και οι δύο θα οδηγούσαν στην ίδια ταξινόμηση.

### 3.3.3 Πλεονεκτήματα και μειονεκτήματα του αφελή ταξινομητή Bayes

Ένα από τα βασικά πλεονεκτήματα του αφελή ταξινομητή Bayes είναι η απλότητά του, οι εύκολοι υπολογισμοί και η καλή απόδοση σε προβλήματα ταξινόμησης. Μάλιστα πολλές φορές η απόδοσή του είναι καλύτερη και από πιο εξελιγμένους ταξινομητές, ακόμα κι αν η υπόθεση ανεξαρτησίας των προγνωστικών παραγόντων δεν μπορεί να ισχύει στην πραγματικότητα. Αυτό το πλεονέκτημα μας εξυπηρετεί ιδιαίτερα όταν έχουμε να κάνουμε με μεγάλο αριθμό προγνωστικών παραγόντων. Ωστόσο, υπάρχουν και κάποια μειονεκτήματα.

Καταρχάς, ο αφελής ταξινομητής του Bayes απαιτεί έναν πολύ μεγάλο αριθμό παρατηρήσεων για να δώσει αξιόπιστα αποτελέσματα. Επίσης, όταν κάποιο από τα χαρακτηριστικά/από τους προγνωστικούς παράγοντες δεν εμφανίζεται στο σύνολο εκπαίδευσης, ο ταξινομητής αυτός υποθέτει ότι μία νέα παρατήρηση με αυτό το χαρακτηριστικό έχει μηδενική πιθανότητα. Αυτό μπορεί να δημιουργήσει πρόβλημα, αν το συγκεκριμένο σπάνιο χαρακτηριστικό είναι σημαντικό. Για παράδειγμα, ας υποθέσουμε ότι ο παράγοντας πρόβλεψης είναι αν ένα άτομο "αγόρασε ασφάλειας ζωής υψηλής αξίας" και ένας από τους προγνωστικούς παράγοντες είναι "έχει ιστιοφόρο" με τιμές 1 αν ισχύει και 0 αν όχι. Αν στο σύνολο εκπαίδευσης δεν υπάρχει καμία παρατήρηση για την οποία ο παράγοντας "έχει ιστιοφόρο"=1, τότε ο αφελής Bayes θα δώσει την τιμή 0 στην πιθανότητα να "αγόρασε ασφάλεια ζωής υψηλής αξίας", παρόλο που στην πραγματικότητα είναι πολύ πιθανό ένα άτομο που διαθέτει ιστιοφόρο να έχει αγοράσει ασφάλεια υψηλής αξίας. Κατά συνέπεια, αν από το σύνολο εκπαίδευσης απουσιάζει αυτό το χαρακτηριστικό, καμία τεχνική της εξόρυξης δεδομένων δε θα ενσωματώσει στο μοντέλο κατηγοριοποίησης αυτήν την ενδεχομένως σημαντική μεταβλητή˙ αντιθέτως, θα την αγνοήσει. Ένας τρόπος για να αποφύγουμε αυτήν τη συνέπεια, είναι να έχουμε στη διάθεσή μας ένα μεγάλο σύνολο εκπαίδευσης, κάνοντας μετασχηματισμό των συνεχών μεταβλητών σε δυαδικές όπου χρειάζεται.

Τέλος, η καλή απόδοση επιτυγχάνεται όταν ο στόχος μας είναι η ταξινόμηση παρατηρήσεων με βάση την πιθανότητα να ανήκουν σε μία συγκεκριμένη κατηγορία. Ωστόσο, όταν στοχεύουμε σε μία εκτίμηση της πιθανότητας αυτής, η μέθοδος του αφελή Bayes δίνει μεροληπτικά αποτελέσματα κι έτσι συνήθως δε χρησιμοποιείται για την εκτίμηση της πιστοληπτικής ικανότητας (credit scoring), δηλαδή για την εκτίμηση της πιθανότητας αθέτησης υποχρεώσεων από το δανειολήπτη.

## 3.4 κ-Κοντινότερος Γείτονας (k-NN)

Η γενική ιδέα της μεθόδου του κοντινότερου γείτονα είναι να καθορίσουμε ποιες από τις παρατηρήσεις του συνόλου εκπαίδευσης μοιάζουν με τη νέα παρατήρηση που θέλουμε να ταξινομήσουμε. Με βάση τις παρόμοιες (ή γειτονικές) παρατηρήσεις, αναθέτουμε τη νέα παρατήρηση στην κατηγορία που κυριαρχεί ανάμεσα στους γείτονες. Έστω ότι $(x_1, x_2, \dots, x_p)$ είναι τα χαρακτηριστικά της νέας παρατήρησης. Θα εντοπίσουμε τις παρατηρήσεις του συνόλου εκπαίδευσης, για τις οποίες οι τιμές των χαρακτηριστικών τους είναι κοντά στις τιμές των $(x_1, x_2, \dots, x_p)$. Στη συνέχεια, βασιζόμενοι στις κατηγορίες που ανήκουν οι γείτονες, τοποθετούμε τη νέα παρατήρηση σε μία από αυτές.

Ο αλγόριθμος του κ-κοντινότερου γείτονα είναι μία μέθοδος ταξινόμησης η οποία δεν κάνει καμία υπόθεση για τη σχέση που θα υπάρχει ανάμεσα στην κατηγορία Y και τους προγνωστικούς παράγοντες $X_1, X_2, \dots, X_p$. Είναι μία μη-παραμετρική μέθοδος αφού δεν εκτιμά τους προγνωστικούς παράγοντες, αλλά αντλεί πληροφορίες από τις ομοιότητες μεταξύ των προγνωστικών παραγόντων στο σύνολο των δεδομένων.

Για να μπορέσουμε να εντοπίσουμε τις γειτονικές παρατηρήσεις, θα πρέπει να βρούμε έναν τρόπο να μετράμε την απόσταση μεταξύ των παρατηρήσεων, βάσει των τιμών των προγνωστικών παραγόντων, και συνήθως χρησιμοποιούμε την Ευκλείδεια απόσταση. Η Ευκλείδεια απόσταση ανάμεσα σε δύο παρατηρήσεις με χαρακτηριστικά $(x_1, x_2, \dots, x_p)$ και $(w_1, w_2, \dots, w_p)$ αντίστοιχα, θα είναι ίση με:

$$\boldsymbol{d(x,w) = \sqrt{(x_1 - w_1)^2 + (x_2 - w_2)^2 + \cdots + (x_p - w_p)^2}} \quad \boldsymbol{(3.3)}$$

Αφού υπολογίσουμε τις ζητούμενες αποστάσεις, θα χρειαστούμε έναν κανόνα για την ανάθεση της νέας παρατήρησης σε μία από τις κατηγορίες των γειτόνων. Η πιο απλή περίπτωση είναι για κ=1 όπου αναζητούμε τον γείτονα που είναι πιο κοντά από όλους και τοποθετούμε τη νέα παρατήρηση στην ίδια κατηγορία με αυτόν. Είναι αξιοσημείωτο το γεγονός ότι αυτή η απλή και διαισθητική μέθοδος είναι πολύ αποδοτική όταν έχουμε πολλές παρατηρήσεις στο σύνολο εκπαίδευσης. Συγκεκριμένα, αποδεικνύεται ότι το ποσοστό του σφάλματος ταξινόμησης του 1-κοντινότερου γείτονα δεν ξεπερνά το διπλάσιο του σφάλματος, όταν γνωρίζουμε τη συνάρτηση πυκνότητας πιθανότητας για την κάθε κατηγορία.

Η ιδέα του 1-κοντινότερου γείτονα γενικεύεται για $\kappa > 1$ με τον εξής αλγόριθμο:

1. Βρες τους κ-κοντινότερους γείτονες στη νέα παρατήρηση.

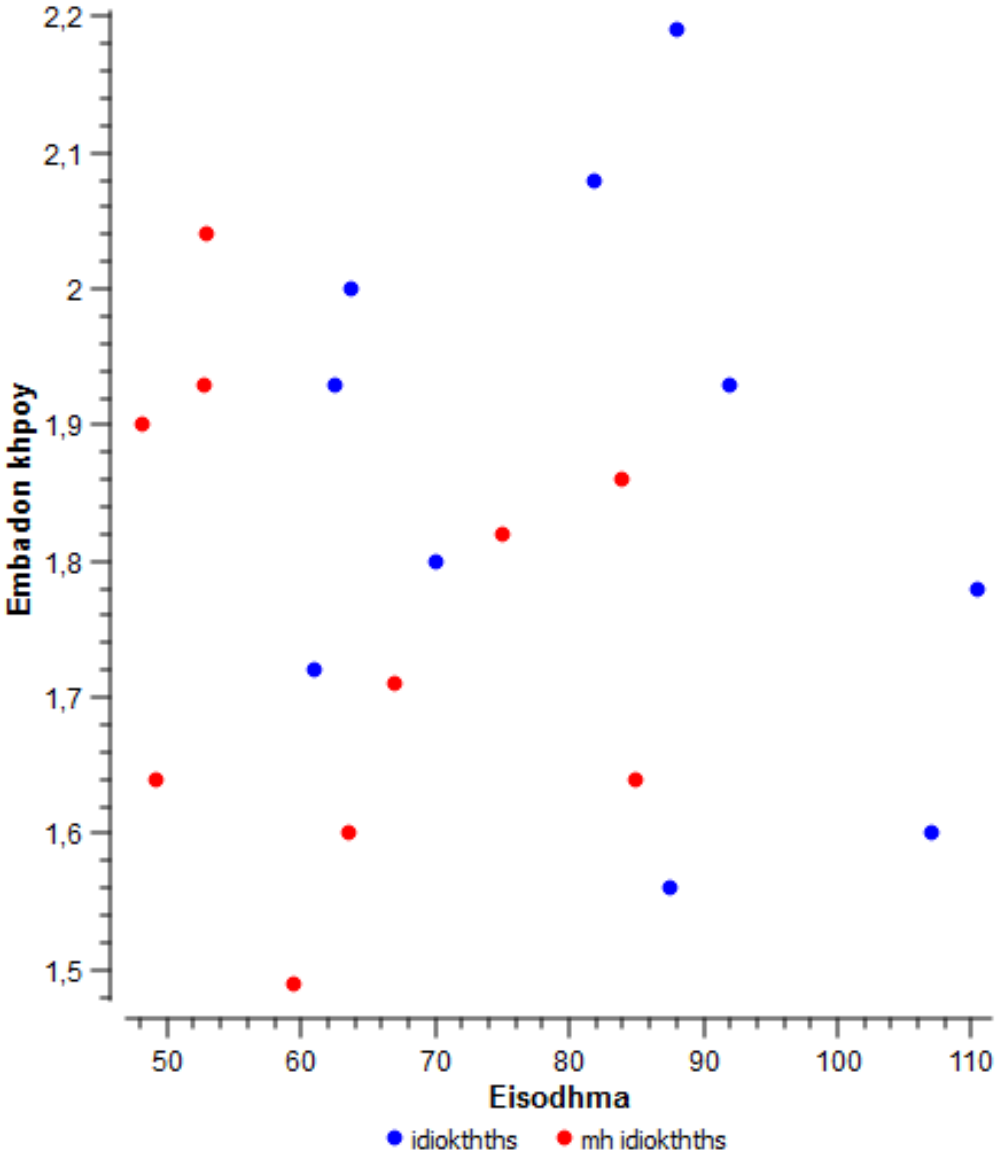
2. Χρησιμοποίησε τον κανόνα ταξινόμησης, σύμφωνα με τον οποίο η νέα παρατήρηση κατατάσσεται στην κατηγορία με τους περισσότερους από τους κ γείτονες.

### 3.4.1 Παράδειγμα 2: Αγορά συστήματος αυτόματου ποτίσματος

Μία εταιρεία κατασκευής συστημάτων αυτόματου ποτίσματος θέλει να βρει έναν τρόπο να ταξινομεί τις οικογένειες μιας πόλης, ανάλογα με τον αν είναι πιο πιθανό να αγοράσουν ή να μην αγοράσουν ένα τέτοιο μηχάνημα. Στον παρακάτω πίνακα φαίνεται ένα δείγμα από 10 οικογένειες που είναι ιδιοκτήτες και 10 οικογένειες που δεν είναι.

| Αριθμός σπιτιού | Εισόδημα (χιλιάδες €/έτος) | Εμβαδόν κήπου (στρέμματα) | Κατάσταση |
|---|---|---|---|
| **1** | 61.0 | 1.72 | ιδιοκτήτης |
| **2** | 87.5 | 1.56 | ιδιοκτήτης |
| **3** | 63.8 | 2.00 | ιδιοκτήτης |
| **4** | 62.5 | 1.93 | ιδιοκτήτης |
| **5** | 88.0 | 2.19 | ιδιοκτήτης |
| **6** | 110.4 | 1.78 | ιδιοκτήτης |
| **7** | 107.0 | 1.60 | ιδιοκτήτης |
| **8** | 81.8 | 2.08 | ιδιοκτήτης |
| **9** | 70.0 | 1.80 | ιδιοκτήτης |
| **10** | 92.0 | 1.93 | ιδιοκτήτης |
| **11** | 53.0 | 2.04 | μη ιδιοκτήτης |
| **12** | 84.0 | 1.86 | μη ιδιοκτήτης |
| **13** | 75.0 | 1.82 | μη ιδιοκτήτης |
| **14** | 52.7 | 1.93 | μη ιδιοκτήτης |
| **15** | 63.6 | 1.60 | μη ιδιοκτήτης |
| **16** | 48.2 | 1.90 | μη ιδιοκτήτης |
| **17** | 85.0 | 1.64 | μη ιδιοκτήτης |
| **18** | 49.2 | 1.64 | μη ιδιοκτήτης |
| **19** | 59.5 | 1.49 | μη ιδιοκτήτης |
| **20** | 67.0 | 1.71 | μη ιδιοκτήτης |

**Πίνακας 3.1. Εισόδημα, Εμβαδόν κήπου και Κατάσταση για 20 σπίτια**

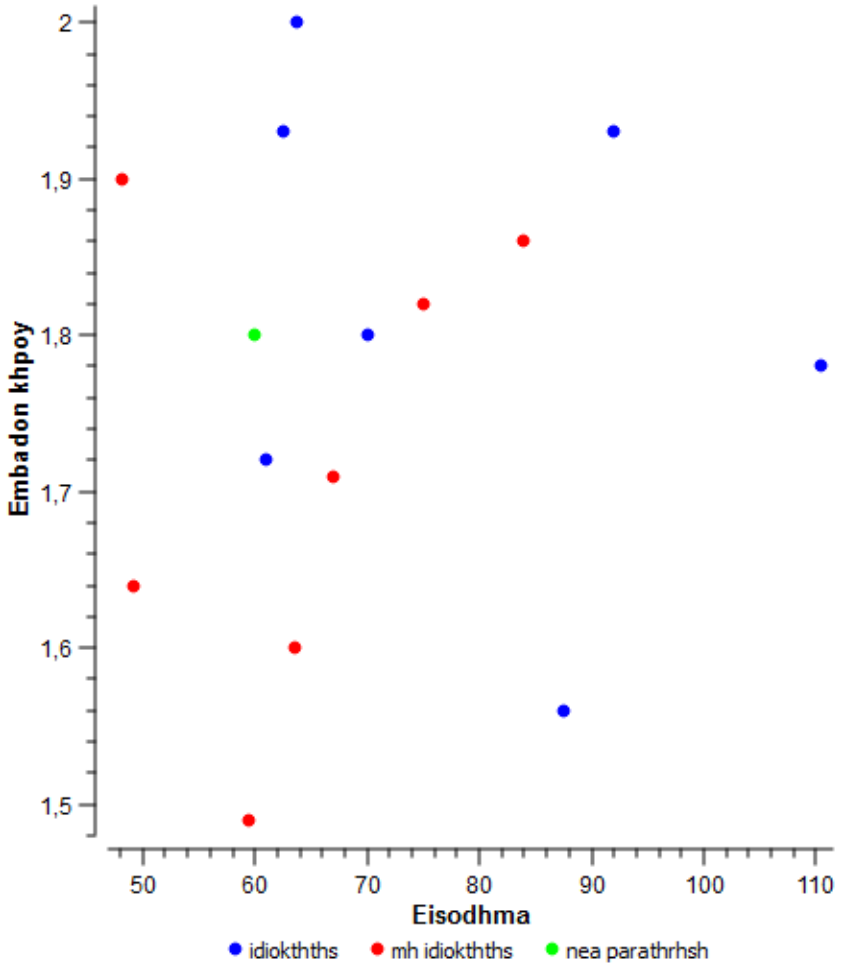



**Σχήμα 3.1. Διάγραμμα διασποράς "Εμβαδόν κήπου" vs "Εισόδημα" για τα 20 σπίτια**

Αρχικά χωρίζουμε τα δεδομένα σε ένα σύνολο εκπαίδευσης με 14 οικογένειες (περίπου τα $^2/_3$ του αρχικού συνόλου) και ένα σύνολο ελέγχου με 6 οικογένειες (περίπου το $^1/_3$ του αρχικού συνόλου). Ας υποθέσουμε τώρα ότι εισάγεται μία νέα παρατήρηση με εισόδημα 60.0 και εμβαδόν αυλής 1.80. Θέλουμε να ταξινομήσουμε τη νέα παρατήρηση με τη μέθοδο του κ-κοντινότερου γείτονα.

| Αριθμός σπιτιού | Εισόδημα (χιλιάδες €/έτος) | Εμβαδόν κήπου (στρέμματα) | Κατάσταση |
|---|---|---|---|
| **1** | 61.0 | 1.72 | ιδιοκτήτης |
| **2** | 87.5 | 1.56 | ιδιοκτήτης |
| **3** | 63.8 | 2.00 | ιδιοκτήτης |
| **4** | 62.5 | 1.93 | ιδιοκτήτης |
| **6** | 110.4 | 1.78 | ιδιοκτήτης |
| **9** | 70.0 | 1.80 | ιδιοκτήτης |
| **10** | 92.0 | 1.93 | ιδιοκτήτης |
| **12** | 84.0 | 1.86 | μη ιδιοκτήτης |
| **13** | 75.0 | 1.82 | μη ιδιοκτήτης |
| **15** | 63.6 | 1.60 | μη ιδιοκτήτης |
| **16** | 48.2 | 1.90 | μη ιδιοκτήτης |
| **18** | 49.2 | 1.64 | μη ιδιοκτήτης |
| **19** | 59.5 | 1.49 | μη ιδιοκτήτης |
| **20** | 67.0 | 1.71 | μη ιδιοκτήτης |
| | 60.0 | 1.80 | νέα παρατήρηση |

**Πίνακας 3.2. Οι 14 παρατηρήσεις του συνόλου ελέγχου (& η νέα παρατήρηση)**




**Σχήμα 3.2. Διάγραμμα διασποράς "Εμβαδόν κήπου" vs "Εισόδημα" για τα 15 σπίτια**

Σημειώνουμε ότι ο εντοπισμός των κοντινότερων γειτόνων δεν μπορεί να γίνει από το διάγραμμα διασποράς (σχήμα 3.2), αφού οι δύο παράγοντες παίρνουν τιμές σε διαφορετική κλίμακα. Έτσι, λόγω της μεγάλης μεταβλητότητας μεταξύ των τιμών των δύο παραγόντων, οι πραγματικές αποστάσεις δεν είναι ευδιάκριτες στο διάγραμμα.

Στον πίνακα που ακολουθεί υπολογίζονται οι Ευκλείδειες αποστάσεις κάθε σημείου του συνόλου εκπαίδευσης από τη νέα παρατήρηση:

| Αριθμός σπιτιού | Εισόδημα (χιλιάδες €/έτος) | Εμβαδόν κήπου (στρέμματα) | Κατάσταση | Απόσταση |
|---|---|---|---|---|
| **1** | 61.0 | 1.72 | ιδιοκτήτης | 1.00 |
| **2** | 87.5 | 1.56 | ιδιοκτήτης | 27.50 |
| **3** | 63.8 | 2.00 | ιδιοκτήτης | 3.81 |
| **4** | 62.5 | 1.93 | ιδιοκτήτης | 2.50 |
| **6** | 110.4 | 1.78 | ιδιοκτήτης | 50.40 |
| **9** | 70.0 | 1.80 | ιδιοκτήτης | 10.00 |
| **10** | 92.0 | 1.93 | ιδιοκτήτης | 12.00 |
| **12** | 84.0 | 1.86 | μη ιδιοκτήτης | 24.00 |
| **13** | 75.0 | 1.82 | μη ιδιοκτήτης | 15.00 |
| **15** | 63.6 | 1.60 | μη ιδιοκτήτης | 3.61 |
| **16** | 48.2 | 1.90 | μη ιδιοκτήτης | 11.80 |
| **18** | 49.2 | 1.64 | μη ιδιοκτήτης | 10.80 |
| **19** | 59.5 | 1.49 | μη ιδιοκτήτης | 0.59 |
| **20** | 67.0 | 1.71 | μη ιδιοκτήτης | 7.00 |

**Πίνακας 3.3. Ευκλείδειες αποστάσεις από τη νέα παρατήρηση**

Από τον πίνακα βλέπουμε ότι το πιο κοντινό σημείο στη νέα παρατήρηση είναι το σπίτι #19, με εισόδημα 59.5 και εμβαδόν αυλής 1.49. Αν, λοιπόν, εφαρμόσουμε τη μέθοδο του 1-κοντινότερου γείτονα, θα ταξινομήσουμε τη νέα παρατήρηση ως "μη ιδιοκτήτη". Αντίστοιχα, για κ=3 παρατηρούμε ότι οι τρεις κοντινότεροι γείτονες είναι τα σπίτια #19, #1 και #4. Πράγματι, υπολογίζοντας αναλυτικά τις Ευκλείδειες αποστάσεις για τους τρεις αυτούς γείτονες από τη σχέση (3.3) είναι:

$$\boldsymbol{d_{19} = \sqrt{(1.80 - 1.49)^2 + (60.0 - 59.5)^2} \approx 0.59}$$

$$\boldsymbol{d_{1} = \sqrt{(1.80 - 1.72)^2 + (60.0 - 61.0)^2} \approx 1.00}$$

$$\boldsymbol{d_{4} = \sqrt{(1.80 - 1.93)^2 + (60.0 - 62.5)^2} \approx 2.50}$$

Τα σπίτια #1 και #4 ανήκουν στην κατηγορία "ιδιοκτήτες" ενώ το #9 είναι "μη ιδιοκτήτης". Επομένως, σύμφωνα με την πλειοψηφία των ψήφων, η νέα παρατήρηση θα ταξινομηθεί ως "ιδιοκτήτης".

### 3.4.2 Επιλογή του κ

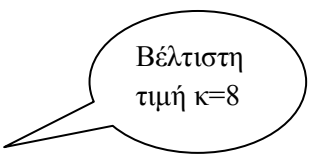
Σε γενικές γραμμές, αν το κ πάρει πολύ μικρή τιμή υπάρχει ο κίνδυνος «μόλυνσης» του μοντέλου από παράγοντες θορύβου στα δεδομένα. Δίνοντας μεγάλες τιμές στο κ αποφεύγουμε τον κίνδυνο της υπερπροσαρμογής, αλλά χάνουμε τη δυνατότητα να αποτυπώσουμε την «τοπική» δομή των δεδομένων. Στην ακραία περίπτωση όπου το κ είναι ίσο με το πλήθος των δεδομένων στο σύνολο εκπαίδευσης, θα αναθέσουμε όλα τα δεδομένα στην κυρίαρχη κατηγορία ανεξάρτητα από τις τιμές των $(x_1, x_2, \dots, x_p)$, και αυτό συμπίπτει με τη μέθοδο του αφελή κανόνα.

Θέλουμε, λοιπόν, να επιλέξουμε το κατάλληλο κ ώστε να υπάρχει ισορροπία, δηλαδή αφενός να αποφύγουμε την υπερπροσαρμογή και αφετέρου να μην αγνοήσουμε τις πληροφορίες από τους προγνωστικούς παράγοντες. Αυτή η ισορροπημένη επιλογή του κ εξαρτάται σε μεγάλο βαθμό από τη φύση των δεδομένων. Όσο πιο περίπλοκη και ακανόνιστη είναι η δομή των δεδομένων, τόσο χαμηλότερη είναι η βέλτιστη τιμή του κ. Συνήθως, οι τιμές κυμαίνονται μεταξύ του 1 και του 20.

Πώς, λοιπόν, επιλέγουμε το κατάλληλο κ; Επιλέγουμε εκείνο το κ που έχει την καλύτερη απόδοση ταξινόμησης. Με βάση το σύνολο εκπαίδευσης κατασκευάζουμε το μοντέλο και στη συνέχεια το χρησιμοποιούμε για να ταξινομήσουμε τα δεδομένα του συνόλου ελέγχου. Έτσι μπορούμε να ελέγξουμε την απόδοσή του, υπολογίζοντας τα ποσοστά λανθασμένης ταξινόμησης για τις διάφορες τιμές του κ.

Στο παράδειγμα που ακολουθεί, βλέπουμε ότι αν επιλέξουμε την τιμή κ=1, το μοντέλο θα είναι προσαρμοσμένο υπερβολικά κοντά στα δεδομένα του συνόλου εκπαίδευσης. Αντίστοιχα, για πολύ μεγάλη τιμή του κ, όπως κ=18, θα προβλέψουμε απλώς την κατηγορία που κυριαρχεί ανάμεσα σε όλα τα δεδομένα του συνόλου. Επομένως, για να βρούμε το κατάλληλο κ θα εντοπίσουμε το χαμηλότερο ποσοστό λανθασμένης ταξινόμησης. Όπως βλέπουμε στον πίνακα, το ποσοστό σφάλματος στο σύνολο ελέγχου ελαχιστοποιείται για κ=8. Επισημαίνουμε, όμως, ότι εφόσον τώρα χρησιμοποιήσαμε και το σύνολο εκπαίδευσης και το σύνολο ελέγχου, θα χρειαστούμε ένα επιπλέον σύνολο ελέγχου για να εκτιμήσουμε την αποδοτικότητα της μεθόδου σε παρατηρήσεις που δεν έχει «ξαναδεί».

| Τιμή κ | %Σφάλμα στο σύνολο εκπαίδευσης | %Σφάλμα στο σύνολο ελέγχου |
|---|---|---|
| **1** | 0.00 | 33.33 |
| **2** | 16.67 | 33.33 |
| **3** | 11.11 | 33.33 |
| **4** | 22.22 | 33.33 |
| **5** | 11.11 | 33.33 |
| **6** | 27.78 | 33.33 |
| **7** | 22.22 | 33.33 |
| **8** | 22.22 | 16.67 |
| **9** | 22.22 | 16.67 |
| **10** | 22.22 | 16.67 |



| | | |
|---|---|---|
| **11** | 16.67 | 33.33 |
| **12** | 16.67 | 16.67 |
| **13** | 11.11 | 33.33 |
| **14** | 11.11 | 16.67 |
| **15** | 5.56 | 33.33 |
| **16** | 16.67 | 33.33 |
| **17** | 11.11 | 33.33 |
| **18** | 50.00 | 50.00 |

### 3.4.3 k-NN για συνεχείς μεταβλητές

Όπως έχουμε ήδη αναφέρει, οι εξαρτημένες και ανεξάρτητες μεταβλητές μπορούν να χαρακτηριστούν είτε ως κατηγορικές είτε ως συνεχείς. Για κατηγορικές εξαρτημένες μεταβλητές χρησιμοποιούμε τη μέθοδο της ταξινόμησης, ενώ για συνεχείς εφαρμόζουμε τη μέθοδο της παλινδρόμησης. Ωστόσο ο κ-κοντινότερος γείτονας μπορεί να εφαρμοστεί και στις δύο αυτές μεθόδους. Η διαφορά είναι ότι σε προβλήματα ταξινόμησης χρησιμοποιείται η πλειοψηφία των ψήφων, ενώ σε προβλήματα παλινδρόμησης οι προβλέψεις βασίζονται στον μέσο όρο των αποτελεσμάτων των κ γειτόνων.

### 3.4.4 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου k-NN

Το βασικό πλεονέκτημα της μεθόδου k-NN είναι η απλότητά της και το ότι δεν κάνει υποθέσεις για τις παραμέτρους. Δουλεύει πολύ καλά όταν έχουμε ένα μεγάλο σύνολο εκπαίδευσης και ειδικότερα όταν η κάθε κατηγορία χαρακτηρίζεται από διάφορους συνδυασμούς των προγνωστικών παραγόντων.

Ωστόσο υπάρχουν δύο δυσκολίες στην πρακτική εφαρμογή της μεθόδου k-NN. Η πρώτη δυσκολία είναι ότι ενώ δεν απαιτείται χρόνος για την εκτίμηση των παραμέτρων από το σύνολο εκπαίδευσης, η διαδικασία εύρεσης των κοντινότερων γειτόνων σε ένα μεγάλο σύνολο μπορεί να είναι χρονοβόρα. Μερικοί τρόποι για να ξεπεράσουμε αυτή τη δυσκολία είναι οι εξής:

1. Να μειώσουμε τη διάσταση των δεδομένων, ώστε να ελαττωθεί ο χρόνος που απαιτείται για τον υπολογισμό των αποστάσεων. Η ανάλυση σε κύριες συνιστώσες είναι μία τεχνική που στοχεύει στη μείωση της διάστασης διατηρώντας τη διακύμανση $\sum_{i=1}^{p} Var(X_i)$ των αρχικών μεταβλητών $X_1, X_2, \dots, X_p$. Συγκεκριμένα, βρίσκουμε λιγότερες μεταβλητές $Y_1, Y_2, \dots, Y_r$ $με$ $r < p$, που να είναι γραμμικοί συνδυασμοί των αρχικών $X_1, X_2, \dots, X_p$, ασυσχέτιστες μεταξύ τους και με διασπορά σχεδόν ίση με αυτή των αρχικών δηλαδή $\sum_{i=1}^{p} Var(X_i) \approx \sum_{i=1}^{r} Var(Y_i)$. Οι μεταβλητές $Y_1, Y_2, \dots, Y_r$ θα αποτελούν τις κύριες

συνιστώσες. Έτσι, αντί να αναλύσουμε τα δεδομένα στον χώρο $\Re^p$, τα αναλύουμε στον $\Re^r$.

2. Να χρησιμοποιήσουμε πιο εξελιγμένες δομές δεδομένων, όπως για παράδειγμα δέντρα αναζήτησης, τα οποία αποτελούν έναν τρόπο αναπαράστασης μιας ταξινομημένης λίστας αντικειμένων. Στόχος είναι να επιταχύνουμε τον εντοπισμό του κοντινότερου γείτονα, ενώ πολλές φορές συμβιβαζόμαστε και με τον «σχεδόν κοντινότερο» γείτονα για ακόμη πιο γρήγορα αποτελέσματα.

3. Να επεξεργαστούμε τα δεδομένα του συνόλου εκπαίδευσης ώστε να αφαιρέσουμε τα σημεία που πλεονάζουν, επιταχύνοντας έτσι τη διαδικασία εύρεσης του κοντινότερου γείτονα. Ένας τρόπος για να το πετύχουμε αυτό, είναι να αφαιρέσουμε τα δεδομένα του συνόλου εκπαίδευσης που περιβάλλονται από δεδομένα της ίδιας κατηγορίας και επομένως δεν έχουν καμία επίδραση στη διαδικασία της ταξινόμησης.

Η δεύτερη δυσκολία της μεθόδου k-NN είναι ότι για να θεωρηθεί ικανοποιητικά μεγάλος ο αριθμός των δεδομένων του συνόλου εκπαίδευσης, θα πρέπει αυτός να αυξάνεται εκθετικά σε σχέση με τον αριθμό των προγνωστικών παραγόντων p. Αυτό συμβαίνει γιατί καθώς αυξάνεται το p, αυξάνεται και η διάσταση των δεδομένων και κατά συνέπεια ο όγκος του χώρου μεγαλώνει τόσο γρήγορα που τα δεδομένα φαίνονται «αραιά». Με άλλα λόγια, η αναμενόμενη απόσταση μέχρι τον κοντινότερο γείτονα ανεβαίνει δραματικά όσο ανεβαίνει το p, εκτός αν αυξήσουμε εκθετικά το μέγεθος του συνόλου εκπαίδευσης σε σχέση με το p. Αυτή είναι η "κατάρα της διαστατικότητας", όπως αναφέρεται και στην παράγραφο 2.5. Η κατάρα της διαστατικότητας είναι ένα θεμελιώδες ζήτημα σε πολλά προβλήματα ταξινόμησης, πρόβλεψης και συσταδοποίησης. Για τον λόγο αυτό, συνήθως επιδιώκουμε να μειώσουμε τη διάσταση των δεδομένων με διάφορες μεθόδους όπως η επιλογή υποσυνόλων των προγνωστικών παραγόντων ή ο συνδυασμός προγνωστικών παραγόντων μέσω της ανάλυσης σε κύριες συνιστώσες (principal components analysis), της αποσύνθεσης μοναδικών τιμών (singular value decomposition), της ανάλυσης παραγόντων (factor analysis) κ.α.

## 3.5 Εφαρμογές

Σε αυτήν την παράγραφο εφαρμόζουμε τις τρεις στατιστικές μεθόδους σε πραγματικά δεδομένα.

### 3.5.1 Εφαρμογή 1: Ζωολογικός κήπος

Διαθέτουμε τα δεδομένα των ζώων που βρέθηκαν σε ένα ζωολογικό κήπο, τα οποία αποτελούνται από 17 χαρακτηριστικά. Το χαρακτηριστικό "τύπος" αποδίδεται ως η ετικέτα κατηγορίας και σκοπός μας είναι να βρούμε έναν ταξινομητή που να αναθέτει το κάθε ζώο σε

μία από τις κατηγορίες θηλαστικό, ψάρι, πτηνό, ασπόνδυλο, έντομο, αμφίβιο ή ερπετό. Οι πληροφορίες για το κάθε χαρακτηριστικό φαίνονται στον ακόλουθο πίνακα:

| Χαρακτηριστικό | Τιμές μεταβλητής |
|---|---|
| **όνομα ζώου** | μοναδικό για κάθε παρατήρηση |
| **τρίχωμα** | δυαδική (1=ναι, 0=όχι) |
| **φτερά** | δυαδική |
| **αυγά** | δυαδική |
| **γάλα** | δυαδική |
| **ικανότητα να πετάει** | δυαδική |
| **ικανότητα να αναπνέει στο νερό** | δυαδική |
| **αρπακτικό** | δυαδική |
| **δόντια** | δυαδική |
| **σπονδυλική στήλη** | δυαδική |
| **αναπνέει** | δυαδική |
| **δηλητηριώδες** | δυαδική |
| **πτερύγιο** | δυαδική |
| **πόδια** | αριθμητική με τιμές {0, 2, 4, 6, 8} |
| **ουρά** | δυαδική |
| **οικόσιτο** | δυαδική |
| **μέγεθος** | δυαδική |
| **τύπος** | αριθμητική με ακέραιες τιμές στο διάστημα [1,7] |

**Πίνακας 3.4. Περιγραφή των μεταβλητών για τα δεδομένα του ζωολογικού κήπου**

Συνολικά έχουμε 101 παρατηρήσεις και στον επόμενο πίνακα παρουσιάζονται τα ονόματα και το πλήθος των ζώων που ανήκουν σε κάθε τύπο:

| Τύπος | Όνομα ζώου | Συνολικό πλήθος |
|---|---|---|
| **1=θηλαστικό** | μυρμηγκοφάγος, αντιλόπη, αρκούδα, αγριόχοιρος, βουβάλι, μοσχάρι, ποντικός της Ν. Αμερικής, γατόπαρδος, ελάφι, δελφίνι, ελέφαντας, νυχτερίδα των φρούτων, καμηλοπάρδαλη, άνθρωπος, τράγος, γορίλας, χάμστερ, λαγός, λεοπάρδαλη, λιοντάρι, λύγκας, βιζόν, τυφλοπόντικας, νυφίτσα, μαρσιπόμυς, γαζέλα, πλατύπους, κουνάβι, πόνυ, φώκαινα, πούμα, γάτα, ρακούν, τάρανδος, φώκια, θαλάσσιος λέοντας, σκίουρος, νυχτερίδα βαμπίρ, αρουραίος, καγκουρό, λύκος | 41 |
| **2=πτηνό** | κότα, κοράκι, περιστέρι, πάπια, φλαμίνγκο, γλάρος, γεράκι, κίβι, κορυδαλλός, στρουθοκάμηλος, παπαγάλος, πιγκουίνος, φασιανός, ρέα, ρύγχωψ, ληστόγλαρος, σπουργίτι, κύκνος, γύπας, τρυποφράχτης | 20 |
| **3=ερπετό** | κροταλίας, φίδι της θάλασσας, τυφλίνος, χελώνα, τουατάρα | 5 |
| **4=ψάρι** | πέρκα, κυπρίνος, γατόψαρο, κέφαλος, σκυλόψαρο, βακαλάος, ρέγκα, λούτσος, πιράνχα, ιππόκαμπος, γλώσσα, σαλάχι, τόνος | 13 |
| **5=αμφίβιο** | κοινός βάτραχος, δηλητηριώδης βάτραχος, σαλαμάνδρα, φρύνος | 4 |
| **6=έντομο** | ψύλλος, σκνίπα, μέλισσα, μύγα, πασχαλίτσα, σκόρος, τερμίτης, σφήκα | 8 |
| **7=ασπόνδυλο** | μύδι, κάβουρας, καραβίδα, αστακός, χταπόδι, σκορπιός, μέδουσα, γυμνοσάλιαγκας, αστερίας, σκουλήκι | 10 |

**Πίνακας 3.5. Αναλυτική περιγραφή των δεδομένων του ζωολογικού κήπου**

**Εικόνα: Φώκαινα (Phocoena phocoena). Απόσπασμα από το έργο "Περί τα ζώια ιστοριών" του Αριστοτέλη (350 π.Χ.)**

Δελφὶς δὲ καὶ φάλαινα καὶ τὰ ἄλλα κήτη, ὅσα μὴ ἔχει βράγχια ἀλλὰ φυσητῆρα, ζῳοτοκοῦσιν, ἔτι δὲ πρίστις καὶ βοῦς· οὐδὲν γὰρ τούτων φαίνεται ἔχον ᾠά, ἀλλ' εὐθέως κύημα, ἐξ οὗ διαρθρουμένου γίνεται τὸ ζῷον, καθάπερ ἄνθρωπος καὶ τῶν τετραπόδων τὰ ζῳοτόκα. Τίκτει δ' ὁ μὲν δελφὶς τὰ μὲν πολλὰ ἕν, ἐνίοτε δὲ καὶ δύο· ἡ δὲ φάλαινα ἢ δύο τὰ πλεῖστα καὶ πλεονάκις, ἢ ἕν. Ὁμοίως δὲ τῷ δελφῖνι καὶ ἡ φώκαινα· καὶ γάρ ἐστιν ὅμοιον δελφῖνι μικρῷ, γίνεται δ' ἐν τῷ Πόντῳ. Διαφέρει δὲ φώκαινα δελφῖνος· ἔστι γὰρ τὸ μέγεθος ἔλαττον, εὐρύτερον δ' ἐκ τοῦ νώτου· τὸ χρῶμα ἔχει κυανοῦν. Πολλοὶ δὲ δελφίνων τι γένος εἶναί φασι τὴν φώκαιναν. Ἀναπνεῖ δὲ πάντα ὅσα ἔχει φυσητῆρα, καὶ δέχεται τὸν ἀέρα· πλεύμονα γὰρ ἔχουσιν. Καὶ ὅ γε δελφὶς ὦπται, ὅταν καθεύδῃ, ὑπερέχων τὸ ῥύγχος, καὶ ῥέγχει καθεύδων. Ἔχει δ' ὁ δελφὶς καὶ ἡ φώκαινα γάλα, καὶ θηλάζονται· καὶ εἰσδέχονται δὲ τὰ τέκνα μικρὰ ὄντα. Τὴν δ' αὔξησιν τὰ τέκνα τῶν δελφίνων ποιοῦνται ταχεῖαν· ἐν ἔτεσι γὰρ δέκα μέγεθος λαμβάνουσι τέλεον. Κύει δὲ δέκα μῆνας. Τίκτει δ' ὁ δελφὶς ἐν τῷ θέρει, ἐν ἄλλῃ δ' ὥρᾳ οὐδεμιᾷ· συμβαίνει δὲ καὶ ἀφανίζεσθαι αὐτὸν ὑπὸ κύνα περὶ τριάκονθ' ἡμέρας. Παρακολουθεῖ δὲ τὰ τέκνα πολὺν χρόνον, καὶ ἔστι τὸ ζῷον φιλότεκνον. Ζῇ δ' ἔτη πολλά· δῆλοι γὰρ ἔνιοι γεγόνασι βιοῦντες οἱ μὲν πλείω ἔτη ἢ πέντε καὶ εἴκοσιν, οἱ δὲ τριάκοντα· ἀποκόπτοντες γὰρ ἐνίων τὸ οὐραῖον οἱ ἁλιεῖς ἀφιᾶσιν, ὥστε τούτῳ γνωρίζουσι τοὺς χρόνους αὐτῶν.Ἡ δὲ φώκη ἐστὶ τῶν ἐπαμφοτεριζόντων ζῴων· οὐ δέχεται μὲν γὰρ τὸ ὕδωρ, ἀλλ' ἀναπνεῖ καὶ καθεύδει καὶ τίκτει ἐν τῇ γῇ μέν, πρὸς αἰγιαλοῖς δέ, ὡς οὖσα τῶν πεζῶν, διατρίβει δὲ τοῦ χρόνου τὸν πολὺν καὶ τρέφεται ἐκ τῆς θαλάττης, διὸ μετὰ τῶν ἐνύδρων περὶ αὐτῆς λεκτέον. Ζῳοτοκεῖ μὲν οὖν εὐθὺς ἐν αὐτῇ, καὶ τίκτει ζῷα, καὶ χόριον καὶ τἆλλα προῖεται ὥσπερ πρόβατον. Τίκτει δ' ἓν ἢ δύο, τὰ δὲ πλεῖστα τρία. Καὶ μαστοὺς δ' ἔχει δύο καὶ θηλάζεται ὑπὸ τῶν τέκνων καθάπερ τὰ τετράποδα. Τίκτει δ' ὥσπερ ἄνθρωπος πᾶσαν ὥραν τοῦ ἔτους, μάλιστα δ' ἅμα ταῖς πρώταις αἰξίν. Ἄγει δὲ περὶ δωδεκαταῖα ὄντα τὰ τέκνα εἰς τὴν θάλατταν πολλάκις τῆς ἡμέρας, συνεθίζουσα κατὰ μικρόν· τὰ δὲ κατάντη φέρεται, ἀλλ' οὐ βαδίζει, διὰ τὸ μὴ δύνασθαι ἀπερείδεσθαι τοῖς ποσίν. Συνάγει δὲ καὶ συστέλλει ἑαυτήν· σαρκῶδες γάρ ἐστι καὶ μαλακόν, καὶ ὀστᾶ χονδρώδη ἔχει. Ἀποκτεῖναι δὲ φώκην χαλεπὸν βιαίως, ἐὰν μή τις πατάξῃ παρὰ τὸν κρόταφον· τὸ γὰρ σῶμα σαρκῶδες αὐτῆς ἐστιν. Ἀφίησι δὲ φωνὴν ὁμοίαν βοΐ. Ἔχει δὲ καὶ τὸ αἰδοῖον ἡ θήλεια ὅμοιον προβάτῳ, πάντα δὲ τἆλλα γυναικί.Περὶ μὲν οὖν τῶν ἐνύδρων καὶ ζῳοτοκούντων ἢ ἐν αὐτοῖς ἢ ἔξω ἡ γένεσις καὶ τὰ περὶ τὸν τόκον τοῦτον ἔχει τὸν τρόπον.

Αρχικά χωρίζουμε το σύνολο των δεδομένων σε ένα σύνολο εκπαίδευσης (61 παρατηρήσεις) και ένα σύνολο ελέγχου (40 παρατηρήσεις), με τυχαίο τρόπο.

## ***Αφελής κανόνας***

Σύμφωνα με τον αφελή κανόνα, θα πρέπει να αναθέσουμε όλες τις παρατηρήσεις στην κυρίαρχη κατηγορία, αγνοώντας τις πληροφορίες για τα χαρακτηριστικά του κάθε ζώου. Στην περίπτωσή μας, το ποσοστό των παρατηρήσεων για τον κάθε τύπο είναι:

| Τύπος | Ποσοστό |
|---|---|
| **1** | $\frac{41}{101} \approx 40\%$ |
| **2** | $\frac{20}{101} \approx 20\%$ |
| **3** | $\frac{5}{101} \approx 5\%$ |
| **4** | $\frac{13}{101} \approx 13\%$ |
| **5** | $\frac{4}{101} \approx 4\%$ |
| **6** | $\frac{8}{101} \approx 8\%$ |
| **7** | $\frac{10}{101} \approx 10\%$ |

Εφόσον λοιπόν η κατηγορία στην οποία ανήκουν οι περισσότερες παρατηρήσεις είναι ο τύπος "θηλαστικό", ο αφελής κανόνας θα χαρακτηρίσει όλες τις παρατηρήσεις του συνόλου ελέγχου ως "θηλαστικό". Η προβλεπόμενη ταξινόμηση θα είναι:

| Όνομα ζώου | Τύπος | Majority |
|---|---|---|
| **μυρμηγκοφάγος** | θηλαστικό | θηλαστικό |
| **αντιλόπη** | θηλαστικό | θηλαστικό |
| **πέρκα** | ψάρι | θηλαστικό |
| **αγριόχοιρος** | θηλαστικό | θηλαστικό |
| **βουβάλι** | θηλαστικό | θηλαστικό |
| **γατόψαρο** | ψάρι | θηλαστικό |
| **μαρσυπόμυς** | θηλαστικό | θηλαστικό |
| **κέφαλος** | ψάρι | θηλαστικό |
| **κάβουρας** | ασπόνδυλο | θηλαστικό |
| **σκυλόψαρο** | ψάρι | θηλαστικό |
| **πάπια** | πτηνό | θηλαστικό |
| **φλαμίνγκο** | πτηνό | θηλαστικό |
| **δηλητηριώδης βάτραχος** | αμφίβιο | θηλαστικό |
| **άνθρωπος** | θηλαστικό | θηλαστικό |
| **κατσίκα** | θηλαστικό | θηλαστικό |
| **γορίλας** | θηλαστικό | θηλαστικό |
| **γλάρος** | πτηνό | θηλαστικό |
| **γεράκι** | πτηνό | θηλαστικό |
| **πασχαλίτσα** | έντομο | θηλαστικό |
| **αστακός** | ασπόνδυλο | θηλαστικό |
| **βιζόν** | θηλαστικό | θηλαστικό |
| **ασβός** | θηλαστικό | θηλαστικό |
| **παπαγάλος** | πτηνό | θηλαστικό |
| **φασιανός** | πτηνό | θηλαστικό |
| **πιράνχα** | ψάρι | θηλαστικό |
| **κροταλίας** | ερπετό | θηλαστικό |
| **πλατύπους** | θηλαστικό | θηλαστικό |
| **ρακούν** | θηλαστικό | θηλαστικό |
| **γλώσσα** | ψάρι | θηλαστικό |
| **σπουργίτι** | πτηνό | θηλαστικό |
| **κύκνος** | πτηνό | θηλαστικό |
| **τερμίτης** | έντομο | θηλαστικό |
| **χελώνα** | ερπετό | θηλαστικό |
| **τουατάρα** | ερπετό | θηλαστικό |
| **τόνος** | ψάρι | θηλαστικό |
| **σκίουρος** | θηλαστικό | θηλαστικό |
| **γύπας** | πτηνό | θηλαστικό |
| **σφήκα** | έντομο | θηλαστικό |
| **σκουλήκι** | ασπόνδυλο | θηλαστικό |
| **τρυποφράχτης** | πτηνό | θηλαστικό |

**Πίνακας 3.6. Προβλεπόμενη ταξινόμηση των δεδομένων του συνόλου ελέγχου με τον αφελή κανόνα**

Ο αφελής κανόνας δίνει πολλές λανθασμένες προβλέψεις και επομένως η αξιοπιστία του είναι πολύ χαμηλή:

$$Classification\ accuracy = \frac{13}{40} = 0.325$$

## ***Αφελής ταξινομητής Bayes***

Χρησιμοποιώντας το πρώτο σύνολο, εκπαιδεύουμε τον αφελή ταξινομητή Bayes και στη συνέχεια τον εφαρμόζουμε στο σύνολο ελέγχου για να διαπιστώσουμε αν η μέθοδος αυτή δίνει αξιόπιστα αποτελέσματα.

Η ταξινόμηση που προβλέφθηκε για τις 40 παρατηρήσεις του συνόλου ελέγχου είναι η εξής:

| Όνομα ζώου | Τύπος | Naive Bayes |
|---|---|---|
| **μυρμηγκοφάγος** | θηλαστικό | θηλαστικό |
| **αντιλόπη** | θηλαστικό | θηλαστικό |
| **πέρκα** | ψάρι | ψάρι |
| **αγριόχοιρος** | θηλαστικό | θηλαστικό |
| **βουβάλι** | θηλαστικό | θηλαστικό |
| **γατόψαρο** | ψάρι | ψάρι |
| **μαρσυπόμυς** | θηλαστικό | θηλαστικό |
| **κέφαλος** | ψάρι | ψάρι |
| **κάβουρας** | ασπόνδυλο | έντομο |
| **σκυλόψαρο** | ψάρι | ψάρι |
| **πάπια** | πτηνό | πτηνό |
| **φλαμίνγκο** | πτηνό | πτηνό |
| **δηλητηριώδης βάτραχος** | αμφίβιο | έντομο |
| **άνθρωπος** | θηλαστικό | θηλαστικό |
| **κατσίκα** | θηλαστικό | θηλαστικό |
| **γορίλας** | θηλαστικό | θηλαστικό |
| **γλάρος** | πτηνό | πτηνό |
| **γεράκι** | πτηνό | πτηνό |
| **πασχαλίτσα** | έντομο | έντομο |
| **αστακός** | ασπόνδυλο | ασπόνδυλο |
| **βιζόν** | θηλαστικό | θηλαστικό |
| **ασβός** | θηλαστικό | θηλαστικό |
| **παπαγάλος** | πτηνό | πτηνό |
| **φασιανός** | πτηνό | πτηνό |
| **πιράνχα** | ψάρι | ψάρι |
| **κροταλίας** | ερπετό | ερπετό |
| **πλατύπους** | θηλαστικό | θηλαστικό |
| **ρακούν** | θηλαστικό | θηλαστικό |
| **γλώσσα** | ψάρι | ψάρι |
| **σπουργίτι** | πτηνό | πτηνό |
| **κύκνος** | πτηνό | πτηνό |
| **τερμίτης** | έντομο | έντομο |
| **χελώνα** | ερπετό | θηλαστικό |
| **τουατάρα** | ερπετό | ερπετό |
| **τόνος** | ψάρι | ψάρι |
| **σκίουρος** | θηλαστικό | θηλαστικό |
| **γύπας** | πτηνό | πτηνό |
| **σφήκα** | έντομο | έντομο |
| **σκουλήκι** | ασπόνδυλο | ασπόνδυλο |
| **τρυποφράχτης** | πτηνό | πτηνό |

**Πίνακας 3.7. Προβλεπόμενη ταξινόμηση των δεδομένων του συνόλου ελέγχου με τον αφελή ταξινομητή Bayes**

Παρατηρούμε ότι ο αφελής ταξινομητής Bayes έδωσε μόνο τρεις λανθασμένες ταξινομήσεις, επομένως είναι αξιόπιστος.

$$Classification\ accuracy = \frac{37}{40} = 0.925$$

## ***K-κοντινότερος γείτονας***

Τώρα εφαρμόζουμε τη μέθοδο k-NN στο σύνολο εκπαίδευσης και δοκιμάζοντας διάφορες τιμές για το κ παρατηρούμε ότι το ελάχιστο ποσοστό σφάλματος στο σύνολο ελέγχου προκύπτει για κ=4. Η ταξινόμηση που προκύπτει για τα δεδομένα του συνόλου ελέγχου είναι:

| Όνομα ζώου | Τύπος | k-NN |
|---|---|---|
| **μυρμηγκοφάγος** | θηλαστικό | θηλαστικό |
| **αντιλόπη** | θηλαστικό | θηλαστικό |
| **πέρκα** | ψάρι | ψάρι |
| **αγριόχοιρος** | θηλαστικό | θηλαστικό |
| **βουβάλι** | θηλαστικό | θηλαστικό |
| **γατόψαρο** | ψάρι | ψάρι |
| **μαρσυπόμυς** | θηλαστικό | θηλαστικό |
| **κέφαλος** | ψάρι | ψάρι |
| **κάβουρας** | ασπόνδυλο | ασπόνδυλο |
| **σκυλόψαρο** | ψάρι | ψάρι |
| **πάπια** | πτηνό | πτηνό |
| **φλαμίνγκο** | πτηνό | πτηνό |
| **δηλητηριώδης βάτραχος** | αμφίβιο | αμφίβιο |
| **άνθρωπος** | θηλαστικό | θηλαστικό |
| **κατσίκα** | θηλαστικό | θηλαστικό |
| **γορίλας** | θηλαστικό | θηλαστικό |
| **γλάρος** | πτηνό | πτηνό |
| **γεράκι** | πτηνό | πτηνό |
| **πασχαλίτσα** | έντομο | έντομο |
| **αστακός** | ασπόνδυλο | ασπόνδυλο |
| **βιζόν** | θηλαστικό | θηλαστικό |
| **ασβός** | θηλαστικό | θηλαστικό |
| **παπαγάλος** | πτηνό | πτηνό |
| **φασιανός** | πτηνό | πτηνό |
| **πιράνχα** | ψάρι | ψάρι |
| **κροταλίας** | ερπετό | ερπετό |
| **πλατύπους** | θηλαστικό | θηλαστικό |
| **ρακούν** | θηλαστικό | θηλαστικό |
| **γλώσσα** | ψάρι | ψάρι |
| **σπουργίτι** | πτηνό | πτηνό |
| **κύκνος** | πτηνό | πτηνό |
| **τερμίτης** | έντομο | έντομο |
| **χελώνα** | ερπετό | πτηνό |
| **τουατάρα** | ερπετό | αμφίβιο |
| **τόνος** | ψάρι | ψάρι |
| **σκίουρος** | θηλαστικό | θηλαστικό |
| **γύπας** | πτηνό | πτηνό |
| **σφήκα** | έντομο | έντομο |
| **σκουλήκι** | ασπόνδυλο | ασπόνδυλο |
| **τρυποφράχτης** | πτηνό | πτηνό |

**Πίνακας 3.8. Προβλεπόμενη ταξινόμηση των δεδομένων του συνόλου ελέγχου με τον ταξινομητή k-NN**

Η αξιοπιστία του ταξινομητή k-NN είναι:

$$Classification\ accuracy = \frac{38}{40} = 0.95$$

Συγκρίνοντας, λοιπόν τους τρεις παραπάνω ταξινομητές βλέπουμε ότι ο πιο αποδοτικός είναι ο k-NN.

### 3.5.3 Εφαρμογή 2: Είδη φυτού της ίριδας

Το σύνολο των δεδομένων περιλαμβάνει 3 κατηγορίες με 50 παρατηρήσεις η κάθε μία, όπου κάθε κατηγορία αναφέρεται σε ένα είδος φυτού της ίριδας. Η μία κατηγορία είναι γραμμικά διαχωρίσιμη από τις άλλες, ενώ οι δύο επόμενες κατηγορίες δεν είναι γραμμικά διαχωρίσιμες μεταξύ τους. Τα δεδομένα χαρακτηρίζονται από 4 παράγοντες (μήκος σέπαλου, πλάτος σέπαλου, μήκος πέταλου και πλάτος πέταλου σε εκατοστά). Ο πέμπτος παράγοντας είναι το είδος του φυτού και αποτελεί την ετικέτα κατηγορίας: “iris-setosa”, “iris-versicolor” ή “iris-virginica”.

Αρχικά χωρίζουμε τα δεδομένα σε ένα σύνολο εκπαίδευσης (90 παρατηρήσεις) και ένα σύνολο ελέγχου (60 παρατηρήσεις). Εφαρμόζουμε τη μέθοδο του κ-κοντινότερου γείτονα και με βάση το χαμηλότερο ποσοστό σφάλματος επιλέγουμε τη βέλτιστη τιμή του κ. Τα αποτελέσματα που προκύπτουν είναι:

**Validation error log for different k**

| Value of k | % Error Training | % Error Validation | |
|---|---|---|---|
| 1 | 0 | 3,333333 | |
| 2 | 3,333333 | 10 | |
| 3 | 5,555556 | 1,666667 | <- Best k |
| 4 | 5,555556 | 6,666667 | |
| 5 | 2,222222 | 1,666667 | |
| 6 | 4,444444 | 5 | |
| 7 | 3,333333 | 3,333333 | |
| 8 | 4,444444 | 5 | |
| 9 | 2,222222 | 1,666667 | |
| 10 | 2,222222 | 1,666667 | |

**Training Data Scoring - Summary Report (for k = 3)**

| Confusion Matrix | | | |
|---|---|---|---|
| | Predicted Class | | |
| Actual Class | Iris-setosa | Iris-versicolor | Iris-virginica |
| Iris-setosa | 29 | 0 | 0 |
| Iris-versicolor | 0 | 29 | 2 |
| Iris-virginica | 0 | 3 | 27 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| Iris-setosa | 29 | 0 | 0 |
| Iris-versicolor | 31 | 2 | 6,451613 |
| Iris-virginica | 30 | 3 | 10 |
| Overall | 90 | 5 | 5,555556 |

**Validation Data Scoring - Summary Report (for k = 3)**

| Confusion Matrix | | | |
|---|---|---|---|
| | Predicted Class | | |
| Actual Class | Iris-setosa | Iris-versicolor | Iris-virginica |
| Iris-setosa | 21 | 0 | 0 |
| Iris-versicolor | 0 | 19 | 0 |
| Iris-virginica | 0 | 1 | 19 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| Iris-setosa | 21 | 0 | 0 |
| Iris-versicolor | 19 | 0 | 0 |
| Iris-virginica | 20 | 1 | 5 |
| Overall | 60 | 1 | 1,666667 |

Το χαμηλότερο ποσοστό σφάλματος στο σύνολο ελέγχου δίνεται για κ=3, επομένως αυτή είναι η βέλτιστη τιμή του κ.

Συγκρίνοντας, τώρα, τα αποτελέσματα που προκύπτουν με τις μεθόδους του αφελή κανόνα, του αφελή ταξινομητή Bayes και του k-NN, βλέπουμε ότι και σε αυτό το σύνολο δεδομένων, ο πιο αποδοτικός ταξινομητής είναι ο k-NN.

| Μέθοδος | Αξιοπιστία |
|---|---|
| Αφελής κανόνας | 0.3000 |
| Αφελής ταξινομητή Bayes | 0.9667 |
| k-NN | 0.9833 |

**Πίνακας 3.9. Σύγκριση της αξιοπιστίας των τριών ταξινομητών για το παράδειγμα της ίριδας**

*Iris setosa*

*Iris versicolor*

*Iris virginica*

# Κεφάλαιο 4. Δέντρα ταξινόμησης και παλινδρόμησης

## 4.1 Περιγραφή

Τα δέντρα ταξινόμησης είναι μία τεχνική που αποδίδει καλά σε ένα μεγάλο εύρος καταστάσεων και δεν απαιτεί μεγάλη προσπάθεια από τον αναλυτή, ενώ συγχρόνως είναι εύκολα κατανοητή από τον χρήστη των αποτελεσμάτων της ανάλυσης. Η μεθοδολογία αυτή αναπτύχθηκε από τους Breiman, Friedman, Olshen και Stone[6] οι οποίοι δημιούργησαν το πρόγραμμα CART για την εφαρμογή αυτής της τεχνικής.

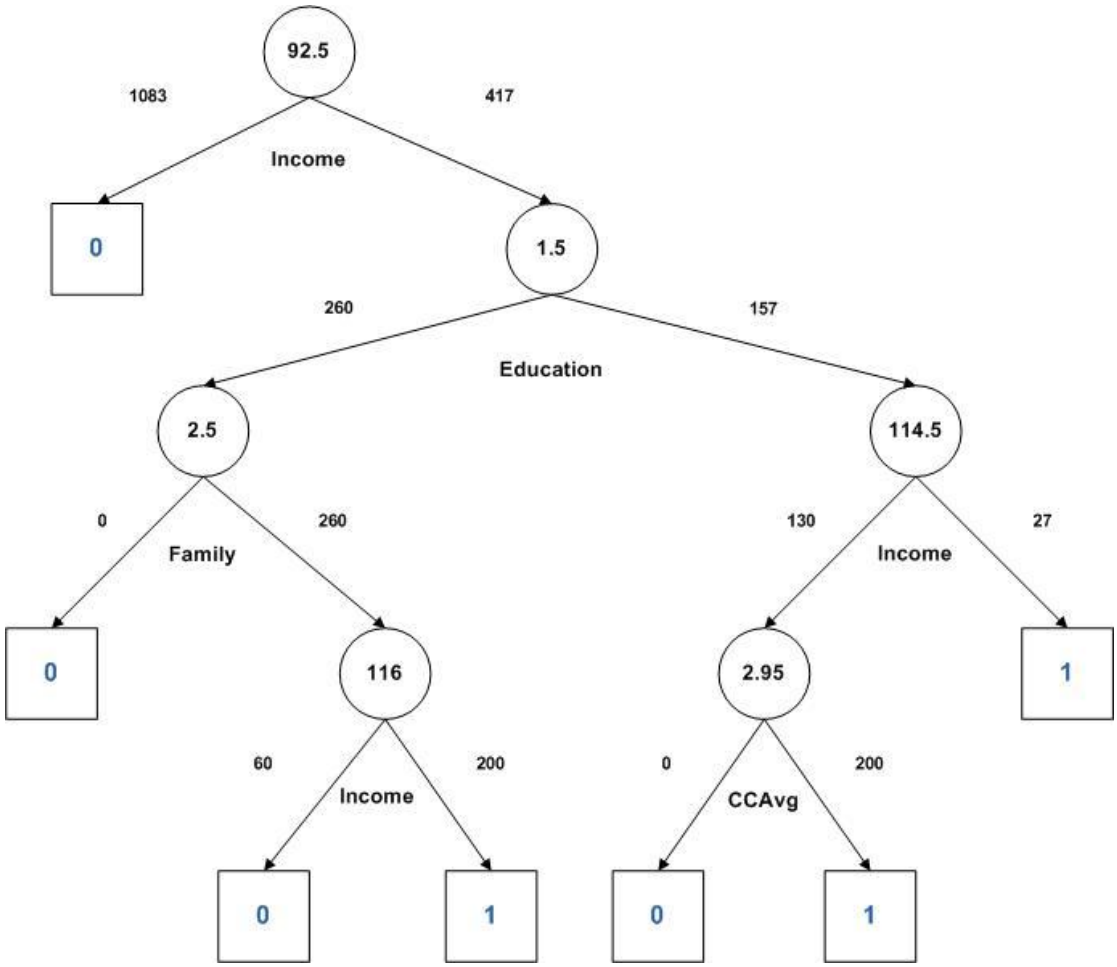
Ο βασικός λόγος που κάνει τα δέντρα ταξινόμησης τόσο δημοφιλή είναι ότι παρέχουν εύκολους και κατανοητούς κανόνες ταξινόμησης. Στο επόμενο γράφημα παρουσιάζεται ένα δέντρο που χωρίζει τους πελάτες μιας τράπεζας ανάμεσα σε δύο κατηγορίες: εκείνους που είναι πιθανότερο να αποδεχθούν μία προσφορά δανείου κι εκείνους που είναι πιθανότερο να

[6] L.Breiman, J.Friedman, R.A.Olshen, C.J.Stone "Classification and Regression Trees" Wadsworth Statistics/Probability (1984)

την απορρίψουν, με βάση κάποια χαρακτηριστικά όπως το εισόδημα, το επίπεδο μόρφωσης και η μέση δαπάνη σε πιστωτικές κάρτες.

Οι ορθογώνιοι τερματικοί κόμβοι περιέχουν την τιμή 0 για τους μη αποδέκτες ή την τιμή 1 για τους αποδέκτες της προσφοράς. Το δέντρο αυτό μας δίνει ένα σύνολο κανόνων για την ταξινόμηση ενός πελάτη. Για παράδειγμα, με βάση τον μεσαίο και αριστερό κόμβο προκύπτει ο παρακάτω κανόνας:

$$IF\ (Income > 92.5)\ AND\ (Education < 1.5)\ AND\ (Family \leq 2.5)\ THEN\ Class = 0$$




Οι κυκλικοί κόμβοι συμβολίζουν κάθε κόμβο που έχει απογόνους. Μέσα σε κάθε κυκλικό κόμβο αναγράφεται η τιμή του διαμερισμού, ενώ κάτω από τον κόμβο είναι το όνομα του χαρακτηριστικού βάσει του οποίου έγινε ο διαμερισμός. Οι αριθμοί πάνω από κάθε αριστερό βέλος δείχνουν το πλήθος των παρατηρήσεων που έχουν τιμές μικρότερες ή ίσες με την τιμή διαμερισμού, και αντίστοιχα σε κάθε δεξιό βέλος είναι το πλήθος των παρατηρήσεων με τιμές μεγαλύτερες από την τιμή διαμερισμού.

Όταν εισάγουμε μία νέα παρατήρηση στο δέντρο ταξινόμησης, ο αλγόριθμος αποφασίζει, σε κάθε κυκλικό κόμβο, ποιο κλαδί θα ακολουθήσει η παρατήρηση μέχρι να καταλήξει σε έναν από τους τερματικούς κόμβους, δηλαδή σε έναν από τους κόμβους που αντιστοιχούν σε κάποια κατηγορία. Οι κυκλικοί κόμβοι λέγονται κόμβοι απόφασης και οι τερματικοί κόμβοι λέγονται φύλλα του δέντρου.

Κάθε φύλλο του δέντρου απεικονίζεται με έναν ορθογώνιο κόμβο και συμβολίζει ένα από τα ορθογώνια που έχουν προκύψει από τον τελικό διαμερισμό του x-χώρου των δεδομένων. Όταν, λοιπόν, η παρατήρηση καταλήξει σε κάποιο από τα φύλλα του δέντρου, θα έχουμε προβλέψει την κατηγορία στην οποία ανήκει. Το όνομα της κατηγορίας αναγράφεται μέσα σε κάθε τερματικό κόμβο.

## 4.2 Δέντρα ταξινόμησης

Η κατασκευή των δέντρων ταξινόμησης βασίζεται στα εξής βήματα:

1. Ξεκινάμε με έναν κόμβο που περιέχει όλα τα δεδομένα.
2. Διαμερίζουμε τον κόμβο, δηλαδή χωρίζουμε τα δεδομένα με βάση έναν κανόνα διαμερισμού για κάποιο από τα χαρακτηριστικά.
3. Επικαλούμαστε αναδρομικά το βήμα 2 σε κάθε κόμβο.
4. Αφού κατασκευάσουμε το δέντρο, κάνουμε κάποιες βελτιστοποιήσεις.

Με άλλα λόγια, σε πρώτη φάση εφαρμόζουμε τον αναδρομικό διαμερισμό του χώρου των ανεξάρτητων μεταβλητών και σε δεύτερη φάση βελτιώνουμε το δέντρο μέσω του κλαδέματος, με τη βοήθεια του συνόλου ελέγχου. Στις επόμενες παραγράφους θα περιγράψουμε αυτές τις δραστηριότητες.

## 4.3 Αναδρομικός Διαμερισμός

Έστω $y$ η εξαρτημένη μεταβλητή (ή μεταβλητή απόκρισης) και $x_1, x_2, \dots, x_p$ οι ανεξάρτητες μεταβλητές (προγνωστικοί παράγοντες ή χαρακτηριστικά). Στα προβλήματα ταξινόμησης η μεταβλητή απόκρισης είναι κατηγορική και το κάθε χαρακτηριστικό X είναι συνεχής, δυαδική ή κατηγορική μεταβλητή. Ο αναδρομικός διαμερισμός χωρίζει τον χώρο διάστασης p των χαρακτηριστικών x σε μη επικαλυπτόμενα πολυδιάστατα ορθογώνια.

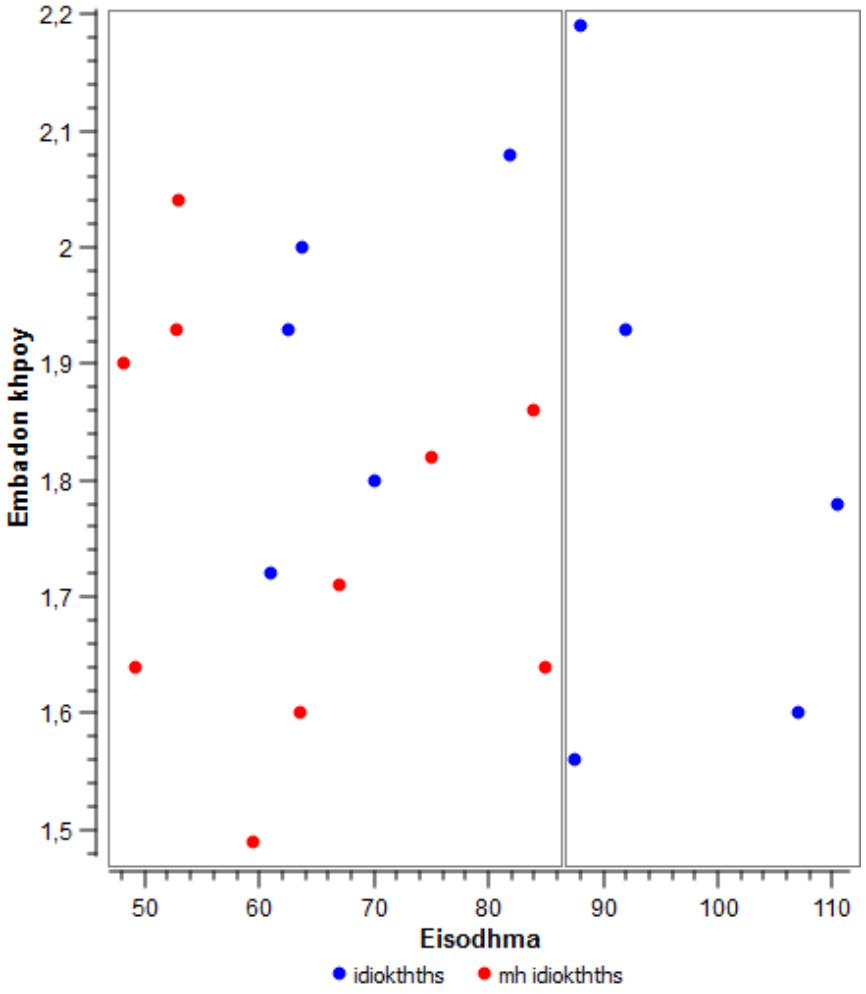
Πρώτα επιλέγουμε μία από τις ανεξάρτητες μεταβλητές, έστω $x_i$, και μία τιμή της $x_i$, έστω η $s_i$, ώστε να διαμερίσουμε τον χώρο των p διαστάσεων σε δύο μέρη: το ένα μέρος θα περιλαμβάνει όλες τις παρατηρήσεις με $x_i \leq s_i$ και το άλλο τις παρατηρήσεις με $x_i > s_i$. Στη συνέχεια διαμερίζουμε το ένα από τα δύο μέρη επιλέγοντας ξανά μία από τις ανεξάρτητες μεταβλητές (την $x_i$ ή κάποια άλλη) και μία τιμή διαμερισμού για τη μεταβλητή αυτή. Με αυτή τη διαδικασία οδηγούμαστε στη δημιουργία τριών πολυδιάστατων ορθογώνιων

περιοχών και αν συνεχίσουμε με τον ίδιο τρόπο καταλήγουμε στη δημιουργία όλο και μικρότερων ορθογώνιων περιοχών.

Η γενική ιδέα είναι ο διαμερισμός ολόκληρου του x-χώρου σε ορθογώνια, ώστε το κάθε ορθογώνιο να είναι όσο το δυνατόν πιο ομοιογενές, δηλαδή να περιέχει παρατηρήσεις που να ανήκουν στην ίδια κατηγορία. Όπως είναι φυσικό, αυτό δεν είναι πάντα εφικτό καθώς σε κάποιες περιπτώσεις μπορεί να υπάρχουν παρατηρήσεις που να ανήκουν σε διαφορετικές κατηγορίες, ακόμη κι αν τα χαρακτηριστικά τους παίρνουν τις ίδιες τιμές.

### 4.3.1 Παράδειγμα 1

Θα χρησιμοποιήσουμε ξανά το παράδειγμα με το σύστημα αυτόματου ποτίσματος που αναφέρεται στο κεφάλαιο 3. Εφαρμόζοντας τη μέθοδο του δέντρου ταξινόμησης, βλέπουμε ότι ο πρώτος διαμερισμός των δεδομένων γίνεται με βάση το χαρακτηριστικό "Εισόδημα" και τιμή διαμερισμού 86.25. Ο χώρος $(x_1, x_2)$ χωρίζεται σε δύο ορθογώνια, ένα με $Εισόδημα < 86.25$ και ένα με $Εισόδημα \geq 86.25$, όπως φαίνεται στο ακόλουθο γράφημα.




**Σχήμα 4.1. Διαμερισμός των 20 παρατηρήσεων με τιμή διαμερισμού 86.25 για το χαρακτηριστικό "Εισόδημα"**

Παρατηρούμε ότι τα δύο ορθογώνια που δημιουργούνται από τον πρώτο διαμερισμό είναι πιο ομοιογενή σε σχέση με το αρχικό ορθογώνιο. Το δεξιό ορθογώνιο περιέχει μόνο παρατηρήσεις που ανήκουν στην κατηγορία "ιδιοκτήτης" (5 ιδιοκτήτες), ενώ το αριστερό ορθογώνιο περιέχει κυρίως "μη ιδιοκτήτες" (5 ιδιοκτήτες και 10 μη ιδιοκτήτες).

Πώς, όμως, επιλέγεται ο συγκεκριμένος διαμερισμός; Ο αλγόριθμος εξετάζει κάθε ένα από τα χαρακτηριστικά (στην περίπτωσή μας "Εισόδημα" και "Εμβαδόν αυλής") και όλες τις πιθανές τιμές διαμερισμού του κάθε χαρακτηριστικού ώστε να εντοπίσει τη βέλτιστη τιμή. Οι πιθανές τιμές διαμερισμού του κάθε χαρακτηριστικού είναι απλά οι μέσες τιμές μεταξύ των ζευγών των διαδοχικών τιμών του χαρακτηριστικού. Για παράδειγμα, οι πιθανές τιμές διαμερισμού για το "Εισόδημα" και το "Εμβαδόν αυλής" αντίστοιχα θα είναι:

| Εισόδημα (διατεταγμένα) | Ζεύγη τιμών | Μέση τιμή (τιμή διαμερισμού) |
|---|---|---|
| **48.2** | (48.2, 49.2) | 48.70 |
| **49.2** | (49.2, 52.7) | 50.95 |
| **52.7** | (52.7, 53.0) | 52.85 |
| **53.0** | (53.0, 59.5) | 56.25 |
| **59.5** | (59.5, 61.0) | 60.25 |
| **61.0** | (61.0, 62.5) | 61.75 |
| **62.5** | (62.5, 63.6) | 63.05 |
| **63.6** | (63.6, 63.8) | 63.70 |
| **63.8** | (63.8, 67.0) | 65.40 |
| **67.0** | (67.0, 70.0) | 68.50 |
| **70.0** | (70.0, 75.0) | 72.50 |
| **75.0** | (75.0, 81.8) | 78.40 |
| **81.8** | (81.8, 84.0) | 82.90 |
| **84.0** | (84.0, 85.0) | 84.50 |
| **85.0** | (85.0, 87.5) | 86.25 |
| **87.5** | (87.5, 88.0) | 87.75 |
| **88.0** | (88.0, 92.0) | 90.00 |
| **92.0** | (92.0, 107.0) | 99.50 |
| **107.0** | (107.0, 110.4) | 108.70 |
| **110.4** | | |

| Εμβαδόν αυλής (διατεταγμένα) | Ζεύγη τιμών | Μέση τιμή (τιμή διαμερισμού) |
|---|---|---|
| **1.49** | (1.49, 1.56) | 1.525 |
| **1.56** | (1.56, 1.60) | 1.580 |
| **1.60** | (1.60, 1.60) | 1.600 |
| **1.60** | (1.60, 1.64) | 1.620 |
| **1.64** | (1.64, 1.64) | 1.640 |
| **1.64** | (1.64, 1.71) | 1.675 |
| **1.71** | (1.71, 1.72) | 1.715 |
| **1.72** | (1.72, 1.78) | 1.750 |
| **1.78** | (1.78, 1.80) | 1.790 |
| **1.80** | (1.80, 1.82) | 1.810 |
| **1.82** | (1.82, 1.86) | 1.840 |
| **1.86** | (1.86, 1.90) | 1.880 |
| **1.90** | (1.90, 1.93) | 1.915 |
| **1.93** | (1.93, 1.93) | 1.930 |
| **1.93** | (1.93, 1.93) | 1.930 |
| **1.93** | (1.93, 2.00) | 1.965 |
| **2.00** | (2.00, 2.04) | 2.020 |
| **2.04** | (2.04, 2.08) | 2.060 |
| **2.08** | (2.08, 2.19) | 2.135 |
| **2.19** | | |

Οι διάφορες τιμές διαμερισμού ιεραρχούνται με βάση το πόσο μειώνουν τη «μη καθαρότητα» (impurity) ή την ετερογένεια στο ορθογώνιο που προκύπτει. Ένα «καθαρό» ορθογώνιο είναι εκείνο που περιλαμβάνει μόνο μία κατηγορία (π.χ. ιδιοκτήτες). Η μείωση της «μη καθαρότητας» ορίζεται ως η διαφορά της «μη καθαρότητας» πριν το διαμερισμό μείον το άθροισμα της «μη καθαρότητας» των δύο ορθογωνίων που προκύπτουν από το διαμερισμό.

## 4.4 Κριτήρια διαμερισμού

Για να μπορέσουμε να επιλέξουμε το χαρακτηριστικό με βάση το οποίο θα χωρίσουμε τις παρατηρήσεις του συνόλου, χρειαζόμαστε κάποια κριτήρια διαμερισμού. Σκοπός είναι να δημιουργήσουμε όσο το δυνατόν πιο «καθαρές» ορθογώνιες περιοχές, δηλαδή με όσο το δυνατόν μεγαλύτερη ομοιογένεια μεταξύ των παρατηρήσεων.

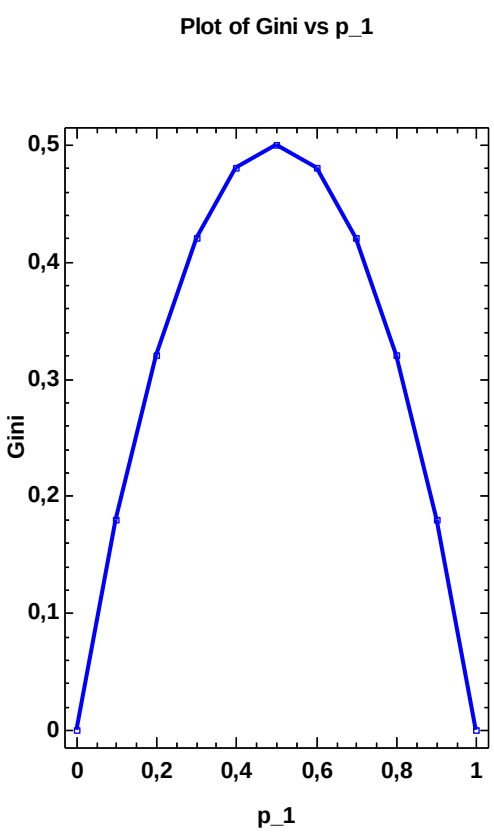
### 4.4.1 Μέτρα «μη καθαρότητας»

Υπάρχουν διάφοροι τρόποι για να υπολογίσουμε τη «μη καθαρότητα». Τα δύο γνωστότερα μέτρα είναι το ευρετήριο Gini (Gini Index) και η εντροπία (entropy). Έστω k η ετικέτα κατηγορίας (ή μεταβλητή απόκρισης) με k=1,2, …, m.

Το ευρετήριο Gini για ένα ορθογώνιο A ορίζεται ως:

$$\boldsymbol{I(A) = 1 - \sum_{k=1}^{m} p_k^2}$$

όπου $p_k$ το ποσοστό των παρατηρήσεων του ορθογωνίου A που ανήκουν στην κατηγορία k. Το μέτρο αυτό παίρνει την τιμή 0, αν όλες οι παρατηρήσεις ανήκουν στην ίδια κατηγορία και την τιμή $\frac{(m-1)}{m}$ όταν όλες οι παρατηρήσεις είναι ομοιόμορφα κατανεμημένες στις m κατηγορίες. Στο επόμενο γράφημα φαίνονται οι διάφορες τιμές του μέτρου Gini συναρτήσει του $p_k$ , για την περίπτωση δύο πιθανών κατηγοριών. Όπως βλέπουμε, η μέγιστη τιμή είναι για $p_k = 0.5$, δηλαδή όταν το ορθογώνιο περιέχει το 50% των παρατηρήσεων της κάθε κατηγορίας. Ο άξονας των x (p_1) συμβολίζει το ποσοστό των παρατηρήσεων που ανήκουν στην κατηγορία 1.




**Σχήμα 4.2. Διάγραμμα "Gini" vs "p_1" για την περίπτωση που έχουμε δύο κατηγορίες**

Ένα δεύτερο μέτρο της «μη καθαρότητας» είναι η εντροπία. Η εντροπία ενός ορθογωνίου Α δίνεται από τη σχέση:

$$\boldsymbol{Entropy(A) = -\sum_{k=1}^{m} p_k \, log_2(p_k)}$$

Η εντροπία παίρνει τιμές μεταξύ του μηδενός (αν όλες οι παρατηρήσεις ανήκουν στην ίδια κατηγορία) και του $log_2(m)$ (όταν οι m κατηγορίες είναι ομοιόμορφα κατανεμημένες). Όπως και για το ευρετήριο Gini, η τιμή της εντροπίας μεγιστοποιείται για $p_k = 0.5$.

Τώρα θα υπολογίσουμε την «μη καθαρότητα» στο παράδειγμα με το σύστημα αυτόματου ποτίσματος, πριν και μετά τον πρώτο διαμερισμό, χρησιμοποιώντας για το χαρακτηριστικό "Εισόδημα" την τιμή διαμερισμού 86.25. Πριν το διαμερισμό το σύνολο δεδομένων περιέχει 10 "ιδιοκτήτες" και 10 "μη ιδιοκτήτες". Έχουμε, λοιπόν, τον ίδιο αριθμό παρατηρήσεων για κάθε μία από τις δύο κατηγορίες, δηλαδή ισχύει $p_1 = p_2 = 0.5$. Επομένως, και τα δύο μέτρα «μη καθαρότητας» παίρνουν τη μέγιστη τιμή τους:

$$Gini = 0.5 \; και \; entropy = log_2(2) = 1$$

Μετά τον πρώτο διαμερισμό, το δεξιό ορθογώνιο περιέχει 5 "ιδιοκτήτες" και 0 "μη ιδιοκτήτες". Η «μη καθαρότητα» για αυτό το ορθογώνιο θα είναι:

$$Gini = 1 - (1^2 + 0^2) = 0$$

και

$$Entropy = -(1log_2(1) + 0) = 0$$

Αντίστοιχα, το αριστερό ορθογώνιο περιέχει 5 "ιδιοκτήτες" και 10 "μη ιδιοκτήτες" , άρα για τη «μη καθαρότητα» θα ισχύει:

$$Gini = 1 - (0.33^2 + 0.67^2) = 0.4422$$

και

$$Entropy = -(0.33log_2(0.33) + 0.67log_2(0.67) = 0.9149$$

Η συνολική «μη καθαρότητα» των δύο ορθογωνίων που προέκυψαν από τον διαμερισμό θα είναι ο σταθμισμένος μέσος όρος των δύο παραπάνω μέτρων, με συντελεστή βαρύτητας το ποσοστό των παρατηρήσεων σε κάθε ορθογώνιο:

$$Gini_{total} = \frac{5}{20}0 + \frac{15}{20}0.4422 = 0.3317$$

και

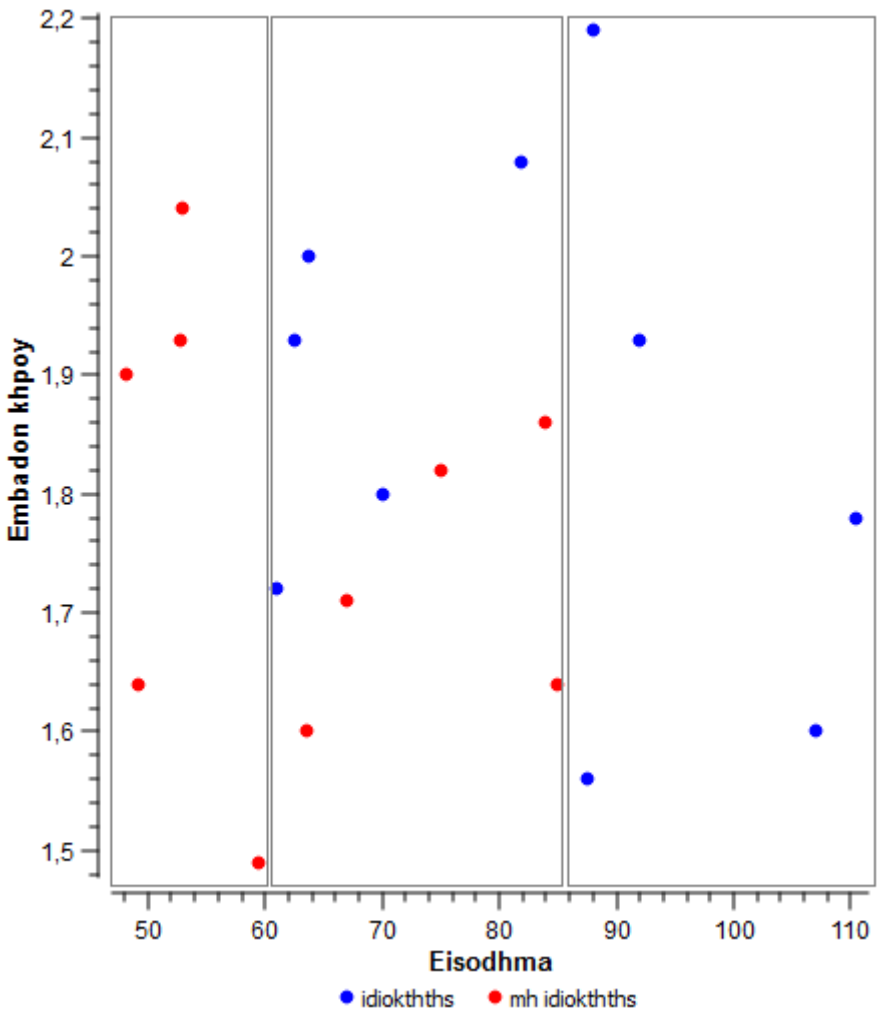
$$Entropy_{total} = \frac{5}{20}0 + \frac{15}{20}0.9149 = 0.6862$$

Παρατηρούμε, λοιπόν, ότι μετά τον διαμερισμό η τιμή Gini μειώθηκε από 0.5 σε 0.3317, ομοίως και η τιμή της εντροπίας από 1 σε 0.6862.

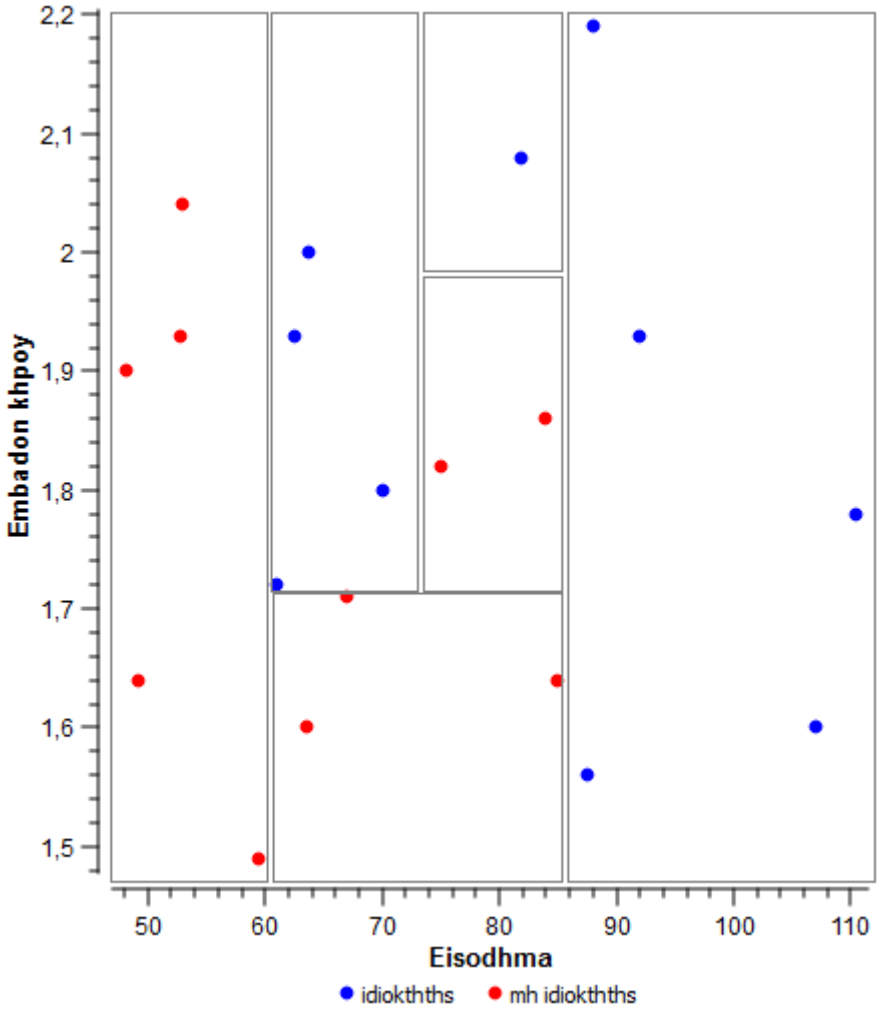
Συνεχίζουμε υπολογίζοντας την μείωση της «μη καθαρότητας» για όλες τις πιθανές τιμές διαμερισμού του κάθε χαρακτηριστικού, με σκοπό να εντοπίσουμε τον διαμερισμό που θα μεγιστοποιεί αυτή τη μείωση. Ο επόμενος διαμερισμός που προκύπτει είναι και πάλι για το χαρακτηριστικό "Εισόδημα" με τιμή 60.25. Όπως φαίνεται στο σχήμα 4.3, ο δεύτερος διαμερισμός των παρατηρήσεων μεγιστοποιεί την «καθαρότητα» στα νέα ορθογώνια που δημιουργούνται. Το δεξιό ορθογώνιο περιέχει τα δεδομένα με $Εισόδημα \geq 86.25$ και όλα τα σημεία είναι "ιδιοκτήτες". Αντίστοιχα, το μεσαίο ορθογώνιο περιλαμβάνει δεδομένα με $60.25 \leq Εισόδημα < 86.25$ και τα σημεία είναι μοιρασμένα στις δύο κατηγορίες, δηλαδή έχουμε 5 "ιδιοκτήτες" και 5 "μη ιδιοκτήτες". Τέλος, το αριστερό ορθογώνιο περιέχει τις

υπόλοιπες 5 παρατηρήσεις με $Εισόδημα < 60.25$ και ανήκουν όλες στην κατηγορία "μη ιδιοκτήτες".

Βλέπουμε, λοιπόν, ότι η διαδικασία του αναδρομικού διαμερισμού δίνει όλο και πιο «καθαρά» ορθογώνια καθώς προχωρά ο αλγόριθμος. Το τελευταίο στάδιο του αναδρομικού διαμερισμού παρουσιάζεται στο σχήμα 4.4.



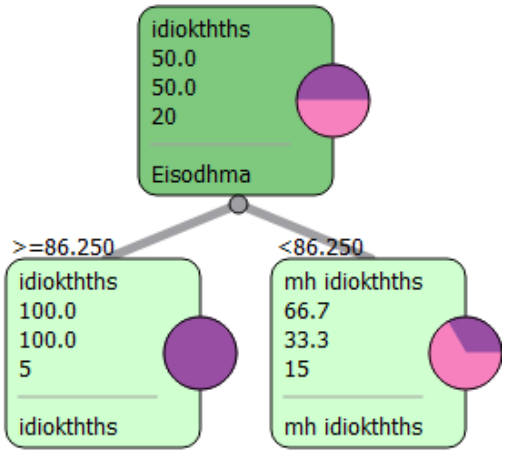

**Σχήμα 4.3. Διαμερισμός των 20 παρατηρήσεων με τιμές διαμερισμού 86.25 και 60.25 για το χαρακτηριστικό "Εισόδημα"**



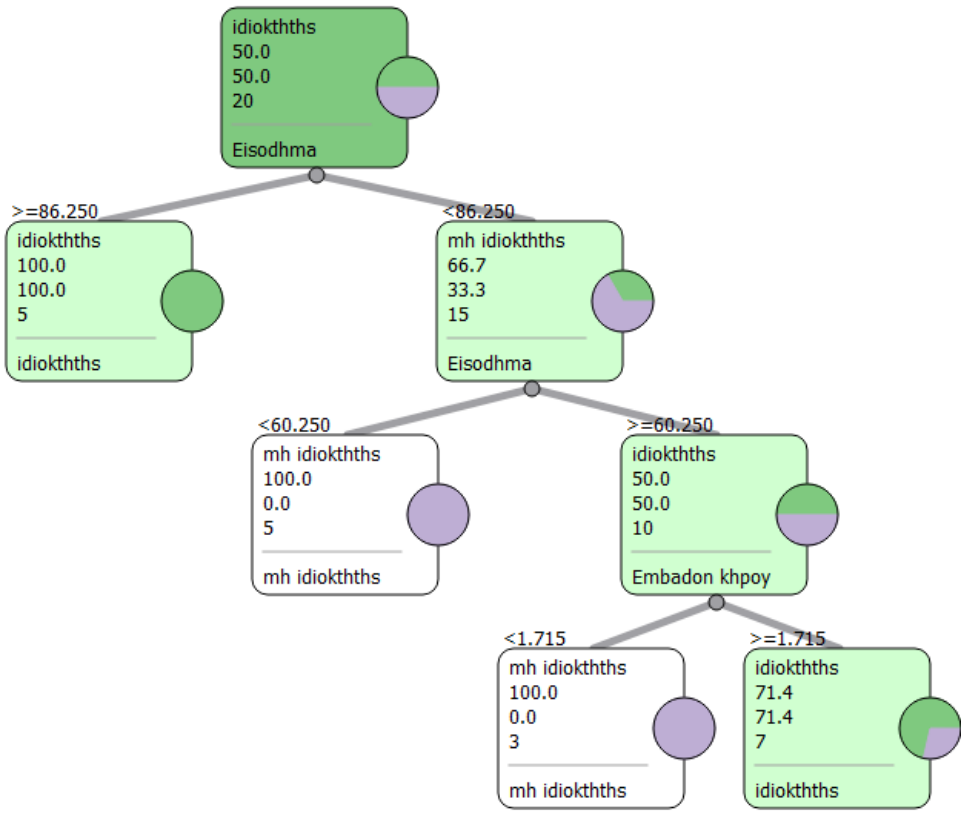

**Σχήμα 4.4. Τελικό στάδιο αναδρομικού διαμερισμού: Κάθε ορθογώνιο περιέχει παρατηρήσεις από μία κατηγορία ("ιδιοκτήτες" ή "μη ιδιοκτήτες").**

Παρατηρούμε ότι πλέον όλα τα ορθογώνια είναι «καθαρά», δηλαδή περιλαμβάνουν μόνο παρατηρήσεις που ανήκουν σε μία από τις δύο κατηγορίες.

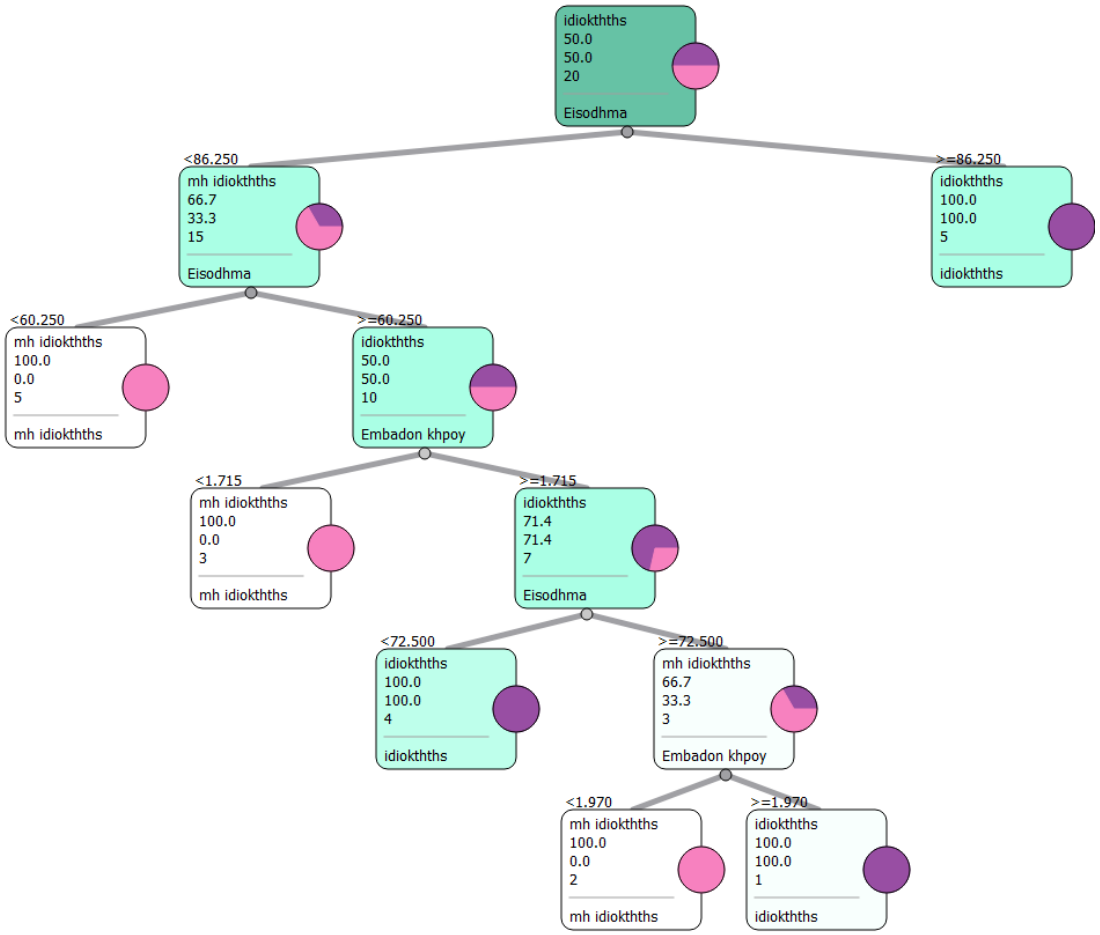
Με τη μέθοδο του δέντρου ταξινόμησης, κάθε διαμερισμός απεικονίζεται ως ο διαμερισμός ενός κόμβου σε δύο νέους κόμβους-απογόνους. Ο πρώτος διαμερισμός φαίνεται στο σχήμα 4.5 ως η διακλάδωση του αρχικού κόμβου ή αλλιώς της ρίζας του δέντρου. Στο σχήμα 4.6 φαίνονται οι διακλαδώσεις που προκύπτουν με τους τρεις πρώτους διαμερισμούς και στο σχήμα 4.7 παρουσιάζεται το τελικό δέντρο.




**Σχήμα 4.5. Διακλάδωση του δέντρου για τον πρώτο διαμερισμό**




**Σχήμα 4.6. Διακλαδώσεις του δέντρου για τους τρεις πρώτους διαμερισμούς**




**Σχήμα 4.7. Τελικό δέντρο ταξινόμησης (κατηγορία-στόχος: "ιδιοκτήτες")**

Αρχικά, για να κατασκευάσουμε το δέντρο επιλέγουμε την κατηγορία-στόχο, βάσει της οποίας θα υπολογιστούν τα στατιστικά χαρακτηριστικά κάθε κόμβου. Η ρίζα του δέντρου ξεκινάει με την κατηγορία στην οποία ανήκουν οι περισσότερες παρατηρήσεις. Στο παράδειγμά μας, οι "ιδιοκτήτες" και οι "μη ιδιοκτήτες" είναι ίσοι σε πλήθος, αλλά αυτό γενικά δεν είναι απαραίτητο.

Σε κάθε κόμβο αναγράφεται το όνομα της κυρίαρχης κατηγορίας, το ποσοστό της επί του συνόλου των παρατηρήσεων που περιέχονται σε αυτόν τον κόμβο, καθώς και το ποσοστό της κατηγορίας-στόχου που έχουμε ορίσει. Επίσης, αναγράφεται το πλήθος των παρατηρήσεων που περιλαμβάνει ο συγκεκριμένος κόμβος, ενώ στο κάτω μέρος φαίνεται το όνομα του χαρακτηριστικού βάσει του οποίου θα γίνει η διακλάδωση. Σε περίπτωση που στο κάτω μέρος του κόμβου αναγράφεται το όνομα κάποιας κατηγορίας, τότε πρόκειται για τερματικό κόμβο και το κλαδί αυτό τερματίζεται. Διαφορετικά, ακολουθεί η επόμενη διακλάδωση του δέντρου.

Επιπλέον, στο κυκλικό γράφημα μπορούμε να δούμε την κατανομή των παρατηρήσεων κάθε κόμβου. Τέλος, σε κάθε κλαδί του δέντρου φαίνεται η τιμή διαμερισμού του χαρακτηριστικού.

Για παράδειγμα, όπως φαίνεται και στο σχήμα 4.5, η ρίζα του δέντρου ξεκινάει με τους "ιδιοκτήτες" που έχουμε ορίσει και ως κατηγορία-στόχο. Περιλαμβάνει συνολικά 20 παρατηρήσεις, αφού ο αρχικός κόμβος ξεκινά πάντα με όλο το σύνολο δεδομένων. Τα στατιστικά χαρακτηριστικά είναι:

$$Ποσοστό\ κυρίαρχης\ κατηγορίας = \frac{10}{20} = 50\%\ (ιδιοκτήτες)$$

$$Ποσοστό\ κατηγορίας - στόχου = \frac{10}{20} = 50\%\ (ιδιοκτήτες)$$

$$Πλήθος\ παρατηρήσεων\ κόμβου = 20\ (ιδιοκτήτες\ και\ μη\ ιδιοκτήτες)$$

$$Όνομα\ χαρακτηριστικού\ επόμενης\ διακλάδωσης = Εισόδημα$$

Στα κλαδιά του σχήματος αναγράφεται η τιμή διαμερισμού για το "Εισόδημα" και είναι 86.25. Ο επόμενος αριστερός κόμβος περιλαμβάνει τις παρατηρήσεις με Εισόδημα $< 86.25$:

$$Ποσοστό\ κυρίαρχης\ κατηγορίας = \frac{10}{15} = 66.7\%\ (μη\ ιδιοκτήτες)$$

$$Ποσοστό\ κατηγορίας - στόχου = \frac{5}{15} = 33.3\%\ (ιδιοκτήτες)$$

$$Πλήθος\ παρατηρήσεων\ κόμβου = 15\ (ιδιοκτήτες\ και\ μη\ ιδιοκτήτες)$$

$$Όνομα\ χαρακτηριστικού\ επόμενης\ διακλάδωσης = Εισόδημα$$

Ο κόμβος αυτός περιλαμβάνει παρατηρήσεις και από τις δύο κατηγορίες, άρα είναι κόμβος απόφασης. Το κλαδί δεν τερματίζεται και θα ακολουθήσει η επόμενη διακλάδωση.

Αντίστοιχα, στον δεξιό κόμβο περιέχονται οι παρατηρήσεις με Εισόδημα $\geq 86.2$:

$$Ποσοστό\ κυρίαρχης\ κατηγορίας = \frac{5}{5} = 100\%\ (ιδιοκτήτες)$$

$$Ποσοστό\ κατηγορίας - στόχου = \frac{5}{5} = 100\%\ (ιδιοκτήτες)$$

$$Πλήθος\ παρατηρήσεων\ κόμβου = 5\ (μόνο\ ιδιοκτήτες)$$

$$Όνομα\ χαρακτηριστικού\ επόμενης\ διακλάδωσης = -$$

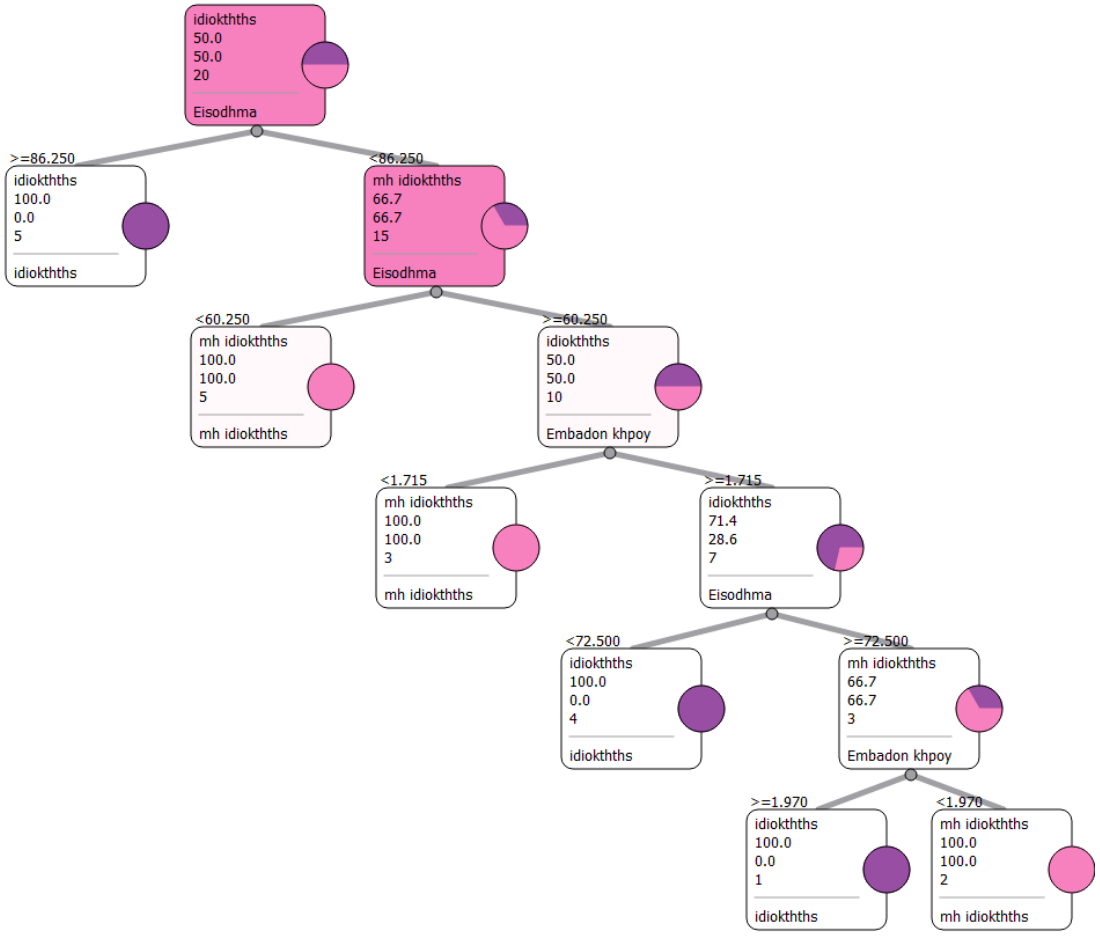
Εφόσον όλες οι παρατηρήσεις ανήκουν σε μία μόνο κατηγορία, ο κόμβος αυτός είναι τερματικός και δεν ακολουθεί διακλάδωση.

**Target class:** idiokththsΗ
**Tree size:** 11 nodes, 6 leaves

| Classification Tree | Class | P(Class) | P(Target) | # Inst |
|---|---|---|---|---|
| | idiokththsΗ | 0.500 | 0.500 | 20 |
| Eisodhma <86.250 | mh idiokththsΗ | 0.667 | 0.333 | 15 |
| Eisodhma <60.250 | mh idiokththsΗ | 1.000 | 0.000 | 5 |
| Eisodhma >=60.250 | idiokththsΗ | 0.500 | 0.500 | 10 |
| Embadon khpoy <1.715 | mh idiokththsΗ | 1.000 | 0.000 | 3 |
| Embadon khpoy >=1.715 | idiokththsΗ | 0.714 | 0.714 | 7 |
| Eisodhma <72.500 | idiokththsΗ | 1.000 | 1.000 | 4 |
| Eisodhma >=72.500 | mh idiokththsΗ | 0.667 | 0.333 | 3 |
| Embadon khpoy <1.970 | mh idiokththsΗ | 1.000 | 0.000 | 2 |
| Embadon khpoy >=1.970 | idiokththsΗ | 1.000 | 1.000 | 1 |
| Eisodhma >=86.250 | idiokththsΗ | 1.000 | 1.000 | 5 |

**Πίνακας 4.1. Στατιστικά χαρακτηριστικά κάθε κόμβου για το παράδειγμα του συστήματος ποτισμού**

Η κατηγορία-στόχος που θα ορίσουμε, δεν παίζει κανένα ρόλο στην κατασκευή του δέντρου, παρά μόνο στον υπολογισμό των στατιστικών χαρακτηριστικών. Δηλαδή, αν ορίσουμε εκ νέου ως κατηγορία-στόχο τους "μη ιδιοκτήτες" το δέντρο που προκύπτει είναι:




**Σχήμα 4.8. Τελικό δέντρο ταξινόμησης (κατηγορία-στόχος: "μη ιδιοκτήτες")**

Το δέντρο ταξινόμησης είναι ακριβώς το ίδιο με το σχήμα 4.7, με μόνη διαφορά τις τιμές των στατιστικών χαρακτηριστικών που αναγράφονται μέσα σε κάθε κόμβο.

Σημειώνουμε ότι στην εργασία αυτή ασχολούμαστε μόνο με δυαδικά δέντρα, δηλαδή δέντρα στα οποία κάθε κόμβος απόφασης χωρίζεται ακριβώς σε δύο κόμβους-απογόνους. Τα δυαδικά δέντρα, σε αντίθεση με τα μη-δυαδικά, είναι ευκολότερο να διαβαστούν και η κατασκευή τους είναι υπολογιστικά λιγότερο δαπανηρή.

### 4.4.2 Άλλα κριτήρια διαμερισμού

Δύο ακόμη κριτήρια που χρησιμοποιούνται για τον διαμερισμό του συνόλου δεδομένων είναι το κέρδος πληροφορίας (Information gain) και ο λόγος κέρδους πληροφορίας (Gain ratio).

Το κέρδος πληροφορίας είναι ένα ακόμη μέτρο «μη καθαρότητας» και βασίζεται στη μείωση της εντροπίας μετά τον διαμερισμό των παρατηρήσεων. Αρχικά υπολογίζεται η εντροπία του κόμβου πριν το διαμερισμό ($Entropy_{before\ split}$). Στη συνέχεια προστίθενται οι εντροπίες των δύο κόμβων-απογόνων που προέκυψαν από το διαμερισμό ($Entropy_{after\ split}$). Το κέρδος πληροφορίας δίνεται από τη σχέση:

$$\boldsymbol{InfoGain = Entropy_{before\ split} - Entropy_{after\ split}}$$

Σε κάθε κόμβο επιλέγεται για διαμερισμό το χαρακτηριστικό που δίνει το μεγαλύτερο κέρδος πληροφορίας. Αν σε κάποιο κόμβο η εντροπία πάρει την τιμή μηδέν, τότε πρόκειται για τερματικό κόμβο, ενώ αν πάρει τιμή μεγαλύτερη του μηδενός αποτελεί κόμβο απόφασης και θα ακολουθήσει ο επόμενος διαμερισμός.

Ωστόσο, η ομοιογένεια των υποσυνόλων είναι πιο πιθανή για μεγάλο αριθμό τιμών στα χαρακτηριστικά. Έτσι, το κριτήριο κέρδους πληροφορίας μεροληπτεί υπέρ της επιλογής τέτοιων χαρακτηριστικών κι αυτό μπορεί να οδηγήσει σε υπερπροσαρμογή. Ο λόγος κέρδους πληροφορίας είναι μία τροποποίηση του κέρδους πληροφορίας που μειώνει αυτή τη μεροληψία. Λαμβάνει υπ' όψιν την εγγενή πληροφορία του διαμερισμού (intrinsic information), δηλαδή την εντροπία της κατανομής των παρατηρήσεων στους κόμβους, κι έτσι διορθώνει το κριτήριο κέρδους πληροφορίας.

$$\boldsymbol{Intrinsic\ Info(S,X) = -\sum_i \frac{|S_i|}{|S|} log \frac{|S_i|}{|S|}}$$

$$\boldsymbol{Gain\ ratio(S,X) = \frac{InfoGain(S,X)}{IntInfo(S,X)}}$$

*όπου X το χαρακτηριστικό που διαμερίζει το σύνολο S σε $S_i$ υποσύνολα.*

## 4.5 Αλγόριθμοι δέντρων ταξινόμησης

### 4.5.1 Αλγόριθμος ID3

Ο ID3 αλγόριθμος ξεκινά με το αρχικό σύνολο, που περιέχει όλες τις παρατηρήσεις, ως ρίζα του δέντρου. Σε κάθε επανάληψη επιλέγει ένα από τα χαρακτηριστικά που δεν έχουν χρησιμοποιηθεί και υπολογίζει το κέρδος πληροφορίας κάθε χαρακτηριστικού. Έπειτα εντοπίζει το μεγαλύτερο κέρδος και γίνεται ο διαμερισμός του συνόλου με βάση το συγκεκριμένο χαρακτηριστικό. Ο αλγόριθμος συνεχίζει με τον ίδιο τρόπο, επιλέγοντας κάθε φορά κάποιο από τα χαρακτηριστικά που δεν έχουν επιλεχθεί ήδη.

Συνοψίζοντας, τα βήματα είναι τα εξής:

1. Υπολόγισε το κέρδος πληροφορίας κάθε χαρακτηριστικού για το σύνολο S.
2. Διαμέρισε το σύνολο S σε υποσύνολα, με βάση το χαρακτηριστικό που έχει το μεγαλύτερο κέρδος πληροφορίας.
3. Κατασκεύασε ένα δέντρο ταξινόμησης που να περιέχει αυτό το χαρακτηριστικό
4. Επανέλαβε τη διαδικασία για κάθε υποσύνολο χρησιμοποιώντας κάποιο από τα υπόλοιπα χαρακτηριστικά.

Ο ID3 δε δίνει απαραίτητα τη βέλτιστη λύση και μπορεί να υπερπροσαρμοστεί στα δεδομένα του συνόλου εκπαίδευσης. Συνήθως, κατασκευάζει δέντρα με λίγα επίπεδα, αλλά δεν κατασκευάζει πάντα το μικρότερο δυνατό δέντρο.

### 4.5.2 Αλγόριθμος C4.5

Ο αλγόριθμος C4.5 είναι μία βελτιωμένη εκδοχή του ID3. Διαχειρίζεται καλύτερα τις ελλιπείς παρατηρήσεις (δηλαδή παρατηρήσεις με ελλιπή πληροφορία αφού απουσιάζουν οι τιμές κάποιων χαρακτηριστικών), καθώς και τα χαρακτηριστικά που παίρνουν συνεχείς τιμές. Σε κάθε κόμβο του δέντρου, ο αλγόριθμος επιλέγει το χαρακτηριστικό με τον μεγαλύτερο λόγο κέρδους πληροφορίας για να κάνει τον διαμερισμό. Μερικές βασικές περιπτώσεις είναι:

*Περίπτωση (i)*:Όταν όλες οι παρατηρήσεις του κόμβου ανήκουν στην ίδια κατηγορία, ο αλγόριθμος δημιουργεί ένα φύλλο του δέντρου.

*Περίπτωση (ii)*:Όταν κανένα χαρακτηριστικό δεν παρέχει κέρδος πληροφορίας, δημιουργεί έναν κόμβο απόφασης ψηλότερα στο δέντρο, χρησιμοποιώντας την αναμενόμενη τιμή της κατηγορίας.

*Περίπτωση (iii)*:Όταν συναντήσει μία παρατήρηση η οποία ανήκει σε κατηγορία που δεν έχει «ξαναδεί», δημιουργεί και πάλι έναν κόμβο απόφασης ψηλότερα στο δέντρο, χρησιμοποιώντας την αναμενόμενη τιμή.

Ο αλγόριθμος ακολουθεί τα παρακάτω βήματα:

1. Κάνε έλεγχο για τις περιπτώσεις (i),(ii) και (iii).
2. Βρες το κέρδος πληροφορίας κάθε χαρακτηριστικού.
3. Εντόπισε το χαρακτηριστικό με τον μεγαλύτερο λόγο κέρδους πληροφορίας.
4. Δημιούργησε έναν κόμβο απόφασης που να διαμερίζεται με βάση το καλύτερο χαρακτηριστικό.
5. Επανέλαβε τη διαδικασία και όρισε ως κόμβους-διαδόχους τα υποσύνολα που προέκυψαν από το διαμερισμό.

### 4.5.3 Αλγόριθμος CART

Ο αλγόριθμος CART δημιουργεί δυαδικά δέντρα ταξινόμησης, δηλαδή κάθε κόμβος έχει ακριβώς δύο κόμβους-διαδόχους, και χρησιμοποιεί ως κριτήριο διαμερισμού την εντροπία ή το ευρετήριο Gini. Τα βήματα είναι τα εξής:

1. Υπολόγισε το Gini σε κάθε κόμβο.
2. Εντόπισε το χαρακτηριστικό με τη μικρότερη τιμή Gini.
3. Κάνε διαμερισμό με βάση αυτό το χαρακτηριστικό.
4. Επανέλαβε τη διαδικασία σε κάθε κόμβο.

## 4.6 Αξιολόγηση της απόδοσης ενός δέντρου ταξινόμησης

Για να μπορέσουμε να αξιολογήσουμε την αξιοπιστία του δέντρου στην ταξινόμηση νέων δεδομένων, χρησιμοποιούμε τις ίδιες μεθόδους και τα ίδια κριτήρια που αναλύονται στην παράγραφο 2.3. Αρχικά χωρίζουμε τα δεδομένα σε ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Με βάση το σύνολο εκπαίδευσης κατασκευάζουμε το δέντρο και με το σύνολο ελέγχου αξιολογούμε την απόδοσή του.

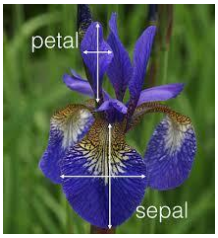
Κάθε παρατήρηση του συνόλου ελέγχου εισάγεται στο δέντρο και ταξινομείται ανάλογα με τον κόμβο στον οποίο καταλήγει. Στη συνέχεια, η κατηγορία που προβλέφθηκε συγκρίνεται με την πραγματική κατηγορία στην οποία ανήκει η παρατήρηση, με τη βοήθεια ενός πίνακα σύγχυσης. Σε περίπτωση που μας ενδιαφέρει μία συγκεκριμένη κατηγορία, χρησιμοποιούμε ένα διάγραμμα ανύψωσης (lift chart) για να αξιολογήσουμε την ικανότητα του μοντέλου να εντοπίζει τις παρατηρήσεις που ανήκουν σε αυτήν.

### 4.6.1 Παράδειγμα 2

Θα χρησιμοποιήσουμε ξανά το παράδειγμα με τα είδη φυτού της ίριδας.

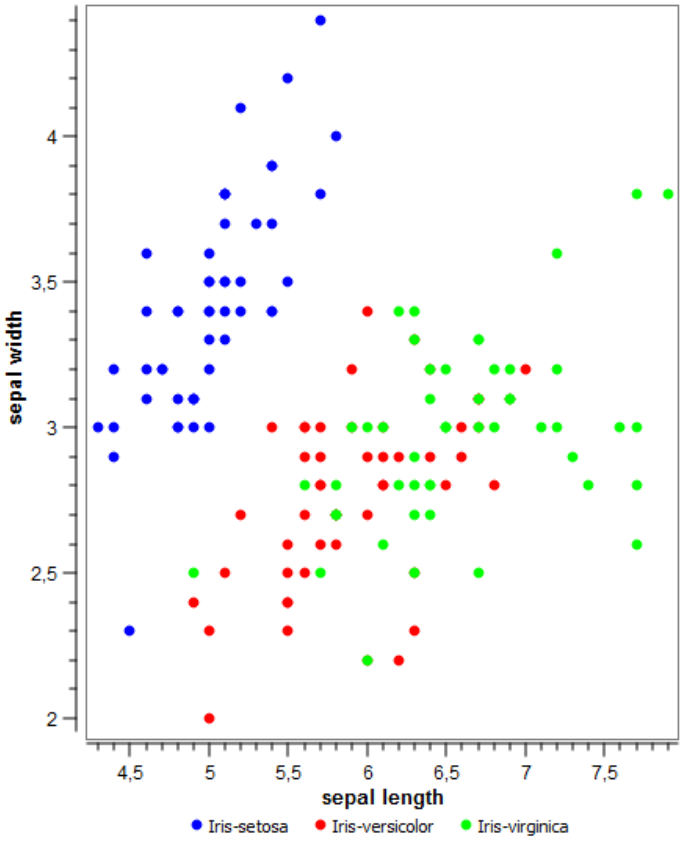
Διαθέτουμε ένα σύνολο δεδομένων με 150 παρατηρήσεις. Η κατηγορία είναι το είδος φυτού της ίριδας και περιγράφεται από τέσσερα χαρακτηριστικά. Στα διαγράμματα διασποράς

φαίνεται ο τρόπος που είναι κατανεμημένες οι παρατηρήσεις με βάση τις τιμές των χαρακτηριστικών “petal width” vs “petal length” και “sepal width” vs “sepal length”.

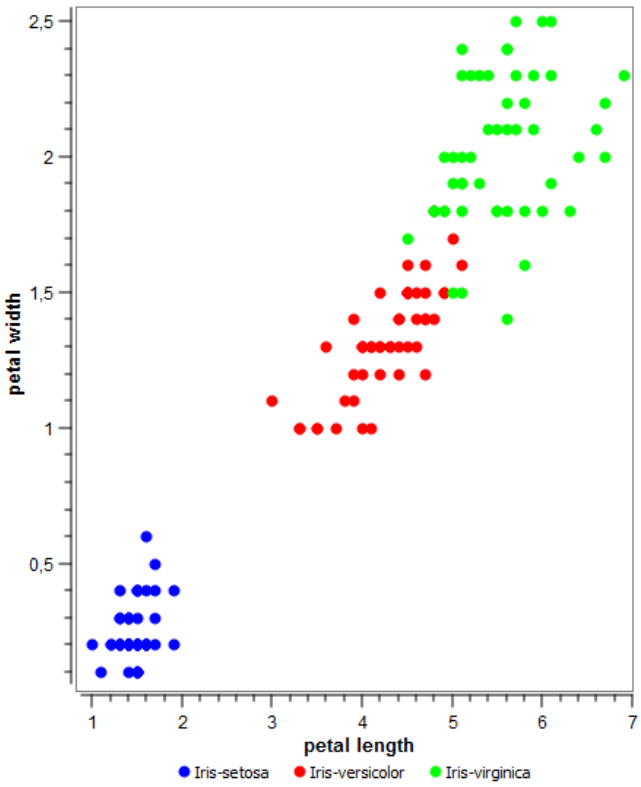



**Εικόνα: Πέταλο και σέπαλο φυτού**

Είναι προφανές ότι ο αλγόριθμος θα επιλέξει τα χαρακτηριστικά “petal width” και “petal length” γιατί η συγκεκριμένη κατανομή των παρατηρήσεων κάνει πιο εύκολο τον διαμερισμό σε ομοιογενείς ορθογώνιες περιοχές.

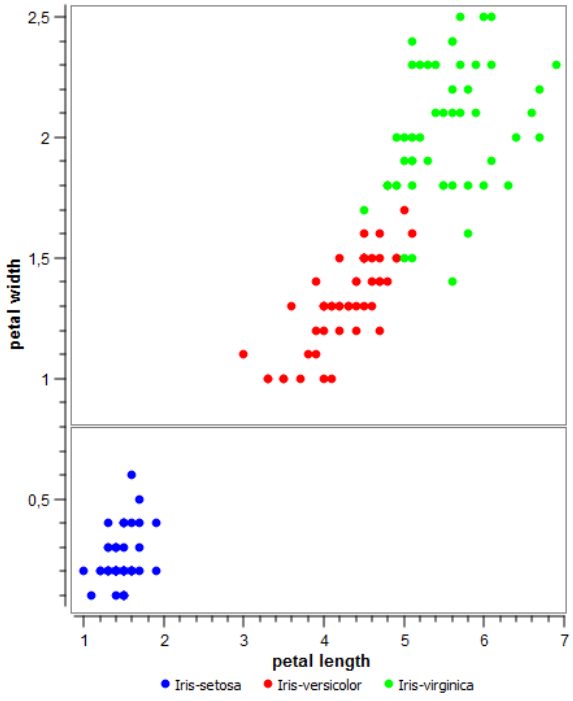



**Σχήμα4.9. Διάγραμμα διασποράς "sepal width" vs "sepal length" για τα τρία είδη φυτού**




**Σχήμα4.10. Διάγραμμα διασποράς "petal width" vs "petal length" για τα τρία είδη φυτού**

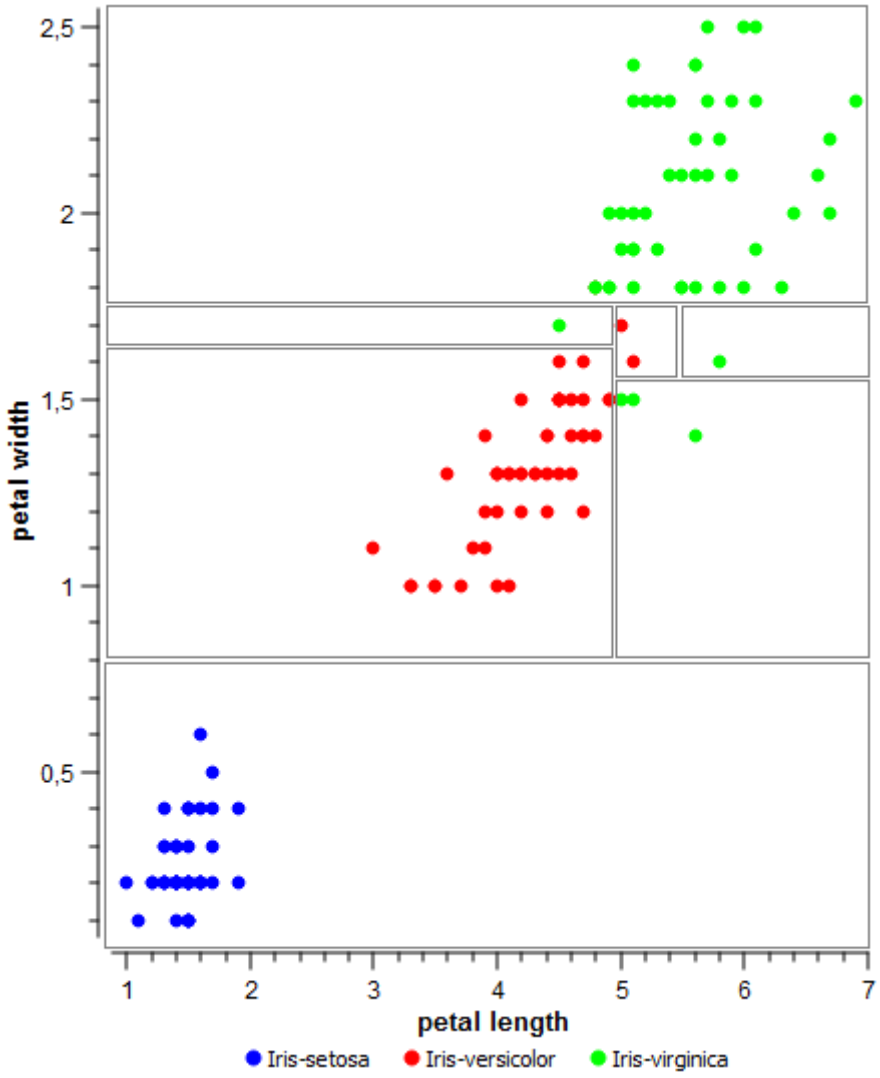
Ο πρώτος διαμερισμός των δεδομένων γίνεται με βάση το χαρακτηριστικό “petal width” με τιμή διαμερισμού 0.8. Ο χώρος $(x_1, x_2)$ χωρίζεται σε δύο ορθογώνια, ένα με $petal\ width \leq 0.8$ και ένα με $petal\ width > 0.8$.




**Σχήμα 4.11. Διαμερισμός των 150 παρατηρήσεων με τιμή διαμερισμού 0.8 για το χαρακτηριστικό "petal width"**

Τα δύο νέα ορθογώνια είναι πιο ομοιογενή σε σχέση με το αρχικό: το πάνω ορθογώνιο περιέχει παρατηρήσεις που ανήκουν στις κατηγορίες “iris-versicolor” και “iris-virginica”, ενώ το κάτω ορθογώνιο περιέχει όλες τις παρατηρήσεις της κατηγορίας “iris-setosa”.

Η διαδικασία του αναδρομικού διαμερισμού επαναλαμβάνεται μέχρι να σχηματιστούν ορθογώνια που να είναι όσο το δυνατόν πιο «καθαρά».




**Σχήμα 4.12. Τελικό στάδιο αναδρομικού διαμερισμού: Κάθε ορθογώνιο περιέχει παρατηρήσεις από μία κατηγορία.**

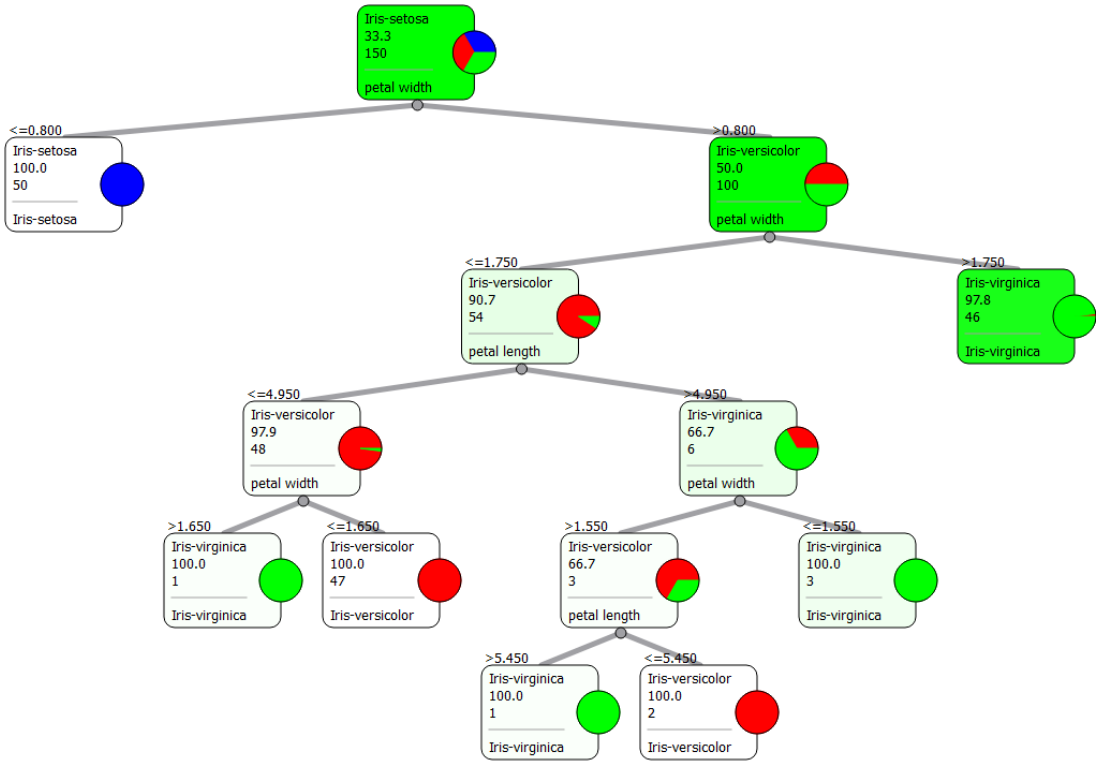
Στον πίνακα που ακολουθεί φαίνονται οι τιμές διαμερισμού, η κατηγορία που κυριαρχεί στο ορθογώνιο που δημιουργείται σε κάθε βήμα, καθώς και το ποσοστό των παρατηρήσεων που ανήκουν στην κυρίαρχη κατηγορία κάθε ορθογωνίου.

Θεωρούμε ως κατηγορία-στόχο την “iris-setosa” και ως κριτήριο διαμερισμού το Gini Index. Το τελικό δέντρο ταξινόμησης που προκύπτει αποτελείται από 13 κόμβους και 7 φύλλα.

**Target class:** Iris-setosa
**Tree size:** 13 nodes, 7 leaves

| Classification Tree | Class | P(Class) | P(Target) |
|---|---|---|---|
| | Iris-setosa | 0.333 | 0.333 |
| petal width <=0.800 | Iris-setosa | 1.000 | 1.000 |
| petal width >0.800 | Iris-versicolor | 0.500 | 0.000 |
| petal width <=1.750 | Iris-versicolor | 0.907 | 0.000 |
| petal length <=4.950 | Iris-versicolor | 0.979 | 0.000 |
| petal width <=1.650 | Iris-versicolor | 1.000 | 0.000 |
| petal width >1.650 | Iris-virginica | 1.000 | 0.000 |
| petal length >4.950 | Iris-virginica | 0.667 | 0.000 |
| petal width <=1.550 | Iris-virginica | 1.000 | 0.000 |
| petal width >1.550 | Iris-versicolor | 0.667 | 0.000 |
| petal length <=5.450 | Iris-versicolor | 1.000 | 0.000 |
| petal length >5.450 | Iris-virginica | 1.000 | 0.000 |
| petal width >1.750 | Iris-virginica | 0.978 | 0.000 |

**Πίνακας 4.2. Στατιστικά χαρακτηριστικά κάθε κόμβου για το παράδειγμα της ίριδας**

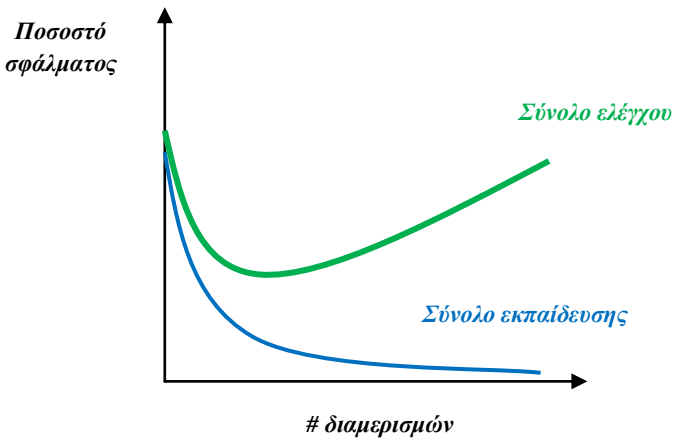



**Σχήμα 4.13. Τελικό δέντρο ταξινόμησης για την ίριδα.**

## 4.7 Αποφυγή της υπερπροσαρμογής

Όταν κατασκευάζουμε ένα δέντρο ταξινόμησης με βάση το σύνολο εκπαίδευσης, οδηγούμαστε σε υπερπροσαρμογή των δεδομένων και αυτό έχει σαν αποτέλεσμα την κακή απόδοση του δέντρου σε νέα δεδομένα. Αν ελέγξουμε το συνολικό σφάλμα στα διάφορα επίπεδα του δέντρου, είναι αναμενόμενο να μειώνεται καθώς προχωράμε στα επίπεδα, μέχρι να φτάσει στο σημείο της υπερπροσαρμογής. Συγκεκριμένα για τα δεδομένα του συνόλου εκπαίδευσης το συνολικό σφάλμα μειώνεται συνεχώς και τελικά μηδενίζεται στο τελευταίο (ή μέγιστο) επίπεδο του δέντρου. Ωστόσο, για νέα δεδομένα, το σφάλμα αναμένεται να μειωθεί μέχρι το σημείο όπου το δέντρο μοντελοποιεί τη σχέση ανάμεσα στην κατηγορία και τα χαρακτηριστικά. Στη συνέχεια, το δέντρο αρχίζει να μοντελοποιεί και τους παράγοντες θορύβου που μπορεί να υπάρχουν στο σύνολο εκπαίδευσης, με αποτέλεσμα να αυξάνεται τελικά το συνολικό σφάλμα στο σύνολο ελέγχου, όπως φαίνεται και στο σχήμα 4.13. Ο λόγος για τον οποίο εμφανίζεται το φαινόμενο της υπερπροσαρμογής στα μεγάλα επίπεδα του δέντρου είναι ότι οι διαμερισμοί γίνονται με βάση έναν πολύ μικρό αριθμό παρατηρήσεων. Έτσι, η λανθασμένη πρόβλεψη της κατηγορίας αποδίδεται, συνήθως, στους παράγοντες θορύβου και όχι στους προγνωστικούς παράγοντες.

Δύο τρόποι για να αποφύγουμε την υπερπροσαρμογή είναι να θέσουμε κανόνες που θα σταματούν την ανάπτυξη του δέντρου ή ,εναλλακτικά, να αφήσουμε το δέντρο να αναπτυχθεί πλήρως και στη συνέχεια να το κλαδέψουμε μέχρι να φτάσουμε στο επίπεδο όπου δεν εμφανίζεται η υπερπροσαρμογή.




**Σχήμα 4.14. Ποσοστό σφάλματος συναρτήσει του πλήθους των διαμερισμών για το σύνολο εκπαίδευσης και το σύνολο ελέγχου: Υπερπροσαρμογή**

### 4.7.1 Διακοπή της ανάπτυξης του δέντρου

Μερικοί κανόνες για να διακόψουμε την ανάπτυξη του δέντρου είναι να θέσουμε ένα όριο στο πλήθος των διαμερισμών, στον ελάχιστο αριθμό παρατηρήσεων για κάθε κόμβο ή στην ελάχιστη μείωση της «μη καθαρότητας». Ωστόσο, δεν είναι εύκολο να καθορίσουμε ποιο θα είναι αυτό το όριο και υπάρχει ο κίνδυνος να σταματήσουμε την ανάπτυξη του δέντρου πολύ νωρίς.

Κάποιες από τις μεθόδους που έχουν αναπτυχθεί βασίζονται στην ιδέα του αναδρομικού διαμερισμού με τη χρήση κανόνων που αποτρέπουν την υπερβολική ανάπτυξη του δέντρου και την υπερπροσαρμογή στα δεδομένα του συνόλου εκπαίδευσης. Μία γνωστή μέθοδος είναι η CHAID (Chi-Squared Automatic Interaction Detection) και αποτελεί μία μέθοδο αναδρομικού διαμερισμού η οποία, χρονολογικά, προηγείται των δέντρων ταξινόμησης και παλινδρόμησης και χρησιμοποιείται μέχρι και σήμερα για την ανάλυση βάσεων δεδομένων στο marketing. Χρησιμοποιεί τον στατιστικό $\chi^2$-έλεγχο ανεξαρτησίας για να εκτιμήσει κατά πόσο ο διαμερισμός ενός κόμβου βελτιώνει την «καθαρότητα» σε στατιστικά σημαντικό βαθμό. Συγκεκριμένα, σε κάθε κόμβο επιλέγεται μία τιμή διαμερισμού για το χαρακτηριστικό που έχει τη μεγαλύτερη συσχέτιση με τη μεταβλητή απόκρισης. Ο βαθμός της συσχέτισης μετριέται από την p- τιμή που δίνει ο $\chi^2$-έλεγχος ανεξαρτησίας. Σε περίπτωση που ο έλεγχος δε δείχνει στατιστικά σημαντική βελτίωση της «καθαρότητας» για το καλύτερο χαρακτηριστικό, δεν πραγματοποιείται ο διαμερισμός και το δέντρο τερματίζεται. Αυτή η μέθοδος είναι καταλληλότερη για κατηγορικές μεταβλητές, αλλά μπορεί να προσαρμοστεί και σε συνεχείς μεταβλητές αν τις μετασχηματίσουμε σε δυαδικές.

### 4.7.2 Κλάδεμα του δέντρου

Μια εναλλακτική και πιο αποδοτική μέθοδος από τη διακοπή της ανάπτυξης του δέντρου, είναι το κλάδεμα του ολοκληρωμένου δέντρου. Σε αυτήν την ιδέα βασίζονται μέθοδοι όπως η CART[7] και η C4.5[8]. Στη μέθοδο C4.5, το σύνολο εκπαίδευσης χρησιμοποιείται και για την ανάπτυξη και για το κλάδεμα του δέντρου, ενώ στη μέθοδο CART το δέντρο αφήνεται εσκεμμένα να αναπτυχθεί πλήρως με βάση το σύνολο εκπαίδευσης και στη συνέχεια κλαδεύεται με βάση το σύνολο ελέγχου.

Η γενική ιδέα πίσω από το κλάδεμα είναι ότι αν ένα δέντρο αποτελείται από πολλά επίπεδα, είναι πολύ πιθανό να προσαρμοστεί υπερβολικά κοντά στα δεδομένα του συνόλου εκπαίδευσης. Κατά συνέπεια, τα κλαδιά που είναι αδύναμα και δε μειώνουν ιδιαίτερα το ποσοστό σφάλματος, θα πρέπει να αποκοπούν.

Στο παράδειγμα με το σύστημα αυτόματου ποτίσματος είδαμε ότι οι τελευταίοι διαμερισμοί προκύπτουν από ορθογώνιες περιοχές που περιέχουν ελάχιστα σημεία˙ μάλιστα υπάρχει ένα ορθογώνιο που περιέχει μόνο ένα σημείο. Από αυτό καταλαβαίνουμε ότι αυτοί οι τελευταίοι διαμερισμοί είναι πιο πιθανό να αποτυπώσουν τους «θορύβους» που ίσως να υπάρχουν στο σύνολο εκπαίδευσης, παρά να αναδείξουν τα κρυμμένα πρότυπα που θα υπάρχουν σε νέα

[7] Breiman et al. (1984)
[8] Quinlan (1994)

δεδομένα, όπως τα δεδομένα του συνόλου ελέγχου. Το κλάδεμα περιλαμβάνει τη διαδοχική επιλογή ενός κόμβου απόφασης και τον επαναπροσδιορισμό του ως φύλλου. Με άλλα λόγια, κόβουμε τα κλαδιά που εκτείνονται πέρα από αυτόν τον κόμβο απόφασης (ή αλλιώς το υπόδεντρό του) κι έτσι μειώνουμε το μέγεθος του δέντρου. Η διαδικασία του κλαδέματος «θυσιάζει» κάποιους από τους κόμβους απόφασης, με σκοπό αφενός να βελτιώσει το ποσοστό λανθασμένης ταξινόμησης στο σύνολο ελέγχου και αφετέρου να προκύψει ένα δέντρο που θα αποτυπώνει τα πρότυπα και όχι τους «θορύβους» του συνόλου εκπαίδευσης.

## 4.8 Μέθοδοι κλαδέματος δέντρου

### 4.8.1 Κόστος πολυπλοκότητας

Επιστρέφοντας στο σχήμα 4.14, θέλουμε να εντοπίσουμε το σημείο στο οποίο η καμπύλη των νέων δεδομένων ξεκινά να αυξάνεται. Για τον εντοπισμό αυτού του σημείου, ο αλγόριθμος CART χρησιμοποιεί το κριτήριο του "κόστους πολυπλοκότητας" (cost complexity) με σκοπό να δημιουργήσει μία ακολουθία δέντρων που θα γίνονται όλο και μικρότερα μέχρι το σημείο που να προκύψει ένα δέντρο με έναν και μοναδικό κόμβο (ρίζα του δέντρου). Αυτό σημαίνει ότι το πρώτο βήμα θα είναι να βρούμε το καλύτερο υπόδεντρο κάθε μεγέθους (1, 2, 3, …). Στη συνέχεια, από αυτήν την ακολουθία υποδέντρων, θα επιλέξουμε εκείνο που ελαχιστοποιεί το ποσοστό λανθασμένης ταξινόμησης στο σύνολο ελέγχου.

Η κατασκευή του καλύτερου υπόδεντρου κάθε μεγέθους βασίζεται στο κριτήριο "κόστους πολυπλοκότητας" (CC), το οποίο ισούται με το άθροισμα του ποσοστού λανθασμένης ταξινόμησης ενός δέντρου (βάσει του συνόλου εκπαίδευσης) συν έναν συντελεστή ποινής για το μέγεθος του δέντρου. Έστω T ένα δέντρο και L(T) τα φύλλα του δέντρου. Το κόστος πολυπλοκότητας του δέντρου θα δίνεται από τη σχέση:

$$\boldsymbol{CC(T) = Err(T) + aL(T)}$$

όπου $Err(T)$ είναι το ποσοστό των παρατηρήσεων του συνόλου εκπαίδευσης που ταξινομήθηκαν λανθασμένα από το δέντρο T και α ο συντελεστής ποινής για το μέγεθος του δέντρου.

Για α=0, δεν υπάρχει ποινή και το κριτήριο "κόστους πολυπλοκότητας" δίνει ως καλύτερο δέντρο το πλήρως αναπτυγμένο δέντρο. Αντίθετα, όταν το α πάρει πολύ μεγάλη τιμή, ο συντελεστής ποινής υπερκαλύπτει το ποσοστό σφάλματος κι έτσι το καλύτερο δέντρο είναι εκείνο που αποτελείται μόνο από τον αρχικό κόμβο (ρίζα). Για να αποφύγουμε αυτήν την κατάσταση, θα ξεκινάμε με το πλήρως αναπτυγμένο δέντρο κι έπειτα θα αυξάνουμε σταδιακά τον συντελεστή ποινής α, μέχρι το σημείο όπου η τιμή του "κόστους πολυπλοκότητας" για το αρχικό δέντρο να ξεπερνά την τιμή του υπόδεντρου. Στο επόμενο βήμα, θα επαναλάβουμε την ίδια διαδικασία χρησιμοποιώντας το υπόδεντρο στη θέση του αρχικού δέντρου κ.ο.κ. Συνεχίζοντας με τον ίδιο τρόπο, θα δημιουργήσουμε μία ακολουθία δέντρων στα οποία θα μειώνεται σταδιακά ο αριθμός των κόμβων μέχρι να καταλήξουμε στο πιο ασήμαντο δέντρο, δηλαδή στο δέντρο που έχει έναν μόνο κόμβο.

Όπως είναι λογικό, από αυτήν την ακολουθία δέντρων, επιλέγουμε εκείνο με το μικρότερο ποσοστό λανθασμένης ταξινόμησης για το σύνολο ελέγχου και λέγεται "δέντρο ελάχιστου σφάλματος".

### 4.8.2 Μειωμένο σφάλμα

Μία απλή μέθοδος για το κλάδεμα των δέντρων ταξινόμησης είναι το μειωμένο σφάλμα (reduced error[9]). Καθώς η διαδικασία διέρχεται από όλους τους κόμβους του δέντρου από κάτω προς τα πάνω, ελέγχει σε κάθε κόμβο αν η αντικατάστασή του με ένα φύλλο, που θα περιέχει την κυρίαρχη κατηγορία, θα μειώσει το σφάλμα ταξινόμησης. Σε περίπτωση που μειώνει το σφάλμα, ο κόμβος αυτός κλαδεύεται. Η διαδικασία συνεχίζεται έως ότου το περαιτέρω κλάδεμα να επηρεάζει αρνητικά την αξιοπιστία του δέντρου. Για τον υπολογισμό του σφάλματος ταξινόμησης στον κόμβο t ισχύει:

$$\boldsymbol{Error(t) = 1 - \max_{class\ i} p(i|t)}$$

Δηλαδή, δίνουμε στον κόμβο t την κυρίαρχη κατηγορία ($max\, p(i|t)$) και υπολογίζουμε το ποσοστό των παρατηρήσεων που τοποθετήθηκαν σε άλλη κατηγορία ($1 - max\, p(i|t)$).

### 4.8.3 Ελάχιστο σφάλμα

Οι Niblett και Bratko[10] πρότειναν μία μέθοδο, από κάτω προς τα πάνω, για τον εντοπισμό του δέντρου που ελαχιστοποιεί το αναμενόμενο σφάλμα σε ένα σύνολο δεδομένων. Θα αναφερθούμε σε μία βελτιωμένη εκδοχή του ελαχίστου σφάλματος (Cestnik, Bratko[11]).

Για ένα πρόβλημα ταξινόμησης με κ-κατηγορίες, η αναμενόμενη πιθανότητα να φτάσει στον κόμβο t μία παρατήρηση που ανήκει στην i-κατηγορία είναι:

$$\boldsymbol{p_i(t) = \frac{n_i(t) + p_{ai} \cdot m}{N(t) + m}}$$

όπου

$n_i(t)$: το πλήθος των παρατηρήσεων του t κόμβου που ταξινομούνται στην i-κατηγορία

$p_{ai}$: η εκ των προτέρων πιθανότητα της i-κατηγορίας

m: παράμετρος της μεθόδου

$N(t)$: το συνολικό πλήθος των παρατηρήσεων που φτάνουν στον κόμβο t

[9] Quinlan (1987)
[10] T.Niblett, I.Bratko "Learning decision rules in noisy domains" Proceedings of Expert Systems 86, Cambridge University Press 1986
[11] B.Cestnik, I.Bratko "On estimating probabilities in tree pruning" Proceedings of the EWSL-91, Berlin Springer-Verlag 1991

Η παράμετρος m καθορίζει τη συμβολή της εκ των προτέρων πιθανότητας της i-κατηγορίας στην εκτίμηση της δεσμευμένης πιθανότητας της i-κατηγορίας, μέσω της σχετικής συχνότητας $^{n_i(t)}/_{N(t)}$. Για λόγους απλότητας, υποθέτουμε ότι η παράμετρος m παίρνει την

ίδια τιμή για όλες τις κατηγορίες. Η $p_i(t)$ πιθανότητα ονομάζεται "m-probability estimate". Όταν ταξινομείται μία νέα παρατήρηση που φτάνει στον κόμβο t, το αναμενόμενο σφάλμα δίνεται από τη σχέση:

$$\boldsymbol{EER(t) = \min_i [1 - p_i(t)] = \min_i \left[\frac{N(t) - n_i(t) + (1 - p_{ai}) \cdot m}{N(t) + m}\right]}$$

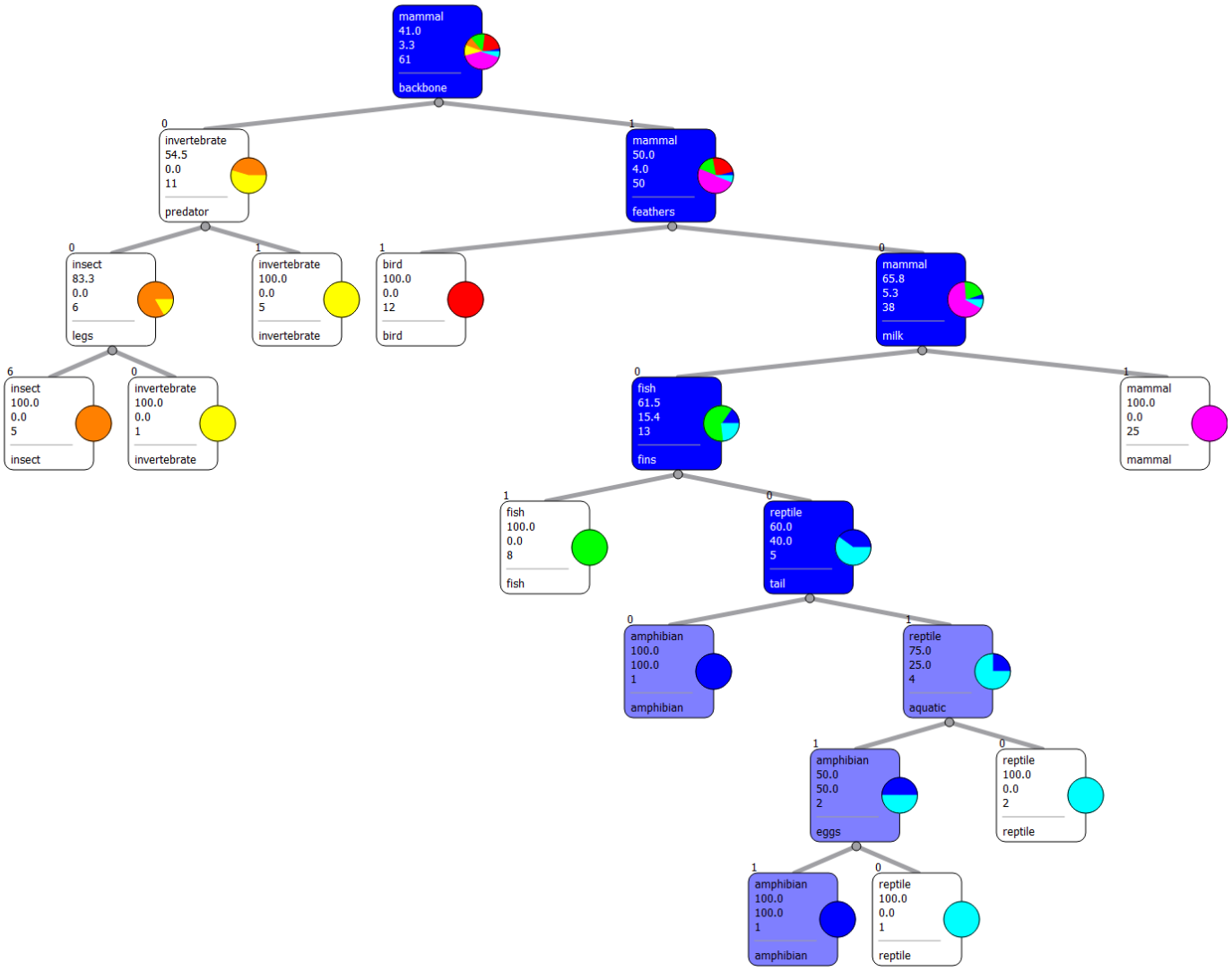
Αυτή η σχέση είναι μία γενίκευση του αναμενόμενου σφάλματος που υπολόγισαν οι Niblett και Bratko. Πράγματι, αν θέσουμε m = κ και $p_{ai} = {}^{1}/_{\kappa}$ η εκ των προτέρων πιθανότητα της ομοιόμορφης κατανομής για όλες τις κατηγορίες (i=1, 2, …, κ), προκύπτει:

$$\boldsymbol{EER(t) = \min_i \left[\frac{N(t) - n_i(t) + k - 1}{N(t) + k}\right]}$$

που είναι η παλαιότερη εκδοχή της μεθόδου.

### 4.8.4 Παράδειγμα

Θα χρησιμοποιήσουμε τα δεδομένα του ζωολογικού κήπου. Διαθέτουμε 101 παρατηρήσεις τις οποίες θα χωρίσουμε σε ένα σύνολο εκπαίδευσης (61 παρατηρήσεις) και ένα σύνολο ελέγχου (40 παρατηρήσεις), με τυχαίο τρόπο. Θέτουμε ως κριτήριο διαμερισμού του Gain Ratio και ως κατηγορία-στόχο τα "αμφίβια". Το πλήρως αναπτυγμένο δέντρο που προκύπτει αν δεν σταματήσουμε την ανάπτυξή του και δεν το κλαδέψουμε θα περιλαμβάνει συνολικά 19 κόμβους και 10 φύλλα:




**Σχήμα 4.15. Πλήρως αναπτυγμένο δέντρο ταξινόμησης για το παράδειγμα του ζωολογικού κήπου**

Χρησιμοποιώντας το σύνολο ελέγχου, κατασκευάζουμε τον πίνακα σύγχυσης για να ελέγξουμε την αξιοπιστία του:

| | amphibian | bird | fish | insect | invertebrate | mammal | reptile | |
|---|---|---|---|---|---|---|---|---|
| **amphibian** | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **2** |
| **bird** | 0 | 8 | 0 | 0 | 0 | 0 | 0 | **8** |
| **fish** | 0 | 0 | 5 | 0 | 0 | 0 | 0 | **5** |
| **insect** | 0 | 0 | 0 | 2 | 1 | 0 | 0 | **3** |
| **invertebrate** | 0 | 0 | 0 | 0 | 4 | 0 | 0 | **4** |
| **mammal** | 0 | 0 | 0 | 0 | 0 | 16 | 0 | **16** |
| **reptile** | 0 | 0 | 0 | 0 | 0 | 0 | 2 | **2** |
| | **2** | **8** | **5** | **2** | **5** | **16** | **2** | **40** |

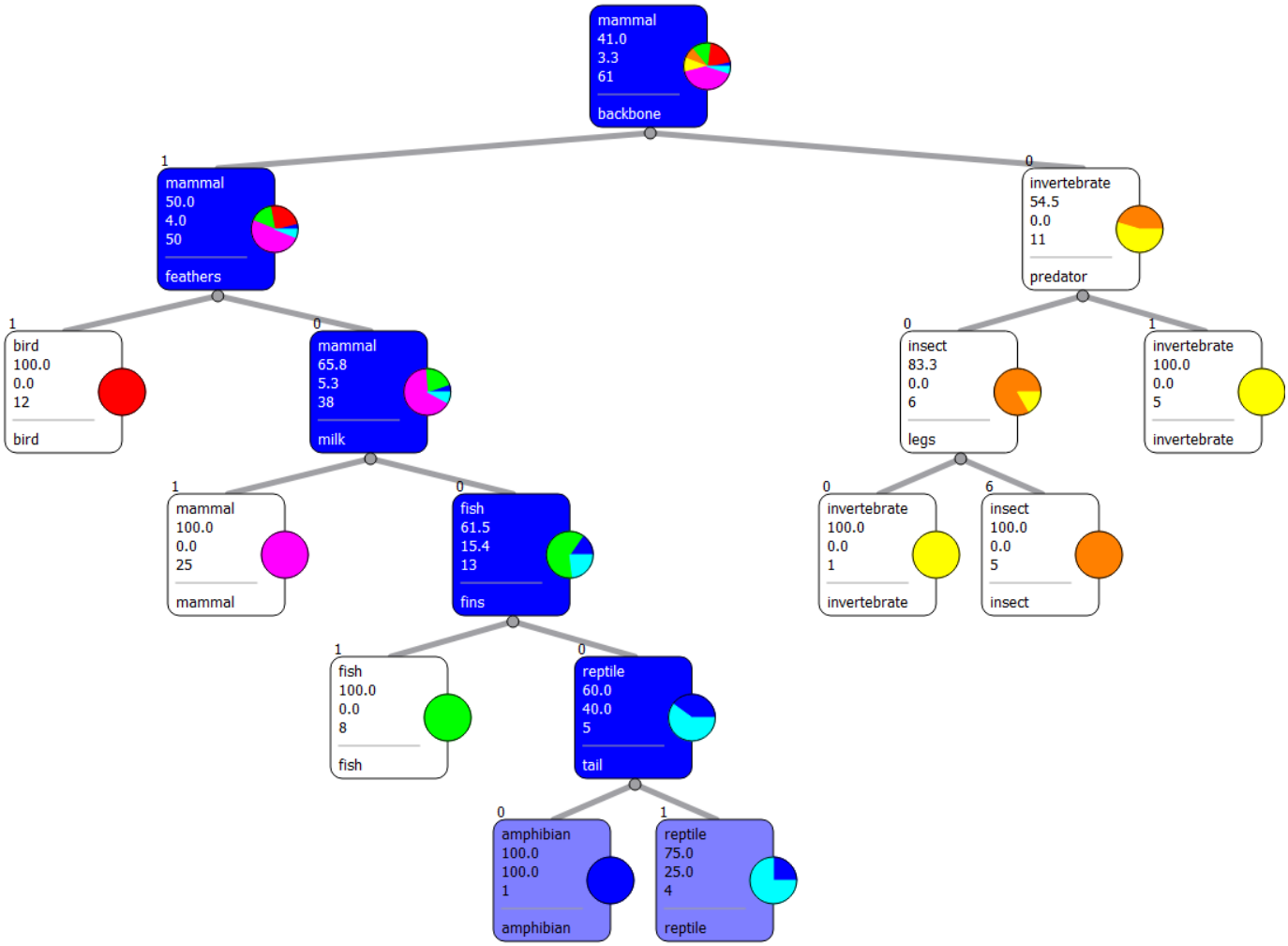
**Note:** columns represent predictions, row represent true classes

**Πίνακας 4.3. Πίνακας σύγχυσης για τα δεδομένα του ζωολογικού κήπου**

Παρατηρούμε, λοιπόν, ότι το δέντρο δίνει μόνο μία λανθασμένη πρόβλεψη, δηλαδή:

$$Classification\ accuracy = \frac{39}{40} = 0.975$$

Τώρα θα κλαδέψουμε το αρχικό δέντρο, εφαρμόζοντας τη μέθοδο "m-probability estimate" (m=2). Το νέο δέντρο που προκύπτει περιλαμβάνει 15 κόμβους και 8 φύλλα. Αποτελείται από λιγότερα επίπεδα, χωρίς όμως να επηρεάζεται αρνητικά η απόδοσή του αφού η αξιοπιστία του παραμένει στο 97.5%.




**Σχήμα 4.16. Κλαδεμένο δέντρο ταξινόμησης για το παράδειγμα του ζωολογικού κήπου**

**Target class:** amphibian
**Tree size:** 15 nodes, 8 leaves

| Classification Tree | Class | P(Class) | P(Target) | # Inst |
|---|---|---|---|---|
| | mammal | 0.410 | 0.033 | 61 |
| backbone = 0 | invertebrate | 0.545 | 0.000 | 11 |
| predator = 0 | insect | 0.833 | 0.000 | 6 |
| legs = 0 | invertebrate | 1.000 | 0.000 | 1 |
| legs = 6 | insect | 1.000 | 0.000 | 5 |
| predator = 1 | invertebrate | 1.000 | 0.000 | 5 |
| backbone = 1 | mammal | 0.500 | 0.040 | 50 |
| feathers = 0 | mammal | 0.658 | 0.053 | 38 |
| milk = 0 | fish | 0.615 | 0.154 | 13 |
| fins = 0 | reptile | 0.600 | 0.400 | 5 |
| tail = 0 | amphibian | 1.000 | 1.000 | 1 |
| tail = 1 | reptile | 0.750 | 0.250 | 4 |
| fins = 1 | fish | 1.000 | 0.000 | 8 |
| milk = 1 | mammal | 1.000 | 0.000 | 25 |
| feathers = 1 | bird | 1.000 | 0.000 | 12 |

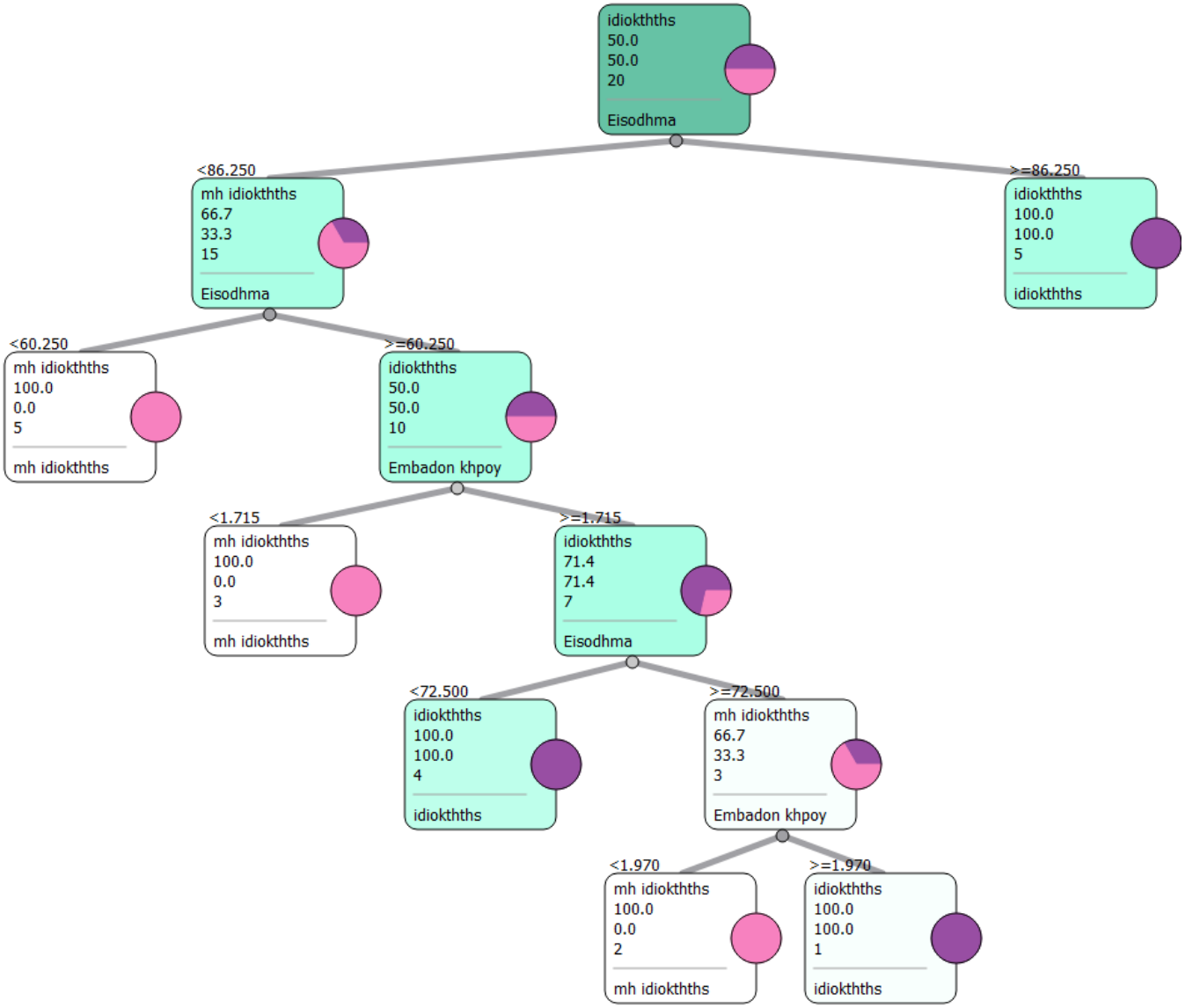
**Πίνακας 4.4. Στατιστικά χαρακτηριστικά κάθε κόμβου για το παράδειγμα του ζωολογικού κήπου**

## 4.9 Κανόνες ταξινόμησης

Όπως αναφέρεται και στην περιγραφή του κεφαλαίου, τα δέντρα ταξινόμησης παρέχουν εύκολα κατανοητούς κανόνες (αν τα δέντρα δεν αποτελούνται από πολλά επίπεδα). Κάθε φύλλο του δέντρου αντιστοιχεί σε έναν κανόνα ταξινόμησης. Επιστρέφοντας στο προηγούμενο παράδειγμα, ο μεσαίος δεξιός κόμβος του κλαδεμένου δέντρου μάς δίνει τον κανόνα:

$$IF\ (backbone = 0) AND\ (predator = 1) THEN\ Class = "invertebrate"$$

Ωστόσο, σε πολλές περιπτώσεις, το πλήθος των κανόνων μπορεί να μειωθεί αφαιρώντας κάποιους περιορισμούς που περισσεύουν. Ας θυμηθούμε το παράδειγμα με το σύστημα αυτόματου ποτίσματος.




Ο μεσαίος αριστερός κόμβος μάς δίνει τον κανόνα

$$IF\ (Eisodhma < 86.250) AND\ (Eisodhma < 60.250) THEN\ Class = "mh\ idiokththsth"$$

ο οποίος μπορεί να απλοποιηθεί ως

$$IF\ (Eisodhma < 60.250) THEN\ Class = "mh\ idiokththsth"$$

Αυτή η διαφάνεια της διαδικασίας και η απλότητα του αλγορίθμου που οδηγεί στην ταξινόμηση μιας παρατήρησης σε μία συγκεκριμένη κατηγορία, μάς βοηθά σε προβλήματα όπου δε μας ενδιαφέρει μόνο η τελική κατηγορία. Οι Berry και Linoff (2000) χρησιμοποιούν το παράδειγμα της ασφάλειας υγείας, όπου ο ασφαλιστής υποχρεούται να αποδείξει ότι η άρνηση της κάλυψης δεν οφείλεται σε διάκριση. Δείχνοντας τους κανόνες που οδήγησαν στην άρνηση, όπως για παράδειγμα το χαμηλό εισόδημα και το χαμηλό πιστωτικό ιστορικό, η εταιρεία μπορεί να αποφύγει τις μηνύσεις. Σε αντίθεση με άλλους ταξινομητές, τα αποτελέσματα που προκύπτουν από τα δέντρα ταξινόμησης γίνονται εύκολα κατανοητά από τους χρήστες των αποτελεσμάτων.

## 4.10 Δέντρα παλινδρόμησης

Η μέθοδος CART μπορεί να εφαρμοστεί και για συνεχείς μεταβλητές απόκρισης. Τα δέντρα παλινδρόμησης για πρόβλεψη λειτουργούν σχεδόν με τον ίδιο όπως και τα δέντρα ταξινόμησης. Η μεταβλητή απόκρισης Υ είναι συνεχής, αλλά οι κανόνες και η διαδικασία είναι ίδια: πραγματοποιούνται διαμερισμοί και σε κάθε κόμβο υπολογίζεται η «μη καθαρότητα» του δέντρου που προέκυψε. Στη συνέχεια, επιλέγεται ο διαμερισμός που ελαχιστοποιεί τα μέτρα «μη καθαρότητας».

### 4.10.1 Πρόβλεψη

Η προβλεπόμενη τιμή της μεταβλητής Υ για μία παρατήρηση, προκύπτει όπως και στα δέντρα ταξινόμησης: χρησιμοποιώντας τις τιμές των χαρακτηριστικών της παρατήρησης, την εισάγουμε στη ρίζα του δέντρου και τελικά καταλήγει σε κάποιο φύλλο. Στα δέντρα ταξινόμησης, η τιμή του κάθε φύλλου (που αποτελεί μία από τις κατηγορίες) καθορίζεται από την πλειοψηφία των παρατηρήσεων που κατέληξαν σε αυτό το φύλλο. Αντίστοιχα, στα δέντρα παλινδρόμησης, η τιμή κάθε φύλλου είναι ο μέσος όρος των τιμών για τις παρατηρήσεις που κατέληξαν σε αυτό.

### 4.10.2 Μέτρα «μη καθαρότητας»

Σε προηγούμενη παράγραφο περιγράψαμε μερικά μέτρα για τον υπολογισμό της «μη καθαρότητας» σε κάθε κόμβο ενός δέντρου ταξινόμησης. Τα μέτρα αυτά αποτελούν συναρτήσεις των αναλογιών μεταξύ των κατηγοριών στις οποίες ανήκουν οι παρατηρήσεις κάθε κόμβου. Για τα δέντρα παλινδρόμησης, ένα τυπικό μέτρο είναι το άθροισμα των τετραγώνων των αποκλίσεων από τον μέσο του κάθε φύλλου (least squared deviation) και δίνεται από τη σχέση:

$$\boldsymbol{R(t) = \frac{1}{N(t)} \sum_{i \in t} (y_i - \bar{y}(t))^2}$$

όπου $N(t)$ το πλήθος των παρατηρήσεων στον κόμβο t, $y_i$ η τιμή της μεταβλητής απόκρισης για την i-παρατήρηση και $\bar{y}(t)$ ο μέσος όρος των τιμών της απόκρισης σε κάθε κόμβο t.

Αυτό είναι ισοδύναμο με το τετραγωνικό σφάλμα, αφού ο μέσος όρος κάθε φύλλου είναι ακριβώς η προβλεπόμενη τιμή. Η χαμηλότερη τιμή της «μη καθαρότητας» είναι το μηδέν, όταν όλες οι τιμές του κόμβου είναι ίσες.

### 4.10.3 Αξιολόγηση της απόδοσης

Όπως ήδη αναφέραμε, οι προβλέψεις γίνονται με βάση τον μέσο όρο των τιμών των αποκρίσεων σε κάθε κόμβο. Η προβλεπτική ισχύς ενός δέντρου παλινδρόμησης υπολογίζεται χρησιμοποιώντας μέτρα όπως το RMSE (root mean squared error), το RRSE (root relative squared error) και ο $R^2$ (coefficient of determination).

### 4.10.4 Παράδειγμα: Πρόβλεψη της αξίας κατοικιών

Η πρώτη ανάλυση του συνόλου δεδομένων έγινε από τους Harrison και Rubinfeld[12] οι οποίοι ήθελαν να εξακριβώσουν αν ο "καθαρός αέρας" επηρεάζει την αξία των κατοικιών μιας πόλης. Το αρχικό σύνολο περιλαμβάνει 506 παρατηρήσεις και στο παράδειγμα αυτό θα χρησιμοποιήσουμε ένα τυχαίο δείγμα 70 παρατηρήσεων για την κατασκευή ενός δέντρου παλινδρόμησης.

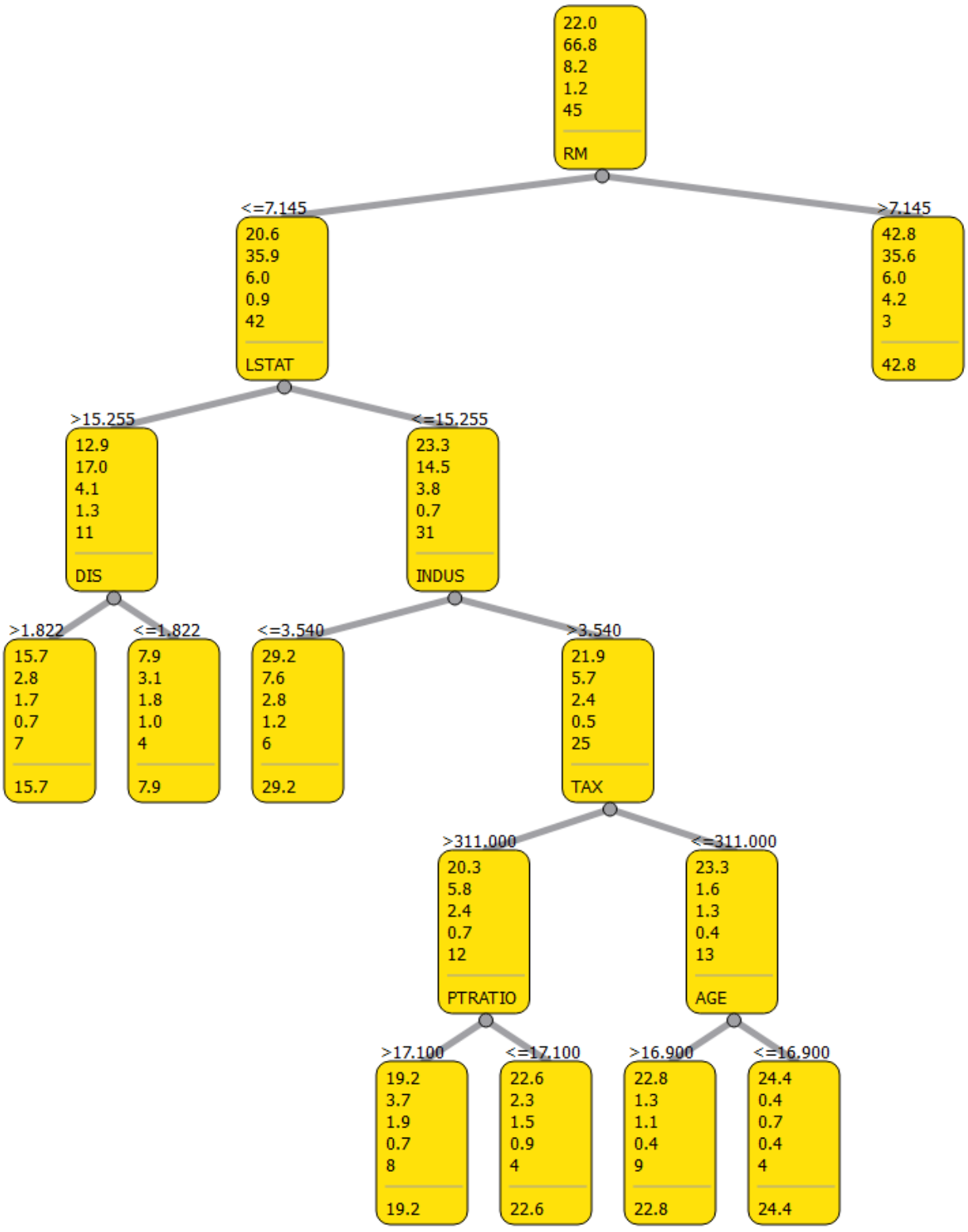
Οι παρατηρήσεις που διαθέτουμε, περιγράφουν κάποια χαρακτηριστικά σχετικά με τις κατοικίες στις γειτονιές μιας πόλης. Για κάθε γειτονιά, δίνεται ένα πλήθος μεταβλητών όπως το ποσοστό εγκληματικότητας, η αναλογία μαθητών/καθηγητών και η μέση αξία της κάθε κατοικίας, που αποτελεί και την μεταβλητή απόκρισης που θέλουμε να προβλέψουμε. Στον πίνακα που ακολουθεί περιγράφονται αναλυτικά τα 14 χαρακτηριστικά.

| Χαρακτηριστικό | Περιγραφή | Είδος μεταβλητής |
|---|---|---|
| **CRIM** | Κατά κεφαλήν ποσοστό εγκληματικότητας ανά γειτονιά | συνεχής, ανεξάρτητη |
| **ZN** | Ποσοστό κατοικιών που ανήκουν σε γειτονιά έκτασης πάνω από 2000$m^2$ | συνεχής, ανεξάρτητη |
| **INDUS** | Ποσοστό της γης που καταλαμβάνεται από επιχειρήσεις μη λιανικού εμπορίου | συνεχής, ανεξάρτητη |
| **CHAS** | Κοντά στο ποτάμι (ναι=1, όχι=0) | ψευδομεταβλητή, δυαδική, ανεξάρτητη |
| **NOX** | Συγκέντρωση οξειδίου του αζώτου (μέρη ανά $10^7$) | συνεχής, ανεξάρτητη |
| **RM** | Μέσος αριθμός δωματίων ανά κατοικία | συνεχής, ανεξάρτητη |
| **AGE** | Ποσοστό κατοικιών που χτίστηκαν πριν το 1940 | συνεχής, ανεξάρτητη |
| **DIS** | Σταθμισμένες αποστάσεις από 5 κέντρα απασχόλησης εργαζομένων της πόλης | συνεχής, ανεξάρτητη |
| **RAD** | Δείκτης προσβασιμότητας σε αυτοκινητόδρομους | συνεχής, ανεξάρτητη |
| **TAX** | Συντελεστής φόρου ιδιοκτησίας ανά 10000€ | συνεχής, ανεξάρτητη |
| **PTRATIO** | Αναλογία μαθητών/καθηγητών ανά γειτονιά | συνεχής, ανεξάρτητη |
| **B** | $1000(Bk-0.63)^2$ όπου Bk το ποσοστό μεταναστών στη γειτονιά | συνεχής, ανεξάρτητη |
| **LSTAT** | %Ελάχιστο ποσοστό πληθυσμού | συνεχής, ανεξάρτητη |
| **MEDV** | Μέση αξία κατοικίας σε χιλιάδες ευρώ | συνεχής, εξαρτημένη |

**Πίνακας 4.5. Αναλυτική περιγραφή των μεταβλητών για το παράδειγμα της πρόβλεψης αξίας κατοικιών**

[12] D.Harrison, D.L.Rubinfeld "Hedonic prices and the demand for clean air",Journal of Environmental Economics and Management (1978)

Αρχικά χωρίζουμε τα δεδομένα σε ένα σύνολο εκπαίδευσης (45 παρατηρήσεις) και ένα σύνολο ελέγχου (25 παρατηρήσεις). Το πλήρως αναπτυγμένο δέντρο παλινδρόμησης που προκύπτει είναι:



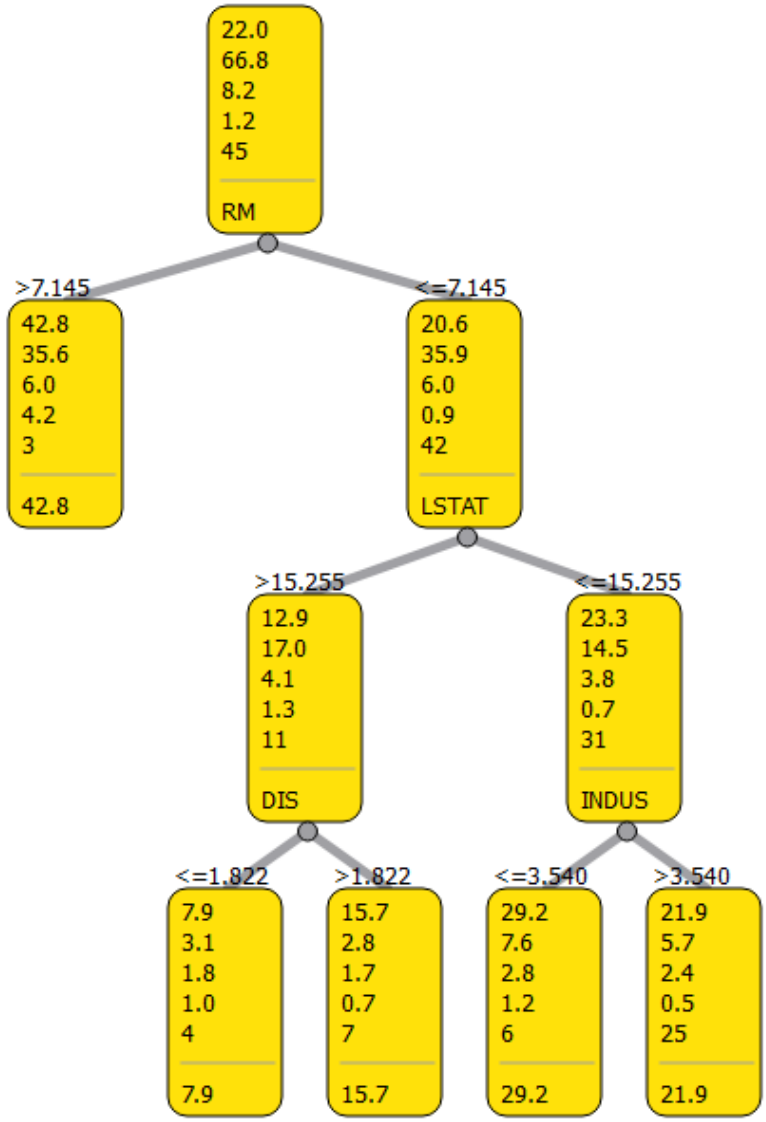

**Σχήμα 4.17. Πλήρως αναπτυγμένο δέντρο ταξινόμησης για το παράδειγμα της αξίας κατοικιών**

Το πλήρες δέντρο αποτελείται από 15 κόμβους και 8 φύλλα. Σε κάθε κόμβο αναγράφεται η προβλεπόμενη τιμή, η διακύμανση, η απόκλιση, το σφάλμα και το πλήθος των παρατηρήσεων. Στους κόμβους απόφασης αναγράφεται επιπλέον το χαρακτηριστικό βάσει του οποίου θα γίνει ο επόμενος διαμερισμός, ενώ στους τερματικούς κόμβους (φύλλα) βλέπουμε την προβλεπόμενη τιμή.

Η αξιολόγηση της απόδοσης του δέντρου γίνεται με χρήση του συνόλου ελέγχου και φαίνεται στον ακόλουθο πίνακα:

| | RMSE | RRSE | R2 |
|---|---|---|---|
| Regression Tree | 4.2213 | 0.4749 | 0.7745 |

Τώρα, θα κλαδέψουμε το αρχικό δέντρο χρησιμοποιώντας το κριτήριο "m-probability estimate" (m=5). Το νέο δέντρο που προκύπτει περιλαμβάνει 9 κόμβους και 5 φύλλα και η προβλεπτική του ισχύς είναι βελτιωμένη σε σχέση με το πλήρες δέντρο.




**Σχήμα 4.18. Κλαδεμένο δέντρο ταξινόμησης για το παράδειγμα της αξίας κατοικιών**

| | RMSE | RRSE | R2 |
|---|---|---|---|
| Regression Tree | 4.0216 | 0.4524 | 0.7953 |

## 4.11 Πλεονεκτήματα, μειονεκτήματα και επεκτάσεις

Τα δέντρα ταξινόμησης και παλινδρόμησης είναι σχετικά απλές μέθοδοι που μπορούμε να χρησιμοποιήσουμε όταν αντιμετωπίζουμε προβλήματα ταξινόμησης και πρόβλεψης. Επιπλέον, εντοπίζουν τα σημαντικότερα χαρακτηριστικά, τα οποία εμφανίζονται στα πρώτα επίπεδα του δέντρου. Σε γενικές γραμμές, τα δέντρα είναι εύκολα στην εφαρμογή καθώς δεν υπάρχει ανάγκη για μετασχηματισμό των μεταβλητών (οποιοσδήποτε μονότονος μετασχηματισμός των μεταβλητών θα δώσει τα ίδια δέντρα) και η επιλογή των υποσυνόλων γίνεται αυτόματα, αφού είναι μέρος της επιλογής των διαμερισμών. Στο παράδειγμα με την πρόβλεψη της τιμής των κατοικιών, το κλαδεμένο δέντρο επέλεξε αυτόματα μόνο 4 (RM, LSTAT, DIS, INDUS) από τα 13 χαρακτηριστικά.

Επίσης, τα δέντρα είναι ισχυρά στην αντιμετώπιση των ακραίων τιμών αφού η επιλογή των διαμερισμών εξαρτάται από τη διάταξη των τιμών των παρατηρήσεων και όχι από τις απόλυτες τιμές τους. Ωστόσο, είναι ευαίσθητα στις αλλαγές των δεδομένων, ακόμα και η παραμικρή αλλαγή κάποιας παρατήρησης μπορεί να οδηγήσει σε τελείως διαφορετικούς διαμερισμούς.

Σε αντίθεση με μοντέλα που υποθέτουν μία συγκεκριμένη σχέση ανάμεσα στη μεταβλητή απόκρισης και τους προγνωστικούς παράγοντες (π.χ. μία γραμμική σχέση όπως στη γραμμική παλινδρόμηση και τη γραμμική διακριτική ανάλυση), τα δέντρα ταξινόμησης και παλινδρόμησης είναι μη-γραμμικά και μη-παραμετρικά μοντέλα. Συνεπώς, επιτρέπουν την ύπαρξη ενός μεγάλου φάσματος σχέσεων μεταξύ της απόκρισης και των προγνωστικών παραγόντων. Ωστόσο, αυτό μπορεί να αποτελέσει και μειονέκτημα: εφόσον κάθε διαμερισμός γίνεται με βάση έναν συγκεκριμένο προγνωστικό παράγοντα, και όχι κάποιον συνδυασμό παραγόντων, είναι πολύ πιθανό το δέντρο να αγνοήσει τις σχέσεις μεταξύ των παραγόντων, και συγκεκριμένα γραμμικές σχέσεις όπως στη γραμμική παλινδρόμηση. Τα δέντρα ταξινόμησης αποδίδουν καλά σε περιπτώσεις όπου ο οριζόντιος και κατακόρυφος διαμερισμός του x-χώρου των δεδομένων διαχωρίζει επαρκώς τις κατηγορίες. Όμως, αν διαθέτουμε ένα σύνολο δεδομένων με δύο προγνωστικούς παράγοντες και δύο κατηγορίες οι οποίες είναι γραμμικά διαχωρίσιμες, το δέντρο ταξινόμησης θα έχει χαμηλότερη απόδοση από μεθόδους όπως η διακριτική ανάλυση. Ένας τρόπος για να βελτιώσουμε την απόδοση είναι η δημιουργία νέων προγνωστικών παραγόντων που να προέρχονται από τους αρχικούς, κι έτσι το δέντρο θα μπορεί να αποτυπώσει πιθανές σχέσεις μεταξύ των παραγόντων.

Ένα άλλο πρόβλημα στην απόδοση των δέντρων ταξινόμησης είναι ότι απαιτούν μεγάλα σύνολα δεδομένων για να δώσουν αξιόπιστα αποτελέσματα. Οι Breiman και Cutler χρησιμοποίησαν τη μέθοδο των "Random Forests"[13], μία επέκταση των δέντρων ταξινόμησης που ξεπερνά αυτό το πρόβλημα. Η γενική ιδέα είναι η κατασκευή πολλών δέντρων

[13] https://www.stat.berkeley.edu/~breiman/RandomForests/

ταξινόμησης από τα δεδομένα και ο συνδυασμός των αποτελεσμάτων τους, ώστε να προκύψει ένας καλύτερος ταξινομητής.

Τέλος, ένα σημαντικό πλεονέκτημα των δέντρων είναι οι εύκολα κατανοητοί κανόνες που παρέχουν και που μπορούν να φανούν χρήσιμοι σε εφαρμογές των επιχειρήσεων.

# Κεφάλαιο 5. Λογιστική παλινδρόμηση

## 5.1 Περιγραφή

Η λογιστική παλινδρόμηση επεκτείνει την ιδέα της γραμμικής παλινδρόμησης στο σημείο όπου η μεταβλητή απόκρισης Y είναι κατηγορική, δηλαδή οι παρατηρήσεις χωρίζονται σε κατηγορίες. Με τη λογιστική παλινδρόμηση μπορούμε να ταξινομήσουμε μία νέα παρατήρηση, για την οποία δε γνωρίζουμε σε ποια κατηγορία ανήκει, σε μία από τις κατηγορίες βασιζόμενοι στις τιμές των χαρακτηριστικών της (classification). Επιπλέον μπορούμε να τη χρησιμοποιήσουμε σε δεδομένα (για τα οποία είναι γνωστή η κατηγορία) ώστε να εντοπίσουμε ομοιότητες μεταξύ των παρατηρήσεων της κάθε κατηγορίας, που να αφορούν τις τιμές των χαρακτηριστικών τους (profiling). Η λογιστική παλινδρόμηση εφαρμόζεται σε περιπτώσεις όπως:

1. Όταν θέλουμε να ταξινομήσουμε τους πελάτες ως "πελάτης που θα επιστρέψει" ή "πελάτης που δε θα επιστρέψει" (classification).
2. Όταν θέλουμε να εντοπίσουμε παράγοντες που διαφοροποιούνται ανάμεσα σε άντρες και γυναίκες (profiling).
3. Όταν θέλουμε να προβλέψουμε την αποδοχή ή απόρριψη ενός δανείου βασιζόμενοι σε πληροφορίες όπως το πιστωτικό ιστορικό (classification).

Σε αυτό το κεφάλαιο θα ασχοληθούμε με την εφαρμογή της λογιστικής παλινδρόμησης σε προβλήματα ταξινόμησης, και συγκεκριμένα σε προβλήματα με δυαδική μεταβλητή απόκρισης, δηλαδή με δύο κατηγορίες. Μερικά παραδείγματα δυαδικής απόκρισης είναι "επιτυχία"/ "αποτυχία", "ναι"/ "όχι", "αγόρασε"/ "μην αγοράσεις" και "επιβίωση"/ "αποβίωση". Για λόγους ευκολίας συνήθως κωδικοποιούμε τη μεταβλητή Y ως 0 και 1.

Σημειώνουμε ότι σε κάποιες περιπτώσεις μπορούμε να επιλέξουμε τον μετασχηματισμό των μεταβλητών, αν είναι συνεχείς ή κατηγορικές με πολλές πιθανές κατηγορίες, ώστε να απλοποιήσουμε το πρόβλημα της λήψης αποφάσεων σε δυαδικό. Όπως και στην πολλαπλή γραμμική παλινδρόμηση, οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_k$ μπορεί να είναι συνεχείς ή κατηγορικές ή ένας συνδυασμός αυτών των δύο τύπων. Ωστόσο, ο στόχος της πολλαπλής γραμμικής παλινδρόμησης είναι να προβλέψουμε την τιμή της συνεχούς μεταβλητής Y για μια νέα παρατήρηση, ενώ με τη λογιστική παλινδρόμηση θέλουμε να προβλέψουμε την κατηγορία στην οποία θα ανήκει. Με άλλα λόγια, θέλουμε να ταξινομήσουμε τη νέα παρατήρηση σε μία από τις πιθανές κατηγορίες.

Στη λογιστική παλινδρόμηση ακολουθούμε δύο βήματα: στο πρώτο βήμα υπολογίζουμε την εκτίμηση των πιθανοτήτων να ανήκει η παρατήρηση σε κάθε μία από τις κατηγορίες. Στην περίπτωση που η μεταβλητή Y είναι δυαδική, υπολογίζουμε την εκτίμηση της $P(Y = 1)$, δηλαδή της πιθανότητας να ανήκει στην κατηγορία 1, και συνεπώς και της πιθανότητας να ανήκει στην κατηγορία 0. Στο δεύτερο βήμα χρησιμοποιούμε ένα όριο για αυτές τις

πιθανότητες, ώστε να ταξινομήσουμε τις παρατηρήσεις σε μία από τις κατηγορίες. Για παράδειγμα, στη δυαδική περίπτωση, ένα όριο 0.5 σημαίνει ότι οι παρατηρήσεις με εκτίμηση πιθανότητας $P(Y = 1) > 0.5$ θα ταξινομηθούν στην κατηγορία 1, ενώ για $P(Y = 1) < 0.5$ θα ταξινομηθούν στην κατηγορία 0. Το όριο αυτό δεν είναι αναγκαίο να πάρει την τιμή 0.5. Όταν εξετάζουμε ένα περιστατικό που έχει μικρή πιθανότητα να συμβεί, το να θέσουμε στο όριο μία τιμή μεγαλύτερη του μέσου όρου ή ακόμα και μικρότερη, μπορεί να αποδειχθεί επαρκές για την ταξινόμηση των παρατηρήσεων στην κατηγορία 1.

## 5.2 Το μοντέλο της λογιστικής παλινδρόμησης

Το μοντέλο της λογιστικής παλινδρόμησης χρησιμοποιείται σε διάφορα πεδία, όταν απαιτείται ένα δομημένο μοντέλο για να εξηγήσει ή να προβλέψει κατηγορικά (και συγκεκριμένα δυαδικά) αποτελέσματα.

Η ιδέα πίσω από τη λογιστική παλινδρόμηση είναι απλή: αντί να χρησιμοποιήσουμε την Y ως την εξαρτημένη μεταβλητή, χρησιμοποιούμε μία συνάρτηση αυτής που λέγεται logit. Για να εξηγήσουμε το logit θα το χωρίσουμε σε δύο στάδια: αρχικά εξετάζουμε το p, που είναι η πιθανότητα να ανήκει μία παρατήρηση στην κατηγορία 1 (ή αλλιώς να μην ανήκει στην κατηγορία 0). Σε αντίθεση με τη μεταβλητή Y που είναι ο αριθμός της κατηγορίας και παίρνει μόνο τις τιμές 0 και 1, το p μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα [0,1]. Ωστόσο, αν εκφράσουμε το p ως γραμμική συνάρτηση των q προγνωστικών παραγόντων θα πάρει τη μορφή:

$$\boldsymbol{p = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q} \quad \mathbf{(5.1)}$$

Όμως, το δεξιό μέλος της σχέσης αυτής δεν είναι απαραίτητο να παίρνει τιμές στο [0,1]. Για να το διορθώσουμε αυτό, χρησιμοποιούμε μία μη-γραμμική συνάρτηση των προγνωστικών παραγόντων της μορφής:

$$\boldsymbol{p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q)}}} \quad \mathbf{(5.2)}$$

Αυτή η συνάρτηση ονομάζεται συνάρτηση λογιστικής απόκρισης. Για οποιεσδήποτε τιμές των $x_1, x_2, \ldots, x_q$ το δεξιό μέλος της σχέσης (5.2) θα δίνει πάντα τιμές στο διάστημα [0,1]. Παρά το γεγονός ότι αυτή η μορφή λύνει το πρόβλημα από μαθηματικής άποψης, κάποιες φορές προτιμούμε να εξετάσουμε τα odds, που είναι ένα διαφορετικό μέτρο για να υπολογίσουμε το ενδεχόμενο "να ανήκει μία παρατήρηση σε μία κατηγορία". Τα odds να ανήκει μία παρατήρηση στην κατηγορία 1 (Y=1) ορίζονται ως το ποσοστό της πιθανότητας να ανήκει στην κατηγορία 1 προς το ποσοστό της πιθανότητας να ανήκει στην κατηγορία 0:

$$\boldsymbol{odds = \frac{p}{1 - p}} \quad \mathbf{(5.3)}$$

Αυτό το μέτρο είναι πολύ γνωστό στον ιππόδρομο, στο πόκερ, στα τυχερά παιχνίδια γενικότερα, αλλά και στην επιδημιολογία και πολλά άλλα πεδία. Αντί να υπολογίζουμε την πιθανότητα νίκης ή εκδήλωσης μιας ασθένειας, υπολογίζουμε τα odds. Ποια είναι, όμως, η

διαφορά; Για παράδειγμα, αν η πιθανότητα νίκης είναι 0.5, τα odds θα είναι αντίστοιχα $0.5/_{0.5} = 1$. Επιπλέον μπορούμε να κάνουμε τον αντίστροφο υπολογισμό: Δεδομένων των odds ενός ενδεχομένου, μπορούμε να υπολογίσουμε την πιθανότητα του ενδεχομένου μετατρέποντας τη σχέση (5.3) ως εξής:

$$\boldsymbol{p = \frac{odds}{1 + odds}} \quad \mathbf{(5.4)}$$

Συνεχίζοντας τις πράξεις η σχέση αυτή μπορεί να γραφτεί ως:

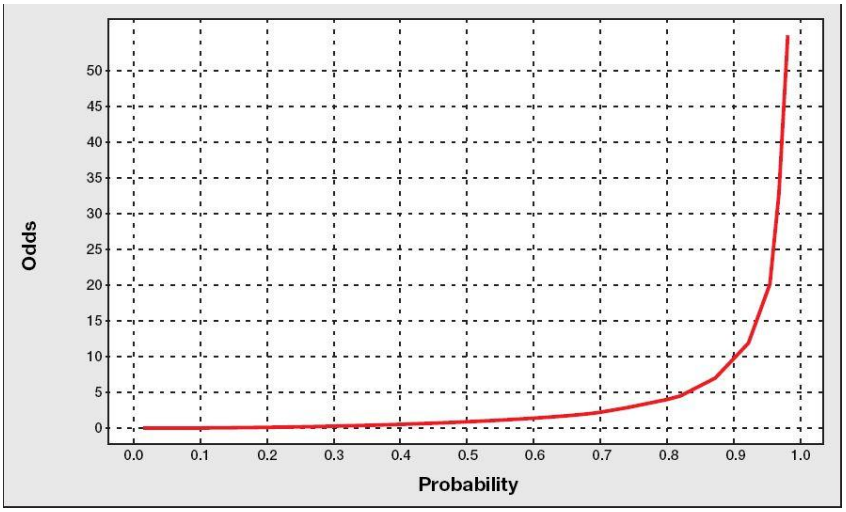
$$\boldsymbol{odds = e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q}} \quad \mathbf{(5.5)}$$

Τώρα, αν λογαριθμήσουμε και τα δύο μέλη της σχέσης (5.5) θα προκύψει η τυπική μορφή του λογιστικού μοντέλου:

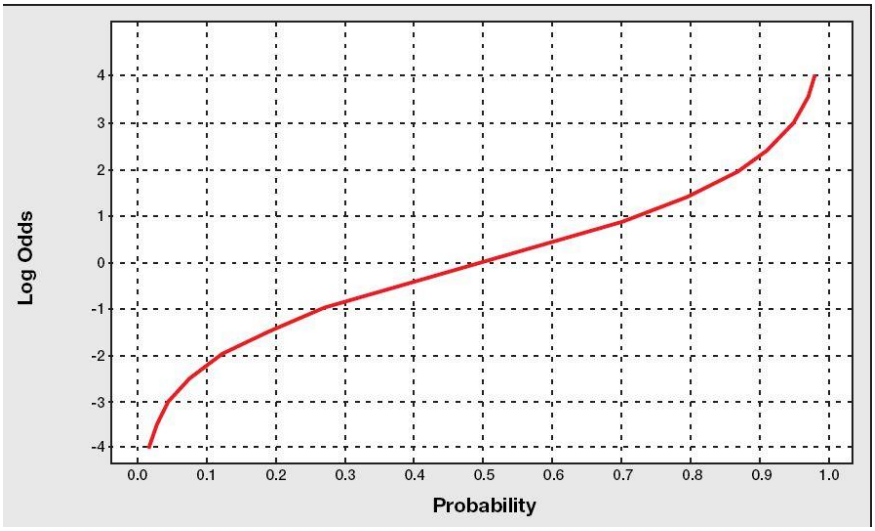
$$\boldsymbol{\log(odds) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q} \quad \mathbf{(5.6)}$$

Το log(odds) ονομάζεται logit και παίρνει τιμές από $-\infty$ έως $+\infty$. Έτσι, η τελική διατύπωση της σχέσης που υπάρχει ανάμεσα στην απόκριση και τους προγνωστικούς παράγοντες χρησιμοποιεί το logit ως την εξαρτημένη μεταβλητή και μοντελοποιεί αυτή τη σχέση ως μία γραμμική συνάρτηση των q προγνωστικών παραγόντων.

Στα γραφήματα που ακολουθούν φαίνεται η σχέση της πιθανότητας με τα odds και το logit. Παρατηρούμε ότι τα odds μπορούν να πάρουν οποιαδήποτε μη-αρνητική τιμή, ενώ το logit παίρνει οποιαδήποτε πραγματική τιμή.




**Σχήμα 5.1. Τα odds ως συνάρτηση της πιθανότητας**




**Σχήμα 5.2. Το logit ως συνάρτηση της πιθανότητας**

### 5.2.1 Παράδειγμα: Πρόβλεψη της ύπαρξης ασθένειας

Η βάση δεδομένων περιλαμβάνει τις μετρήσεις 345 ανδρών, καθένας από τους οποίους υποβλήθηκε σε 5 διαφορετικές εξετάσεις αίματος. Οι εξετάσεις αυτές θεωρούνται ευαίσθητες σε ηπατικές διαταραχές που μπορεί να προκύψουν από την υπερβολική κατανάλωση αλκοόλ και αποτελούν τις ανεξάρτητες μεταβλητές. Η εξαρτημένη μεταβλητή είναι η "ύπαρξη ηπατικής διαταραχής" με δυαδική απόκριση "ναι" ή "όχι". Στον πίνακα που ακολουθεί περιγράφονται αναλυτικά όλες οι μεταβλητές:

| Μεταβλητή | Περιγραφή | Χαρακτηρισμός μεταβλητής |
|---|---|---|
| **mcv** | μέση πυκνότητα αιμοσφαιρίνης | συνεχής |
| **alkphos** | αλκαλική φωσφατάση | συνεχής |
| **sgpt** | αμινοτρανσφεράση της αλανίνης | συνεχής |
| **sgot** | αμινοτρανσφεράση του ασπαρτικού οξέος | συνεχής |
| **gammagt** | γ-γλουταμυλτρανσπεπτιδάση | συνεχής |
| **drinks** | ποσότητα αλκοολούχων ποτών τη μέρα | συνεχής |
| **selector(ετικέτα κατηγορίας)** | ύπαρξη ηπατικής διαταραχής | δυαδική $\{0 = όχι, 1 = ναι\}$ |

**Πίνακας 5.1. Αναλυτική περιγραφή των μεταβλητών για το παράδειγμα της ηπατικής διαταραχής**

Ξεκινάμε χωρίζοντας τα δεδομένα σε ένα σύνολο εκπαίδευσης (60%) και ένα σύνολο ελέγχου (40%).

### 5.2.2 Μοντέλο με έναν προγνωστικό παράγοντα

Αρχικά θα κατασκευάσουμε ένα μοντέλο λογιστικής παλινδρόμησης με μία και μοναδική ανεξάρτητη μεταβλητή. Αυτό είναι ανάλογο του μοντέλου της γραμμικής παλινδρόμησης στο οποίο χρησιμοποιούμε μία ευθεία γραμμή για να αποτυπώσουμε τη σχέση ανάμεσα στην εξαρτημένη μεταβλητή Y και την ανεξάρτητη μεταβλητή X.

Στο παράδειγμα αυτό, θα εφαρμόσουμε τη λογιστική παλινδρόμηση με μοναδικό προγνωστικό παράγοντα τον "sgot" με σκοπό την ταξινόμηση των ατόμων σε "ναι" ή "όχι" όσον αφορά την ύπαρξη ηπατικής διαταραχής. Η εξίσωση που συνδέει την εξαρτημένη με την ανεξάρτητη μεταβλητή, σε σχέση με την πιθανότητα θα είναι:

$$Prob(selector = 1|sgot = x) = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

ή ισοδύναμα, σε σχέση με τα odds

$$Odds(selector = 1) = e^{b_0 + b_1 x} \quad (5.7)$$

**Logistic Regression - selector**

Dependent variable: selector
Factors:
sgot

**Estimated Regression Model (Maximum Likelihood)**

| | | *Standard* | *Estimated* |
|---|---|---|---|
| *Parameter* | *Estimate* | *Error* | *Odds Ratio* |
| CONSTANT | 0,518045 | 0,425325 | |
| sgot | -0,0370308 | 0,0172464 | 0,963646 |

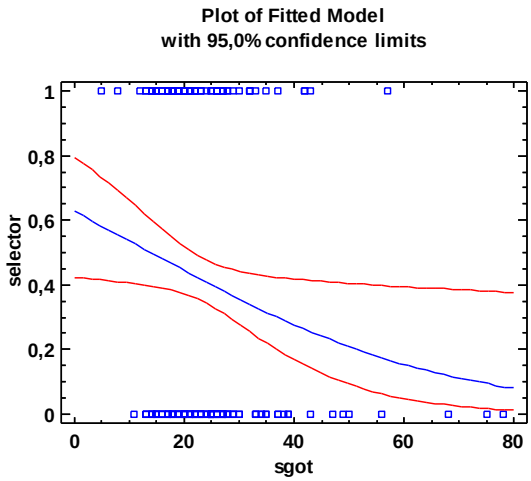
Το μοντέλο που προκύπτει είναι:

$$logit = 0.518054 - 0.0370308 \cdot sgot$$

Οι εκτιμήτριες της μέγιστης πιθανοφάνειας για τους συντελεστές του μοντέλου είναι $b_0 = 0.518$ και $b_1 = -0.037$. Έτσι, το προσαρμοσμένο μοντέλο που προκύπτει είναι:

$$Prob(selector = 1|sgot = x) = \frac{1}{1 + e^{-0.518 + 0.037x}} \quad (5.8)$$

Παρά το γεγονός ότι η λογιστική παλινδρόμηση μπορεί να εφαρμοστεί για πρόβλεψη, με την έννοια της πρόβλεψης της πιθανότητας μιας κατηγορικής απόκρισης, τις περισσότερες φορές χρησιμοποιείται για ταξινόμηση. Η διαφορά μεταξύ των δύο είναι ότι στην πρώτη περίπτωση θέλουμε να προβλέψουμε την πιθανότητα ενός ατόμου να πάσχει από ηπατική διαταραχή, ενώ στη δεύτερη περίπτωση θέλουμε να ταξινομήσουμε τα άτομα ως πάσχοντες και μη πάσχοντες. Στο επόμενο διάγραμμα βλέπουμε ότι η εξαρτημένη μεταβλητή "selector" παίρνει τιμές στο διάστημα $[0, 1]$. Για να μπορέσουμε να ταξινομήσουμε τα άτομα στις κατηγορίες 1 ή 0 (δηλαδή αν πάσχει από ηπατική διαταραχή ή όχι) χρειαζόμαστε μία τιμή διαχωρισμού (cutoff value). Αυτό ισχύει και στην περίπτωση που έχουμε περισσότερους προγνωστικούς παράγοντες.




**Σχήμα 5.3. Διάγραμμα των παρατηρήσεων ("selector" vs "sgot") και η προσαρμοσμένη λογιστική καμπύλη**

## Η τιμή διαχωρισμού

Δεδομένων των τιμών των χαρακτηριστικών, μπορούμε να προβλέψουμε την πιθανότητα κάθε παρατήρηση να ανήκει στην κατηγορία 1. Το επόμενο βήμα είναι να θέσουμε μία τιμή διαχωρισμού σε αυτές τις πιθανότητες, ώστε η κάθε παρατήρηση να ταξινομείται σε μία από τις δύο κατηγορίες. Έτσι, θέτουμε μία τιμή διαχωρισμού, έστω c, και στη συνέχεια ταξινομούμε τις παρατηρήσεις με πιθανότητα μεγαλύτερη της τιμής c ως "κατηγορία 1" και αντίστοιχα, τις παρατηρήσεις με πιθανότητα μικρότερη της c ως "κατηγορία 0".

Στο παράδειγμά μας, θα ταξινομήσουμε τα άτομα ως πάσχοντες ή μη πάσχοντες, με βάση την πληροφορία από το χαρακτηριστικό "sgot", δηλαδή εισάγοντας την τιμή x του χαρακτηριστικού στην εξίσωση (5.8). Το αποτέλεσμα της εξίσωσης θα μας δώσει, για κάθε άτομο, την εκτίμηση της πιθανότητας να ανήκει στους πάσχοντες. Το κάθε άτομο ταξινομείται ως πάσχων αν η πιθανότητα να πάσχει είναι μεγαλύτερη από την τιμή διαχωρισμού. Αντίστοιχα, αν θέλουμε να υπολογίσουμε τα odds της ύπαρξης ασθένειας και όχι την πιθανότητα, μία ισοδύναμη μέθοδος είναι να εφαρμόσουμε τον τύπο (5.7) και να συγκρίνουμε τα odds με την τιμή $^{c}/_{(1-c)}$. Αν τα odds δίνουν μεγαλύτερη τιμή, το άτομο ταξινομείται ως πάσχων, ενώ για μικρότερη τιμή ταξινομείται ως μη πάσχων. Για παράδειγμα, τα odds της ύπαρξης ηπατικής διαταραχής για ένα άτομο με sgot=33 εκτιμώνται από το μοντέλο ως:

$$Odds(selector = 1) = e^{-0.518+0.037 \cdot 33} = 2.020 \quad (5.9)$$

Για τιμή διαχωρισμού c=0.5, θα συγκρίνουμε τα odds με την τιμή $\frac{0.5}{1-0.5} = 1$. Εναλλακτικά, μπορούμε να υπολογίσουμε την πιθανότητα να πάσχει ένα άτομο από τη σχέση:

$$p = \frac{odds}{1 + odds} = \frac{2.020}{3.020} \approx 0.669$$

και να τη συγκρίνουμε απευθείας με την τιμή διαχωρισμού c=0.5. Και με τους δύο τρόπους, θα ταξινομούσαμε το άτομο ως πάσχων.

Οι διάφορες τιμές διαχωρισμού οδηγούν σε διαφορετικές ταξινομήσεις και κατά συνέπεια σε διαφορετικούς πίνακες σύγχυσης. Υπάρχουν διάφορες προσεγγίσεις για να καθορίσουμε τη βέλτιστη τιμή διαχωρισμού: Η τιμή που χρησιμοποιείται συνήθως για την περίπτωση που έχουμε δύο κατηγορίες είναι 0.5. Η βασική ιδέα είναι να αναθέσουμε μία παρατήρηση στην κατηγορία που είναι πιο πιθανό να ανήκει. Επιπλέον, μπορούμε να επιλέξουμε μία τιμή διαχωρισμού που να μεγιστοποιεί τη συνολική αξιοπιστία της ταξινόμησης. Η συνολική αξιοπιστία υπολογίζεται για τις διάφορες τιμές διαχωρισμού και στη συνέχεια επιλέγεται η τιμή που δίνει τη μεγαλύτερη αξιοπιστία. Φυσικά, υπάρχει ο κίνδυνος της υπερπροσαρμογής.

Στον ακόλουθο πίνακα βλέπουμε τη συνολική αξιοπιστία για τις διάφορες τιμές διαχωρισμού. Αρχικά, το μοντέλο (5.8) εφαρμόζεται για να προβλέψουμε την τιμή απόκρισης κάθε παρατήρησης χρησιμοποιώντας την πληροφορία από την ανεξάρτητη μεταβλητή. Αν η προβλεπόμενη τιμή είναι μεγαλύτερη από την τιμή διαχωρισμού, η παρατήρηση ταξινομείται στην κατηγορία 1 (TRUE). Αντίθετα, αν η προβλεπόμενη τιμή είναι μικρότερη ταξινομείται στην κατηγορία 0 (FALSE). Ο πίνακας δείχνει επίσης το ποσοστό των παρατηρήσεων που ταξινομήθηκαν στην σωστή κατηγορία για τις διάφορες τιμές διαμερισμού. Για παράδειγμα, για τιμή διαχωρισμού 0.45 ταξινομήθηκε σωστά το 40% των παρατηρήσεων που ανήκουν στην κατηγορία 1 και το 73.77% των παρατηρήσεων που ανήκουν στην κατηγορία 0, με συνολική αξιοπιστία 59.9%. Εντοπίζοντας τη βέλτιστη τιμή διαχωρισμού, μπορούμε να εξασφαλίσουμε πιο έγκυρα αποτελέσματα στην ταξινόμηση νέων δεδομένων.

**Prediction Performance - Percent Correct**

| *Cutoff* | *TRUE* | *FALSE* | *Total* |
|---|---|---|---|
| 0,0 | 100,00 | 0,00 | 41,06 |
| 0,05 | 100,00 | 0,00 | 41,06 |
| 0,1 | 100,00 | 1,64 | 42,03 |
| 0,15 | 100,00 | 2,46 | 42,51 |
| 0,2 | 98,82 | 3,28 | 42,51 |
| 0,25 | 98,82 | 5,74 | 43,96 |
| 0,3 | 94,12 | 10,66 | 44,93 |
| 0,35 | 88,24 | 16,39 | 45,89 |
| 0,4 | 65,88 | 47,54 | 55,07 |
| 0,45 | 40,00 | 73,77 | 59,90 |
| 0,5 | 5,88 | 96,72 | 59,42 |
| 0,55 | 2,35 | 100,00 | 59,90 |
| 0,6 | 0,00 | 100,00 | 58,94 |
| 0,65 | 0,00 | 100,00 | 58,94 |
| 0,7 | 0,00 | 100,00 | 58,94 |
| 0,75 | 0,00 | 100,00 | 58,94 |
| 0,8 | 0,00 | 100,00 | 58,94 |
| 0,85 | 0,00 | 100,00 | 58,94 |
| 0,9 | 0,00 | 100,00 | 58,94 |
| 0,95 | 0,00 | 100,00 | 58,94 |
| 1,0 | 0,00 | 100,00 | 58,94 |

**Πίνακας 5.2. Συνολική αξιοπιστία για τις διάφορες τιμές διαχωρισμού**

### 5.2.3 Εκτιμήσεις των παραμέτρων του λογιστικού μοντέλου

Στη λογιστική παλινδρόμηση, η σχέση ανάμεσα στην εξαρτημένη μεταβλητή Y και τις παραμέτρους b είναι μη γραμμική. Γι' αυτό, οι παράμετροι b δεν εκτιμώνται με τη μέθοδο των ελαχίστων τετραγώνων (όπως στην πολλαπλή παλινδρόμηση), αλλά με τη μέθοδο της μέγιστης πιθανοφάνειας. Η γενική ιδέα είναι να βρούμε τις εκτιμήσεις που μεγιστοποιούν την πιθανότητα να λάβουμε τα δεδομένα που έχουμε στην πραγματικότητα. Αυτό απαιτεί πολλές επαναλήψεις με τη χρήση υπολογιστικού προγράμματος. Σε γενικές γραμμές, οι εκτιμήτριες της μέγιστης πιθανοφάνειας είναι:

- Συνεπείς: Η πιθανότητα να διαφέρει η εκτιμήτρια από την πραγματική τιμή, πλησιάζει το μηδέν καθώς αυξάνεται το μέγεθος του δείγματος.
- Ασυμπτωτικά αποτελεσματικές: Η διακύμανση μεταξύ των εκτιμητριών είναι η μικρότερη δυνατή.
- Ασυμπτωτικά κανονικά-κατανεμημένες: Αυτό μάς επιτρέπει να υπολογίσουμε διαστήματα εμπιστοσύνης και να κάνουμε στατιστικούς ελέγχους με ανάλογο τρόπο, όπως και στην ανάλυση των μοντέλων γραμμικής πολλαπλής παλινδρόμησης, εφόσον το δείγμα είναι αρκετά μεγάλο.

Οι αλγόριθμοι που χρησιμοποιούνται για να υπολογισθούν οι συντελεστές των εκτιμητριών είναι επαναληπτικοί και λιγότερο ισχυροί από τους αλγορίθμους της γραμμικής παλινδρόμησης. Οι εκτιμήτριες που προκύπτουν είναι γενικά αξιόπιστες για σύνολα δεδομένων που συμπεριφέρονται «καλά», δηλαδή το πλήθος των παρατηρήσεων με μεταβλητή απόκρισης 0 και 1 είναι αρκετά μεγάλο, η αναλογία τους δεν είναι πολύ κοντά ούτε στο μηδέν ούτε στο 1 και το πλήθος των συντελεστών του μοντέλου της λογιστικής παλινδρόμησης είναι μικρό σε σχέση με το μέγεθος του δείγματος (περίπου 10%). Όπως και στη γραμμική παλινδρόμηση, η συγγραμμικότητα ( δηλαδή η υψηλή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών) μπορεί να οδηγήσει σε υπολογιστικές δυσκολίες. Ωστόσο, έχουν αναπτυχθεί υπολογιστικά απαιτητικοί αλγόριθμοι οι οποίοι παρακάμπτουν αυτές τις δυσκολίες.

Επιστρέφοντας στο παράδειγμα, θα κατασκευάσουμε το λογιστικό μοντέλο που έχει προσαρμοστεί στο σύνολο εκπαίδευσης, κρατώντας αυτή τη φορά και τις 6 ανεξάρτητες μεταβλητές. Στους επόμενους πίνακες φαίνονται οι συντελεστές του μοντέλου καθώς και τα αποτελέσματα του στατιστικού ελέγχου για κάθε μεταβλητή:

**Logistic Regression - selector**

Dependent variable: selector
Factors:
mcv
alkphos
sgpt
sgot
gammagt
drinks

**Estimated Regression Model (Maximum Likelihood)**

| | | *Standard* | *Estimated* |
|---|---|---|---|
| *Parameter* | *Estimate* | *Error* | *Odds Ratio* |
| CONSTANT | -6,27811 | 3,54763 | |
| mcv | 0,0658308 | 0,0396366 | 1,06805 |
| alkphos | 0,0219666 | 0,00897654 | 1,02221 |
| sgpt | 0,0571799 | 0,0158672 | 1,05885 |
| sgot | -0,128575 | 0,0330611 | 0,879347 |
| gammagt | -0,0116186 | 0,00645234 | 0,988449 |
| drinks | 0,065244 | 0,058356 | 1,06742 |

**Likelihood Ratio Tests**

| *Factor* | *Chi-Square* | *Df* | *P-Value* |
|---|---|---|---|
| mcv | 2,82223 | 1 | 0,0930 |
| alkphos | 6,299 | 1 | 0,0121 |
| sgpt | 14,7579 | 1 | 0,0001 |
| sgot | 17,7393 | 1 | 0,0000 |
| gammagt | 3,78217 | 1 | 0,0518 |
| drinks | 1,26035 | 1 | 0,2616 |

Αγνοώντας τις p-τιμές για τις μεταβλητές, ένα μοντέλο που να περιλαμβάνει όλες τις μεταβλητές περιγράφεται από την εξίσωση:

$$logit = -6.278 + 0.066 \cdot mcv + 0.022 \cdot alkphos + 0.057 \cdot sgpt - 0.129 \cdot sgot$$
$$-0.012 \cdot gammagt + 0.065 \cdot drinks \qquad (5.10)$$

Οι θετικοί συντελεστές για τις μεταβλητές “mcv”, “alkphos”, “sgpt” και “drinks” δείχνουν ότι οι υψηλές τιμές στα συγκεκριμένα χαρακτηριστικά είναι πιθανότερο να συνδέονται με την ύπαρξη ηπατικής διαταραχής. Αντιθέτως, για τις μεταβλητές “sgot” και “gammagt” με τους αρνητικούς συντελεστές συμπεραίνουμε ότι οι υψηλές τιμές αυτών των χαρακτηριστικών δίνουν μικρότερη πιθανότητα ύπαρξης της διαταραχής.

Για τα odds, μπορούμε να χρησιμοποιήσουμε την αντίστοιχη στήλη του παραπάνω πίνακα κι έτσι προκύπτει η εξίσωση:

$$Odds(selector = 1) = e^{-6.278}(1.068)^{mcv}(1.022)^{alphos}(1.059)^{sgpt}$$
$$(0.879)^{sgot}(0.988)^{gammagt}(1.067)^{drinks} \qquad (5.11)$$

Όπως παρατηρούμε, οι θετικοί συντελεστές του λογιστικού μοντέλου μετατρέπονται σε συντελεστές μεγαλύτερους της μονάδας στο μοντέλο των odds, και αντίστοιχα οι αρνητικοί συντελεστές γίνονται μικρότεροι της μονάδας στα odds.

Επιπλέον, μπορούμε να εφαρμόσουμε την εξίσωση (5.2), η οποία χρησιμεύει στην εκτίμηση της πιθανότητας να πάσχει ένα άτομο, δεδομένων των τιμών των 6 χαρακτηριστικών.

### 5.2.4 Ερμηνεία των αποτελεσμάτων σε σχέση με τα odds

Υπενθυμίζεται ότι τα odds δίνονται από τη σχέση:

$$\boldsymbol{odds = e^{b_0+b_1x_1+b_2x_2+\cdots+b_qx_q}}$$

Στο παράδειγμα του μοντέλου με έναν προγνωστικό παράγοντα, χρησιμοποιήσαμε την εξίσωση:

$$Odds(selector = 1) = e^{b_0+b_1 \cdot sgot}$$

όπου μοντελοποιήσαμε την πιθανότητα να πάσχει ένα άτομο από την ασθένεια ως συνάρτηση της τιμής που έδωσε η εξέταση "sgot" για το άτομο αυτό.

Μπορούμε να δούμε το μοντέλο αυτό ως ένα πολλαπλασιαστικό μοντέλο των odds. Τα odds για ένα άτομο με τιμή sgot=0 να πάσχει από την ασθένεια εκτιμώνται από τη σχέση $e^{-0.518+0.037 \cdot 0} = 0.596$ και αυτή είναι η βασική περίπτωση των odds. Είναι προφανές ότι σε αυτό το παράδειγμα δεν έχει νόημα να θέσουμε μηδενική τιμή σε κάποια από τις μεταβλητές, ωστόσο έχει νόημα να εξετάσουμε τον πολλαπλασιαστικό παράγοντα που προκύπτει για τις διάφορες τιμές του sgot. Για παράδειγμα, για ένα άτομο με τιμή sgot=50, τα odds να πάσχει από την ασθένεια θα αυξηθούν κατά τον πολλαπλασιαστικό παράγοντα $e^{0.037 \cdot 50} = 6.360$. Δηλαδή, για την περίπτωση sgot=50, τα odds θα αυξηθούν κατά 6.360 φορές σε σχέση με τη βασική περίπτωση. Έτσι, τα odds που προκύπτουν θα είναι $e^{-0.518+0.037 \cdot 50} = 3.789$.

Για να γενικεύσουμε το σκεπτικό αυτό στην περίπτωση που έχουμε πολλούς προγνωστικούς παράγοντες, θα χρησιμοποιήσουμε όλες τις ανεξάρτητες μεταβλητές του παραδείγματος. Τα odds για ένα άτομο να πάσχει από την ασθένεια δίνονται στη σχέση (5.11) ως συνάρτηση των 6 μεταβλητών.

Ας υποθέσουμε ότι η τιμή της μεταβλητής "sgot", έστω $x_1$, αυξάνεται κατά μία μονάδα, δηλαδή γίνεται $x_1 + 1$, ενώ οι υπόλοιπες μεταβλητές με τις τιμές $x_2, x_3, \dots, x_6$ αντίστοιχα, παραμένουν σταθερές. Η αναλογία των odds (odds ratio) θα είναι:

$$\frac{\boldsymbol{odds(x_1, x_2, \dots, x_6)}}{\boldsymbol{odds(x_1 + 1, x_2, \dots, x_6)}} = \frac{\boldsymbol{e^{b_0+b_1(x_1+1)+b_2x_2+\cdots+b_6x_6}}}{\boldsymbol{e^{b_0+b_1x_1+b_2x_2+\cdots+b_6x_6}}} = \boldsymbol{e^{b_1}}$$

Από αυτό συμπεραίνουμε ότι η αύξηση της $x_1$ έστω και κατά μία μόνο μονάδα, κρατώντας τις $x_2, x_3, \dots, x_6$ σταθερές, οδηγεί σε αύξηση της τιμής των odds κατά έναν παράγοντα $e^{b_1}$. Με άλλα λόγια, ο $b_1$ είναι ο πολλαπλασιαστικός παράγοντας κατά τον οποίο αυξάνονται τα odds (να ανήκει ένα άτομο στην κατηγορία 1) όταν η τιμή της $x_1$ αυξάνεται κατά μία μονάδα και οι υπόλοιπες μεταβλητές παραμένουν σταθερές. Για $b_1 < 0$, η αύξηση της $x_1$ οδηγεί σε μείωση των odds να ανήκει στην κατηγορία 1, ενώ για $b_1 > 0$ σημειώνεται αύξηση των odds.

Το πλεονέκτημα της ερμηνείας των αποτελεσμάτων που προκύπτουν με βάση τα odds, σε αντίθεση με τις πιθανότητες, είναι ότι το συμπέρασμα στο οποίο καταλήξαμε παραπάνω, ισχύει για οποιαδήποτε τιμή της $x_1$. Αυτό οφείλεται στο γεγονός ότι το μοντέλο των odds είναι πολλαπλασιαστικό και δεν ισχύει αν ερμηνεύσουμε τα αποτελέσματα με βάση τις πιθανότητες. Η επίδραση της αύξησης στην τιμή της $x_1$ δεν είναι η ίδια, εκτός αν η $x_1$ είναι δυαδική κατηγορική μεταβλητή, και αυτό γιατί στις πιθανότητες παίζει ρόλο ποια είναι η

πραγματική τιμή της $x_1$. Για παράδειγμα, αν αυξήσουμε την $x_1$ από 3 σε 4 θα έχουμε διαφορετικά αποτελέσματα στις πιθανότητες, σε σχέση με αύξηση από 30 σε 31. Συνεπώς, η αλλαγή που θα προκύψει στην πιθανότητα p για αύξηση μιας συγκεκριμένης μεταβλητής κατά μία μονάδα, κρατώντας τις υπόλοιπες μεταβλητές ίδιες, δεν είναι σταθερή, αλλά εξαρτάται από τις τιμές των μεταβλητών. Έτσι, συνήθως χρησιμοποιούμε τις πιθανότητες όταν εξετάζουμε συγκεκριμένες παρατηρήσεις.

## 5.3 Γιατί είναι ακατάλληλη η χρήση της γραμμικής παλινδρόμησης σε προβλήματα ταξινόμησης

Η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης δεν μπορεί να εφαρμοστεί σε οποιοδήποτε σύνολο δεδομένων. Για να εφαρμοστεί θα πρέπει να ισχύουν ορισμένες υποθέσεις, όπως: α) Η εξαρτημένη μεταβλητή Y πρέπει να είναι συνεχής και να ακολουθεί την κανονική κατανομή με σταθερή διακύμανση, β) Η κάθε ανεξάρτητη μεταβλητή πρέπει να συνδέεται γραμμικά με την εξαρτημένη μεταβλητή, γ) Τα υπόλοιπα πρέπει να είναι τυχαία κατανεμημένα και με σταθερή διακύμανση κ.ά.

Πρακτικά, μπορούμε να εφαρμόσουμε τη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης σε αυτό το παράδειγμα, αν αντιμετωπίσουμε την εξαρτημένη μεταβλητή Y ως συνεχή. Φυσικά, θα πρέπει να κωδικοποιήσουμε την Y με αριθμούς (αν η απόκριση είναι της μορφής "ναι/όχι" θα κωδικοποιηθεί ως 1 και 0 αντίστοιχα). Παρά το γεγονός ότι η μέθοδος θα δώσει αποτέλεσμα, αν παρατηρήσουμε προσεκτικά θα δούμε ότι υπάρχουν αρκετές ανωμαλίες:

1. Αν χρησιμοποιήσουμε το μοντέλο για να προβλέψουμε την τιμή της Y για κάθε παρατήρηση (ή για να ταξινομήσουμε κάθε παρατήρηση), η προβλεπόμενη τιμή δε θα είναι απαραίτητα 0 ή 1.
2. Αν παρατηρήσουμε το ιστόγραμμα ή το διάγραμμα πιθανότητας των υπολοίπων, θα δούμε ότι παραβιάζεται η υπόθεση της κανονικότητας των υπολοίπων. Όπως είναι αναμενόμενο, αν η Y παίρνει μόνο τις τιμές 0 ή 1 δεν μπορεί να ακολουθεί την κανονική κατανομή. Συγκεκριμένα, η καταλληλότερη κατανομή για τις τιμές 1 στο σύνολο δεδομένων είναι η διωνυμική με $p = P(Y = 1)$.
3. Επιπλέον, παραβιάζεται η υπόθεση ότι η διακύμανση της Y είναι σταθερή ανάμεσα σε όλες τις κατηγορίες. Εφόσον η Y ακολουθεί τη διωνυμική κατανομή, η διασπορά της θα ισούται με $np(1-p)$. Αυτό σημαίνει ότι η διασπορά θα είναι μεγαλύτερη για τις κατηγορίες όπου η πιθανότητα $p$ να συμβούν θα είναι κοντά στο 0.5, παρά αν είναι πιο κοντά στο 0 ή στο 1.

Παρακάτω βλέπουμε το μοντέλο που προκύπτει αν χρησιμοποιήσουμε την πολλαπλή γραμμική παλινδρόμηση στα δεδομένα του παραδείγματος, κρατώντας μόνο 3 από τις ανεξάρτητες μεταβλητές:

## Multiple Regression - selector

Dependent variable: selector
Independent variables:
mcv
sgpt
gammagt

| | | *Standard* | *T* | |
|---|---|---|---|---|
| *Parameter* | *Estimate* | *Error* | *Statistic* | *P-Value* |
| CONSTANT | -0,710923 | 0,732607 | -0,970401 | 0,3330 |
| mcv | 0,0127014 | 0,00820134 | 1,54869 | 0,1230 |
| sgpt | 0,00247651 | 0,00205309 | 1,20624 | 0,2291 |
| gammagt | -0,00253701 | 0,0010631 | -2,38642 | 0,0179 |

**Analysis of Variance**

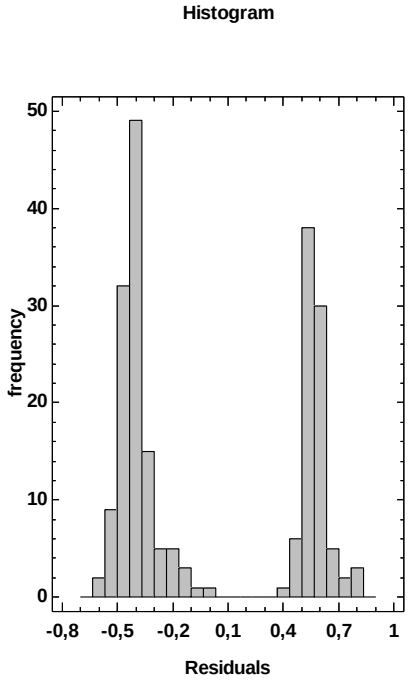
| *Source* | *Sum of Squares* | *Df* | *Mean Square* | *F-Ratio* | *P-Value* |
|---|---|---|---|---|---|
| Model | 1,67478 | 3 | 0,558259 | 2,34 | 0,0745 |
| Residual | 48,4218 | 203 | 0,238531 | | |
| Total (Corr.) | 50,0966 | 206 | | | |

$$selector = -0.711 + 0.013 \cdot mcv + 0.002 \cdot sgpt - 0.003 \cdot gammagt$$

Για να προβλέψουμε αν ένα άτομο πάσχει από ηπατική διαταραχή ($selector = 1$) ή όχι ($selector = 0$), θα δώσουμε κάποιες τιμές στις 3 μεταβλητές. Για παράδειγμα, για ένα άτομο με mcv=90, sgpt=28 και gammagt=30, το μοντέλο θα πάρει τη μορφή:

$$selector = -0.711 + 0.013 \cdot 90 + 0.002 \cdot 28 - 0.003 \cdot 30 = 0.425$$

Είναι προφανές ότι αυτή η τιμή απόκρισης δεν είναι έγκυρη για το αν το άτομο πάσχει ή όχι. Επιπλέον, όπως φαίνεται και στο ακόλουθο ιστόγραμμα, τα υπόλοιπα δεν ακολουθούν την κανονική κατανομή. Συμπεραίνουμε, λοιπόν, ότι η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης είναι ακατάλληλη σε προβλήματα ταξινόμησης, αφού το προσαρμοσμένο μοντέλο βασίζεται σε υποθέσεις που δεν ισχύουν κι έτσι τα αποτελέσματα δεν είναι αξιόπιστα.




**Σχήμα 5.4. Ιστόγραμμα των υπολοίπων του μοντέλου πολλαπλής γραμμικής παλινδρόμησης**

## 5.4 Αξιολογώντας την απόδοση ενός λογιστικού μοντέλου ταξινόμησης

Δύο από τα βασικότερα μέτρα που χρησιμοποιούμε για να αξιολογήσουμε την απόδοση ενός λογιστικού μοντέλου ταξινόμησης είναι οι πίνακες σύγχυσης και το διάγραμμα ανύψωσης. Στόχος είναι, όπως και στις άλλες μεθόδους, να εντοπίσουμε το μοντέλο που να ταξινομεί τις παρατηρήσεις στη σωστή κατηγορία, με βάση τις πληροφορίες από τους προγνωστικούς παράγοντες. Μία παραλλαγή αυτού του στόχου είναι να βρούμε το μοντέλο που θα αποδίδει καλύτερα στον εντοπισμό των παρατηρήσεων που ανήκουν στην κατηγορία που μας ενδιαφέρει. Εφόσον η κατασκευή του μοντέλου γίνεται με βάση το σύνολο εκπαίδευσης, είναι αναμενόμενο να αποδίδει καλά σε αυτά τα δεδομένα και γι' αυτό προτιμούμε τη χρήση ενός συνόλου ελέγχου για να ελέγξουμε την απόδοσή του. Υπενθυμίζεται ότι το σύνολο ελέγχου δε συμμετέχει στη διαδικασία κατασκευής του μοντέλου κι έτσι το χρησιμοποιούμε για να αξιολογήσουμε την ικανότητα του μοντέλου να ταξινομεί δεδομένα που δεν έχει «ξαναδεί».

### 5.4.1 Πίνακας σύγχυσης

Για να δημιουργήσουμε έναν πίνακα σύγχυσης από την ανάλυση της λογιστικής παλινδρόμησης, χρησιμοποιούμε την εξίσωση πρόβλεψης της πιθανότητας να ανήκει σε μία κατηγορία κάθε παρατήρηση του συνόλου ελέγχου, και στη συνέχεια τη συγκρίνουμε με την τιμή διαχωρισμού για να αποφασίσουμε σε ποια κατηγορία θα τοποθετηθεί. Έπειτα συγκρίνουμε αυτές τις ταξινομήσεις με την πραγματική κατηγορία στην οποία ανήκει κάθε παρατήρηση. Στο παράδειγμά μας, εφαρμόζουμε το μοντέλο για να προβλέψουμε την ταξινόμηση των παρατηρήσεων του συνόλου ελέγχου (138 παρατηρήσεις). Πρακτικά, χρησιμοποιούμε την εξίσωση (5.10) για να προβλέψουμε το logit και μετά υπολογίζουμε τις πιθανότητες $p$ μέσω της σχέσης $p = \frac{e^{logit}}{1+e^{logit}}$. Στη συνέχεια, συγκρίνουμε αυτές τις πιθανότητες με τη βέλτιστη τιμή διαχωρισμού και η διαδικασία ολοκληρώνεται ταξινομώντας τις παρατηρήσεις του συνόλου ελέγχου σε μία από τις δύο κατηγορίες.

Ο πίνακας σύγχυσης που προκύπτει για τα δεδομένα του συνόλου ελέγχου είναι:

| **Πραγματική κατηγορία** | **Προβλεπόμενη κατηγορία** | | |
|---|---|---|---|
| | 0 | 1 | |
| **0** | 65 | 24 | 89 |
| **1** | 16 | 33 | 49 |
| **Σύνολο** | 81 | 57 | 138 |

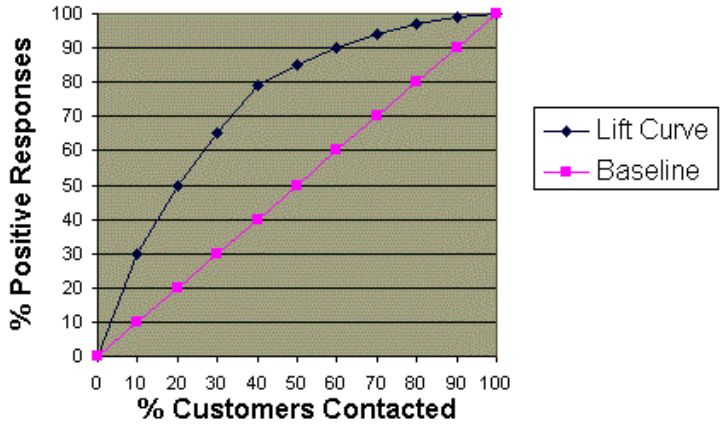
### 5.4.2 Διάγραμμα ανύψωσης

Ένα άλλο χρήσιμο εργαλείο για την αξιολόγηση της απόδοσης του μοντέλου ταξινόμησης είναι το διάγραμμα ανύψωσης. Για να περιγράψουμε τη χρήση του διαγράμματος ανύψωσης και τις πληροφορίες που μπορούμε να πάρουμε από αυτό, θα χρησιμοποιήσουμε ένα παράδειγμα. Ας υποθέσουμε ότι το τμήμα μάρκετινγκ μιας εταιρείας θέλει να ξεκινήσει μία στοχευμένη καμπάνια για να αυξήσει τις πωλήσεις, στέλνοντας διαφημιστικά φυλλάδια μέσω ταχυδρομείου. Η εταιρεία γνωρίζει ότι το ποσοστό των ατόμων που συνήθως ανταποκρίνονται σε μία τέτοιου είδους καμπάνια είναι περίπου 10%, και διαθέτει έναν κατάλογο με τα στοιχεία 10000 υποψήφιων πελατών. Με βάση το παραπάνω ποσοστό, αναμένεται να ανταποκριθούν περίπου 1000 άτομα. Ωστόσο, η εταιρεία έχει την οικονομική δυνατότητα να στείλει τα διαφημιστικά φυλλάδια μόνο σε 5000 άτομα, ενώ συγχρόνως θέλει να βελτιώσει το ποσοστό ανταπόκρισης. Υπάρχουν, λοιπόν, δύο επιλογές: Η πρώτη είναι να επιλέξει τυχαία 5000 άτομα από τον κατάλογο και η δεύτερη να χρησιμοποιήσει ένα μοντέλο ταξινόμησης ώστε να προβλέψει ποια άτομα είναι πιθανότερο να ανταποκριθούν. Με το διάγραμμα ανύψωσης μπορούμε να συγκρίνουμε τα αναμενόμενα αποτελέσματα και για τις δύο περιπτώσεις. Αν η εταιρεία επιλέξει τυχαία τα 5000 άτομα, αναμένεται να ανταποκριθούν περίπου 500, και αυτή είναι η περίπτωση που αναπαριστά η διαγώνια γραμμή

στο διάγραμμα (random line). Στη δεύτερη περίπτωση αναμένεται να αυξηθεί το ποσοστό ανταπόκρισης αφού το μοντέλο ταξινόμησης θα εντοπίσει τα κατάλληλα άτομα, και οι προβλέψεις του μοντέλου αποτυπώνονται στην καμπύλη του διαγράμματος. Οποιαδήποτε βελτίωση σε σχέση με τη διαγώνια γραμμή, θεωρείται «ανύψωση».

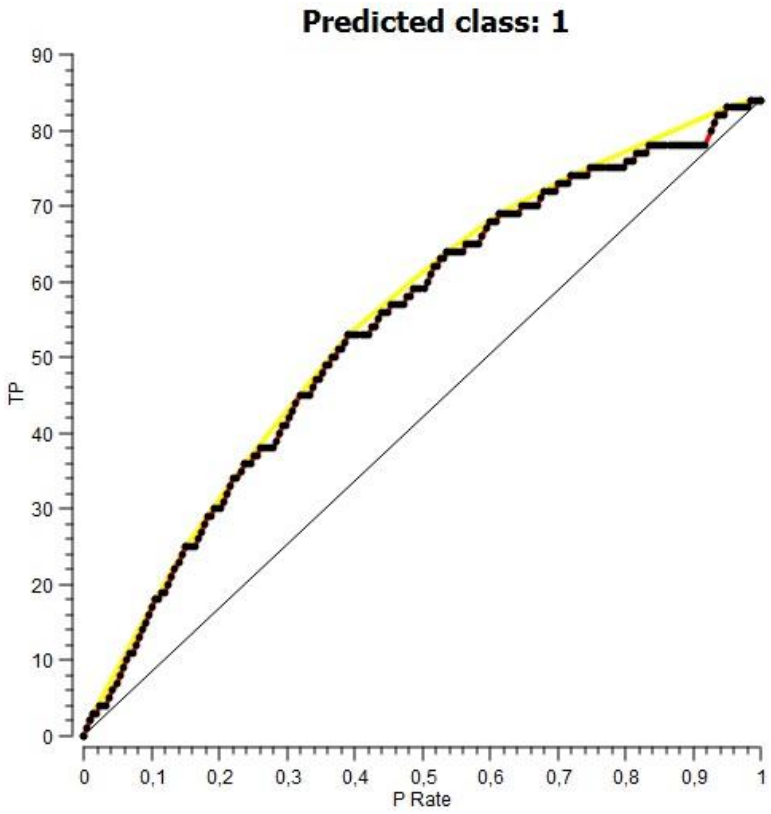
Έτσι, λοιπόν, με τη βοήθεια του μοντέλου υπολογίζεται για κάθε άτομο του καταλόγου η πιθανότητα να ανήκει στην κατηγορία των πελατών που θα ανταποκριθούν. Στη συνέχεια, τα άτομα ιεραρχούνται κατά τη φθίνουσα τιμή αυτής της πιθανότητας. Ο οριζόντιος άξονας δείχνει το ποσοστό του αρχικού συνόλου των 10000 ατόμων, τα οποία όμως πλέον είναι ιεραρχημένα. Ο κάθετος άξονας απεικονίζει το ποσοστό των ατόμων που προβλέπεται ότι θα ανταποκριθούν, με μέγιστη τιμή τα 1000 άτομα (αν η εταιρεία στείλει τα διαφημιστικά φυλλάδια σε όλο τον κατάλογο).

Για παράδειγμα:




Σε αυτό το διάγραμμα ανύψωσης βλέπουμε ότι αν τα διαφημιστικά αποσταλούν στο 10% του ιεραρχημένου καταλόγου, δηλαδή στα 1000 πρώτα άτομα, αναμένεται να ανταποκριθούν $\frac{30}{100} \cdot 1000 = 300$ άτομα. Σύμφωνα με την οικονομική δυνατότητα της εταιρείας, αν προσεγγίσει το 50% των ατόμων του ιεραρχημένου καταλόγου (τα 5000 πρώτα άτομα) προβλέπεται να ανταποκριθούν περίπου $\frac{85}{100} \cdot 1000 = 850$ άτομα. Παρατηρούμε, λοιπόν, ότι χρησιμοποιώντας τις προβλέψεις του μοντέλου το ποσοστό ανταπόκρισης βελτιώνεται σημαντικά σε σχέση με το αναμενόμενο ποσοστό που ήταν 10%.

Επιστρέφοντας στο αρχικό παράδειγμα πρόβλεψης της ασθένειας, θέτουμε ως στόχο την κατηγορία 1 (δηλαδή το άτομο πάσχει από ηπατική διαταραχή) και το διάγραμμα ανύψωσης που προκύπτει είναι:




**Σχήμα 5.5. Διάγραμμα ανύψωσης για το παράδειγμα της ηπατικής διαταραχής**

Παρατηρώντας το διάγραμμα, μπορούμε να δούμε, για ένα συγκεκριμένο ποσοστό του αρχικού συνόλου (x-άξονας), πόσα παραπάνω άτομα που πάσχουν εντοπίστηκαν από το μοντέλο σε σχέση με την τυχαία επιλογή. Για παράδειγμα, για το 40% του αρχικού συνόλου δηλαδή $\frac{40}{100} \cdot 138 \approx 55$ άτομα, η τυχαία επιλογή προβλέπει $\frac{30}{100} \cdot 33 \approx 10$ άτομα που πράγματι πάσχουν ενώ το μοντέλο προβλέπει $\frac{53}{100} \cdot 33 \approx 18$ άτομα που πράγματι πάσχουν.

### 5.4.3 Επιλογή των ανεξάρτητων μεταβλητών

Η επιλογή των ανεξάρτητων μεταβλητών συνδέεται με την εύρεση εναλλακτικών μοντέλων. Μπορούμε να κατασκευάσουμε πιο πολύπλοκα μοντέλα που να αποτυπώνουν την αλληλεπίδραση μεταξύ των ανεξάρτητων μεταβλητών. Για παράδειγμα, αν υποθέσουμε ότι υπάρχει αλληλεπίδραση μεταξύ των μεταβλητών “mcv” και “drinks”, θα προσθέσουμε στο μοντέλο τον παράγοντα “mcv*drinks”. Η επιλογή του καταλληλότερου από το σύνολο των εναλλακτικών μοντέλων γίνεται με βάση το πόσο καλά αποδίδει στο σύνολο ελέγχου. Για μοντέλα που αποδίδουν σχεδόν το ίδιο καλά, συνήθως επιλέγουμε το πιο απλό. Σημειώνουμε, ωστόσο, ότι δεν αρκεί η καλή απόδοση στα δεδομένα του συνόλου ελέγχου για να μπορούμε να είμαστε σίγουροι ότι το μοντέλο θα «δουλεύει» καλά σε οποιαδήποτε νέα δεδομένα. Πρέπει να λάβουμε υπ’ όψιν το γεγονός ότι η επιλογή του τελικού μοντέλου εφαρμόζοντάς το στο σύνολο ελέγχου, σημαίνει ότι μπορεί να έχει καλή απόδοση στα συγκεκριμένα δεδομένα, αλλά είναι πιθανό να έχει ενσωματώσει και κάποιες από τις ιδιαιτερότητες των

δεδομένων αυτών κι έτσι να παραπλανηθούμε σχετικά με το ποιο είναι τελικά το καλύτερο μοντέλο. Μπορεί, πράγματι, το τελικό μοντέλο να είναι το βέλτιστο μεταξύ όλων αλλά αυτό δε σημαίνει ότι θα έχει πάντα καλή απόδοση σε δεδομένα που δεν έχει «ξαναδεί». Γι' αυτό πρέπει να εξετάζουμε και πρακτικά θέματα όπως το κόστος συλλογής μεταβλητών, η τάση για σφάλματα (error-proneness) και η πολυπλοκότητα του μοντέλου πριν από την τελική επιλογή.

## 5.5 Έλεγχος καλής προσαρμογής

Η αξιολόγηση της ικανότητας του μοντέλου να προσαρμόζεται καλά στα δεδομένα είναι πολύ σημαντική όταν ο σκοπός την ανάλυσης είναι να εντοπίσουμε σε ποια σημεία διαφέρουν μεταξύ τους οι κατηγορίες με βάση τις πληροφορίες από τους προγνωστικούς παράγοντες (profiling), και όχι τόσο όταν σκοπός είναι η ταξινόμηση. Για παράδειγμα, αν μας ενδιαφέρει να χαρακτηρίσουμε ένα σύνολο ατόμων ως "ασθενείς" και "μη ασθενείς" με βάση τα αποτελέσματα των εξετάσεων, θέλουμε να βρούμε ένα μοντέλο που να προσαρμόζεται καλά στα δεδομένα. Ωστόσο, υπάρχει πάντα ο κίνδυνος της υπερπροσαρμογής και γι' αυτό ένα μοντέλο με υπερβολικά καλή προσαρμογή θα πρέπει να μας προβληματίσει. Επιπλέον, ερωτήματα σχετικά με τη χρησιμότητα συγκεκριμένων μεταβλητών μπορεί να μας απασχολήσουν ακόμη και σε προβλήματα ταξινόμησης. Έτσι, χρειαζόμαστε κάποια μέτρα για τον έλεγχο της καλής προσαρμογής του μοντέλου (goodness-of-fit) και η αξιολόγηση γίνεται με βάση το σύνολο εκπαίδευσης.

**Συνολική προσαρμογή**

Αρχικά θα αξιολογήσουμε τη συνολική προσαρμογή του μοντέλου στα δεδομένα, προτού εξετάσουμε τον κάθε προγνωστικό παράγοντα, και θα ελέγξουμε αν το μοντέλο που περιλαμβάνει όλες τις μεταβλητές είναι καλύτερο από ένα αφελές μοντέλο (δηλαδή το μοντέλο που αναθέτει τις παρατηρήσεις στην κυρίαρχη κατηγορία).

Η ελεγχοσυνάρτηση deviance (D) είναι ένα μέτρο για τον έλεγχο της καλής προσαρμογής. Συγκρίνουμε την αποκλίνουσα συμπεριφορά D του μοντέλου μας, με την $D_0$ του αφελούς μοντέλου. Αν η μείωση στην αποκλίνουσα συμπεριφορά είναι στατιστικά σημαντική (κάτι που φαίνεται από τη χαμηλή p-τιμή ή αντίστοιχα από την υψηλή τιμή του συντελεστή προσδιορισμού $R^2$), θεωρούμε ότι το μοντέλο είναι καλά προσαρμοσμένο.

Τέλος, με τον πίνακα σύγχυσης και το διάγραμμα ανύψωσης μπορούμε να εκτιμήσουμε την ικανότητα του μοντέλου να ταξινομεί τα δεδομένα. Αν το μοντέλο προσαρμόζεται καλά στα δεδομένα, περιμένουμε ότι θα αναθέτει τα δεδομένα στη σωστή κατηγορία. Όμως, όπως έχουμε ήδη αναφέρει, αυτό δεν εγγυάται την καλή απόδοσή του σε νέα δεδομένα, αφού ο πίνακας σύγχυσης και το διάγραμμα ανύψωσης βασίζονται στα ίδια δεδομένα που χρησιμοποιήσαμε για την κατασκευή του μοντέλου. Έτσι, χρησιμοποιούμε αυτά τα δύο μέτρα κυρίως για να ελέγξουμε αν υπάρχει υπερπροσαρμογή (που φαίνεται αν τα

αποτελέσματα είναι υπερβολικά καλά) και διάφορα τεχνικά προβλήματα ( που φαίνονται αν τα αποτελέσματα είναι υπερβολικά κακά) όπως λάθη στην εισαγωγή των δεδομένων κ.ά.

**Η επίδραση ενός προγνωστικού παράγοντα**

Κάθε προγνωστικός παράγοντας $X_i$ χαρακτηρίζεται από έναν συντελεστή $b_i$ και ένα σχετικό τυπικό σφάλμα $e_i$. Η p-τιμή κάθε παράγοντα δείχνει αν είναι στατιστικά σημαντικός ή αν συμβάλλει στο μοντέλο περισσότερο από τους άλλους παράγοντες. Ο λόγος $b_i/e_i$ χρησιμοποιείται για τον έλεγχο των υποθέσεων:

$$H_0: b_i = 0$$

$$H_a: b_i \neq 0$$

Ένας ισοδύναμος έλεγχος υποθέσεων με βάση τα odds είναι:

$$H_0: e^{b_i} = 1$$

$$H_a: e^{b_i} \neq 1$$

Στην περίπτωση που το σύνολο δεδομένων είναι πολύ μεγάλο, όλες οι p-τιμές θα είναι μικρές. Αν ένας προγνωστικός παράγοντας θεωρείται σημαντικός, μπορούμε να δούμε την πραγματική του επίδραση στο μοντέλο ελέγχοντας τα odds. Επίσης, συγκρίνοντας τα odds διαφορετικών παραγόντων, μπορούμε να ελέγξουμε ποιοι από αυτούς έχουν μεγαλύτερη και ποιοι μικρότερη επίδραση.

## 5.6 Λογιστική παλινδρόμηση για περισσότερες από δύο κατηγορίες

Το μοντέλο της λογιστικής παλινδρόμησης με δυαδική απόκριση μπορεί να επεκταθεί και για περισσότερες από δύο κατηγορίες. Έστω ότι έχουμε $m$ κατηγορίες. Με χρήση του λογιστικού μοντέλου, θα έχουμε για κάθε παρατήρηση συνολικά $m$ πιθανότητες να ανήκει σε μία από τις $m$ κατηγορίες. Δεδομένου ότι το άθροισμα των m πιθανοτήτων θα ισούται με τη μονάδα, αρκεί να εκτιμήσουμε μόνο τις $m - 1$ πιθανότητες.

### 5.6.1 Τακτικές κατηγορίες

Οι τακτικές κατηγορίες είναι κατηγορίες στις οποίες παίζει ρόλο η ιεράρχησή τους. Για παράδειγμα, στην αγοραπωλησία μετοχών, οι τρεις κατηγορίες "πούλα", "περίμενε" και "αγόρασε" μπορούν να αντιμετωπιστούν ως τακτικές. Ένας απλός κανόνας, λοιπόν, είναι ότι αν έχει νόημα να ιεραρχηθούν οι κατηγορίες, τότε θεωρούνται τακτικές. Όταν το πλήθος των κατηγοριών είναι μεγάλο (τυπικά μεγαλύτερο του 5), μπορούμε να αντιμετωπίσουμε την εξαρτημένη μεταβλητή ως συνεχή και να εφαρμόσουμε τη μέθοδο της πολλαπλής γραμμικής παλινδρόμησης. Για $m = 2$ χρησιμοποιείται το μοντέλο που περιγράφεται στις προηγούμενες παραγράφους. Έτσι, για την περίπτωση που έχουμε ένα μικρό πλήθος τακτικών κατηγοριών $(3 \leq m \leq 5)$, χρειαζόμαστε μία επέκταση του λογιστικού μοντέλου. Μία μέθοδος για να επεκτείνουμε την περίπτωση της δυαδικής κατηγορίας είναι το αθροιστικό λογιστικό μοντέλο (cumulative logit).

Για λόγους απλότητας στην ερμηνεία και τους υπολογισμούς, θα υπολογίσουμε τις αθροιστικές πιθανότητες ανάθεσης σε μία κατηγορία. Στο παράδειγμα της αγοραπωλησίας μετοχών, θα κωδικοποιήσουμε τις τρεις κατηγορίες ως 1="πούλα", 2="περίμενε" και "αγόρασε". Οι πιθανότητες που εκτιμώνται από το μοντέλο είναι $P(Y \leq 1)$ (όταν συμφέρει η πώληση μετοχών) και $P(Y \leq 2)$ (όταν συμφέρει η πώληση ή αναμονή). Οι τρεις μη αθροιστικές πιθανότητες ανάθεσης σε μία από τις κατηγορίες μπορούν να υπολογιστούν από τις δύο αθροιστικές πιθανότητες:

$$P(Y = 1) = P(Y \leq 1)$$

$$P(Y = 2) = P(Y \leq 2) - P(Y \leq 1)$$

$$P(Y = 3) = 1 - P(Y \leq 2)$$

Στη συνέχεια, θα φτιάξουμε το λογιστικό μοντέλο κάθε κατηγορίας ως συνάρτηση των προγνωστικών παραγόντων. Σε κάθε μία από τις $m - 1$ πιθανότητες αντιστοιχεί ένα $logit$. Έτσι, θα είναι:

$$logit(\alpha\gamma\acute{o}\rho\alpha\sigma\varepsilon) = log\frac{P(Y \leq 1)}{1 - P(Y \leq 1)}$$

$$logit(\alpha\gamma\acute{o}\rho\alpha\sigma\varepsilon \text{ ή } \pi\varepsilon\rho\acute{\iota}\mu\varepsilon\nu\varepsilon) = log\frac{P(Y \leq 2)}{1 - P(Y \leq 2)}$$

Έπειτα κάθε ένα από τα $logits$ μοντελοποιείται ως γραμμικής συνάρτηση των προγνωστικών παραγόντων (όπως και στην περίπτωση με τις δύο κατηγορίες). Αν στην αγοραπωλησία μετοχών υπάρχει μία ανεξάρτητη μεταβλητή, έστω x, θα πάρουμε τις δύο εξισώσεις:

$$logit(\alpha\gamma\acute{o}\rho\alpha\sigma\varepsilon) = a_0 + b_1 x$$

$$logit(\alpha\gamma\acute{o}\rho\alpha\sigma\varepsilon \text{ ή } \pi\varepsilon\rho\acute{\iota}\mu\varepsilon\nu\varepsilon) = b_0 + b_1 x$$

Αυτό σημαίνει ότι και οι δύο ευθείες έχουν την ίδια κλίση αλλά διαφορετικά σημεία τομής. Όταν εκτιμήσουμε τους συντελεστές $a_0, b_0, b_1$ , θα μπορούμε να υπολογίσουμε τις πιθανότητες ανάθεσης στην κάθε κατηγορία χρησιμοποιώντας τις $logit$ εξισώσεις. Θα είναι:

$$P(Y=1) = P(Y \leq 1) = \frac{1}{1+e^{-(a_0+b_1x)}}$$

$$P(Y=2) = P(Y \leq 2) - P(Y \leq 1) = \frac{1}{1+e^{-(b_0+b_1x)}} - \frac{1}{1+e^{-(a_0+b_1x)}}$$

$$P(Y=3) = 1 - P(Y \leq 2) = 1 - \frac{1}{1+e^{-(b_0+b_1x)}}$$

όπου $a_0, b_0, b_1$ είναι οι συντελεστές που εκτιμώνται από το σύνολο εκπαίδευσης.

Τώρα, γνωρίζουμε για κάθε παρατήρηση τις τρεις εκτιμημένες πιθανότητες ανάθεσης σε κάθε μία από τις τρεις κατηγορίες. Το τελευταίο βήμα είναι η ταξινόμηση των δεδομένων που γίνεται αναθέτοντας την κάθε παρατήρηση στην κατηγορία που είναι πιθανότερο να ανήκει, δηλαδή στην εκτιμημένη πιθανότητα με τη μεγαλύτερη τιμή.

### 5.6.2 Ονοματικές κατηγορίες

Όταν υπάρχουν διαφορετικές κατηγορίες που δεν μπορούν να ιεραρχηθούν, λέγονται ονοματικές κατηγορίες. Ένα τέτοιο παράδειγμα είναι η ταξινόμηση στα τρία είδη φυτών της ίριδας που αναφέρεται σε προηγούμενα κεφάλαια. Γενικά, έχει νόημα να μιλάμε για ονοματικές κατηγορίες όταν μπορούμε να τις χαρακτηρίσουμε ως A, B, C κ.ο.κ. Στο παράδειγμα με τα φυτά της ίριδας θα κωδικοποιήσουμε τις κατηγορίες ως A="iris-setosa", B="iris-versicolor" και C="iris-virginica" και θα εκτιμήσουμε τις πιθανότητες $P(Y = A), P(Y = B)$ και $P(Y = C)$. Όπως και προηγουμένως, αν γνωρίζουμε τις δύο πιθανότητες, μπορούμε να υπολογίσουμε και την τρίτη.

Σκοπός είναι και πάλι να φτιάξουμε ένα μοντέλο ως συνάρτηση των προγνωστικών παραγόντων. Έτσι, στο παράδειγμά μας, θέλουμε να τοποθετήσουμε τα φυτά σε μία από τις τρεις κατηγορίες ανάλογα με το μήκος του πετάλου, που είναι η ανεξάρτητη μεταβλητή x. Στη συνέχεια, γράφουμε τις $m-1$ $logit$ εξισώσεις που συνδέονται γραμμικά με την ανεξάρτητη μεταβλητή και είναι:

$$logit(A) = log\frac{P(Y=A)}{P(Y=C)} = a_0 + a_1x$$

$$logit(B) = log\frac{P(Y=B)}{P(Y=C)} = b_0 + b_1x$$

Μόλις υπολογίσουμε τους 4 συντελεστές από το σύνολο εκπαίδευσης, θα μπορούμε να εκτιμήσουμε τις πιθανότητες ανάθεσης σε κάθε κατηγορία ως εξής:

$$P(Y=A) = \frac{e^{a_0+a_1x}}{1+e^{a_0+a_1x}+e^{b_0+b_1x}}$$

$$P(Y=B) = \frac{e^{b_0+b_1x}}{1+e^{a_0+a_1x}+e^{b_0+b_1x}}$$

και επειδή $P(Y = A) + P(Y = B) + P(Y = C) \Leftrightarrow P(Y = C) = 1 - P(Y = A) - P(Y = B)$. Τέλος, η κάθε παρατήρηση τοποθετείται στην κατηγορία με τη μεγαλύτερη πιθανότητα.

## 5.7 Παράδειγμα: Ταξινόμηση ατόμων σε πολιτικές παρατάξεις

Το σύνολο δεδομένων αποτελείται από 435 παρατηρήσεις που αφορούν τον τρόπο που ψήφισαν τα μέλη του Κογκρέσου πάνω σε διάφορα ζητήματα (ανεξάρτητες μεταβλητές). Οι τιμές των 16 ανεξάρτητων μεταβλητών περιγράφουν 9 διαφορετικούς τύπους ψήφων: "voted for", paired for" και "announced for" οι οποίες απλοποιούνται ως "ναι", "voted against", "paired against" και "announced against" οι οποίες απλοποιούνται ως "όχι", "voted present", "voted present to avoid conflict of interest" και "did not vote" οι οποίες συμβολίζονται ως "?". Η εξαρτημένη μεταβλητή είναι δίτιμη με απόκριση "republicans" ή "democrats" ανάλογα με την παράταξη που ανήκει το κάθε άτομο. Σκοπός της ανάλυσης είναι η κατασκευή ενός μοντέλου που να προβλέπει την παράταξη στην οποία ανήκει κάθε μέλος ανάλογα με το πώς ψήφισε. Στον επόμενο πίνακα περιγράφονται όλες οι μεταβλητές του συνόλου δεδομένων:

| Όνομα μεταβλητής | Συμβολισμός | Είδος μεταβλητής | Τιμές μεταβλητής |
|---|---|---|---|
| **handicapped-infants** | hi | κατηγορική | {yes, no} |
| **water project cost sharing** | wpcs | κατηγορική | {yes, no} |
| **adoption of the budget resolution** | abs | κατηγορική | {yes, no} |
| **physician fee freeze** | pff | κατηγορική | {yes, no} |
| **el Salvador aid** | esa | κατηγορική | {yes, no} |
| **religious groups in schools** | rgs | κατηγορική | {yes, no} |
| **anti satellite test ban** | astb | κατηγορική | {yes, no} |
| **aid to Nicaraguan contras** | anc | κατηγορική | {yes, no} |
| **mx missile** | mxm | κατηγορική | {yes, no} |
| **immigration** | imm | κατηγορική | {yes, no} |
| **synfuels corporation cutback** | scc | κατηγορική | {yes, no} |
| **education spending** | es | κατηγορική | {yes, no} |
| **superfund right to sue** | srs | κατηγορική | {yes, no} |
| **crime** | cr | κατηγορική | {yes, no} |
| **duty free exports** | dfe | κατηγορική | {yes, no} |
| **export administration act South Africa** | eaasa | κατηγορική | {yes, no} |
| **class name(party)** | party | κατηγορική | {republicans, democrats} |

**Πίνακας 5.3. Αναλυτική περιγραφή των μεταβλητών για το παράδειγμα με τις πολιτικές παρατάξεις**

## Προεπεξεργασία των δεδομένων

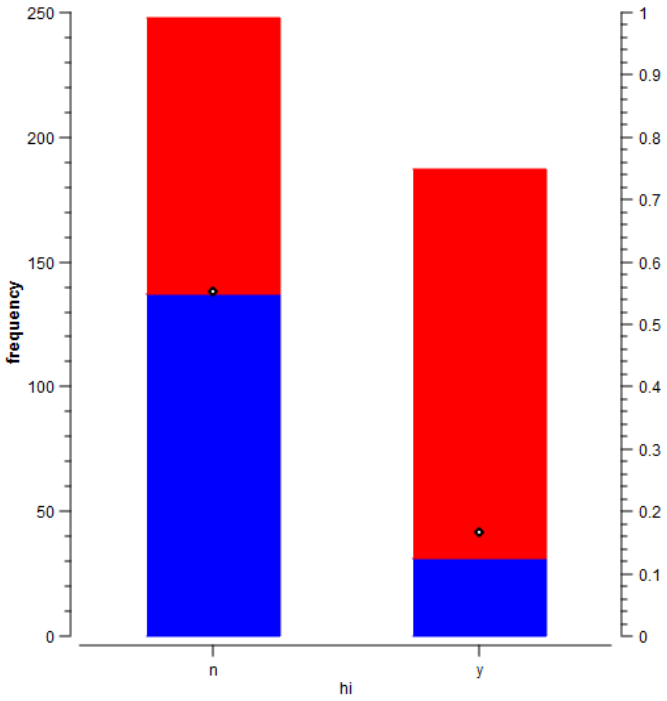
Όπως αναφέρεται παραπάνω, στο σύνολο δεδομένων υπάρχουν 203 παρατηρήσεις με την τιμή "?", η οποία ερμηνεύεται ως αγνοούμενη τιμή. Για να μπορούμε, λοιπόν, να προχωρήσουμε στην ανάλυση των δεδομένων θα πρέπει είτε να αφαιρέσουμε τις ελλιπείς παρατηρήσεις, είτε να υπολογίσουμε τις αγνοούμενες τιμές. Δεδομένου ότι αν αφαιρέσουμε τις παρατηρήσεις αυτές θα χάσουμε ένα μεγάλο μέρος των δεδομένων, επιλέγουμε να τις κρατήσουμε. Έτσι, αντικαθιστούμε τις αγνοούμενες τιμές με την τιμή που εμφανίζεται συχνότερα για το κάθε χαρακτηριστικό.

**Imputed values**
**Method:** Average/Most frequent
**Impute class values:** No

**handicapped-infants:** n
**water-project-cost-sharing:** y
**adoption-of-the-budget-resolution:** y
**physician-fee-freeze:** n
**el-salvador-aid:** y
**religious-groups-in-schools:** y
**anti-satellite-test-ban:** y
**aid-to-nicaraguan-contras:** y
**mx-missile:** y
**immigration:** y
**synfuels-corporation-cutback:** n
**education-spending:** n
**superfund-right-to-sue:** y
**crime:** y
**duty-free-exports:** n
**export-administration-act-south-africa:** y

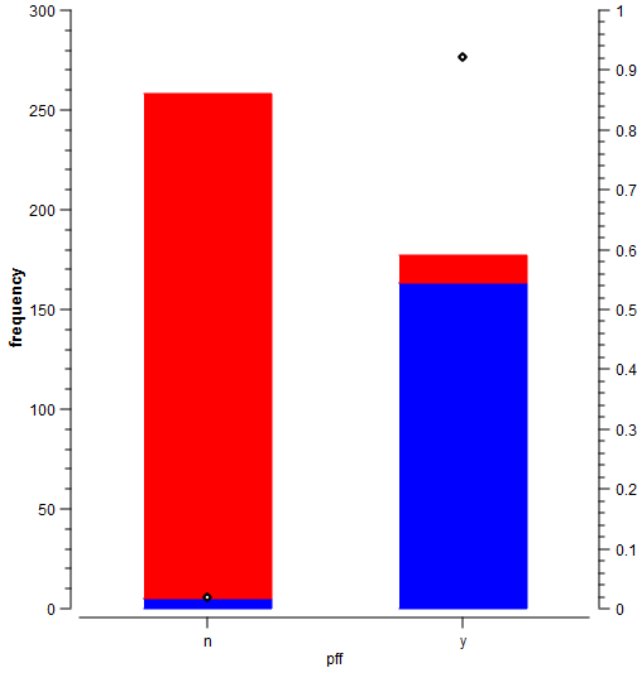
## Στατιστικά χαρακτηριστικά των δεδομένων

Χρησιμοποιώντας τα ιστογράμματα μπορούμε να δούμε πώς κατανέμονται οι ψήφοι των μελών των δύο παρατάξεων για κάθε χαρακτηριστικό. Για παράδειγμα, για την ανεξάρτητη μεταβλητή "hi" ο αριθμός των αρνητικών ψήφων είναι μοιρασμένος στις δύο παρατάξεις, ενώ θετική ψήφο έδωσαν περισσότερα μέλη των "democrats".



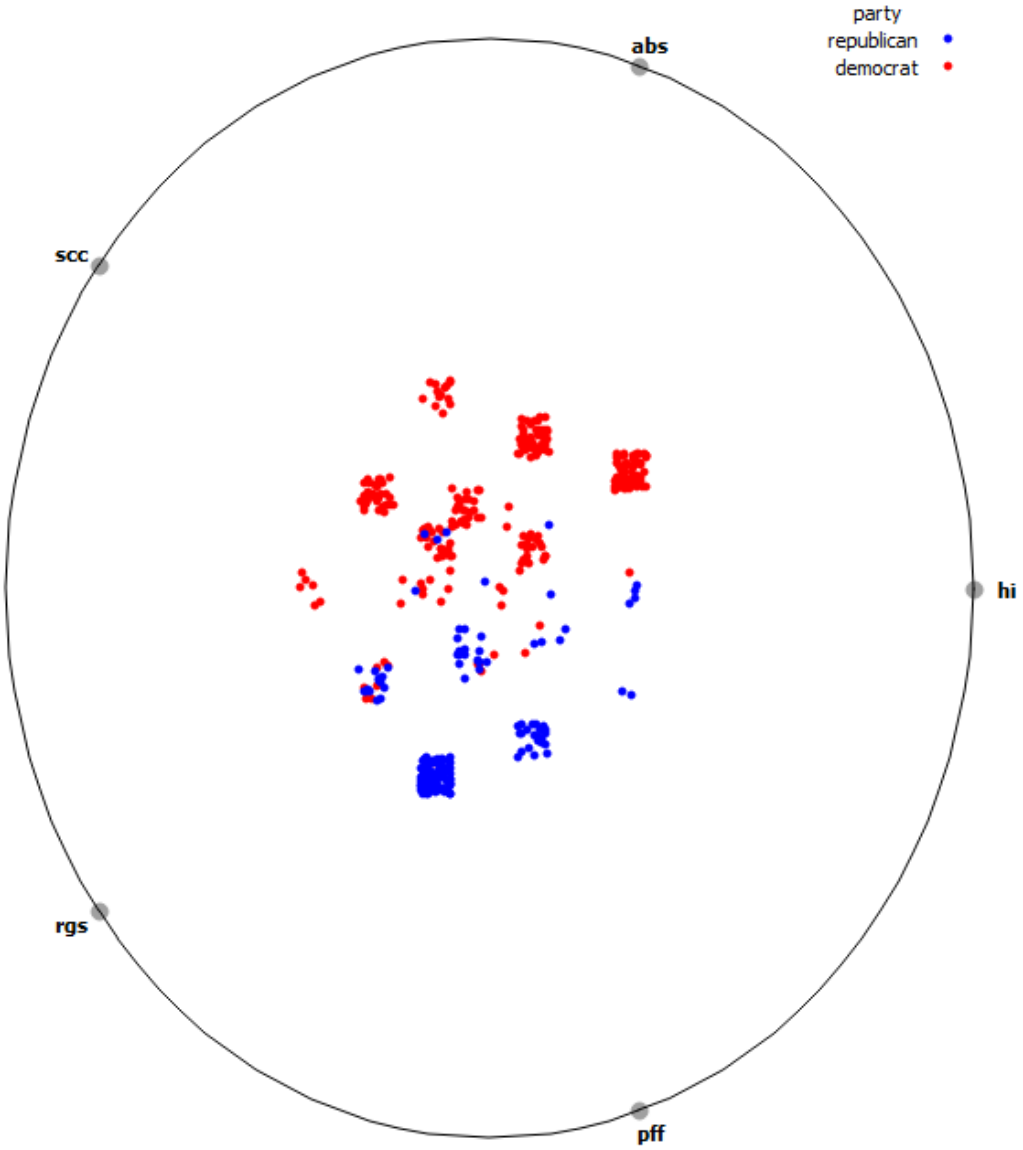

**Σχήμα 5.6. Συχνότητα ψήφων για το χαρακτηριστικό "hi". Μπλε χρώμα για τους "republicans" και κόκκινο χρώμα για τους "democrats".**

Αντίστοιχα, παρατηρώντας το ιστόγραμμα για το χαρακτηριστικό "pff" βλέπουμε ότι κάποιες μεταβλητές πολώνουν περισσότερο τα μέλη των δύο παρατάξεων.




**Σχήμα 5.7. Συχνότητα ψήφων για το χαρακτηριστικό "hi". Μπλε χρώμα για τους "republicans" και κόκκινο χρώμα για τους "democrats".**

Επιπλέον, μέσω του διαγράμματος RadViz[14] (radiate coordinal visualization) βλέπουμε ότι μπορούμε πράγματι να προβλέψουμε την κατηγορία που ανήκει το κάθε μέλος ( "republicans" ή "democrats") βασιζόμενοι στις τιμές των χαρακτηριστικών.




**Σχήμα 5.8. Διάγραμμα RadViz για 5 ανεξάρτητες μεταβλητές του παραδείγματος με τις πολιτικές παρατάξεις**

[14] Ankerst et. al (1996)

## Εφαρμογή της λογιστικής παλινδρόμησης

Αρχικά χωρίζουμε τα δεδομένα σε ένα σύνολο εκπαίδευσης (261 παρατηρήσεις) και ένα σύνολο ελέγχου (174 παρατηρήσεις). Στη συνέχεια, κατασκευάζουμε το λογιστικό μοντέλο με χρήση του συνόλου εκπαίδευσης και εντοπίζουμε το βέλτιστο μοντέλο με τη μέθοδο backward selection: το στατιστικό πρόγραμμα ξεκινά με το μοντέλο που περιέχει όλες τις μεταβλητές, και σταδιακά αφαιρεί μία μεταβλητή σε κάθε βήμα μέχρι να καταλήξει στο καλύτερα προσαρμοσμένο μοντέλο.

**Logistic Regression - party**

Dependent variable: party
Factors: hi, wpcs, abs, pff, esa, rgs, astb, anc, mxm, imm, scc, es, srs, cr, dfe, eaasa

**Estimated Regression Model (Maximum Likelihood)**

| | | *Standard* | *Estimated* |
|---|---|---|---|
| *Parameter* | *Estimate* | *Error* | *Odds Ratio* |
| CONSTANT | 4.71631 | 1.44978 | |
| abs=n | -2.56282 | 0.892901 | 0.0770869 |
| pff=n | 6.13274 | 1.19171 | 460.695 |
| scc=n | -5.5746 | 1.41409 | 0.003793 |
| dfe=n | -2.72465 | 1.12736 | 0.0655693 |

**Analysis of Deviance**

| *Source* | *Deviance* | *Df* | *P-Value* |
|---|---|---|---|
| Model | 288.321 | 4 | 0.0000 |
| Residual | 51.6429 | 256 | 1.0000 |
| Total (corr.) | 339.964 | 260 | |

Percentage of deviance explained by model = 84.8093
Adjusted percentage = 81.8678

**Likelihood Ratio Tests**

| *Factor* | *Chi-Square* | *Df* | *P-Value* |
|---|---|---|---|
| abs | 9.45695 | 1 | 0.0021 |
| pff | 72.5247 | 1 | 0.0000 |
| scc | 36.794 | 1 | 0.0000 |
| dfe | 8.68183 | 1 | 0.0032 |

Stepwise factor selection
Method: backward selection
P-to-enter: 0.05
P-to-remove: 0.05

Step 0:
16 factors in the model. 244 d.f. for error.
Percentage of deviance explained = 87.82% Adjusted percentage = 77.82%

Step 1:
Removing factor cr with P-to-remove = 0.996261
15 factors in the model. 245 d.f. for error.
Percentage of deviance explained = 87.82% Adjusted percentage = 78.41%

.
.
.

Step 12:
Removing factor mxm with P-to-remove = 0.100011
4 factors in the model. 256 d.f. for error.
Percentage of deviance explained = 84.81% Adjusted percentage = 81.87%

Final model selected.

Το βέλτιστο μοντέλο περιέχει 4 από τις 16 ανεξάρτητες μεταβλητές και είναι:

party = exp(eta)/(1+exp(eta))

όπου

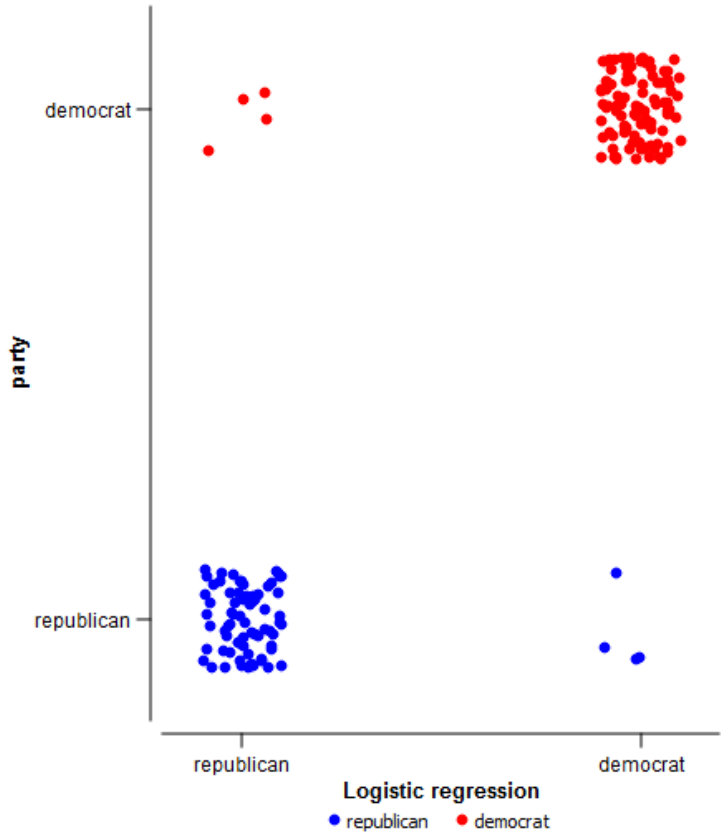
eta = 4.71631 - 2.56282*abs=n + 6.13274*pff=n - 5.5746*scc=n - 2.72465*dfe=n

Όπως παρατηρούμε από τον πίνακα "Analysis of Deviance", το μοντέλο εξηγεί το 84.8% της αποκλίνουσας συμπεριφοράς και αυτό σημαίνει ότι είναι καλά προσαρμοσμένο.

Εφαρμόζοντας, τώρα, το μοντέλο αυτό στα δεδομένα του συνόλου ελέγχου, μπορούμε να εκτιμήσουμε την προβλεπτική του ισχύ. Στο διάγραμμα, καθώς και στον πίνακα σύγχυσης που ακολουθούν, βλέπουμε ότι το μοντέλο ταξινομεί σωστά τα περισσότερα άτομα στις δύο κατηγορίες.

| **Πραγματική κατηγορία** | **Προβλεπόμενη κατηγορία** | | |
|---|---|---|---|
| | **republicans** | **democrats** | |
| **republicans** | 71 | 4 | 75 |
| **democrats** | 4 | 95 | 99 |
| | 75 | 99 | 174 |

**Πίνακας 5.4. Πίνακας σύγχυσης του λογιστικού μοντέλου για το παράδειγμα με τις πολιτικές παρατάξεις**




**Σχήμα 5.9. Προβλεπόμενη ταξινόμηση των ατόμων στις κατηγορίες "republicans" και "democrats": Το λογιστικό μοντέλο ταξινομεί σωστά τις περισσότερες παρατηρήσεις.**

## Σύγκριση με άλλους ταξινομητές

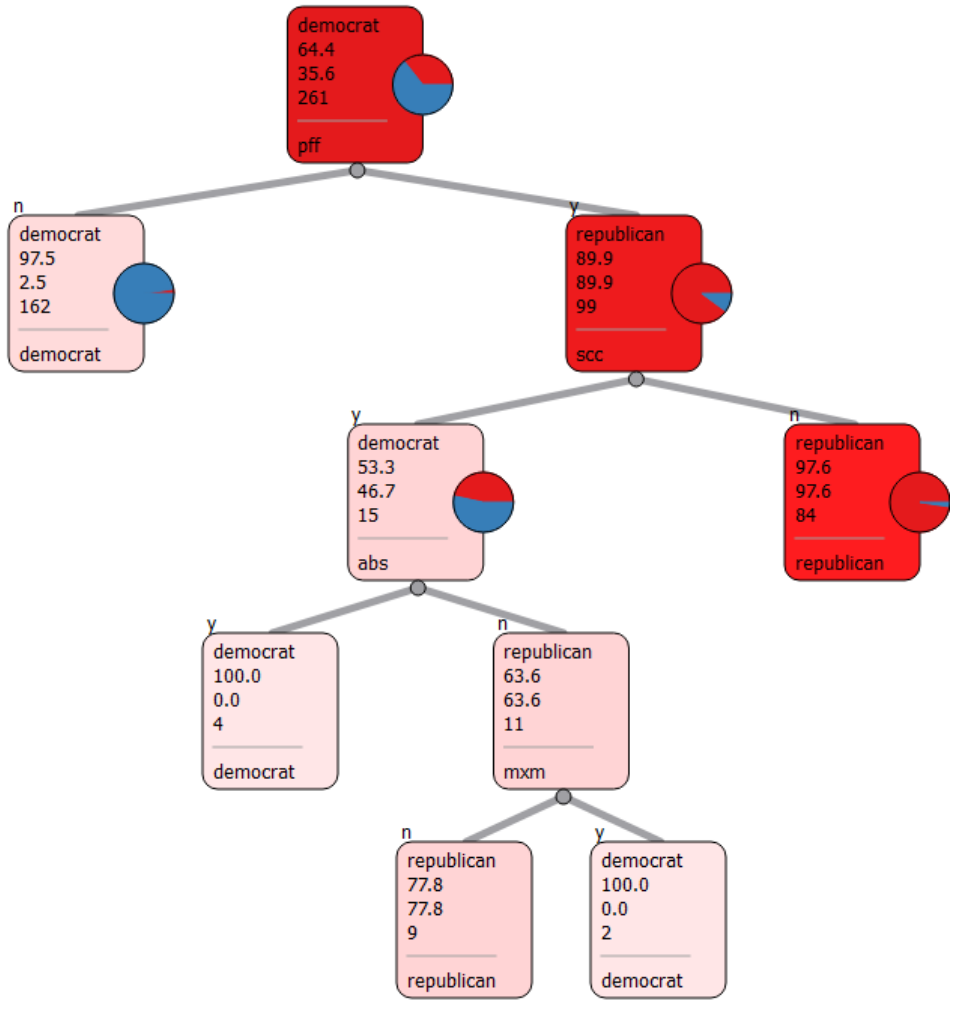
Χρησιμοποιώντας το ίδιο σύνολο εκπαίδευσης θα εφαρμόσουμε τον αφελή ταξινομητή Bayes και επιπλέον θα κατασκευάσουμε ένα δέντρο ταξινόμησης για να συγκρίνουμε την αξιοπιστία των τριών μοντέλων.

Ο πίνακας σύγχυσης για τον αφελή ταξινομητή Bayes είναι:

| Πραγματική κατηγορία | Προβλεπόμενη κατηγορία | | |
|---|---|---|---|
| | republicans | democrats | |
| republicans | 70 | 5 | 75 |
| democrats | 12 | 87 | 99 |
| | 82 | 92 | 174 |

**Πίνακας 5.5. Πίνακας σύγχυσης του αφελή ταξινομητή Bayes για το παράδειγμα με τις πολιτικές παρατάξεις**

Αντιστοίχως, το δέντρο ταξινόμησης που προκύπτει είναι:




**Σχήμα 5.10. Κλαδεμένο δέντρο ταξινόμησης για το παράδειγμα με τις πολιτικές παρατάξεις**

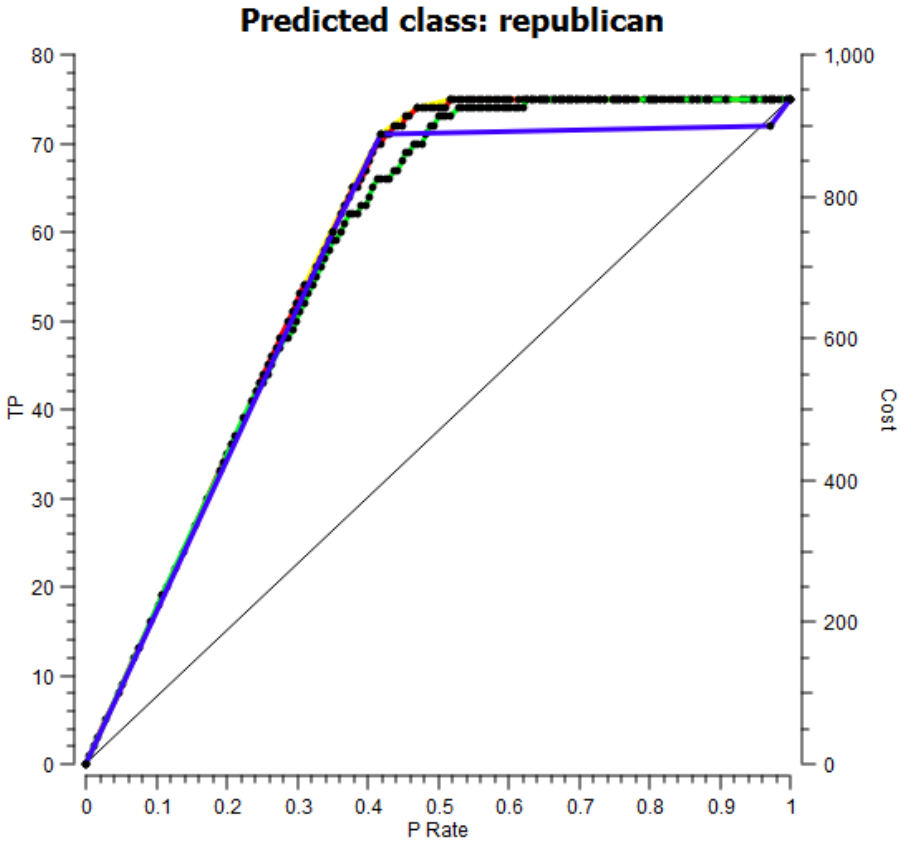
| Πραγματική κατηγορία | Προβλεπόμενη κατηγορία | | |
|---|---|---|---|
| | republicans | democrats | |
| republicans | 71 | 4 | 75 |
| democrats | 2 | 97 | 99 |
| | 73 | 101 | 174 |

**Πίνακας 5.6. Πίνακας σύγχυσης του δέντρου ταξινόμησης για το παράδειγμα με τις πολιτικές παρατάξεις**

Συγκρίνοντας, λοιπόν, τη συνολική αξιοπιστία των τριών ταξινομητών συμπεραίνουμε ότι το καλύτερο μοντέλο προκύπτει από το δέντρο ταξινόμησης.

| Μέθοδος | Classification accuracy |
|---|---|
| **Λογιστική παλινδρόμηση** | **0.954** |
| **Αφελής ταξινομητής Bayes** | **0.9023** |
| **Δέντρο ταξινόμησης** | **0.9655** |

**Πίνακας 5.7. Σύγκριση της αξιοπιστίας των τριών ταξινομητών για το παράδειγμα με τις πολιτικές παρατάξεις**




**Σχήμα 5.11. Διάγραμμα ανύψωσης: Σύγκριση των τριών ταξινομητών (μπλε χρώμα =δέντρο ταξινόμησης, κόκκινο χρώμα = λογιστική παλινδρόμηση, πράσινο χρώμα = Naive Bayes)**

# Γενικά συμπεράσματα

Η εξόρυξη δεδομένων προσφέρει πολλούς τρόπους για την ανακάλυψη κρυφών προτύπων μέσα από μεγάλες βάσεις δεδομένων. Ωστόσο, η τεχνική που θα χρησιμοποιηθεί για να πάρουμε χρήσιμες πληροφορίες πρέπει να επιλεχθεί προσεκτικά. Όλες οι τεχνικές εξόρυξης πληροφορίας από δεδομένα μπορούν να είναι τόσο αποδοτικές, όσο τους επιτρέπουν τα ίδια τα δεδομένα.

Για να προκύψουν ασφαλή και αξιόπιστα συμπεράσματα θα πρέπει να έχουμε «καλά» δεδομένα. Για τον λόγο αυτό, είναι αναγκαία η προεπεξεργασία τους με σκοπό να γίνει καθαρισμός ή μετασχηματισμός όπου αυτός θεωρείται απαραίτητος.

Αν, λοιπόν, υποθέσουμε ότι τα δεδομένα που διαθέτουμε είναι «καλά», το επόμενο βήμα είναι η επιλογή της κατάλληλης μεθόδου. Δεν υπάρχει συνταγή για το ποια μέθοδος θα αποδώσει καλύτερα˙ αυτό είναι κάτι που εξαρτάται καθαρά από τον τύπο του προβλήματος και τη φύση των δεδομένων.

Τέλος, ο εντοπισμός του βέλτιστου μοντέλου προκύπτει μέσα από πολλές επαναλήψεις και απαιτεί τη δοκιμή διαφορετικών αλγορίθμων και τεχνικών. Έτσι, τις περισσότερες φορές, ο αναλυτής χρειάζεται να συγκρίνει ή ακόμα και να συνδυάσει διαφορετικές τεχνικές για να εξάγει τα καλύτερα δυνατά αποτελέσματα.

# Βιβλιογραφία


Berry, M.J.A. and Linoff, G.S. (2011), *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management* (3rd ed.), Wiley Publishing

Bratko, I. and Niblett, T. (1986), *Learning decision rules in noisy domains*, Cambridge University Press

Bratko, I. and Cestnik, B. (1991), *On estimating probabilities in tree pruning*, Springer-Verlang

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Chapman & Hall

Demsar, J. and Stajdohar, M. (2013), *Interactive Network Exploration with Orange*, Journal of Statistical Software

Elkan, C. (2013), *Predictive analytics and data mining*, http://cseweb.ucsd.edu/~elkan/255/dm.pdf

Ghosh, A.K. (2005), *On optimal choice of k in nearest neighbor classification*, Computational Statistics and Data Analysis Journal

Han, J. and Kamber, M. (2006), *Data Mining: Concepts and Techniques* (2nd ed.), Morgan Kaufmann Publishers

Hand, D., Mannila, H. and Smyth, P. (2001), *Principles of Data Mining*, The MIT Press

Harrison, D. and Rubinfeld, D.L. (1978), *Hedonic prices and the demand for clean air*, Journal of Environmental Economics and Management

Kumar, V., Steinbach, M. and Tan, P.N. (2005), *Introduction to Data Mining*, Addison-Wesley

Leskovec, J., Rajaraman A. and Ullman J.D.(2012), *Mining of Massive Datasets*, Cambridge University Press

Maimon, O. and Rokach, L. (2010), *Data Mining and Knowledge Discovery Handbook* (2nd ed.), Springer

North, M. (2012), *Data Mining for the Masses*, The Global Text Project

Shmueli, G., Patel N.R. and Bruce P.C. (2005), *Data Mining In Excel: Lecture Notes and Cases*,http://www.researchgate.net/publication/242384346_Data_Mining_In_Excel_Lecture_Notes_and_Cases

Stanton, J. (2012), *An Introduction to Data Science*, Syracuse University https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

Witten, I.H., Frank, E. and Hall, M. (2005), *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.), Morgan Kaufmann Publishers

Zaki, M.J. and Meira, W. (2014), *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press

Zhao, Y. (2012), *R and Data Mining: Examples and Case Studies*, Elsevier

**Statistical Software**

Orange http://orange.biolab.si/

Statgraphics http://www.statgraphics.com/

XLMiner http://www.solver.com/xlminer-data-mining