ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ
ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΝΑΤΑΛΙΑ ΣΑΜΑΡΑ

«ΔΕΝΔΡΑ ΑΠΟΦΑΣΕΩΝ»

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

Δημήτριος Φουσκάκης, Καθηγητής (Επιβλέπων)
Λουλάκης Μιχαήλ, Αναπληρωτής Καθηγητής
Βόντα Φιλία, Καθηγήτρια

ΑΘΗΝΑ, Σεπτέμβριος 2022

Στην μνήμη της αγαπημένης μου μητέρας,
Άρτεμις Κοτσώλη.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή, κ. Δημήτριο Φουσκάκη για την βοήθεια και τον χρόνο που αφιέρωσε για να με καθοδηγήσει, καθώς και για την εμπιστοσύνη που μου έδειξε καθ᾽ όλη τη διάρκεια της διπλωματικής μου εργασίας.

Στην συνέχεια, θα ήθελα να ευχαριστήσω τον κ. Μιχαήλ Λουλάκη και την κ. Φιλία Βόντα που δέχθηκαν να συμμετέχουν στην τριμελή επιτροπή για την αξιολόγηση της παρούσας διπλωματικής εργασίας.

Τέλος, δεν θα μπορούσα να παραλείψω τον καταλυτικό ρόλο που έπαιξε η οικογένειά μου, αλλά και οι φίλοι μου, οι οποίοι με στήριξαν και πάντα με υποστηρίζουν σε κάθε προσπάθειά μου.

Περίληψη

Στην παρούσα διπλωματική εργασία αναλύεται η στατιστική μέθοδος της δενδρικής παλινδρόμησης και της δενδρικής ταξινόμησης, καθώς και οι αντίστοιχες επεκτάσεις αυτών. Στην συνέχεια, εφαρμόζεται στο οικονομικό πρόβλημα της υπόθεσης του κύκλου ζωής (πρόβλημα παλινδρόμησης) και στο πρόβλημα της αναζήτησης εξωπλανητών μέσω του δορυφόρου Kepler της NASA (πρόβλημα ταξινόμησης).

Οι μέθοδοι της δενδρικής παλινδρόμησης και ταξινόμησης βασίζονται στον αλγόριθμο CART, ο οποίος κατασκευάζει δένδρα παλινδρόμησης και ταξινόμησης ή, πιο γενικά, δένδρα αποφάσεων. Τα τελευταία αποτελούν στατιστικά μοντέλα πρόβλεψης, όπου σε κάθε κλαδί του δένδρου πραγματοποιούνται αποφάσεις που αφορούν τα χαρακτηριστικά (επεξηγηματικές μεταβλητές) από τα οποία εξαρτάται η μεταβλητή απόκρισης που μας ενδιαφέρει να προβλέψουμε. Η κάθε απόφαση στηρίζεται σε μία συνθήκη τμήσης ή, αλλιώς, σε ένα κριτήριο διαμέρισης, το οποίο οφείλει να είναι το βέλτιστο σε κάθε τμήση, προκειμένου η τελική πρόβλεψη να είναι όσον το δυνατόν πιο ακριβής, συνοδευόμενη από ένα χαμηλό σφάλμα. Ως επεκτάσεις (βελτιώσεις) της μεθόδου της δενδρικής παλινδρόμησης (και ταξινόμησης) εισάγονται κάποιες τεχνικές συνόλου, όπως είναι η Ενσάκιση, τα Τυχαία Δάση και η Ενίσχυση, οι οποίες στοχεύουν σε μία ακόμα πιο εύστοχη πρόβλεψη και στην δυνατότητα αυτή να παραμένει σχετικά αναλλοίωτη σε τυχόν μεταβολές στα υπάρχοντα δεδομένα. Τέλος, με αφορμή δύο στατιστικά προβλήματα, η μέθοδος της δενδρικής παλινδρόμησης συγκρίνεται με την κλασική μέθοδο της πολλαπλής γραμμικής παλινδρόμησης, όπως και η μέθοδος της δενδρικής ταξινόμησης εξετάζεται σε σχέση με την γνωστή μέθοδο της λογιστικής παλινδρόμησης. Η διαδικασία αυτή αποσκοπεί στο να αποφανθούμε, τελικά, ποια μέθοδος αποδίδει καλύτερα ανάλογα με το πρόβλημα που διαθέτουμε και ποια είναι αυτή που παρέχει τις ορθότερες, χαμηλότερου σφάλματος προβλεψεις και εκτιμήσεις για την εκάστοτε μεταβλητή ενδιαφέροντος.

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Η επιστήμη της Στατιστικής αποτελεί κλάδο των εφαρμοσμένων μαθηματικών, ο οποίος ασχολείται με την συλλογή, ανάλυση, επεξεργασία, ερμηνεία και παρουσίαση διαφόρων ειδών δεδομένων με σκοπό την εξαγωγή έγκυρων συμπερασμάτων για λήψη ορθών, τεκμηριωμένων αποφάσεων.

Τα τελευταία έτη γίνεται όλο και πιο συχνή η χρήση της έννοιας των 'Μεγάλων Δεδομένων' (Big Data), μία σχετικά καινούργια έννοια, το περιεχόμενο της οποίας, ωστόσο, φαίνεται να μετατρέπεται σε απαραίτητη και προαπαιτούμενη γνώση, κυρίως, στον επαγγελματικό τομέα. Η εν λόγω έννοια αντιπροσωπεύει τον τεράστιο όγκο δεδομένων που καλούνται να χειριστούν καθημερινά άτομα με εξειδικευμένες γνώσεις προγραμματισμού, πληροφορικής και στατιστικής, προκειμένου να εξάγουν έγκυρα συμπεράσματα και προβλέψεις που θα συμβάλλουν στην ομαλότερη και αποδοτικότερη λειτουργία μίας διαδικασίας (π.χ. μίας επιχείρησης). Για τον λόγο αυτό, εύλογη προϋπόθεση αποτελεί μία ακριβής και στοχευμένη επεξεργασία όλης αυτής της πληροφορίας.

Η επεξεργασία των δεδομένων πραγματοποιείται με τη βοήθεια στατιστικών μεθόδων, όπως είναι για παράδειγμα η πολλαπλή γραμμική παλινδρόμηση, η παλινδρόμηση Poisson και η λογιστική παλινδρόμηση. Μία εξαιρετικά εύχρηστη και χρήσιμη μέθοδος επεξεργασίας δεδομένων είναι η μέθοδος της δενδρικής παλινδρόμησης και ταξινόμησης (CART), η οποία βασίζεται στην θεωρία των δένδρων αποφάσεων. Σε περιπτώσεις όπου το πλήθος των υπό μελέτη δεδομένων είναι μεγάλο, η συγκεκριμένη μέθοδος, συνήθως, υπερτερεί έναντι των υπολοίπων, λόγω του γεγονότος ότι είναι εύκολη στη χρήση, άμεση και κατανοητή όσον αφορά την ερμηνεία. Η τεχνική CART προτιμάται πολλές φορές για προβλήματα του 'πραγματικού' κόσμου, όπου οι ερευνητές έρχονται αντιμέτωποι με μεγάλο όγκο στοιχείων και καλούνται μέσω της ταξινόμησης και αξιοποίησης αυτών να εξάγουν κάποιο συμπέρασμα ή κάποια πρόβλεψη για το μέλλον.

1.1 Προβλέψεις στην Στατιστική

Αυτό που μας ενδιαφέρει ουσιαστικά σε μία στατιστική μελέτη είναι οι μετρήσεις κάποιων χαρακτηριστικών, όπως ο ημερήσιος ρυθμός παραγωγής ενός προϊόντος, η ποιότητα ζωής ενός ασθενή, η διάρκεια ζωής μίας μπαταρίας, κλπ. Οι μετρήσεις αυτές των χαρακτηριστικών στηρίζονται στην μελέτη και την επεξεργασία κάποιων άλλων μεταβλητών.

Σε κάθε στατιστικό πρόβλημα διακρίνονται δύο τύποι μεταβλητών: οι ανεξάρτητες ή επεξηγηματικές μεταβλητές που θα συμβολίζονται ως X_p με $p \geq 1$ και η εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης που θα δηλώνεται με Y. Στόχος του προβλήματος είναι η πρόβλεψη, μέσω της βοήθειας των επεξηγηματικών μεταβλητών, των τιμών της μεταβλητής απόκρισης αν αυτή είναι ποσοτική (αν λαμβάνει μόνο αριθμητικές τιμές) ή η πρόβλεψη της κατηγορίας ή κλάσης της μεταβλητής απόκρισης αν αυτή είναι ποιοτική ή, αλλιώς, κατηγορική (αν αποτελείται αποκλειστικά από κλάσεις). Για την διάκριση των μεταβλητών σε ποσοτικές και κατηγορικές θα γίνει λόγος στο επόμενο κεφάλαιο.

Στην παρούσα φάση, αξίζει να σημειωθεί ότι η επεξηγηματική μεταβλητή X (ή οι επεξηγηματικές μεταβλητές X_p με $p \geq 1$) μπορεί να πάρει και την μορφή διανύσματος και τότε θα αναφερόμαστε σε αυτή ως τυχαίο διάνυσμα $\boldsymbol{X}$ (ή τυχαία διανύσματα $\boldsymbol{X_p}$ με $\boldsymbol{p} \geq 1$). Ένα τυχαίο διάνυσμα διάστασης $n > 1$ είναι ένα διάνυσμα αποτελούμενο από $n > 1$ τυχαίες μεταβλητές και, συνεπώς,

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_n \end{pmatrix}, \tag{1.1}$$

όπου $X_1, X_2, ..., X_n$ είναι τυχαίες μεταβλητές.
Αντίστοιχα, η παραπάνω περίπτωση μπορεί να γενικευτεί και για την περίπτωση των p επεξηγηματικών μεταβλητών X_p με $p \geq 1$, όπου είναι πιθανό να έχουμε p τυχαία διανύσματα διάστασης $n > 1$, αντί για p τυχαίες μεταβλητές σε κάποια προβλήματα. Τα τυχαία αυτά διανύσματα δηλώνονται με παρόμοιο τρόπο ως εξής:

$$\left[\boldsymbol{X_1} = \begin{pmatrix} X_{11} \\ X_{21} \\ . \\ . \\ . \\ X_{n1} \end{pmatrix}, \boldsymbol{X_2} = \begin{pmatrix} X_{12} \\ X_{22} \\ . \\ . \\ . \\ X_{n2} \end{pmatrix}, \ldots, \boldsymbol{X_p} = \begin{pmatrix} X_{1p} \\ X_{2p} \\ . \\ . \\ . \\ X_{np} \end{pmatrix} \right], \tag{1.2}$$

με $p \geq 1$.

Γενικά, σε μία στατιστική μελέτη, πρωταρχικό βήμα αποτελεί η συλλογή παρατηρήσεων για μία τυχαία μεταβλητή Y που μας ενδιαφέρει (μεταβλητή απόκρισης). Έστω $y_1, y_2, ..., y_n$ οι τιμές των παρατηρήσεων του δείγματος για την μεταβλητή Y που μελετάμε.
Αν η μεταβλητή απόκρισης Y λαμβάνει αριθμητικές τιμές, είναι, δηλαδή, ποσοτική τυχαία μεταβλητή, τότε για την πρόβλεψη της Y, ένας προφανής συλλογισμός θα ήταν ο υπολογισμός της μέσης τιμής των παρατηρήσεων αυτής. Για την μέση τιμή μ μίας τυχαίας μεταβλητής Y, η καλύτερη εκτίμηση είναι η δειγματική μέση τιμή ή μέσος όρος της Y που συμβολίζεται με $\bar{Y}$ και είναι το 'κέντρο ισορροπίας' των δεδομένων [5]. Η τιμή της δειγματικής μέσης τιμής ορίζεται ως

$$\bar{y} = \frac{y_1 + y_2 + ... + y_n}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{1.3}$$

Για ποιοτικά δεδομένα, δηλαδή όταν η μεταβλητή απόκρισης Y είναι κατηγορική, η πρόβλεψη αυτής πραγματοποιείται μέσα από τον υπολογισμό της συχνότητας εμφάνισης f_i στις παρατηρήσεις $y_1, y_2, ..., y_n$ $(1 \leq i \leq n)$ της κάθε διακεκριμένης τιμής (κατηγορίας) a_i για k διακεκριμένες τιμές $a_1, a_2, ..., a_k$ [5]. Η σχετική συχνότητα εμφάνισης ή, αλλιώς, το ποσοστό p_i ορίζεται ως

$$p_i = \frac{f_i}{n}. \tag{1.4}$$

Γενικά, όταν η μεταβλητή που πρόκειται να εκτιμηθεί είναι ποσοτική, συνηθίζεται να χρησιμοποιείται η τεχνική της δενδρικής παλινδρόμησης, ενώ, όταν πρόκειται για πρόβλεψη κάποιας κατηγορικής μεταβλητής, η τεχνική που επιλέγεται είναι αυτή της δενδρικής ταξινόμησης. Ο διαχωρισμός και η περιγραφή των δύο τεχνικών θα γίνουν περισσότερο κατανοητά στα επόμενα κεφάλαια.

1.2 Εισαγωγή στα δένδρα αποφάσεων

Η θεωρία των δένδρων αποφάσεων είναι μία από τις πιο δημοφιλείς θεωρίες στον κλάδο της στατιστκής επιστήμης σε παγκόσμιο επίπεδο. Η μεγάλη της φήμη οφείλεται, κυρίως, στο γεγονός ότι σε περιπτώσεις μεγάλων όγκων δεδομένων, φαινόμενο αρκετά συχνό στην σημερινή εποχή της πληροφορίας και της τεχνολογίας, όπου ερευνητές και στατιστικοί έχουν να διαχειριστούν πολλή πληροφορία σε μικρό χρονικό διάστημα, η συγκεκριμένη θεωρία έχει ξεχωρίσει ανάμεσα σε άλλες, χάρη στην εύκολη και γρήγορη υλοποίηση και ερμηνεία αυτής, αλλά και λόγω της σύγχρονης 'ευελιξίας' που διαθέτει. Το τελευταίο χαρακτηριστικό αναδιατυπώνεται στο γεγονός ότι υπάρχουν διαθέσιμες, πλέον, διάφορες

επεκτάσεις και βελτιώσεις με αποτέλεσμα η αρχική θεωρία των δένδρων αποφάσεων να μπορεί να προσαρμόζεται στο εκάστοτε στατιστικό πρόβλημα για πιο ακριβή και αξιόπιστα συμπεράσματα και προβλέψεις που είναι, άλλωστε, και ο απώτερος σκοπός μίας στατιστικής έρευνας.

Τα δένδρα αποφάσεων είναι μία μη παραμετρική, αποτελεσματική μέθοδος, η οποία μπορεί να εφαρμοστεί τόσο σε προβλήματα παλινδρόμησης όσο και σε προβλήματα ταξινόμησης. Έστω, αρχικά, ότι η μεταβλητή απόκρισης του προβλήματος που μας ενδιαφέρει είναι η Y και ότι οι επεξηγηματικές μεταβλητές που έχουμε στην διάθεσή μας είναι οι $X_1, X_2, ..., X_p$, όπου $p \geq 1$. Ορίζουμε ως χώρο πρόβλεψης τον χώρο που αποτελείται από το σύνολο των πιθανών τιμών για τις επεξηγηματικές μεταβλητές $X_1, X_2, ..., X_p$, όπου $p \geq 1$. Ένα δένδρο αποφάσεων κατασκευάζεται από το δείγμα εκπαίδευσης (training sample) και, ύστερα, με ένα ανεξάρτητο δείγμα ελέγχου (testing sample) ελέγχεται η ακρίβεια του δένδρου.

↪ Δένδρα παλινδρόμησης:

Η διαδικασία πρόβλεψης μίας μεταβλητής απόκρισης Y, χρησιμοποιώντας ένα δένδρο παλινδρόμησης, ακολουθεί την εξής βασική ιδέα [7]:

(i) Διαίρεση του χώρου πρόβλεψης σε J διακεκριμένες, μη επικαλυπτόμενες περιοχές $R_1, R_2, ..., R_J$.

(ii) Για κάθε παρατήρηση από το δείγμα ελέγχου που ανήκει στην περιοχή R_j με $1 \leq j \leq J$, η τελική πρόβλεψη για την μεταβλητή απόκρισης Y θα είναι κάθε φορά η μέση τιμή όλων των παρατηρήσεων εκπαίδευσης (παρατηρήσεις που προέρχονται από το δείγμα εκπαίδευσης) που βρίσκονται εντός αυτής της περιοχής.

Η διαδικασία που μόλις περιγράφτηκε φαίνεται συνοπτικά με ένα απλό παράδειγμα: Για οποιαδήποτε j και s, υποθέτουμε ότι ο χώρος πρόβλεψης διαιρείται σε δύο περιοχές της μορφής

$$R_1 = R_1(j, s) = \{X | X_j < s\}, \; R_2 = R_2(j, s) = \{X | X_j \geq s\}. \tag{1.5}$$

Έστω ότι η μέση τιμή των παρατηρήσεων εκπαίδευσης στην περιοχή R_1 είναι 20 και στην περιοχή R_2 είναι 25. Τότε, για δοσμένη παρατήρηση $X = x$ από το δείγμα ελέγχου, αν ισχύει ότι $x \in R_1$, η πρόβλεψη για την μεταβλητή απόκρισης Y θα είναι 20. Αντίστοιχα, αν $x \in R_2$, τότε η μεταβλητή απόκρισης Y προβλέπεται να παίρνει την τιμή 25. Τα κριτήρια με τα οποία γίνεται η διαίρεση των παρατηρήσεων εκπαίδευσης των $X_1, X_2, ..., X_p$, όπου $p \geq 1$, σε περιοχές R_j με $1 \leq j \leq J$, θα αναφερθούν λεπτομερώς στο επόμενο κεφάλαιο.

↪ Δένδρα ταξινόμησης:

Η διαδικασία πρόβλεψης μίας μεταβλητής απόκρισης Y, μέσω ενός δένδρου ταξινόμησης, είναι παρόμοια με αυτήν στα δένδρα παλινδρόμησης. Η μόνη διαφορά

έγκειται στο γεγονός ότι, πλέον, ως πρόβλεψη της Y, θεωρείται να είναι εκείνη η κλάση που υπερτερεί ανάμεσα στις παρατηρήσεις του δείγματος εκπαίδευσης που ανήκουν στην περιοχή R_j, όπου $1 \leq j \leq J$ και όχι η μέση τιμή αυτών, όπως ίσχυε στα δένδρα παλινδρόμησης. Ακόμα, απαιτείται ιδιαίτερη προσοχή κατά την διαίρεση του χώρου πρόβλεψης, καθώς τα κριτήρια τμήσης διαφέρουν από αυτά των δένδρων παλινδρόμησης, όπως θα διαπιστωθεί στο κεφάλαιο που ακολουθεί.

Τα δένδρα αποφάσεων αποτελούν τη βάση για την ευρείας χρήσης μέθοδο των 'Τυχαίων Δασών' (Random Forests) που είναι ένας από τους ισχυρότερους διαθέσιμους αλγορίθμους του Machine Learning σήμερα. Ακόμα, η μέθοδος των δένδρων αποφάσεων μπορεί να εφαρμοστεί σχεδόν σε όλα τα είδη προβλημάτων και είναι ικανή να χειρίζεται μικρά, αλλά και μεγάλα σε μέγεθος μελετούμενα δείγματα. Ένα άλλο σημαντικό πλεονέκτημα αποτελεί το γεγονός ότι με τα δένδρα αποφάσεων δεν είναι απαραίτητη η κατασκευή εικονικών μεταβλητών ή, αλλιώς, ψευδομεταβλητών (dummy variables) για την πρόβλεψη μίας κατηγορικής μεταβλητής Y, όπως θα χρειαζόταν στην (πολλαπλή) γραμμική παλινδρόμηση για παράδειγμα, γεγονός που πιστοποιεί την απλότητα της δενδρικής ταξινόμησης.

Επιπρόσθετα, όπως αναφέρθηκε προηγουμένως, υπάρχουν ορισμένες επεκτάσεις της μεθόδου της δενδρικής παλινδρόμησης και ταξινόμησης με απώτερο σκοπό τη βέλτιστη πρόβλεψη (σε ποσοτικές εξαρτημένες μεταβλητές) ή τη βέλτιστη ταξινόμηση (σε κατηγορικές εξαρτημένες μεταβλητές) τις οποίες θα αναλύσουμε περαιτέρω στο Κεφάλαιο 3.

Κεφάλαιο 2

Δένδρα Αποφάσεων, Δενδρική Παλινδρόμηση και Ταξινόμηση

Στο κεφάλαιο αυτό εισάγονται οι βασικοί ορισμοί των διάφορων ειδών μεταβλητών που εμφανίζονται σε εφαρμογές στατιστικής, οι οποίοι είναι απαραίτητοι για την κατανόηση του υποβάθρου και των ζητούμενων του εκάστοτε προβλήματος, έτσι ώστε να είμαστε σε θέση να αποφανθούμε ποια στατιστική μεθοδολογία απαιτείται κάθε φορά για την επίλυση αυτού. Επίσης, εισάγεται η έννοια της δενδρικής παλινδρόμησης και της δενδρικής ταξινόμησης ανάλογα με το είδος κάθε φορά της μεταβλητής απόκρισης του στατιστικού προβλήματος. Επιπροσθέτως, γίνεται αναφορά μιας χρήσιμης και αρκετά συνηθισμένης μεθόδου επεξεργασίας δεδομένων, η οποία βασίζεται στον λεγόμενο αλγόριθμο CART που θα αναλυθεί λεπτομερώς στο επόμενο κεφάλαιο. Η ύπαρξη μεγάλου αριθμού παρατηρήσεων και περισσότερων από μίας επεξηγηματικών μεταβλητών σε ένα πρόβλημα στατιστικής καθιστούν αναγκαίο τον προσδιορισμό κάποιων κριτηρίων διαμέρισης, προκειμένου να ταξινομηθούν οι μελετούμενες παρατηρήσεις σε ομάδες για μία απλούστερη και πιο αποτελεσματική διαχείριση των δεδομένων. Τα κριτήρια αυτά διαφέρουν ανάλογα με το είδος του δενδρικού προβλήματος (πρόβλημα δενδρικής παλινδρόμησης ή δενδρικής ταξινόμησης). Με αυτό τον τρόπο, η τελική πρόβλεψη είναι πιο ακριβής, με μικρότερα σφάλματα και λιγότερο κόστος.

2.1 Είδη Μεταβλητών

Σε κάθε στατιστικό πρόβλημα, η εξαρτημένη μεταβλητή ή, αλλιώς, η μεταβλητή απόκρισης που θέλουμε να εκτιμήσουμε βάσει κάποιων ανεξάρτητων ή, αλλιώς, επεξηγηματικών μεταβλητών μπορεί να είναι ποσοτική (numerical) ή κατηγορική (categorical). **Ποσοτική** ονομάζεται μία μεταβλητή που παίρνει μόνο

ποσοτικές τιμές, εκφράζει, δηλαδή, ποσότητα, όπως για παράδειγμα το βάρος ενός ανθρώπου ή ο αριθμός των ζώων υπό εξαφάνιση μίας ηπείρου, ενώ **κατηγορική** λέγεται μία μεταβλητή, η οποία εκφράζει κατηγορία ή τύπο, όπως είναι το φύλο ή ο τύπος μαλλιών ενός ατόμου (κοντά, μακριά).

Οι ποσοτικές μεταβλητές διακρίνονται σε [1]:

- Συνεχείς (Continuous): Είναι οι μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή εντός ενός συνεχούς διαστήματος, όπως για παράδειγμα το βάρος, το ύψος, ο μηνιαίος μισθός, κλπ.

- Διακριτές (Discrete): Είναι οι μεταβλητές που παίρνουν μόνο διακεκριμένες τιμές από ένα πεπερασμένο ή το πολύ αριθμήσιμο σύνολο, όπως για παράδειγμα ο αριθμός των φοιτητών σε κάποιο τμήμα, ο αριθμός των παιδιών μίας οικογένειας, κλπ.

Οι κατηγορικές μεταβλητές διακρίνονται σε [1]:

- Ονομαστικές (Nominal): Οι κατηγορίες της μεταβλητής δεν εμφανίζουν κάποια διάταξη, όπως για παράδειγμα το φύλο, ο τύπος αίματος, το χρώμα ματιών, κλπ.

- Διάταξης (Ordinal): Οι κατηγορίες της μεταβλητής εμφανίζουν διάταξη, όπως είναι το επίπεδο εκπαίδευσης (δημοτικό, γυμνάσιο, λύκειο, πανεπιστήμιο), η επίδοση του μαθητή σε κάποιο διαγώνισμα (αδύναμη, μέτρια, καλή, άριστη), κλπ.

Ο σωστός διαχωρισμός των μεταβλητών ή η εισαγωγή και αξιοποίηση των κατάλληλων μεταβλητών σε ένα πρόβλημα στατιστικής αποτελούν βασικά εργαλεία για την εξαγωγή έγκυρων και επιστημονικά ορθών αποτελεσμάτων και συμπερασμάτων και ανήκουν στις βασικές γνώσεις που μαθαίνει κάποιος όταν εισέρχεται στον συγκεκριμένο επιστημονικό κλάδο.

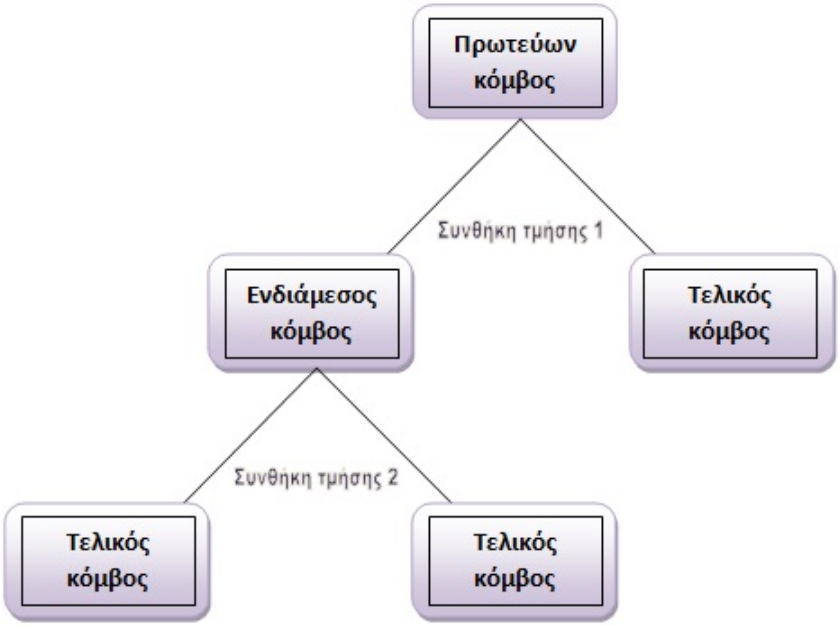
Μέσω των στατιστικών μεθόδων γίνεται ανάλυση και επεξεργασία των δεδομένων που έχουμε στην διάθεσή μας κάθε φορά, προκειμένου να καταλήξουμε σε κάποιο συμπέρασμα για μία βραχυπρόθεσμη κατάσταση (π.χ. η πρόβλεψη του καιρού τις επόμενες ημέρες) ή για ένα μακρυπρόθεσμο γεγονός (π.χ. τα κέρδη μίας εταιρίας τον επόμενο χρόνο).

2.2 Δένδρα Παλινδρόμησης και Ταξινόμησης

Η μέθοδος της **δενδρικής παλινδρόμησης** και της **δενδρικής ταξινόμησης** είναι ιδιαίτερα διαδεδομένη όσον αφορά την διαχείριση και μελέτη

πολύπλοκων και πολυάριθμων σειρών δεδομένων και βασίζεται στον αλγόριθμο CART που δημιουργήθηκε από τους Breimen et al. (1984), τους Loh and Vanichestakul (1988) και Loh and Shih (1997). Η συγκεκριμένη τεχνική κατασκευάζει δέντρα, τα λεγόμενα δένδρα αποφάσεων (decision trees), υπό την μορφή διακλαδώσεων. Τα τελευταία αποτελούνται από τον πρωτεύοντα κόμβο ή, αλλιώς, την ρίζα (root) που είναι ο κόμβος από τον οποίο ξεκινά η πρώτη διακλάδωση, από τους ενδιάμεσους κόμβους ή, απλά, κόμβους (internal nodes) και τους τελικούς κόμβους ή φύλλα (terminal nodes or leaves), οι οποίοι δεν μπορούν να τμηθούν περαιτέρω, καθώς η τμήση αυτών δεν συνεισφέρει επιπλέον στην ακρίβεια της πρόβλεψης. Κάθε διακλάδωση βασίζεται σε μία συνθήκη τμήσης ανάμεσα στις ανεξάρτητες μεταβλητές. Εκείνες που την ικανοποιούν κατατάσσονται αριστερά της διαμέρισης, ενώ οι υπόλοιπες ανήκουν στο δεξί τμήμα αυτής. Με άλλα λόγια, η διάταξη αυτή των μεταβλητών στηρίζεται στην σχέση 'αν-τότε' για την ταξινόμηση των παρατηρήσεών τους σε ομάδες.

Στο Σχήμα 2.1 απεικονίζεται ένα απλό δείγμα ενός δένδρου αποφάσεων, το οποίο αποτελείται από έναν μόνο ενδιάμεσο κόμβο και τρεις τελικούς κόμβους:



Σχήμα 2.1: Παράδειγμα ενός δένδρου αποφάσεων.

► Αν επιθυμούμε να προβλέψουμε τις τιμές μίας εξαρτημένης ποσοστικής μεταβλητής, χρησιμοποιώντας μία ή περισσότερες ανεξάρτητες μεταβλητές, οι οποίες μπορεί να είναι ποσοτικές ή και κατηγορικές, τότε εφαρμόζουμε δενδρική

παλινδρόμηση. Σ᾽ αυτή την περίπτωση, κάθε κόμβος θα αντιπροσωπεύει μία αριθμητική τιμή.

► Αν, από την άλλη, στόχος μας είναι να κατατάξουμε μία εξαρτημένη κατηγορική μεταβλητή σε κάποια κατηγορία, μέσω μίας ή περισσοτέρων ανεξάρτητων μεταβλητών, ποσοτικών ή και κατηγορικών, τότε χρησιμοποιούμε δενδρική ταξινόμηση.

Γενικά, η ιδέα της κατασκευής δένδρων παλινδρόμησης ή ταξινόμησης είναι η εξής:

Αρχικά, κατασκευάζουμε ένα δένδρο μεγάλου μεγέθους με τη βοήθεια της μεθόδου του Αναδρομικού Δυαδικού Διαχωρισμού (Recursive Binary Splitting) και αναζητούμε την καλύτερη δυνατή διαμέριση των παρατηρήσεων εκπαίδευσης. Έπειτα, γίνεται προέκταση του δένδρου έως ότου όλοι οι τελικοί κόμβοι να έχουν το πολύ n παρατηρήσεις ή μέχρι όλα τα μέλη κάποιου κόμβου να δίνουν (σχεδόν) τις ίδιες εκτιμήσεις. Το n επιλέγεται από τον χρήστη εκ των προτέρων. Με τον όρο 'καλύτερη διαμέριση' εννοούμε εκείνη την διαμέριση που θα αποφέρει την μεγαλύτερη μείωση στο λεγόμενο μέτρο νοθείας των παρατηρήσεων, το οποίο θα ορίσουμε παρακάτω. Στην συνέχεια, μέσω του Κλαδέματος του Δένδρου (Tree Pruning) δημιουργούμε μία (εμφυτευμένη στο αρχικό δένδρο) ακολουθία (υπο)δένδρων, μειώνοντας μ᾽ αυτόν τον τρόπο την πολυπλοκότητα του αρχικού, ολόκληρου δένδρου. Η τελευταία τακτική αποσκοπεί σε μία βελτιωμένη, πιο ακριβή πρόβλεψη ή ταξινόμηση της μεταβλητής απόκρισης, ενώ, παράλληλα, αποφεύγεται η εμφάνιση πιθανών προβλημάτων 'υπερπροσαρμογής' (overfitting), δηλαδή προβλήματα όπου η απόδοση του μοντέλου, όταν αυτό εφαρμόζεται σε νέα δεδομένα από κάποιο δείγμα ελέγχου, δεν είναι τόσο καλή.

Τις τεχνικές του Αναδρομικού Δυαδικού Διαχωρισμού και του Κλαδέματος θα τις δούμε αναλυτικά στο επόμενο κεφάλαιο.

Ένα πρόβλημα ονομάζεται δυαδικό πρόβλημα ταξινόμησης όταν η συνθήκη τμήσης που αφορά ένα δυαδικό χαρακτηριστικό, δηλαδή ένα χαρακτηριστικό που παίρνει μόνο δύο τιμές ή έχει μόνο δύο πιθανές κλάσεις ταξινόμησης, μπορεί να προκαλέσει δύο πιθανά αποτελέσματα. Ένα παράδειγμα τέτοιου είδους συνθήκης τμήσης αποτελεί το φύλο, έχοντας ως πιθανά αποτελέσματα 'αρσενικό', 'θηλυκό'.

Παραπάνω αναφέραμε ότι απαραίτητη προϋπόθεση στην διαδικασία διαχωρισμού των παρατηρήσεων είναι η εύρεση της βέλτιστης διαμέρισης σε κάθε κόμβο

με την έννοια της πιο αποδοτικής τμήσης των δεδομένων σε επίπεδο ακρίβειας, ώστε τα αποτελέσματα να είναι όσον το δυνατόν πιο έγκυρα και αληθή.

Ωστόσο, σε περίπτωση ύπαρξης πολλών ανεξάρτητων μεταβλητών, είναι απαραίτητη και η εύρεση, επιπλέον, της καταλληλότερης ανεξάρτητης μεταβλητής που θα χρησιμοποιείται για την τμήση του εκάστοτε κόμβου, μία μεταβλητή που θα υπερτερεί έναντι των υπολοίπων έχοντας το μικρότερο κόστος.

Με ποια κριτήρια, λοιπόν, διαχωρίζουμε τα δεδομένα κατά την κατασκευή ενός δένδρου αποφάσεων και πώς επιλέγουμε ποια ανεξάρτητη μεταβλητή να χρησιμοποιήσουμε για μία συνθήκη τμήσης όταν έχουμε τουλάχιστον δύο επεξηγηματικές μεταβλητές στην διάθεσή μας;

2.2.1 Κριτήρια διαμέρισης

Ο προσδιορισμός των κριτηρίων διαμέρισης των δεδομένων σε ένα στατιστικό πρόβλημα αποτελεί πρωταρχικό και βασικό βήμα για την κατασκευή ενός δένδρου αποφάσεων. Το ζητούμενο είναι μία ακριβής πρόβλεψη, μία πρόβλεψη με όσο το δυνατόν μικρότερο κόστος, δηλαδή μικρότερο δυνατό ποσοστό λανθασμένης ταξινόμησης των παρατηρήσεων στην περίπτωση της δενδρικής ταξινόμησης ή ελάχιστη διακύμανση στην περίπτωση της δενδρικής παλινδρόμησης.

Εν συνεχεία, με βάση τα κριτήρια διαμέρισης γίνεται η επιλογή των επεξηγηματικών μεταβλητών, ανάλογων για τις τμήσεις. Σε κάθε κόμβο, ο διαχωρισμός εκτελείται σύμφωνα με το εκάστοτε χαρακτηριστικό (μεταβλητή). Ως **μεταβλητή τμήσης** ορίζεται εκείνη η μεταβλητή που τελικά επικρατεί για τη διαμόρφωση της κατάλληλης συνθήκης τμήσης, δηλαδή η μεταβλητή που θα προσφέρει τη μεγαλύτερη μείωση στο μέτρο νοθείας των παρατηρήσεων. Η επιλογή της μεταβλητής τμήσης γίνεται κάθε φορά με την βοήθεια των κριτηρίων διαμέρισης, τα οποία θα αναλυθούν στην συνέχεια.

Το 'μέτρο νοθείας' είναι, γενικά, μία συνάρτηση με τη βοήθεια της οποίας μετρείται η ακρίβεια της πρόβλεψης σε κάθε τελικό κόμβο και δείχνει κατά πόσο οι παρατηρήσεις είναι ομοιογενείς σ᾽ αυτό τον κόμβο [2]. Με άλλα λόγια, δείχνει αν οι παρατηρήσεις στους τελικούς κόμβους ανήκουν σε μία μοναδική κλάση. Αν συμβαίνει αυτό, τότε λέμε ότι η νοθεία ή, αλλιώς, η 'ακαθαρσία' του συγκεκριμένου κόμβου είναι μικρή, η ομοιογένεια πολύ μεγάλη και, συνεπώς, προκύπτει μία βέλτιστη πρόβλεψη.

Στην δενδρική παλινδρόμηση, ως μέτρο νοθείας χρησιμοποιείται το άθροισμα των τετραγώνων των αποκλίσεων της εξαρτημένης μεταβλητής από τον δειγματικό μέσο των παρατηρήσεων εκπαίδευσης στον τελικό κόμβο ή, αλλιώς, **το**

άθροισμα των τετραγώνων των υπολοίπων (SSR). Συγκεκριμένα, έστω ότι έχουμε J τελικούς κόμβους (φύλλα) σε ένα δένδρο παλινδρόμησης με $J \geq 1$ και ότι ο j-οστός τελικός κόμβος συμβολίζεται με R_j. Τότε το άθροισμα τετραγώνων των υπολοίπων δίνεται από τον τύπο

$$SSR = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{2.1}$$

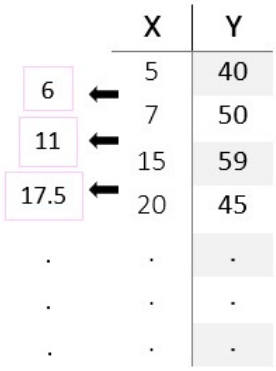
όπου $\hat{y}_{R_j}$ είναι ο δειγματικός μέσος των παρατηρήσεων εκπαίδευσης που ανήκουν στον j-οστό (τελικό) κόμβο, δηλαδή στην περιοχή R_j και είναι ίσος με

$$\hat{y}_{R_j} = \frac{1}{n_j} \sum_{x_i \in R_j} y_i, \tag{2.2}$$

δεδομένου ότι n_j αποτελεί το πλήθος των παρατηρήσεων εκπαίδευσης στον j-οστό τελικό κόμβο.

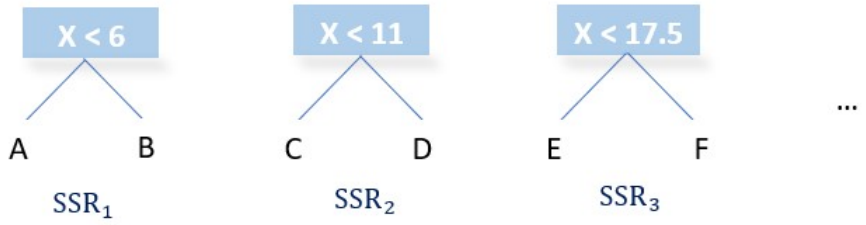
Στην πράξη, όσο πιο χαμηλές τιμές λαμβάνει το SSR σε ένα δένδρο παλινδρόμησης, τόσο πιο ακριβής θα είναι και η τελική εκτίμηση της εξαρτημένης μεταβλητής στο μοντέλο.

Παράδειγμα 2.2.1. Έστω ότι ενδιαφερόμαστε να κατασκευάσουμε ένα δένδρο παλινδρόμησης, έτσι ώστε να εκτιμήσουμε την τιμή της ποσοτικής τυχαίας μεταβλητής Y μέσω της βοήθειας της ανεξάρτητης μεταβλητής X. Ενδεικτικά, κάποιες από τις τιμές αυτών φαίνονται στον πίνακα του Σχήματος 2.2. Πρώτο βήμα στην ανάπτυξη ενός δένδρου παλινδρόμησης είναι η ταξινόμηση των παρατηρήσεων της X κατά αύξουσα σειρά και, έπειτα, η εύρεση της μέσης τιμής κάθε 2 γειτονικών παρατηρήσεων:



Σχήμα 2.2: Δεδομένα του παραδείγματος 2.2.1.

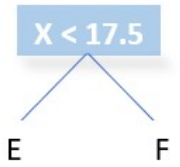
Στην συνέχεια, χρησιμοποιούμε την κάθε μέση τιμή ως όριο (threshold) για την συνθήκη τμήσης κάθε φορά των τιμών της X. Τα A, B, C, D, E, F, ... στο



Σχήμα 2.3: Πιθανά όρια για την συνθήκη τμήσης της ρίζας του δένδρου.

Σχήμα 2.3 εκφράζουν τις αντίστοιχες μέσες τιμές των παρατηρήσεων της Y που ανήκουν στην εκάστοτε (υπό)περιοχή. Για κάθε μία περίπτωση υπολογίζεται το SSR και επιλέγεται εκείνο το όριο που δίνει το χαμηλότερο SSR ως όριο για την πρώτη τμήση (ρίζα του δένδρου).

Έστω ότι η βέλτιστη ρίζα του δένδρου είναι αυτή του Σχήματος 2.4.



Σχήμα 2.4: Ρίζα του δένδρου παλινδρόμησης για το παράδειγμα 2.2.1.

Η διαδικασία που περιγράφτηκε επαναλαμβάνεται, έτσι ώστε να βρεθεί το επόμε-

νο όριο που δίνει το χαμηλότερο SSR και η επόμενη τμήση να βασίζεται σε αυτό. Ωστόσο, για να αποφύγουμε την περίπτωση όπου το δένδρο μας προσαρμόζεται πολύ καλά στα εν λόγω δεδομένα και όχι τόσο καλά σε νέα δεδομένα, συνήθως, χωρίζουμε τις παρατηρήσεις μόνο όταν αυτές υπερβαίνουν κάποιο ελάχιστο αριθμό που καθορίζει ο χρήστης. Για παράδειγμα, έστω ότι αυτός ο αριθμός είναι το 10. Αν η περιοχή E περιέχει πάνω από 10 παρατηρήσεις (της μεταβλητής Y), μπορούμε να τις διαχωρίσουμε σε 2 ομάδες, ακολουθώντας την ίδια διαδικασία με πριν. Διαφορετικά, η E μετατρέπεται σε τελικό κόμβο (φύλλο).

Έστω, τώρα, ότι οι ανεξάρτητες μεταβλητές είναι X_p με $p > 1$. Για την κατασκευή του νέου δένδρου παλινδρόμησης επαναλαμβάνουμε την ίδια διαδικασία με πριν για κάθε ένα από τα διαθέσιμα X_p. Πιο αναλυτικά, γίνεται δοκιμή διαφορετικών ορίων (thresholds) για την κάθε μεταβλητή X_p και υπολογίζεται το SSR σε κάθε περίπτωση. Έπειτα, για κάθε ανεξάρτητη μεταβλητή X_p επιλέγεται το όριο που δίνει το χαμηλότερο SSR και θεωρείται ως υποψήφιο για την ρίζα του δένδρου.

Αν κάποια μεταβλητή από τις X_p είναι κατηγορική τυχαία μεταβλητή, έστω δύο κλάσεων, τότε έχουμε μόνο ένα πιθανό όριο και γίνεται χρήση αυτού του ορίου για τον υπολογισμό του SSR (και αυτό μετατρέπεται ως υποψήφιο για την ρίζα).

Έπειτα, συγκρίνουμε όλα τα SSR για κάθε υποψήφιο όριο για την ρίζα, όπως προέκυψαν από κάθε ένα X_p ($p > 1$) και επιλέγεται τελικά εκείνο με το μικρότερο SSR. Αναπτύσσουμε το δένδρο με αυτόν τον τρόπο, συγκρίνοντας κάθε φορά το χαμηλότερο SSR από κάθε X_p ($p > 1$) με τα υπόλοιπα και το όριο στο οποίο αντιστοιχεί το μικρότερο SSR επιλέγεται για την τμήση. Όταν ένα φύλλο περιλαμβάνει λιγότερες από κάποιο ελάχιστο αριθμό παρατηρήσεις, η διαδικασία διαχωρισμού των παρατηρήσεων διακόπτεται.

Στην δενδρική ταξινόμηση, τα συνηθέστερα μέτρα νοθείας αναφέρονται παρακάτω:

✓ Το **Ποσοστό του Σφάλματος Ταξινόμησης (Classification Error Rate)**. Είναι μία εναλλακτική του SSR στην δενδρική ταξινόμηση και εκφράζει το ποσοστό των παρατηρήσεων εκπαίδευσης της εξαρτημένης μεταβλητής που δεν ανήκουν στην κλάση στην οποία έχουν ταξινομηθεί συναρτήσει του ολικού αριθμού των παρατηρήσεων. Το ποσοστό σφάλματος ταξινόμησης υπολογίζεται από τη σχέση

$$C \;=\; 1 \;-\; \max_{k \in K}(\hat{p}_{jk}), \tag{2.3}$$

με K το σύνολο των διαφορετικών κατηγοριών που υπάρχουν στο πρόβλημα ταξινόμησης και $\hat{p}_{jk}$ το ποσοστό των παρατηρήσεων (εκπαίδευσης) στον j-οστό κόμβο που ανήκει στην κατηγορία k (j=1,...,J).

Συνήθως, οι αλγόριθμοι ταξινόμησης αναζητούν μοντέλα με χαμηλό ποσοστό σφάλματος ταξινόμησης όταν αυτοί εφαρμόζονται σε νέα δεδομένα, διαφορετικά από αυτά που χρησιμοποιήθηκαν για την κατασκευή του μοντέλου, δηλαδή από τα δεδομένα εκπαίδευσης (training data). Τα καινούργια, λοιπόν, δεδομένα, χρήσιμα για τον έλεγχο της απόδοσης και της αποτελεσματικότητας του μοντέλου καλούνται δεδομένα ελέγχου (testing data).

✓ Ο **Δείκτης Νοθείας του Gini (Gini Index)** χαρακτηρίζεται ως η διαφορά του αθροίσματος των τετραγώνων των πιθανοτήτων κάθε κλάσης (κατηγορίας) από τη μονάδα και συμβολίζεται ως

$$G \;=\; 1 \;-\; \sum_{k=1}^{K} (\hat{p}_{jk})^2 \;=\; \sum_{k=1}^{K} \hat{p}_{jk}\,(1-\hat{p}_{jk}), \tag{2.4}$$

j=1,...,J, αφού $\sum_{k=1}^{K} \hat{p}_{jk} = 1$.

Ομοίως με πριν, το K δηλώνει το σύνολο των διαφορετικών κατηγοριών που εμφανίζονται στο εκάστοτε πρόβλημα ταξινόμησης, ενώ $\hat{p}_{jk}$ είναι το ποσοστό των παρατηρήσεων εκπαίδευσης στον j-οστό κόμβο που προέρχεται από την κατηγορία k.

Ένας μηδενικός συντελεστής Gini δηλώνει την 'καθαρότητα' του κόμβου, δηλαδή υποδεικνύει ότι όλες οι παρατηρήσεις του δείγματος εκπαίδευσης που περιέχονται στον εν λόγω κόμβο ανήκουν στην ίδια κλάση.

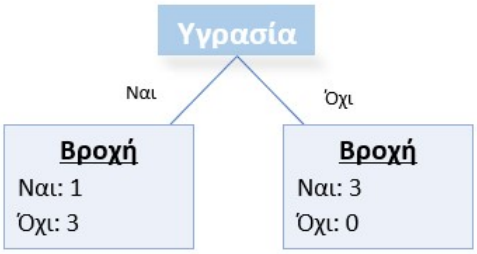
Στην πράξη, το χαρακτηριστικό με τον μικρότερο δείκτη Gini επιλέγεται για την τμήση του κόμβου κάθε φορά. Στην πλειοψηφία των περιπτώσεων, ο αλγόριθμος CART που θα περιγράψουμε στο επόμενο κεφάλαιο χρησιμοποιεί το δείκτη νοθείας του Gini, ως κριτήριο διαχωρισμού των παρατηρήσεων, έτσι ώστε να κατασκευάσει δένδρα αποφάσεων στην δενδρική ταξινόμηση.

Παράδειγμα 2.2.2. Έστω ότι ενδιαφερόμαστε να κατασκευάσουμε ένα δένδρο ταξινόμησης, προκειμένου να εκτιμήσουμε την κλάση της κατηγορικής τυχαίας μεταβλητής Y μέσω της βοήθειας των τιμών των ανεξάρτητων μεταβλητών X_p ($p \geq 1$). Για το συγκεκριμένο παράδειγμα και για λόγους ευκολίας στην κατανόηση υποθέτουμε ότι οι ανεξάρτητες μεταβλητές που έχουμε στην διάθεσή μας είναι οι εξής: η κατηγορική μεταβλητή X_1 που εκφράζει το ενδεχόμενο υγρασίας στην ατμόσφαιρα κάποιας περιοχής, η κατηγορική μεταβλητή X_2 που δηλώνει την παρουσία ανέμων στην ίδια περιοχή και η ποσοτική μεταβλητή X_3 που εκφράζει την θερμοκρασία της περιοχής, μετρημένη σε °C. Η μεταβλητή ενδιαφέροντός μας είναι η Y και δηλώνει το ενδεχόμενο βροχής στην συγκεκριμένη περιοχή. Τα δεδομένα φαίνονται στο Σχήμα 2.5.

Υγρασία $(\mathbf{X_1})$	Άνεμος $(\mathbf{X_2})$	Θερμοκρασία $(\mathbf{X_3})$	Βροχή $(\mathbf{Y})$
Όχι	Ναι	13°C	Ναι
Ναι	Όχι	14°C	Ναι
Ναι	Ναι	15°C	Όχι
Ναι	Όχι	17°C	Όχι
Όχι	Ναι	21°C	Ναι
Ναι	Όχι	22°C	Όχι
Όχι	Ναι	25°C	Ναι

Σχήμα 2.5: Δεδομένα του παραδείγματος 2.2.2.

Πρωταρχικό βήμα αποτελεί η εύρεση εκείνης της (ανεξάρτητης) μεταβλητής που θα χρησιμοποιηθεί για την ρίζα του δένδρου. Αυτό πραγματοποιείται ελέγχοντας πόσο καλά προβλέπουν τις κλάσεις της μεταβλητής απόκρισης Y κάθε μία από τις ανεξάρτητες μεταβλητές X_1, X_2 και X_3 μέσω της βοήθειας του δείκτη νοθείας του Gini (στο συγκεκριμένο παράδειγμα). Αρχικά, όσον αφορά την μεταβλητή X_1 κατασκευάζουμε το απλό δένδρο του Σχήματος 2.6, σύμφωνα με τις τιμές αυτής και της εξαρτημένης μεταβλητής Y.



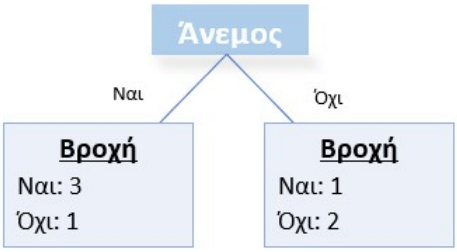
Σχήμα 2.6: Δένδρο για την μεταβλητή 'Υγρασία' (X_1).

Στην συνέχεια, υπολογίζεται ο δείκτης νοθείας του Gini για την εν λόγω μεταβλητή:
Δείκτης Gini για το $1^ο$ φύλλο $= 1 -$ (πιθανότητα του 'Ναι')$^2 -$ (πιθανότητα του 'Όχι')$^2 = 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375$.
Δείκτης Gini για το $2^ο$ φύλλο $= 1 -$ (πιθανότητα του 'Ναι')$^2 -$ (πιθανότητα του 'Όχι')$^2 = 1 - (\frac{3}{3})^2 - (\frac{0}{3})^2 = 0$.

Παρατηρούμε ότι ο δείκτης νοθείας του Gini για το δεύτερο φύλλο είναι μηδέν, το οποίο και αναμέναμε, καθώς όλες οι παρατηρήσεις της Y στο εν λόγω φύλλο ανήκουν στην ίδια κλάση (καθαρός κόμβος). Ωστόσο, επειδή ο αριθμός των παρατηρήσεων διαφέρει στα 2 φύλλα, πρέπει να υπολογίσουμε τον σταθμισμένο μέσο όρο των δεικτών νοθείας του Gini για τα 2 φύλλα (G_1):

$$G_1 = (\frac{4}{4+3}) \times 0.375 + (\frac{3}{4+3}) \times 0 = 0.214.$$

Για την δεύτερη κατηγορική μεταβλητή X_2, δουλεύουμε με παρόμοιο τρόπο και το αποτέλεσμα φαίνεται στο Σχήμα 2.7.



Σχήμα 2.7: Δένδρο για την μεταβλητή 'Άνεμος' (X_2).

Δείκτης Gini για το $1^ο$ φύλλο = 1 − (πιθανότητα του 'Ναι')^2 − (πιθανότητα του 'Όχι')^2 = $1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = 0.375$.
Δείκτης Gini για το $2^ο$ φύλλο = 1 − (πιθανότητα του 'Ναι')^2 − (πιθανότητα του 'Όχι')^2 = $1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = 0.44$.
Συνεπώς, ο σταθμισμένος μέσος όρος των δεικτών Gini για την μεταβλητή X_2 (G_2) είναι:

$$G_2 = (\frac{4}{4+3}) \times 0.375 + (\frac{3}{4+3}) \times 0.44 = 0.403.$$

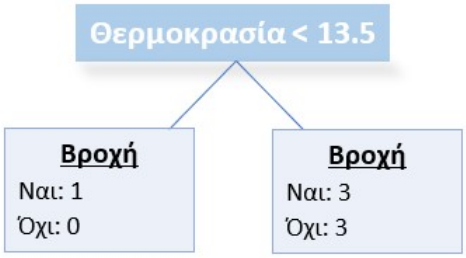
Όσον αφορά την ποσοστική μεταβλητή X_3, η διαδικασία υπολογισμού του δείκτη νοθείας του Gini είναι διαφορετική σε αυτήν την περίπτωση, λόγω του γεγονότος ότι λαμβάνει αριθμητικές τιμές. Όπως και στο προηγούμενο παράδειγμα, ταξινομούμε τις παρατηρήσεις της X_3 κατά αύξουσα σειρά και, έπειτα, υπολογίζουμε την μέση τιμή κάθε 2 γειτονικών παρατηρήσεων, οι οποίες φαίνονται στο Σχήμα 2.8.

Θερμοκρασία (σε °C)	Βροχή
13	Ναι
14	Ναι
15	Όχι
17	Όχι
21	Ναι
22	Όχι
25	Ναι

13.5 ⬅ 14.5 ⬅ 16 ⬅ 19 ⬅ 21.5 ⬅ 23.5 ⬅

Σχήμα 2.8: Μέσες τιμές κάθε 2 γειτονικών παρατηρήσεων της μεταβλητής 'Θερμοκρασία' (X_3).

Στην συνέχεια, χρησιμοποιούμε την κάθε μέση τιμή ως όριο (threshold) για την συνθήκη τμήσης κάθε φορά των τιμών της X_3. Για την πρώτη μέση τιμή, 13.5, λαμβάνουμε το απλό δένδρο του Σχήματος 2.9.



Σχήμα 2.9: Δένδρο για την μεταβλητή 'Θερμοκρασία' κάνοντας χρήση της πρώτης μέσης τιμής (13.5).

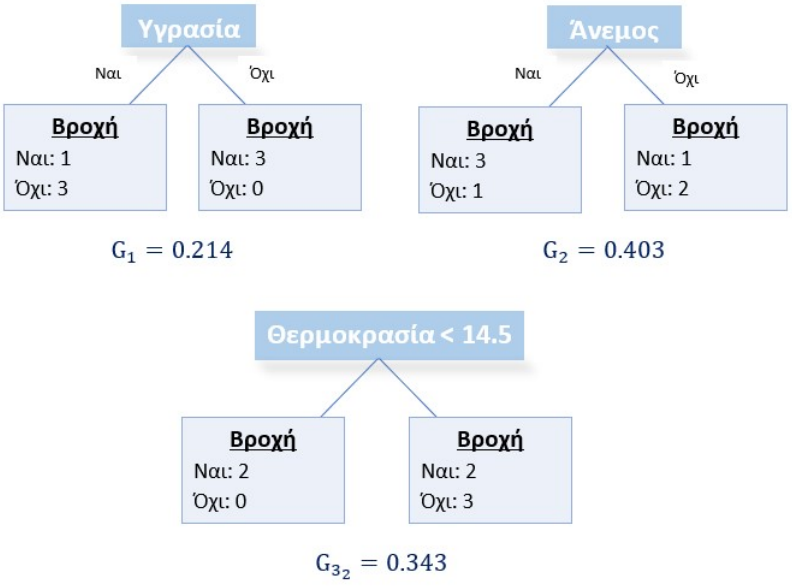
Ομοίως με πριν, γίνεται υπολογισμός του δείκτη νοθείας του Gini για το κάθε φύλλο του δένδρου:
Δείκτης Gini για το $1^ο$ φύλλο $= 1 - (\text{πιθανότητα του 'Ναι'})^2 - (\text{πιθανότητα του 'Όχι'})^2 = 1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0$.
Δείκτης Gini για το $2^ο$ φύλλο $= 1 - (\text{πιθανότητα του 'Ναι'})^2 - (\text{πιθανότητα του 'Όχι'})^2 = 1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$.
Συνεπώς, ο σταθμισμένος μέσος όρος των δεικτών νοθείας του Gini για το εν λόγω δένδρο που αφορά την μεταβλητή X_3 (G_{3_1}) είναι:

$$G_{3_1} = (\frac{1}{1+6}) \times 0 + (\frac{6}{1+6}) \times 0.5 = 0.423.$$

Έτσι, υπολογίζουμε τον δείκτη νοθείας του Gini για όλες τις υπόλοιπες πιθανές τιμές για το όριο της τιμήσης των παρατηρήσεων της X_3 και οι οποίοι φαίνονται στο Σχήμα 2.10.

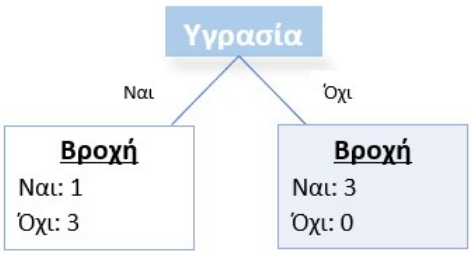
Σχήμα 2.10: Δείκτες νοθείας του Gini για όλα τα πιθανά όρια τμήσης (μέσες τιμές) των παρατηρήσεων της μεταβλητής 'Θερμοκρασία'.

Από τους δείκτες που προκύπτουν επιλέγουμε αυτόν με την χαμηλότερη τιμή που σ᾽ αυτήν την περίπτωση είναι 0.343. Συνεπώς, επιλέγεται και η αντίστοιχη μέση τιμή (14.5) ως υποψήφιο όριο για την ρίζα του ζητούμενου δένδρου.



Σχήμα 2.11: Πιθανά όρια τμήσης για την ρίζα του δένδρου ταξινόμησης.

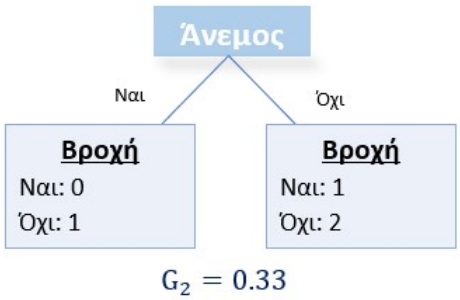
Συγκρίνουμε τον δείκτη Gini του συγκεκριμένου ορίου με τους αντίστοιχους δείκτες των X_1, X_2, όπως προέκυψαν προηγουμένως (Σχήμα 2.11) και, τελικά, επιλέγουμε τον μικρότερο δείκτη για την ρίζα του δένδρου. Επομένως, το δένδρο ταξινόμησης που θέλουμε να κατασκευάσουμε θα έχει την αρχική μορφή του Σχήματος 2.12.



Σχήμα 2.12: Ρίζα του δένδρου για το παράδειγμα 2.2.2.

Παρατηρούμε ότι ο κόμβος στα δεξιά είναι καθαρός, συνεπώς δεν απαιτείται περαιτέρω διαχωρισμός των παρατηρήσεων. Από την άλλη πλευρά, ο κόμβος στα αριστερά έχει τη δυνατότητα να τμηθεί ξανά, έχοντας ως υποψήφιες μεταβλητές τμήσεις τον 'Άνεμο' και την 'Θερμοκρασία'. Επαναλαμβάνουμε την ίδια διαδικασία με πριν, υπολογίζοντας τον δείκτη νοθείας του Gini για κάθε μεταβλητή και λαμβάνοντας υπόψη μόνο τις παρατηρήσεις για τις οποίες υπάρχει υγρασία στην

περιοχή. Το δένδρο που προκύπτει για την μεταβλητή 'Άνεμος' είναι αυτό του Σχήματος 2.13. Οι παρατηρήσεις για τις οποίες υπάρχει υγρασία στην περιοχή υπάρχουν στο σύνολο δεδομένων του Σχήματος 2.14.



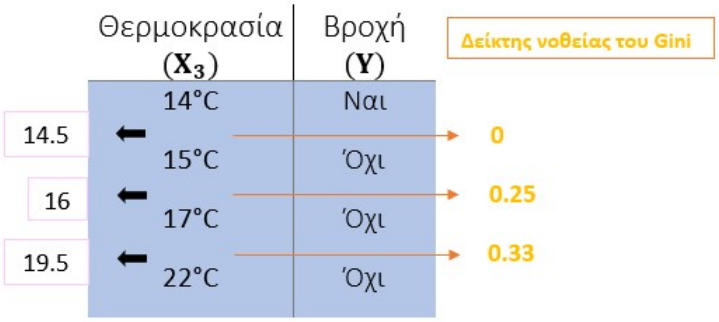
Σχήμα 2.13: Δένδρο για την μεταβλητή 'Άνεμος' κάνοντας χρήση μόνο των παρατηρήσεων για τις οποίες υπάρχει υγρασία στην περιοχή.

Υγρασία ($\mathbf{X_1}$)	Άνεμος ($\mathbf{X_2}$)	Θερμοκρασία ($\mathbf{X_3}$)	Βροχή ($\mathbf{Y}$)
Όχι	Ναι	13°C	Ναι
Ναι	Όχι	14°C	Ναι
Ναι	Ναι	15°C	Όχι
Ναι	Όχι	17°C	Όχι
Όχι	Ναι	21°C	Ναι
Ναι	Όχι	22°C	Όχι
Όχι	Ναι	25°C	Ναι

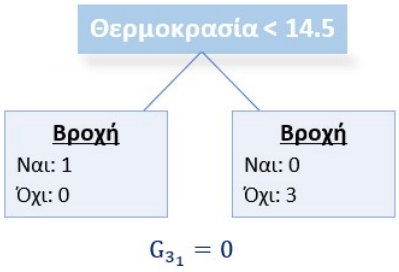
Σχήμα 2.14: Παρατηρήσεις για τις οποίες υπάρχει υγρασία στην περιοχή.

Συνεπώς, για την 'Θερμοκρασία' θα έχουμε τα αποτελέσματα του Σχήματος 2.15.

Όπως προηγουμένως, διαλέγουμε τον μικρότερο δείκτη Gini και έχουμε το δένδρο του Σχήματος 2.16 για την μεταβλητή 'Θερμοκρασία'.



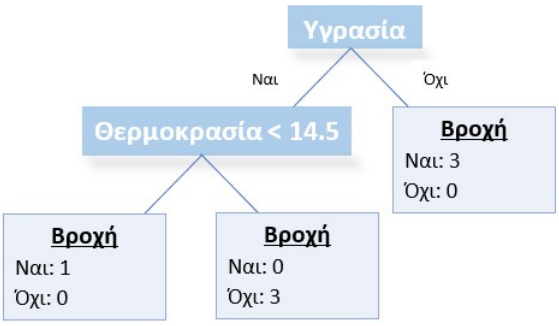
Σχήμα 2.15: Σύνολο δεδομένων και δείκτες νοθείας του Gini για την μεταβλητή 'Θερμοκρασία', λαμβάνοντας υπόψη μόνο τις παρατηρήσεις για τις οποίες υπάρχει υγρασία στην περιοχή.



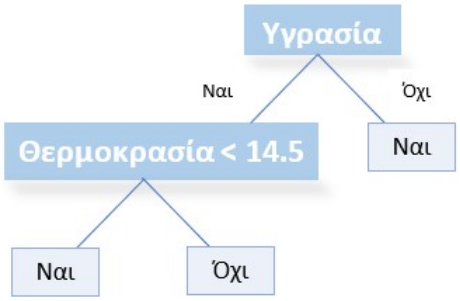
Σχήμα 2.16: Δένδρο για την μεταβλητή 'Θερμοκρασία' κάνοντας χρήση μόνο των παρατηρήσεων για τις οποίες υπάρχει υγρασία στην περιοχή.

Από τα παραπάνω προτιμάται η 'Θερμοκρασία' ως μεταβλητή για την επόμενη τμήση με όριο το 14.5 και, έτσι, προκύπτει το δένδρο του Σχήματος 2.17.

Οι κόμβοι που διαμορφώθηκαν είναι καθαροί, συνεπώς σταματάει η διαδικασία διαχωρισμού των παρατηρήσεων. Εναλλακτικά, όπως και στο προηγούμενο παράδειγμα, θα μπορούσε να οριστεί κάποιος ελάχιστος αριθμός παρατηρήσεων ως ένα άλλο κριτήριο τερματισμού. Τέλος, ως τελική πρόβλεψη κάθε φύλλου θεωρείται η κλάση με τις περισσότερες ψήφους ("majority vote") και, άρα το δένδρο ταξινόμησης για τα εν λόγω δεδομένα θα είναι αυτό του Σχήματος 2.18.



Σχήμα 2.17: Δένδρο ταξινόμησης για το παράδειγμα 2.2.2.



Σχήμα 2.18: Τελικό δένδρο ταξινόμησης κάνοντας χρήση της ψήφου πλειοψηφίας για το παράδειγμα 2.2.2.

✓ Η **Εντροπία (Entropy)** αποτελεί ένα άλλο μέτρο της 'καθαρότητας' ενός κόμβου, δηλαδή έναν άλλο τρόπο κατανόησης αν ένας κόμβος περιέχει παρατηρήσεις ίδιας κλάσης (καθαρός κόμβος) ή όχι. Η εντροπία δίνεται από τον τύπο

$$E = \sum_{k=1}^{K} -(\hat{p}_{jk}) \log \hat{p}_{jk}, \tag{2.5}$$

j=1,...,J, με K το πλήθος των κατηγοριών του προβλήματος ταξινόμησης και $\hat{p}_{jk}$ το ποσοστό των παρατηρήσεων εκπαίδευσης του j-οστού κόμβου και κατηγορίας k.

Η εντροπία δίνει προσεγγιστικά παρόμοιες τιμές με εκείνες του δείκτη νοθείας του Gini.

Ισχύει ότι $0 \leq \hat{p}_{jk} \leq 1$, αφού $\hat{p}_{jk}$ είναι, ουσιαστικά, μία πιθανότητα και, δεδομένου ότι οι λογάριθμοι των κλασμάτων που ανήκουν στο διάστημα $[0, 1]$ παίρνουν αρνητικές τιμές, συνεπάγεται ότι $0 \leq -\hat{p}_{jk} \log \hat{p}_{jk}$.

Επομένως, οι τιμές της εντροπίας βρίσκονται μεταξύ του μηδενός και του ένα με προτιμότερη την πιο χαμηλή. Αν ένας κόμβος περιέχει τον ίδιο αριθμό παρατηρήσεων εκπαίδευσης σε κάθε κλάση, τότε η εντροπία παίρνει την τιμή 1, ενώ αν όλες οι παρατηρήσεις εκπαίδευσης σε κάποιο κόμβο ανήκουν στην ίδια κλάση, η εντροπία μηδενίζεται.

Παράδειγμα 2.2.3. Αξιοποιώντας τα δεδομένα του παραδείγματος 2.2.2., μπορούμε εξίσου να κατασκευάσουμε ένα δένδρο ταξινόμησης με χρήση του ποσοστού του σφάλματος ταξινόμησης ως συνάρτηση νοθείας, αντί για τον δείκτη νοθείας του Gini. Η διαδικασία παραμένει ακριβώς η ίδια, αλλάζοντας μονάχα το μέτρο νοθείας σε κάθε τμήση. Για να γίνει καλύτερα κατανοητό αυτό στον αναγνώστη, θα υπολογίσουμε το ποσοστό σφάλματος ταξινόμησης για τα δύο πρώτα φύλλα της μεταβλητής X_1 ('Υγρασία') του Σχήματος 2.6.
Σύμφωνα με τον τύπο (2.3), το ποσοστό του σφάλματος ταξινόμησης για τα δύο αυτά φύλλα είναι:
Ποσοστό του σφάλματος ταξινόμησης για το $1^ο$ φύλλο $= 1 - \max$ [(πιθανότητα του 'Ναι'), (πιθανότητα του 'Όχι')] $= 1 - \max\left(\frac{1}{4}, \frac{3}{4}\right) = 1 - \frac{3}{4} = 0.25$.
Ποσοστό του σφάλματος ταξινόμησης για το $2^ο$ φύλλο $= 1 - \max$ [(πιθανότητα του 'Ναι'), (πιθανότητα του 'Όχι')] $= 1 - \max\left(\frac{3}{3}, \frac{0}{3}\right) = 1 - \frac{3}{3} = 0$.
($k = 2$, καθώς, για τα εν λόγω δεδομένα, οι κλάσεις της μεταβλητής 'Υγρασία' είναι δύο: 'Ναι' και 'Όχι'.)
Ο σταθμισμένος μέσος όρος των ποσοστών των σφαλμάτων ταξινόμησης για τα 2 φύλλα είναι:

$$\mathrm{C}_1 = \left(\frac{4}{4+3}\right) \times 0.25 + \left(\frac{3}{4+3}\right) \times 0 = 0.143.$$

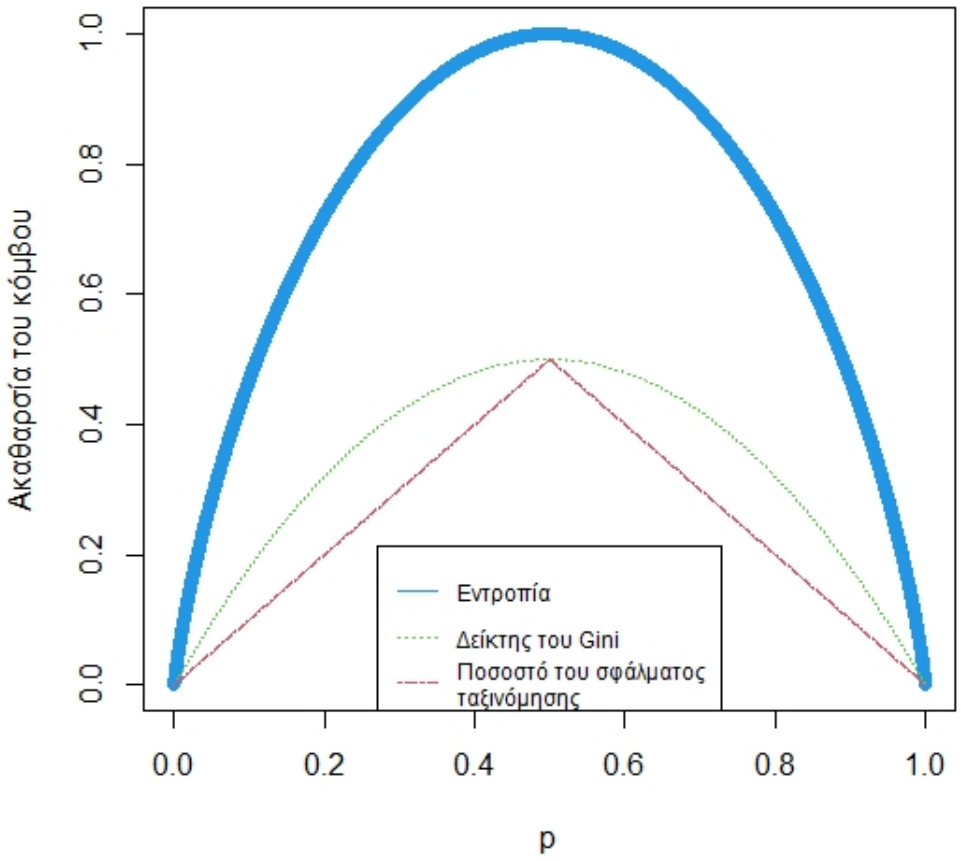
Με παρόμοιο τρόπο υπολογίζεται το ποσοστό του σφάλματος ταξινόμησης και για τις υπόλοιπες μεταβλητές, προτιμώντας πάντα εκείνο με την χαμηλότερη τιμή για την εκάστοτε τμήση στο δένδρο ταξινόμησης.

Σε περίπτωση που θέλαμε να θεωρήσουμε την εντροπία ως συνάρτηση νοθείας, υπολογίζουμε ενδεικτικά την εντροπία για την ίδια μεταβλητή X_1 ('Υγρασία') και, όπως πριν, η διαδικασία συνεχίζεται με τον ίδιο τρόπο. Σύμφωνα με τον τύπο της εντροπίας (2.5), ισχύουν τα εξής:
Εντροπία για το $1^ο$ φύλλο $=$ − (πιθανότητα του 'Ναι') log(πιθανότητα του 'Ναι') − (πιθανότητα του 'Όχι') log(πιθανότητα του 'Όχι') $= -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.811$.
Εντροπία για το $2^ο$ φύλλο $=$ − (πιθανότητα του 'Ναι') log(πιθανότητα του 'Ναι') − (πιθανότητα του 'Όχι') log(πιθανότητα του 'Όχι') $= -\frac{3}{3}\log\frac{3}{3} -$

$0 = 0$.
($k = 2$, καθώς, για τα εν λόγω δεδομένα, οι κλάσεις της μεταβλητής 'Υγρασία' είναι δύο: 'Ναι' και 'Όχι'.)

Ο σταθμισμένος μέσος όρος των εντροπίων για τα 2 φύλλα είναι:

$$\mathrm{E}_1 = (\frac{4}{4+3}) \times 0.811 + (\frac{3}{4+3}) \times 0 = 0.463.$$



Σχήμα 2.19: Γράφημα σύγκρισης του ποσοστού του σφάλματος ταξινόμησης, του δείκτη νοθείας του Gini και της εντροπίας για το δυαδικό πρόβλημα ταξινόμησης.

Ωστόσο, η συνάρτηση της εντροπίας μπορεί να λάβει και τιμές μεγαλύτερες της μονάδας. Έτσι, καθώς αυξάνεται η εντροπία αυξάνεται και το επίπεδο της

μη καθαρότητας του κόμβου, μειώνοντας με αυτό τον τρόπο την καθαρότητά του (οι παρατηρήσεις εκπαίδευσης παύουν να ανήκουν σε μία μόνο κλάση).

Στο Σχήμα 2.19 περιγράφεται η σύγκριση των τριων μέτρων νοθείας μέσω των αντίστοιχων γραφημάτων τους. Το p στον οριζόντιο άξονα αναφέρεται στο ποσοστό των παρατηρήσεων εκπαίδευσης που ανήκουν στην μία από τις δύο κλάσεις. Αυτό το p είναι το $\hat{p}_{jk}$ στους αντίστοιχους τύπους των μέτρων νοθείας, όπου το $k = 1, 2$ για το δυαδικό πρόβλημα ταξινόμησης.

Μπορούμε, λοιπόν, να παρατηρήσουμε ότι και τα τρία μέτρα νοθείας λαμβάνουν τις ελάχιστες τιμές τους όταν όλες οι παρατηρήσεις εκπαίδευσης ανήκουν στην ίδια κλάση ($p = 0, p = 1$), ενώ όταν υπάρχει ίσος αριθμός παρατηρήσεων (εκπαίδευσης) και στις δύο κλάσεις, τότε τα μέτρα νοθείας μεγιστοποιούνται ($p = 0.5$).

Συνεπώς, ως μεταβλητή τμήσης ξεχωρίζει, μεταξύ των άλλων ανεξάρτητων μεταβλητών, εκείνη με την μεγαλύτερη μείωση στην εντροπία σε σχέση με την (αρχική) εντροπία όλων των παρατηρήσεων (εκπαίδευσης). Πώς, όμως, υπολογίζεται στην πράξη αυτή η μείωση; Την απάντηση δίνει η επόμενη μετρική.

✓ Το **Κέρδος Πληροφορίας (Information Gain)** αποτελεί μία προέκταση της συνάρτησης νοθείας της εντροπίας, καθώς περιγράφεται σαν την μείωση στην εντροπία. Συγκεκριμένα, όπως φαίνεται και στον παρακάτω τύπο, είναι η διαφορά της εντροπίας μετά την τμήση από την εντροπία πριν την τμήση, δηλαδή

$$I(Y, X) \;=\; E(Y) \;-\; \sum_{x \in Values(X)} \frac{|Y_x|}{|Y|}\, E(Y_x), \tag{2.6}$$

όπου Χ είναι η ανεξάρτητη μεταβλητή, οι τιμές της οποίας συμβολίζονται ως '$Values(X)$' , βάσει της οποίας επιχειρείται ο επόμενος διαχωρισμός, x μία από τις δυνατές τιμές της Χ και $E(Y_x)$ η εντροπία στην περίπτωση που $Y = x$. Γενικά, αν Y είναι ένα σύνολο, τότε με $|Y|$ συμβολίζουμε τον πληθάριθμο ή, αλλιώς, την πληθικότητα (μέγεθος) του συνόλου Y. Το $|Y_x|$ δίνει το πλήθος των παρατηρήσεων της εξαρτημένης μεταβλητής Y που παίρνουν την τιμή (ή που ανήκουν στην κλάση) x της X και το $|Y|$ είναι το συνολικό πλήθος των παρατηρήσεων της εξαρτημένης μεταβλητής Y. Ο όρος $E(Y)$ δηλώνει την εντροπία πριν από την τμήση, δηλαδή την εντροπία για την ρίζα του κάθε υποδένδρου που κατασκευάζεται για την αναζήτηση της κατάλληλης τμήσης. Ο δεύτερος όρος από το δεξί μέλος της (2.6) είναι η εντροπία της μεταβλητής Y που θέλουμε να ταξινομήσουμε μετά την τμήση, σύμφωνα με την τιμή του χαρακτηριστικού X και αποτελείται από την εντροπία στην περίπτωση που $Y = x$ επί την πιθανότητα του ενδεχομένου η εξαρτημένη μεταβλητή Y να λαμβάνει την τιμή x, όπου

x ανήκει στο σύνολο $Values(X)$ που είναι οι πιθανές τιμές της ανεξάρτητης μεταβλητής X.

Όσο μεγαλύτερες είναι οι τιμές του κέρδους πληροφορίας, τόσο καλύτερη είναι η συνθήκη τμήσης. Συνήθως, οι αλγόριθμοι κατασκευής δέντρων αποφάσεων (δενδροδιαγραμμάτων) προτιμούν συνθήκες που μεγιστοποιούν το κέρδος πληροφορίας. Δεδομένου ότι η εντροπία πριν από την τμήση είναι η ίδια για όλες τις συνθήκες τμήσης, η μεγιστοποίηση του κέρδους πληροφορίας ισοδυναμεί με την ελαχιστοποίηση της εντροπίας μετά την τμήση, δηλαδή με την ελαχιστοποίηση του αθροίσματος της εξίσωσης (2.6).

Παράδειγμα 2.2.4. Χρησιμοποιώντας το ίδιο σύνολο δεδομένων με το παράδειγμα 2.2.2., υπολογίζεται στην συνέχεια το κέρδος πληροφορίας για τα δύο φύλλα της μεταβλητής X_1 ('Υγρασία') του Σχήματος 2.6. Η εντροπία για το κάθε φύλλο έχει ήδη υπολογιστεί από το προηγούμενο παράδειγμα και είναι ίση με:
Εντροπία για το 1^o φύλλο $= 0.811$.
Εντροπία για το 2^o φύλλο $= 0$.
Η εντροπία για την ρίζα του εν λόγω δένδρου είναι
Εντροπία(Υγρασία) $= E(\text{Υγρασία}) = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.985$.
Επομένως, από την σχέση (2.6), το κέρδος πληροφορίας υπολογίζεται ως εξής:

I = $E(\text{Υγρασία})$ − (Πιθανότητα να υπάρχει υγρασία στην περιοχή) × (Εντροπία για το 1^o φύλλο) − (Πιθανότητα να μην υπάρχει υγρασία στην περιοχή) × (Εντροπία για το 2^o φύλλο) $= 0.985 - (\frac{4}{7} \times 0.811 + \frac{3}{7} \times 0) = 0.522$.

Συνεπώς, ένα από τα πιο απαιτητικά και κρίσιμα στάδια κατά την κατασκευή των δένδρων αποφάσεων και, ίσως, αυτό για το οποίο διαθέτουμε τον περισσότερο χρόνο είναι η εύρεση της ιδανικότερης τμήσης, χρησιμοποιώντας κάθε φορά την μεταβλητή που φέρνει μεγαλύτερη μείωση στα παραπάνω μέτρα νοθείας, ανάλογα με το πρόβλημα. Για το τελικό μοντέλο, ζητούμενα είναι η βέλτιστη πρόβλεψη, συνοδευόμενη, εάν είναι εφικτό, με την λιγότερη δυνατή πολυπλοκότητα. Η τελευταία παράμετρος υπολογίζεται από τον συνολικό αριθμό των ενδιάμεσων κόμβων, από τον συνολικό αριθμό των φύλλων και από το πλήθος των μεταβλητών που χρησιμοποιούνται στο δένδρο. Η έννοια της πολυπλοκότητας θα γίνει καλύτερα κατανοητή στο τρίτο κεφάλαιο.

Κεφάλαιο 3

Ο Αλγόριθμος CART και οι προεκτάσεις του

Η μέθοδος της επαγωγικής κατασκευής δένδρων παλινδρόμησης και δένδρων ταξινόμησης μέσω διακλαδώσεων (αλγόριθμος CART), αποτελεί μία από τις πιο δημοφιλείς και εύχρηστες τεχνικές στον κλάδο της Μηχανικής Μάθησης (Maching Learning) και της Επιστήμης των Δεδομένων (Data Science), χάρη στην ικανότητά της να χειρίζεται μεγάλο αριθμό δεδομένων, δίχως παραμέτρους, όπως και στην διαφάνεια και αμεσότητα αυτής όσον αφορά την ερμηνεία της. Η συγκεκριμένη μέθοδος περιέχει στρατηγικές, όπως ο Αναδρομικός Δυαδικός Σχεδιασμός και το γνωστό Κλάδεμα Δένδρων, η συμβολή των οποίων αποτελεί αρωγό στην κατασκευή του καταλληλότερου δένδρου για τα εκάστοτε στοιχεία. Ωστόσο, η σύγχρονη ανάγκη για βέλτιστη και όσον το δυνατόν πιο ρεαλιστική πρόβλεψη ή ταξινόμηση που λαμβάνουμε από ένα δένδρο αποφάσεων με το επιπλέον χαρακτηριστικό να παραμένει αναλλοίωτη σε τυχόν μεταβολές στα δεδομένα αποτέλεσε κίνητρο για αναζήτηση νέων, βελτιωμένων μορφών της εν λόγω μεθόδου. Οι πιο γνωστές επεκτάσεις, οι οποίες χρησιμοποιούνται ευρέως στην επιστημονική κοινότητα και αποσκοπούν στην αποδοτικότητα και την ευρωστία του μοντέλου είναι η λεγόμενη Ενσάκιση (Bagging), τα Τυχαία Δάση (Random Forests) και η Ενίσχυση (Boosting).

3.1 Αλγόριθμος CART

Προτού περιγράψουμε αναλυτικά τα βήματα του αλγορίθμου κατασκευής δένδρων παλινδρόμησης και δένδρων ταξινόμησης, θα εισάγουμε κάποιες βασικές έννοιες και στατιστικές μεθόδους, οι οποίες είναι απαραίτητες για την κατανόηση και εφαρμογή του αλγορίθμου.

3.1.1 Αναδρομικός Δυαδικός Διαχωρισμός

Ο **Αναδρομικός Δυαδικός Διαχωρισμός** (Recursive Binary Splitting) πρόκειται για μία στρατηγική που ξεκινά από την κορυφή του δένδρου, δηλαδή από εκείνο το τμήμα που περιέχει όλες τις παρατηρήσεις, διαμορφώνοντας, στην συνέχεια, τους διαχωρισμούς προς τα κάτω. Κάθε διαχωρισμός συνοδεύεται από δύο νέα κλαδιά (branches) και ο βέλτιστος υπολογίζεται κάθε φορά τοπικά, σε κάθε κόμβο, δίχως να λαμβάνεται υπόψιν το ενδεχόμενο προνόησης ενός διαχωρισμού που να οδηγεί σε καλύτερα αποτελέσματα αργότερα στην διαδικασία κατασκευής του δένδρου, δηλαδή η λήψη της βέλτιστης απόφασης πραγματοποιείται κάθε φορά στον εν λόγω κόμβο (greedy approach). Η επιλογή του βέλτιστου διαχωρισμού γίνεται με βάση κάποιου κριτηρίου τμήσης σε ένα από τα χαρακτηριστικά (επεξηγηματικές μεταβλητές), τα οποία αναφέρθηκαν λεπτομερώς στο προηγούμενο κεφάλαιο.

Η στρατηγική αυτή καλείται αναδρομική, διότι αναπτύσσει αναδρομικά το δένδρο αποφάσεων, ξεκινώντας από τον κόμβο που περιέχει όλα τα δεδομένα (αρχικός κόμβος) και καταλήγοντας στους τελικούς κόμβους (φύλλα), οι οποίοι δεν μπορούν να τμηθούν περαιτέρω, καθώς η τμήση αυτών δεν προσφέρει παραπάνω ακρίβεια για την πρόβλεψη.

Αφού βρούμε την κατάλληλη διαμέριση για τον πρώτο κόμβο σε περίπτωση μίας ανεξάρτητης μεταβλητής, διαφορετικά, σε περίπτωση ύπαρξης πολλών ανεξάρτητων μεταβλητών, είναι απαραίτητη και η εύρεση, επιπλέον, της καταλληλότερης ανεξάρτητης μεταβλητής που θα χρησιμοποιηθεί για την τμήση του αρχικού κόμβου, επαναλαμβάνουμε την διαδικασία για τους δύο επόμενους κόμβους. Η διαδικασία συνεχίζεται μέχρι να ικανοποιηθεί κάποιο κριτήριο διακοπής, έτσι ώστε ο εκάστοτε κόμβος να μην μπορεί να τμηθεί περαιτέρω και να μετατραπεί σε τελικό κόμβο για το δένδρο.

Ορισμένα κριτήρια για το πότε πρέπει να τερματιστεί η διαδικασία τμήσεων των κόμβων αναφέρονται παρακάτω:

- Ο προσδιορισμός του ελάχιστου αριθμού παρατηρήσεων που πρέπει να περιέχει ένας τελικός κόμβος (δενδρική παλινδρόμηση).
- Η επιλογή του ελάχιστου αριθμού των παρατηρήσεων που ανήκουν σε μία συγκεκριμένη κλάση που οφείλει να περιέχει ένας τελικός κόμβος (δενδρική ταξινόμηση).

Επίσης, η διαδικασία του διαχωρισμού των παρατηρήσεων διακόπτεται όταν όλα τα στοιχεία του κόμβου ανήκουν στην ίδια κλάση (κατηγορία) ή την στιγμή που πετυχαίνεται το όριο του ελάχιστου αριθμού παρατηρήσεων που πρέπει να διαθέτει ο κόμβος, προκειμένου να γίνει ο επόμενος διαχωρισμός.

Τα όρια του ελάχιστου αριθμού παρατηρήσεων, όπως και η επιλογή του ελάχιστου αριθμού των παρατηρήσεων που πρέπει να ανήκουν σε μία κλάση καθορίζονται από τον χρήστη εκ των προτέρων.

Μόλις κατασκευαστεί το δένδρο αποφάσεων σύμφωνα με την παραπάνω διαδικασία, πραγματοποιούνται σε αυτό κάποιες βελτιστοποιήσεις, έτσι ώστε να επιτευχθεί η ακριβέστερη πρόβλεψη ή ταξινόμηση που είναι και ο στόχος της δενδρικής παλινδρόμησης και της δενδρικής ταξινόμησης αντίστοιχα.

Μία σημαντική μορφή βελτιστοποίησης, η εφαρμογή της οποίας συμβάλλει σημαντικά στην βελτίωση της απόδοσης του μοντέλου (δένδρου αποφάσεων) είναι το λεγόμενο Κλάδεμα Δένδρων (Tree Pruning).

3.1.2 Κλάδεμα Δένδρου

Η διαδικασία που περιγράφτηκε παραπάνω έχει ως αποτέλεσμα ένα πλήρες δένδρο του οποίου οι προβλέψεις πιθανόν να είναι αρκετά καλές όταν αυτό εφαρμόζεται στα δεδομένα εκπαίδευσης, δηλαδή στα δεδομένα που στηρίχθηκε ο αναδρομικός δυαδικός διαχωρισμός για την κατασκευή του (training data). Ωστόσο, αν το εν λόγω δένδρο αποφάσεων δοκιμαστεί σε νέα, διαφορετικά από τα προηγούμενα, δεδομένα ελέγχου (testing data), τότε ενδέχεται η τελική απόδοση να είναι χαμηλή και οι συνολικές προβλέψεις να μην παρουσιάζουν μεγάλο βαθμό εγκυρότητας και ευστοχίας. Συνήθως, το γεγονός αυτό συμβαίνει, διότι το τελικό δένδρο είναι υπερβολικά μεγάλο σε μέγεθος (με πολλά κλαδιά), όπως και υπερβολικά πολύπλοκο (δύσκολο στην ερμηνεία).

Για τον λόγο αυτό, εισάγεται η έννοια του **Κλαδέματος του Δένδρου** (Tree Pruning), προκειμένου να προσφέρει μία καλύτερη, πιο απλοποιημένη μορφή στο δένδρο παλινδρόμησης (ή ταξινόμησης), έτσι ώστε οι τελικές εκτιμήσεις να αντιπροσωπεύουν και δεδομένα, πέρα των δεδομένων εκπαίδευσης. Ένα μικρότερο μεγέθους δένδρο (με λιγότερα κλαδιά) πιθανόν να οδηγήσει σε χαμηλότερη διασπορά, μειώνοντας, έτσι, τα σφάλματα πρόβλεψης (ή ταξινόμησης), ενώ, ταυτόχρονα, διευκολύνεται η ερμηνεία των αποτελεσμάτων, έχοντας, ίσως, ως μοναδικό τίμημα την εμφάνιση κάποιας μεροληψίας στο μοντέλο. Αξίζει να σημειωθεί ότι το ζητούμενο δεν είναι ένα υπερβολικά μεγάλο δένδρο (με πολλαπλούς κλάδους) που να δυσχεραίνει την εξήγηση των αποτελεσμάτων, αλλά ούτε και ένα εξαιρετικά μικρό σε μέγεθος δένδρο, το οποίο δεν αξιοποιεί όλες τις απαραίτητες πληροφορίες, χρήσιμες για μία ακριβής πρόβλεψη.

Προτού αναλύσουμε τη συγκεκριμένη μέθοδο, θα περιγράψουμε την στρατηγική της 'Διασταυρωτικής Επικύρωσης' και της 'Διασταυρωτικής Επικύρωσης V ομάδων' [2]. Η τελευταία λαμβάνει μέρος στην διαδικασία του κλαδέματος,

όπως θα δούμε στην συνέχεια.
Οι δύο προαναφερθείσες στρατηγικές έχουν ένα κοινό χαρακτηριστικό: και στις δύο, ένα μέρος των διαθέσιμων δεδομένων αξιοποιείται για την ανάλυση της δενδρικής παλινδρόμησης (για την κατασκευή του δένδρου), ενώ το υπόλοιπο χρησιμοποιείται για την εκτίμηση και την αξιολόγηση της πρόβλεψης.

- **Διασταυρωτική Επικύρωση (Cross Validation):**

  Στην μέθοδο της Διασταυρωτικής Επικύρωσης (Cross Validation), το δένδρο διαμορφώνεται με βάση μίας ομάδας παρατηρήσεων γνωστής και ως δείγμα εκπαίδευσης (training sample). Η αξιοπιστία και η εγκυρότητα του καταλληλότερου δένδρου εξετάζεται μέσω ενός άλλου, ανεξάρτητου δείγματος παρατηρήσεων γνωστό και σαν δείγμα ελέγχου (testing sample).

- **Διασταυρωτική Επικύρωση V ομάδων (V-fold Cross Validation):**

  Με βάση τη μέθοδο της Διασταυρωτικής Επικύρωσης V ομάδων (V-fold Cross Validation), το αρχικό δείγμα των παρατηρήσεων χωρίζεται σε V μικρότερα υποδείγματα, τα οποία επιλέγονται με τυχαίο τρόπο το καθένα. Έστω V_i ένα τέτοιο υποδείγμα με $i = 1, ..., V$. Έπειτα, γίνεται εφαρμογή του Αναδρομικού Δυαδικού Διαχωρισμού σε όλα τα υποδείγματα, εκτός του V_i. Ύστερα, εκτιμάται η ποιότητα και η εγκυρότητα του δένδρου που δημιουργήθηκε, κάνοντας χρήση του V_i υποδείγματος, ώστε να βρεθεί μία προσέγγιση του μέσου τετραγωνικού σφάλματος πρόβλεψης. Η διαδικασία αυτή επαναλαμβάνεται για κάθε υποδείγμα V_i με $i = 1, ..., V$.

Σε πρακτικό επίπεδο, η τεχνική του κλαδέματος αφαιρεί την συνθήκη τμήσης ενός κόμβου που παράγει δύο τελικούς κόμβους (φύλλα) και αντικαθιστά τους δύο αυτούς κόμβους με ένα. Το κλάδεμα ενός πλήρους δένδρου αποφάσεων πραγματοποιείται μέσω μίας τεχνικής που ονομάζεται 'κλάδεμα της πολυπλοκότητας του κόστους' (cost complexity pruning), η οποία διευκολύνει στην επιλογή εκείνου του συνόλου από υπόδενδρα που θα τεθεί προς κλάδεμα. Η συγκεκριμένη τεχνική είναι γνωστή και ως 'κλάδεμα του πιο αδύναμου κρίκου'. Αντί, λοιπόν, να λαμβάνεται υπόψιν κάθε δυνατό υπόδενδρο, επιλέγεται μία ακολουθία από δένδρα που υποδεικνύει μία μη αρνητική παράμετρος, η οποία συμβολίζεται με α και ονομάζεται παράμετρος πολυπλοκότητας.

Για κάθε τιμή του α, υπάρχει ένα υπόδενδρο T (του αρχικού, πλήρους δένδρου) που δίνει τον αντίστοιχο **βαθμό δένδρου** (Tree Score), σύμφωνα με τον τύπο:

$$\text{Tree Score } = \sum_{j=1}^{|T|} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \; + \; a|T|. \tag{3.1}$$

Στην σχέση (3.1), το $|T|$ δηλώνει τον αριθμό των τελικών κόμβων στο δένδρο T, ή, αλλιώς, τον αριθμό των φύλλων στο δένδρο T, R_j είναι το σύνολο των παρατηρήσεων εκπαίδευσης που αντιστοιχούν στον j-οστό τελικό κόμβο του δένδρου T και $\hat{y}_{R_j}$ αποτελεί την εκτιμημένη πρόβλεψη που σχετίζεται με το σύνολο R_j που είναι στην ουσία η μέση τιμή των παρατηρήσεων εκπαίδευσης στο σύνολο R_j.

Παρατηρούμε ότι ο πρώτος όρος στο δεξί μέλος της (3.1) είναι στην πραγματικότητα το άθροισμα τετραγώνων των υπολοίπων (SSR), το οποίο οφείλει να είναι όσον το δυνατόν μικρότερο. Ο δεύτερος όρος καλείται 'ποινή πολυπλοκότητας του δένδρου' (tree complexity penalty) και εύκολα διαπιστώνουμε ότι αυξάνεται, καθώς αυξάνεται ο αριθμός των τελικών κόμβων στο δένδρο.

Στην πράξη, υπερέχει εκείνος ο βαθμός δένδρου που έχει την μικρότερη τιμή έναντι των άλλων βαθμών.

Έστω ότι ασχολούμαστε με ένα πρόβλημα δενδρικής παλινδρόμησης. Τότε, η μέθοδος του κλαδέματος πολυπλοκότητας του κόστους περιγράφεται στα βήματα που ακολουθούν:

1) Κατασκευή του αρχικού, πλήρους δένδρου, έστω T_0, κάνοντας χρήση όλων των δεδομένων με τη βοήθεια της τεχνικής του Αναδρομικού Δυαδικού Διαχωρισμού και επιλέγοντας μία αρχική τιμή για το α (a_0). Ο βαθμός του δένδρου T_0 θα είναι ίσος με το SSR, όταν το $a = 0$.

2) Κλάδεμα του δένδρου T_0 μία φορά, ενώ, ταυτόχρονα, αύξηση του α και υπολογισμός του βαθμού του εν λόγω κλαδεμένου δένδρου. Έπειτα, σταδιακή αύξηση του α μαζί με κλάδεμα κάθε φορά του τελευταίου (κλαδεμένου) δένδρου έως ότου να λάβουμε τον χαμηλότερο βαθμό δένδρου και το τελικό δένδρο που προκύπτει από την παραπάνω διαδικασία να μην μπορεί να κλαδευτεί περαιτέρω (η διαδικασία διακόπτεται μόλις φτάσουμε στην ρίζα του δένδρου). Με άλλα λόγια, καθώς το α αυξάνεται, τα κλαδιά του εκάστοτε δένδρου κλαδεύονται με ένθετο (προβλέψιμο) και διαδοχικό τρόπο (κλάδεμα από κάτω προς τα πάνω) και με αυτόν τον τρόπο λαμβάνουμε μία ακολουθία υποδένδρων ως συνάρτηση της παραμέτρου πολυπλοκότητας α.

   Η γενική μορφή της σχέσης (3.1) είναι η εξής:

$$\text{Tree Score} \;:=\; \text{Tree Score}(T) \;=\; R(T) \;+\; a|T|, \qquad (3.2)$$

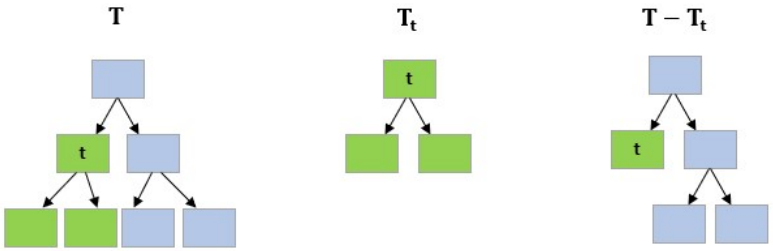
   όπου $T \subset T_0$ υποδένδρο του αρχικού, ολικού δένδρου T_0, $|T|$ ο αριθμός των φύλλων (τελικών κόμβων) στο δένδρο T και $R(T)$ η συνάρτηση

σφάλματος των παρατηρήσεων εκπαίδευσης. Για ένα πρόβλημα δενδρικής παλινδρόμησης, η συνάρτηση $R(T)$ χαρακτηρίζεται ως το άθροισμα των τετραγώνων των υπολοίπων (SSR). Σε κάθε βήμα, επιλέγεται εκείνο το υποδένδρο T_t που ελαχιστοποιεί την μείωση στην συνάρτηση (3.1), δηλαδή ελαχιστοποιεί την διαφορά

$$\begin{aligned} &\text{Tree Score}(T - T_t) \ - \ \text{Tree Score}(T) \ = \\ &SSR(T - T_t) \ + \ a|T - T_t| \ - \ (SSR(T) \ + \ a|T|) \ = \\ &SSR(T - T_t) \ - \ SSR(T) \ + \ a\,(|T - T_t| - |T|), \end{aligned} \tag{3.3}$$

όπου $T - T_t$ είναι το κλαδεμένο δένδρο, αν από το αρχικό δένδρο, έστω T, αφαιρέσουμε το υποδένδρο T_t.

Στην συνέχεια, παρατίθεται το Σχήμα 3.1, το οποίο θα βοηθήσει τον αναγνώστη να ακολουθήσει τους επόμενους συλλογισμούς.



Σχήμα 3.1: Ενδεικτικά δένδρα αποφάσεων για την μέθοδο του κλαδέματος. Το δένδρο T είναι το αρχικό, ολόκληρο δένδρο αποφάσεων, το T_t είναι το υποδένδρο που πρόκειται να κλαδευτεί από το δένδρο T και το δένδρο $T - T_t$ θα είναι το τελικό, κλαδεμένο δένδρο.

Το άθροισμα τετραγώνων των υπολοίπων, λοιπόν, για το κλαδεμένο δένδρο $T - T_t$ θα είναι ίσο με
$SSR(T - T_t) \ = \ SSR(T) \ - \ SSR(T_t) + \ SSR(t)$
και, αντίστοιχα, θα ισχύει ότι
$|T - T_t| \ = \ |T| \ - \ |T_t| \ + 1,$
με το $SSR(t)$ να δηλώνει το άθροισμα των τετραγώνων των υπολοίπων στον κόμβο t.

Επομένως, η σχέση (3.3) παίρνει την μορφή

$$\begin{aligned} SSR(T) - SSR(T_t) + SSR(t) - SSR(T) + a(|T| - |T_t| + 1 - |T|) \ = \\ SSR(t) \ - \ SSR(T_t) \ + \ a\,(1 \ - \ |T_t|). \end{aligned} \tag{3.4}$$

Η σχέση (3.4) λαμβάνει μη αρνητικές τιμές, επομένως ελαχιστοποιείται όταν παίρνει την τιμή 0 και αυτό ισχύει όταν το

$$a \;=\; \frac{SSR(t) - SSR(T_t)}{|T_t| - 1}. \tag{3.5}$$

Επομένως, η ελαχιστοποίηση της διαφοράς
Tree Score$(T - T_t)$ $-$ Tree Score(T) ισοδυναμεί με την ελαχιστοποίηση της ποσότητας (3.5). Συνεπώς, αρχίζοντας με το αρχικό, ολόκληρο δένδρο, έστω T^0, ο αλγόριθμος του κλαδέματος της πολυπλοκότητας του κόστους, σε κάθε επανάληψη s:
▷ Διαλέγει τον κόμβο t που ελαχιστοποιεί την ποσότητα $\frac{SSR(t)-SSR(T_t^{s-1})}{|T_t^{s-1}|-1}$

και, έπειτα
▷ Θέτει $T^s \;=\; T^{s-1} \;-\; T_t$ και $a_s \;=\; \frac{SSR(t)-SSR(T_t^{s-1})}{|T_t^{s-1}|-1}$

μέχρι να φτάσουμε στην ρίζα του δένδρου.

Ως αποτέλεσμα παίρνουμε, όπως αναφέρθηκε προηγουμένως, μία ακολουθία υποδένδρων $T^0 \supseteq T^1 \supseteq ... \supseteq T^n \supseteq ... \supseteq T^r$ συναρτήσει των τιμών της παραμέτρου πολυπλοκότητας $0 = a_0 \leq a_1 \leq ... \leq a_n \leq ...$ με T^r η ρίζα του δένδρου T^0.

3) Τυχαίος διαχωρισμός του αρχικού συνόλου δεδομένων σε σύνολο δεδομένων εκπαίδευσης και σε σύνολο δεδομένων ελέγχου.

4) Στο σύνολο δεδομένων εκπαίδευσης, με βάση τις προηγούμενες τιμές του α, κατασκευάζεται ένα ολόκληρο δένδρο μαζί με μία ακολουθία υποδένδρων που να ελαχιστοποιούν τον αντίστοιχο βαθμό.

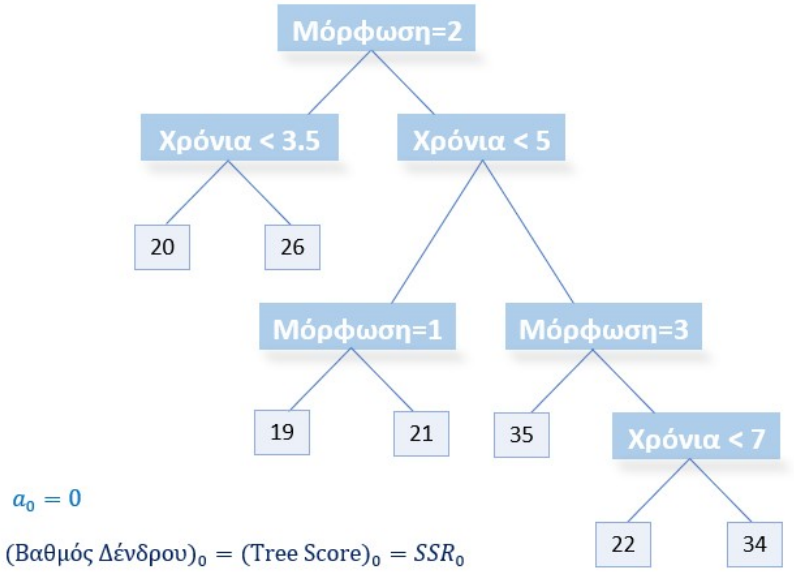
5) Υπολογισμός του SSR για κάθε νέο δένδρο, χρησιμοποιώντας μόνο το σύνολο δεδομένων ελέγχου και επιλογή εκείνου με το μικρότερο SSR.

6) Επανάληψη των βημάτων (3), (4) και (5) μέχρι να επιτευχθεί η διασταυρωτική επικύρωση V ομάδων με το V να καθορίζεται από τον ερευνητή. Η τιμή του α που, κατά μέσο όρο, δίνει το χαμηλότερο SSR με το σύνολο δεδομένων ελέγχου είναι και το τελικό α.

Επιστρέφοντας στα δένδρα που προέκυψαν από όλα τα δεδομένα στο βήμα (2), διαλέγουμε εκείνο που αντιστοιχεί στο α που βρήκαμε κατά το τελευταίο βήμα. Αυτό το (υπο)δένδρο θα είναι το τελικό, κλαδεμένο δένδρο.

Παράδειγμα 3.1.1. Έστω η τυχαία μεταβλητή Y που αντιπροσωπεύει τον ετήσιο μισθό (σε χιλιάδες €) των υπαλλήλων μίας ελληνικής εταιρίας, οι τιμές της οποίας προκύπτουν σύμφωνα με τα χρόνια υπηρεσίας των υπαλλήλων στην συγκεκριμένη εταιρία (μεταβλητή X_1) και το μέγιστο μορφωτικό επίπεδο των υπαλλήλων (μεταβλητή X_2). Η τελευταία είναι μία κατηγορική μεταβλητή με κλάσεις 1 (απολυτήριο λυκείου), 2 (απόφοιτος ΑΕΙ) και 3 (κάτοχος μεταπτυχιακού).

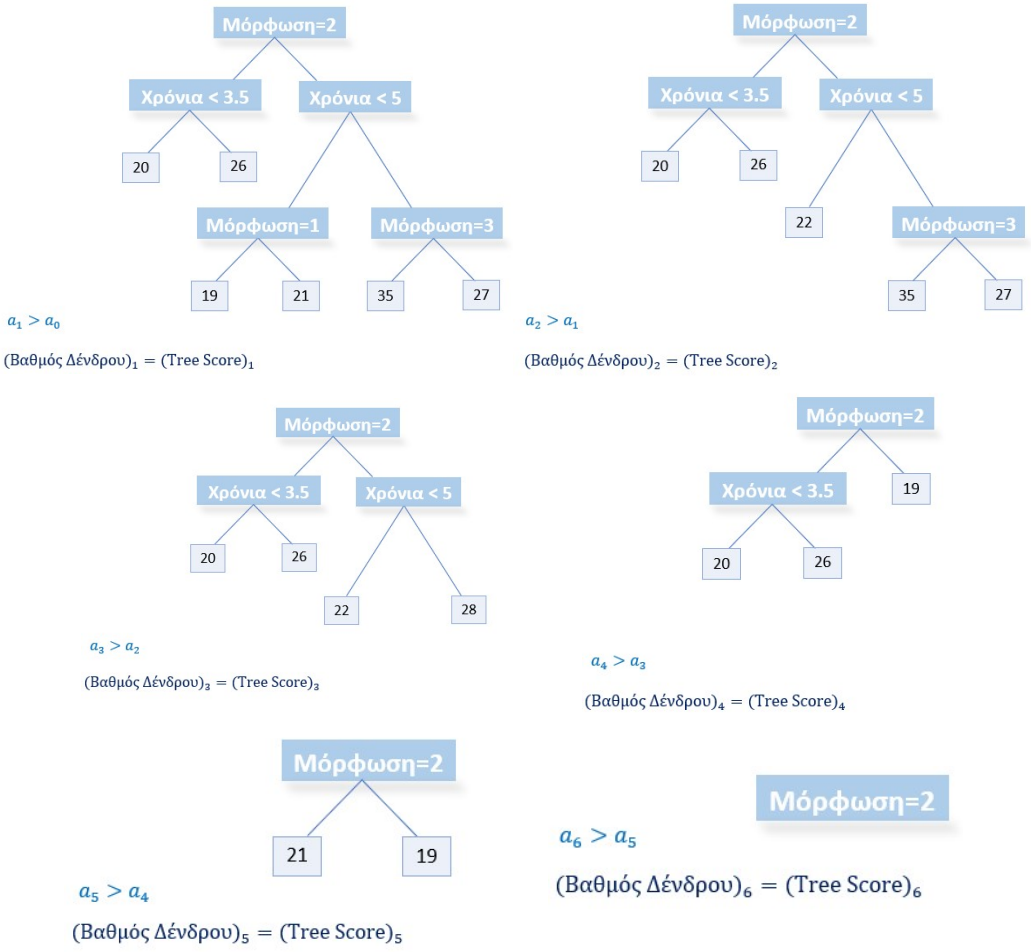
Έστω ότι το δένδρο παλινδρόμησης για το εν λόγω πρόβλημα έχει την μορφή του Σχήματος 3.2 (δένδρο T^0):



Σχήμα 3.2: Δένδρο παλινδρόμησης (T^0) για το παράδειγμα 3.1.1.

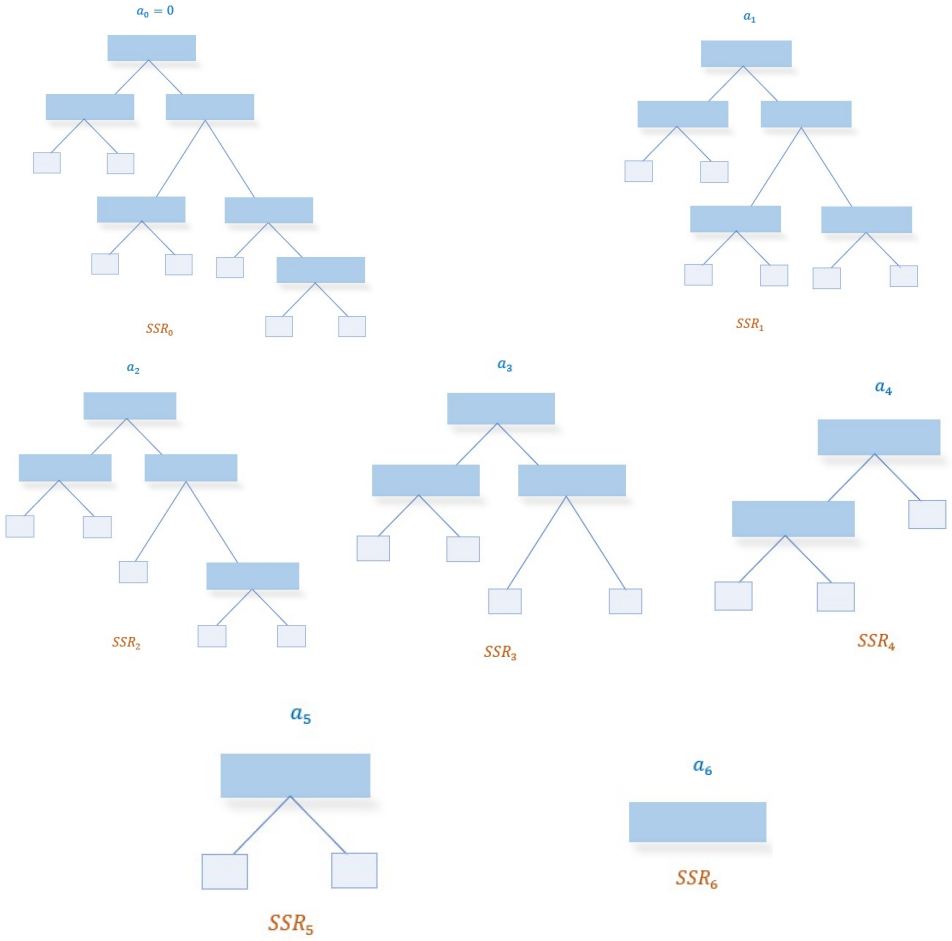
Για αποφυγή πιθανών προβλημάτων υπερπροσαρμογής, προχωράμε σε κλάδεμα του δένδρου. Το αρχικό δένδρο T^0 κατασκευάστηκε από όλα τα διαθέσιμα δεδομένα, μέσω της τεχνικής του Αναδρομικού Δυαδικού Διαχωρισμού και δίνει τον χαμηλότερο βαθμό δένδρου (Tree Score) για $a = 0 := a_0$.

Στην συνέχεια, πραγματοποιείται κλάδεμα του δένδρου T^0 με διαδοχικό τρόπο, όπως περιγράφεται στο βήμα (2) της παραπάνω διαδικασίας του κλαδέματος δένδρων και, έστω ότι προκύπτει η εξής ακολουθία υποδένδρων ως συνάρτηση του α (Σχήμα 3.3), τα οποία προκύπτουν σύμφωνα με τον αλγόριθμο που περιγράφτηκε στο βήμα (2).



Σχήμα 3.3: Ακολουθία υποδένδρων συναρτήσει της παραμέτρου πολυπλοκότητας α.

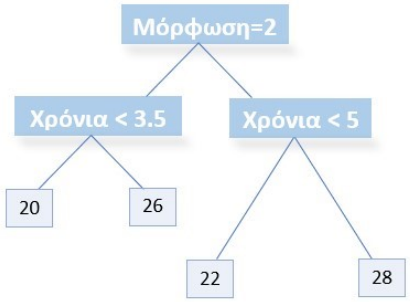
Έπειτα, για τις τιμές του α που βρήκαμε, αφού διαχωρίσουμε τα δεδομένα σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου, γίνεται κατασκευή ενός ολόκληρου δένδρου, καθώς και μίας ακολουθίας υποδένδρων χρησιμοποιώντας μόνο τα δεδομένα εκπαίδευσης. Σε κάθε (υπο)δένδρο, στόχος είναι η ελαχιστοποίηση του αντίστοιχου βαθμού. Ομοίως με πριν, το αρχικό, ολόκληρο δένδρο θα παρουσιάζει τον μικρότερο βαθμό για $a = 0$. Στην συνέχεια, κάνοντας χρήση των δεδομένων ελέγχου, υπολογίζουμε το άθροισμα των τετραγώνων των υπολοίπων (SSR) για κάθε δένδρο (Σχήμα 3.4).



Σχήμα 3.4: Εικονική ακολουθία υποδένδρων χρησιμοποιώντας τα δεδομένα εκπαίδευσης.

Τελικά, διαλέγουμε το α στο οποίο αντιστοιχεί το (υπό)δένδρο με το χαμηλότερο SSR. Ξαναχωρίζουμε το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου και επαναλαμβάνουμε την ίδια διαδικασία με πριν μέχρι να επιτευχθεί η διασταυρωτική επικύρωση V ομάδων με το V να καθορίζεται από τον ερευνητή. Η τιμή του α που, κατά μέσο όρο, δίνει το χαμηλότερο SSR με το σύνολο δεδομένων ελέγχου είναι και το τελικό α. Έστω ότι αυτό το α είναι το a_3.

Συνεπώς, επιστρέφοντας στα δένδρα που προέκυψαν από όλα τα δεδομένα, το (υπό)δένδρο που θα αποτελεί το τελικό, κλαδεμένο δένδρο θα είναι το δένδρο του Σχήματος 3.5.



Σχήμα 3.5: Τελικό, κλαδεμένο δένδρο για το παράδειγμα 3.1.1.

Σε περίπτωση δενδρικής ταξινόμησης, επαναλαμβάνεται η ίδια διαδικασία με μόνη διαφορά την αντικατάσταση του αθροίσματος τετραγώνων των υπολοίπων (SSR) με το μέτρο νοθείας του Gini

$$G \;=\; 1 \;-\; \sum_{k=1}^{K} (\hat{p}_{jk})^2 \;=\; \sum_{k=1}^{K} \hat{p}_{jk}\,(1-\hat{p}_{jk}), \tag{3.6}$$

j=1,...,J ή με οποιαδήποτε άλλη μετρική που χρησιμοποιεί το δένδρο για να αποφασίσει πως θα χωρίσει τα δεδομένα (ποσοστό του σφάλματος ταξινόμησης, εντροπία, κέρδος πληροφορίας).

Γενικά, η μέθοδος του κλαδέματος των δένδρων μπορεί να γίνει και με άλλους τρόπους, πέραν αυτού που αναφέρθηκε παραπάνω, δίχως να χρησιμοποιηθεί η παράμετρος α. Συγκεκριμένα, σε περίπτωση που η εξαρτημένη μεταβλητή είναι κατηγορική, το κλάδεμα του δένδρου ταξινόμησης, ενδεχομένως, να μπορεί να πραγματοποιηθεί, για παράδειγμα, με βάση το κέρδος πληροφορίας (μετρική που αναφέρθηκε στο προηγούμενο κεφάλαιο). Το κλάδεμα αυτής της μορφής αποσκοπεί στην αφαίρεση τμημάτων του δένδρου που προσφέρουν το μικρότερο κέρδος πληροφορίας και στηρίζεται στην πληροφορία που έχει υπολογιστεί, ήδη, από την κατασκευή του δένδρου, βάσει των δεδομένων εκπαίδευσης. Έτσι, τα βήματα που ακολουθούμε είναι τα εξής:

1) Εντοπισμός των κόμβων που καταλήγουν σε φύλλα.

2) Σημείωση του συνολικού αριθμού των φύλλων στο δένδρο.

3) Όσο ο αριθμός των φύλλων στο δένδρο υπερβαίνει τον επιθυμητό αριθμό:

   (α′) Εύρεση του κόμβου (που καταλήγει σε φύλλα) με το μικρότερο κέρδος πληροφορίας.

   (β′) Διαγραφή των φύλλων του εν λόγω κόμβου.

(γ′) Μετατροπή του εν λόγω κόμβου σε φύλλο.

(δ′) Ανανέωση του αριθμού των φύλλων.

Ο επιθυμητός αριθμός των φύλλων σε ένα δένδρο αποφάσεων δηλώνεται από τον χρήστη στην αρχή της διαδικασίας.

Εφόσον, λοιπόν, διευκρινίστηκαν κάποιες ουσιαστικές και εύχρηστες μέθοδοι στην στατιστική, είμαστε σε θέση τώρα να ορίσουμε τον αλγόριθμο κατασκευής δένδρων παλινδρόμησης και ταξινόμησης (αλγόριθμος CART), όπως φαίνεται ακολούθως [7]:

1) Με Αναδρομικό Δυαδικό Διαχωρισμό αναπτύσσεται ένα πλήρες δένδρο, βάσει των δεδομένων εκπαίδευσης. Η διαδικασία διακόπτεται μόνο όταν κάθε τελικός κόμβος περιέχει λιγότερες από έναν καθορισμένο ελάχιστο αριθμό παρατηρήσεων (καθορίζεται από τον χρήστη εκ των προτέρων).

2) Εφαρμογή της τεχνικής του κλαδέματος της πολυπλοκότητας κόστους στο αρχικό δένδρο, έτσι ώστε να ληφθεί μία ακολουθία από τα βέλτιστα υποδένδρα, ως συνάρτηση της παραμέτρου πολυπλοκότητας α.

3) Με διασταυρωτική επικύρωση V ομάδων υπολογίζεται το α. Εν συνεχεία, επιλέγεται εκείνο το α που ελαχιστοποιεί το μέσο σφάλμα πρόβλεψης (ή ταξινόμησης).

4) Το τελικό δένδρο θα είναι το υποδένδρο μεταξύ αυτών του βήματος (2) που αντιστοιχεί στην τιμή του α που επιλέχθηκε.

Ο εν λόγω αλγόριθμος αποσκοπεί στην πρόβλεψη της τιμής μίας ποσοτικής εξαρτημένης μεταβλητής ή στην ταξινόμηση σε κάποια κλάση μίας κατηγορικής εξαρτημένης μεταβλητής. Τα δένδρα αποφάσεων που προκύπτουν διαχειρίζονται μη γραμμικά δεδομένα με αρκετά αποτελεσματικό τρόπο.

Συμπερασματικά, ο αλγόριθμος CART δεν απαιτεί μεγάλη υπολογιστική δύναμη, το οποίο σημαίνει γρήγορη κατασκευή των μοντέλων.

Ωστόσο, ένα σημαντικό μειονέκτημα του συγκεκριμένου αλγορίθμου είναι ότι μία μικρή αλλαγή στα αρχικά δεδομένα πιθανόν να προκαλέσει μεγάλη μεταβολή στο τελικό δένδρο, δηλαδή να προκύψει ένα τελείως διαφορετικό μοντέλο παλινδρόμησης (ή ταξινόμησης). Αυτό συμβαίνει, διότι τα δένδρα αποφάσεων που κατασκευάζει αυτή η μέθοδος παρουσιάζουν αρκετά υψηλή διασπορά με αποτέλεσμα να μην είναι ανεκτικά σε μικρές μεταβολές των δεδομένων. Αυτή η αυξημένη διακύμανση των μοντέλων σημαίνει μία λανθασμένη και, μάλλον, άστοχη τελική πρόβλεψη.

Εν τούτοις, η ένταξη και αξιοποίηση πολλών δένδρων αποφάσεων στο μοντέλο συμβάλλει σε μεγάλο βαθμό στην βελτίωση της προβλεπτικής ικανότητας των δένδρων. Για αυτόν ακριβώς τον λόγο, διάφορες μέθοδοι, γνωστές ως 'Τεχνικές Συνόλου' ("Ensemble Techniques"), καθώς συνδυάζουν πολλαπλά μοντέλα, όπως η Ενσάκιση (Bagging), τα Τυχαία Δάση (Random Forests) και η Ενίσχυση (Boosting) χρησιμοποιούνται ως επεκτάσεις του αλγορίθμου κατασκευής των δένδρων παλινδρόμησης και ταξινόμησης, δημιουργώντας πολλαπλά δένδρα αποφάσεων για ισχυρότερα και πιο ικανά μοντέλα πρόβλεψης.

3.2 Ενσάκιση (Bagging)

Τα δένδρα αποφάσεων, όπως έγινε λόγος νωρίτερα, αποτελούν μοντέλα μεγάλης διασποράς, δηλαδή μικρές αλλαγές στα δεδομένα εκπαίδευσης μπορούν να επιφέρουν διαφορετική εκτίμηση όσον αφορά το τελικό δένδρο παλινδρόμησης (ή ταξινόμησης). Επιπλέον, συχνά παρατηρείται το πρόβλημα της υπερπροσαρμογής των δένδρων (overfitting). Πιο αναλυτικά, τα δένδρα αποφάσεων αδυνατούν να δώσουν τις ίδιες εκτιμήσεις όταν αυτά εφαρμόζονται σε νέα δεδομένα ελέγχου, δηλαδή σε δεδομένα που δεν χρησιμοποιούνται για την κατασκευή αυτών, δηλαδή σε δεδομένα διαφορετικά από τα δεδομένα εκπαίδευσης και, έτσι, λέμε ότι έχουμε πρόβλημα υπερπροσαρμογής.

Μία μέθοδο που μπορούμε να χρησιμοποιήσουμε για την επίλυση των προαναφέροντων προβληματικών καταστάσεων είναι η λεγόμενη **Ενσάκιση** ή, αλλιώς, γνωστή με τον αγγλικό όρο ως **Bagging**.

Η ιδέα της τεχνικής της ενσάκισης αναπτύχθηκε από τον Leo Breiman το 1994 και περιγράφεται ως εξής:

Αρχικά, διαχωρίζουμε το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και σε δεδομένα ελέγχου, με τέτοιον τρόπο, ώστε να είναι περίπου το 60% και το 40% των αρχικών (συνολικών) δεδομένων αντίστοιχα. Έστω ότι το σύνολο των δεδομένων εκπαίδευσης περιέχει n σε πλήθος στοιχεία (παρατηρήσεις). Στην συνέχεια, εφαρμόζοντας δειγματοληψία με επανάθεση (sampling with replacement), παίρνουμε M διαφορετικά υποσύνολα από το σύνολο των δεδομένων εκπαίδευσης (60% των συνολικών δεδομένων), γνωστά και ως M "Bootstrapped" σύνολα δεδομένων με n στοιχεία (παρατηρήσεις) το καθένα. Υπενθυμίζεται στον αναγνώστη ότι η μέθοδος της δειγματοληψίας με επανάθεση είναι η τυχαία επιλογή παρατηρήσεων από το σύνολο δεδομένων που επιτρέπει το ίδιο στοιχείο να επαναλαμβάνεται στο δείγμα.

Τελικά, καταλήγουμε να έχουμε M διαφορετικά "bootstrapped" υποσύνολα δεδομένων. Έπειτα, εφαρμόζοντας τον αλγόριθμο CART στο i-οστό υποσύνολο - σάκο και κάνοντας χρήση μίας παρατήρησης από το δείγμα ελέγχου, λαμ-

βάνουμε την εκτίμηση $\hat{Y}_i$ για $i = 1, ..., M$ αντίστοιχα. Τέλος, επαναλαμβάνουμε την διαδικασία για όλες τις παρατηρήσεις του δείγματος ελέγχου.

Τελικά, η ζητούμενη εκτίμηση θα προκύπτει από τον μέσο όρο όλων των εκτιμήσεων, δηλαδή θα είναι ίση με

$$\hat{Y}_{bag} = \frac{1}{M} \sum_{i=1}^{M} \hat{Y}_i \tag{3.7}$$

για πρόβλεψη της τιμής μίας ποσοτικής μεταβλητής Y.

Για αυτόν τον λόγο, η ενσάκιση (όπως και τα τυχαία δάση που θα αναλυθούν στην συνέχεια) μπορεί να αναφερθεί και με την ονομασία "Bootstrap Aggregation".

Αν η μεταβλητή Y είναι κατηγορική, τότε για δοσμένη παρατήρηση από το δείγμα ελέγχου, καταγράφουμε την κλάση που προκύπτει από το i-οστό δένδρο - σάκο ($i = 1, ..., M$) και, τελικά, επιλέγουμε την **ψήφο πλειοψηφίας** (majority vote), δηλαδή ως τελική πρόβλεψη επιλέγουμε εκείνη την κλάση της οποίας η συχνότητα είναι εντονότερη ανάμεσα στις κλάσεις των υπόλοιπων M προβλέψεων.

3.2.1 Εκτίμηση του Out-Of-Bag (OOB) Σφάλματος

Κάθε δένδρο που προκύπτει από την διαδικασία της ενσάκισης χρησιμοποιεί, κατά μέσον όρο, τα 2/3 των παρατηρήσεων του εκάστοτε προβλήματος. Το υπόλοιπο 1/3 των παρατηρήσεων που δεν επιλέγεται όταν δημιουργούνται τα υποσύνολα δεδομένων και, έτσι, δεν αξιοποιείται για την προσαρμογή του μοντέλου αναφέρονται ως **out-of-bag (OOB)** παρατηρήσεις [7].

Το σύνολο αυτό των παρατηρήσεων χρησιμοποιείται για την εκτίμηση της j-οστής παρατήρησης της μεταβλητής απόκρισης με χρήση των δένδρων στα οποία αυτή η παρατήρηση ήταν OOB και, επομένως, προκύπτουν στο σύνολο $\frac{1}{3} M = \frac{M}{3}$ προβλέψεις για την εν λόγω j-οστή παρατήρηση. Παίρνοντας τον μέσο όρο αυτών (σε περίπτωση παλινδρόμησης) ή την ψήφο πλειοψηφίας αυτών (σε περίπτωση ταξινόμησης), λαμβάνεται μία μοναδική (Out-Of-Bag) πρόβλεψη για αυτή την παρατήρηση. Με παρόμοιο τρόπο μπορούμε να λάβουμε αντίστοιχες προβλέψεις για όλες τις n παρατηρήσεις και, στην συνέχεια, να υπολογίσουμε το ολικό Out-Of-Bag μέσο τετραγωνικό σφάλμα για ένα πρόβλημα παλινδρόμησης ή το ολικό Out-Of-Bag σφάλμα ταξινόμησης για ένα πρόβλημα ταξινόμησης.

Συνεπώς, η εκτίμηση του σφάλματος ελέγχου για ένα μοντέλο ενσάκισης ισοδυναμεί με την εκτίμηση του Out-Of-Bag σφάλματος, όπως περιγράφτηκε παραπάνω.

Ο συγκεκριμένος τρόπος εύρεσης του σφάλματος είναι ιδιαίτερα χρήσιμος και βολικός σε περιπτώσεις που καλούμαστε να εφαρμόσουμε τη μέθοδο της ενσάκισης σε υπερβολικά μεγάλα σύνολα δεδομένων, όπου η τακτική της διασταυρωτικής επικύρωσης καθιστάται πλέον αδύνατη.

Παράδειγμα 3.2.1. Θεωρούμε το σύνολο δεδομένων του Σχήματος 3.6, το οποίο αποτελείται από τρεις ανεξάρτητες μεταβλητές (X_1, X_2, X_3) και μία ποσοτική εξαρτημένη μεταβλητή Y. Χωρίζουμε τα δεδομένα σε δεδομένα εκπαίδευσης

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	Y
Δεδομένα εκπαίδευσης	110	5	Ναι	22
	98	7	Ναι	19
	125	12	Όχι	25
	114	9	Ναι	23
Δεδομένα ελέγχου	102	10	Ναι	21
	109	14	Όχι	15

Σχήμα 3.6: Δεδομένα του παραδείγματος 3.2.1.

και δεδομένα ελέγχου σε αναλογία, περίπου, 60% και 40% και κατασκευάζουμε M διαφορετικά υποσύνολα από το σύνολο των δεδομένων εκπαίδευσης με πλήθος παρατηρήσεων όσες είναι και οι παρατηρήσεις που ανήκουν στο σύνολο εκπαίδευσης $(n = 4)$. Έστω ότι $M = 6$. Συνεπώς, εφαρμόζοντας δειγματοληψία με επανάθεση προκύπτουν 6 Bootstrapped σύνολα δεδομένων, τα οποία φαίνονται στο Σχήμα 3.7.

X_1	X_2	X_3	Y
98	7	Ναι	19
114	9	Ναι	23
110	5	Ναι	22
110	5	Ναι	22

X_1	X_2	X_3	Y
125	12	Όχι	25
114	9	Ναι	23
114	9	Ναι	23
98	7	Ναι	19

X_1	X_2	X_3	Y
125	12	Όχι	25
110	5	Ναι	22
125	12	Όχι	25
114	9	Ναι	23

X_1	X_2	X_3	Y
98	7	Ναι	19
125	12	Όχι	25
98	7	Ναι	19
110	5	Ναι	22

X_1	X_2	X_3	Y
114	9	Ναι	23
110	5	Ναι	22
110	5	Ναι	22
98	7	Ναι	19

X_1	X_2	X_3	Y
110	5	Ναι	22
114	9	Ναι	23
98	7	Ναι	19
98	7	Ναι	19

Σχήμα 3.7: Bootstrapped σύνολα δεδομένων για το παράδειγμα 3.2.1.

Στην συνέχεια, με τη βοήθεια του αλγορίθμου CART προσαρμόζεται ένα δένδρο παλινδρόμησης για κάθε υποσύνολο - σάκο, με αποτέλεσμα να προκύψουν 6 δένδρα παλινδρόμησης. Έπειτα, προσαρμόζουμε τα δένδρα αυτά σε κάποια παρατήρηση από τα δεδομένα ελέγχου, έστω την πρώτη, και λαμβάνουμε την εκτίμηση $\hat{Y}_i$ για $i = 1, ..., 6$ αντίστοιχα, όπως φαίνεται στο Σχήμα 3.8.

Τέλος, η ζητούμενη εκτίμηση προκύπτει υπολογίζοντας τον μέσο όρο όλων των εκτιμήσεων και είναι ίση με

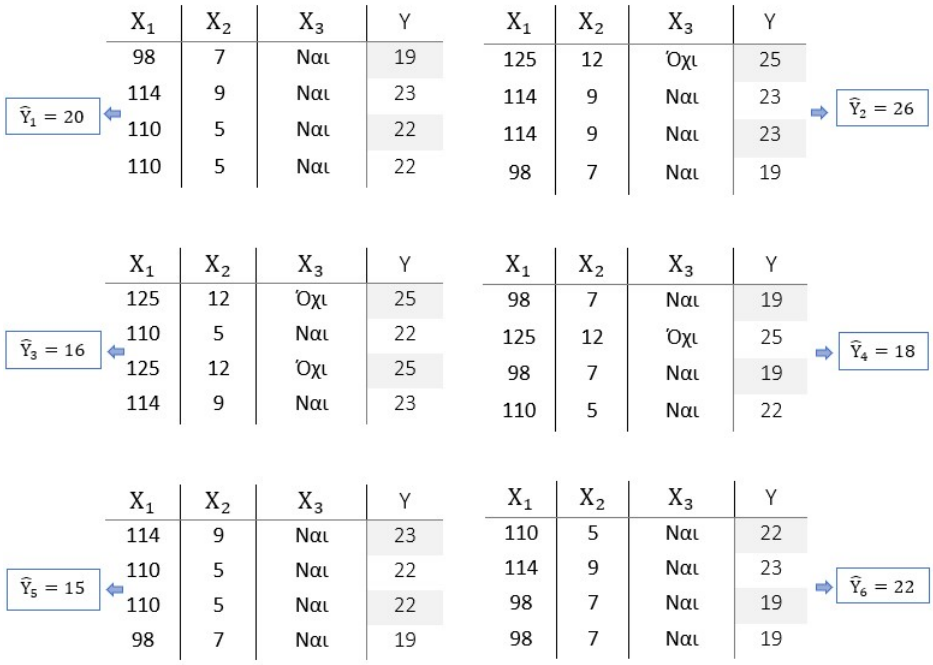
$$\hat{Y}_{bag} = \frac{1}{6}\sum_{i=1}^{6} \hat{Y}_i = \frac{1}{6}(20 + 26 + 16 + 18 + 15 + 22) = 19.5.$$

Για κάθε δένδρο που κατασκευάζεται με βάση το κάθε Bootstrapped δείγμα, το σφάλμα υπολογίζεται από τα αχρησιμοποίητα δείγματα για το εκάστοτε Bootstrapped δείγμα ("Out-Of-Bag" δείγματα). Η μέση τιμή των σφαλμάτων δίνει το Out-Of-Bag σφάλμα. Στην συνέχεια, για λόγους ευκολίας θα αναφερόμαστε στο δένδρο που αντιστοιχεί στο i-οστό Bootstrapped δείγμα ως δένδρο i $(i = 1, ..., 6)$.

• Χρησιμοποιώντας την πρώτη παρατήρηση (του συνόλου εκπαίδευσης) στο δένδρο 2, λαμβάνουμε την εκτίμηση $\hat{Y}_1 = 25$.

• Χρησιμοποιώντας την δεύτερη παρατήρηση (του συνόλου εκπαίδευσης) στο δένδρο 3, λαμβάνουμε την εκτίμηση $\hat{Y}_2 = 18$.

• Χρησιμοποιώντας την τρίτη παρατήρηση (του συνόλου εκπαίδευσης) στα δέν-



Σχήμα 3.8: Εκτιμήσεις των τιμών της εξαρτημένης μεταβλητής του προβλήματος, όπως αυτές προκύπτουν από το παραγόμενο μοντέλο της ενσάκισης.

δρα 1, 5 και 6, λαμβάνουμε αντίστοιχα τις τιμές 27, 29 και 24. Οπότε, η τελική εκτίμηση για το εν λόγω δείγμα θα είναι η μέση τιμή αυτών, δηλαδή $\hat{Y}_3 = 26.7$.
• Χρησιμοποιώντας την τέταρτη παρατήρηση (του συνόλου εκπαίδευσης) στο δένδρο 4, λαμβάνουμε την εκτίμηση $\hat{Y}_4 = 21$.

Επομένως, το Out-Of-Bag σφάλμα του μοντέλου θα είναι το μέσο τετραγωνικό σφάλμα

$$\text{OOB Σφάλμα} = \text{MSE} = \frac{1}{4}\sum_{i=1}^{4}(Y_i - \hat{Y}_i)^2 = \frac{1}{4}\left[(22-25)^2 + (19-18)^2 + (25-26.7)^2 + (23-21)^2\right] = 4.2225.$$

Παράδειγμα 3.2.2. Έστω, τώρα, ότι για το ίδιο σύνολο δεδομένων με αυτό του προηγούμενου παραδείγματος, η εξαρτημένη μεταβλητή Y είναι κατηγορική με δύο κλάσεις ('Ναι', 'Όχι'), όπως φαίνεται στον πίνακα του Σχήματος 3.9.

Ομοίως με πριν, διαχωρίζουμε τα δεδομένα σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου και κατασκευάζουμε M Bootstrapped σύνολα δεδομένων, μέσω δειγματοληψίας με επανάθεση (έστω και εδώ ότι $M = 6$). Με τη βοήθεια του αλγορίθμου CART προσαρμόζεται ένα δένδρο ταξινόμησης για κάθε

	X_1	X_2	X_3	Y
Δεδομένα εκπαίδευσης	110	5	Ναι	Ναι
	98	7	Ναι	Όχι
	125	12	Όχι	Όχι
	114	9	Ναι	Όχι
Δεδομένα ελέγχου	102	10	Ναι	Ναι
	109	14	Όχι	Ναι

Σχήμα 3.9: Δεδομένα του παραδείγματος 3.2.2.

υποσύνολο - σάκο, με αποτέλεσμα να προκύψει ένα μοντέλο με 6 δένδρα ταξινόμησης. Στην συνέχεια, εφαρμόζουμε το μοντέλο αυτό στην πρώτη παρατήρηση από τα δεδομένα ελέγχου και λαμβάνουμε την πρόβλεψη $\hat{Y}_i$ για $i = 1, ..., 6$ αντίστοιχα (Σχήμα 3.10).

Ως τελική πρόβλεψη για το εν λόγω δείγμα παίρνουμε την ψήφο πλειοψηφίας μεταξύ των προβέψεων $\hat{Y}_i$ $(i = 1, ..., 6)$, συνεπώς η μεταβλητή απόκρισης του δείγματος θα ταξινομηθεί στην κλάση 'Ναι'.

Ομοίως με πριν, για κάθε δένδρο που κατασκευάζεται με βάση το κάθε Bootstrapped δείγμα, το σφάλμα υπολογίζεται από τα αχρησιμοποίητα δείγματα για το εκάστοτε Bootstrapped δείγμα. Το ποσοστό των Out-Of-Bag δειγμάτων που ταξινομήθηκαν λανθασμένα είναι το Out-Of-Bag σφάλμα. Για λόγους ευκολίας και σε αυτήν την περίπτωση, θα αναφερόμαστε στο δένδρο που αντιστοιχεί στο i-οστό Bootstrapped δείγμα ως δένδρο i $(i = 1, ..., 6)$.

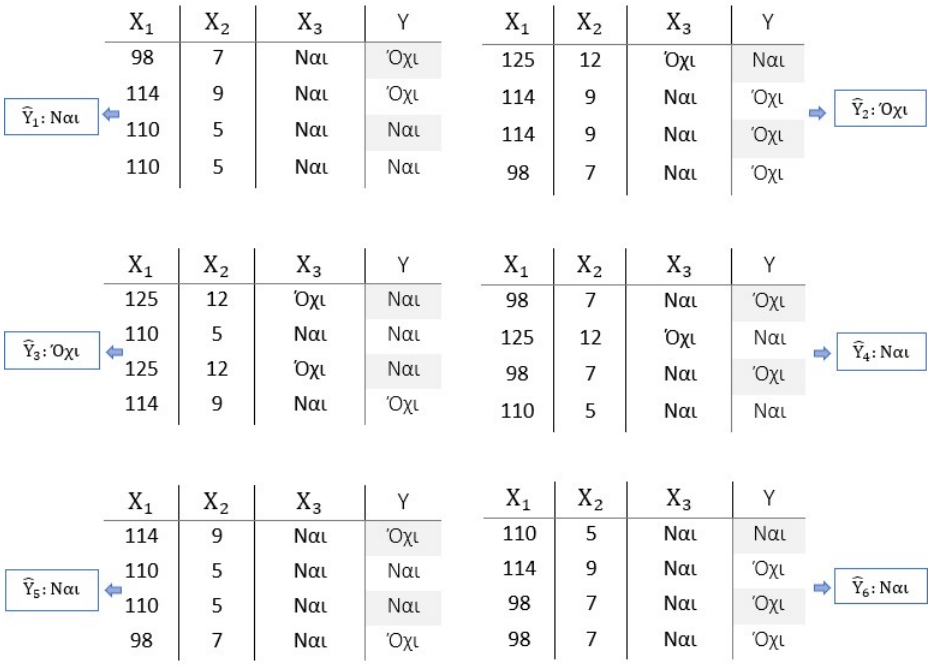
• Χρησιμοποιώντας την πρώτη παρατήρηση (του συνόλου εκπαίδευσης) στο δένδρο 2, λαμβάνουμε την πρόβλεψη $\hat{Y}_1$: 'Ναι'.

• Χρησιμοποιώντας την δεύτερη παρατήρηση (του συνόλου εκπαίδευσης) στο δένδρο 3, λαμβάνουμε την πρόβλεψη $\hat{Y}_2$: 'Ναι'.

• Χρησιμοποιώντας την τρίτη παρατήρηση (του συνόλου εκπαίδευσης) στα δένδρα 1, 5 και 6, λαμβάνουμε αντίστοιχα τις προβλέψεις 'Όχι', 'Όχι' και 'Ναι'. Συνεπώς, η τελική πρόβλεψη για το εν λόγω δείγμα θα είναι η ψήφος πλειοψηφίας ανάμεσα σε αυτές, δηλαδή $\hat{Y}_3$: 'Όχι'.

• Χρησιμοποιώντας την τέταρτη παρατήρηση (του συνόλου εκπαίδευσης) στο δένδρο 4, λαμβάνουμε την πρόβλεψη $\hat{Y}_4$: 'Ναι'.

Επομένως, το Out-Of-Bag σφάλμα του μοντέλου θα είναι το ποσοστό των λανθασμένα ταξινομημένων δειγμάτων, δηλαδή θα είναι ίσο με:



Σχήμα 3.10: Προβλέψεις των τιμών της εξαρτημένης μεταβλητής του προβλήματος, όπως αυτές προκύπτουν από το παραγόμενο μοντέλο της ενσάκισης.

$$\text{OOB Σφάλμα} = \frac{2}{4} = 0.5.$$

Όπως είδαμε παραπάνω, η μέθοδος της ενσάκισης χρησιμοποιεί έναν μεγάλο αριθμό δένδρων προκειμένου να καταλήξει σε μία τελική πρόβλεψη για το εκάστοτε πρόβλημα. Είναι αδύνατον, βασιζόμενοι σε αυτή την μέθοδο, να εξάγουμε κάποιο συμπέρασμα, δίχως να κατασκευάσουμε τουλάχιστον δύο δένδρα αποφάσεων. Το γεγονός αυτό δυσχεραίνει σε μεγάλο βαθμό την ερμηνεία των αποτελεσμάτων, καθώς δεν υπάρχει πλέον η πολυτέλεια του ενός μόνο δενδρικού διαγράμματος, όπως υπήρχε στις προηγούμενες ενότητες, το οποίο απλοποιούσε σημαντικά την κατανόηση και την εξήγηση των συλλογισμών που ακολουθήθηκαν μέχρι να προκύψει η τελική εκτίμηση σε κάποιον τρίτο.

Επομένως, μπορεί η μέθοδος της ενσάκισης να βελτιώνει την προβλεπτική ικανότητα και την ακρίβεια του μοντέλου, αλλά, ταυτόχρονα, υστερεί στην ερμηνεία.

3.3 Τυχαία Δάση (Random Forests)

Τα **Τυχαία Δάση (Random Forests)** αποτελούν μία βελτίωση στην μεθοδολογία της ενσάκισης και στηρίζονται περίπου στην ίδια γενική ιδέα.

Έστω ότι συνολικά υπάρχουν p χαρακτηριστικά, δηλαδή p ανεξάρτητες μεταβλητές. Όπως στην ενσάκιση, έτσι και στα τυχαία δάση, τα δεδομένα εκπαίδευσης διαιρούνται σε M διαφορετικά υποσύνολα γνωστά και ως "Bootstrapped" σύνολα δεδομένων με n στοιχεία το καθένα, βάσει των οποίων κατασκευάζονται M σε πλήθος δένδρα αποφάσεων. Πριν από κάθε τμήση, ωστόσο, επιλέγεται τυχαίο δείγμα m χαρακτηριστικών (ανεξάρτητων μεταβλητών) ως υποψήφιες μεταβλητές τμήσης από τις οποίες μόνο μία, τελικά, θα χρησιμοποιηθεί για την συγκεκριμένη τμήση. Το τυχαίο αυτό δείγμα των m χαρακτηριστικών επιλέγεται από το ολικό σύνολο των p χαρακτηριστικών. Συνήθως, το m προτιμάται να είναι περίπου ίσο με την τετραγωνική ρίζα του συνολικού αριθμού των χαρακτηριστικών του προβλήματος, δηλαδή, συνήθως, ισχύει ότι $m \approx \sqrt{p}$. Σημειώνεται εδώ ότι ένα χαρακτηριστικό μπορεί να εμφανίζεται περισσότερες από μία φορές στο δένδρο, καθώς υπάρχει η δυνατότητα επιλογής του ίδιου χαρακτηριστικού όταν διαλέγεται το τυχαίο δείγμα από τον συνολικό αριθμό χαρακτηριστικών p, ακόμη και αν αυτό έχει ξαναχρησιμοποιηθεί σε προηγούμενο δείγμα. Ακολουθώντας την μεθολογία που μόλις αναφέρθηκε, δημιουργείται σταδιακά ένα τυχαίο δάσος αποτελούμενο από πολλαπλά και διαφορετικά μεταξύ τους δένδρα αποφάσεων.

Αντιθέτως, στην μέθοδο της ενσάκισης, ως m θεωρείται το σύνολο όλων των χαρακτηριστικών, δηλαδή έχουμε ότι $m = p$. Έτσι, σε κάθε δένδρο, προτού πραγματοποιηθεί μία τμήση, λαμβάνονται υπόψιν όλες οι ανεξάρτητες μεταβλητές και προτιμάται η ισχυρότερη, δηλαδή εκείνη που δίνει τα καλύτερα μέτρα νοθείας (αναλύθηκαν λεπτομερώς στο Κεφάλαιο 2). Συνεπώς, στην ενσάκιση υπάρχει μία συσχέτιση μεταξύ των δένδρων αποφάσεων που κατασκευάζονται και αυτό, διότι πάντα θα επιλέγεται ως μεταβλητή τμήσης η πιο ισχυρή μεταξύ των υπολοίπων.

Από την άλλη πλευρά, στα τυχαία δάση, το τυχαίο δείγμα που επιλέγεται δεν αφορά όλα τα χαρακτηριστικά, επομένως μεταβλητή τμήσης δεν είναι πάντα η ισχυρότερη όλων των χαρακτηριστικών, αλλά η ισχυρότερη μεταξύ εκείνων που υπάρχουν στο εκάστοτε τυχαίο δείγμα. Άρα, η μέθοδος αυτή 'ευνοεί' την απουσία ή τη μείωση της συσχέτισης στα δένδρα αποφάσεων που προκύπτουν σε μεγάλο βαθμό.

Ιδανικά, τα δένδρα αποφάσεων θα πρέπει να είναι ασυσχέτιστα μεταξύ τους, έτσι ώστε να εμφανίζεται μία αισθητή μείωση στην διασπορά. Όταν υπολογίζεται ο μέσος όρος πολλών ποσοτήτων, οι οποίες σχετίζονται μεταξύ τους, αυτό δεν συμβαίνει. Για αυτόν ακριβώς τον λόγο, τα τυχαία δάση υπερτερούν έναντι της ενσάκισης, καθώς δεν λογαριάζουν πάντα την ισχυρότερη μεταβλητή όταν έρθει

η στιγμή της τμήσης.

Όταν το πρόβλημα είναι πρόβλημα ταξινόμησης, κάθε δένδρο 'ψηφίζει' για μία κλάση. Στην συνέχεια, το διαμορφώμενο δάσος 'διαλέγει' αυτή με τις περισσότερες ψήφους μεταξύ όλων των δένδρων στο δάσος (η γνωστή ψήφος πλειοψηφίας).

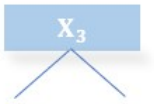
Παράδειγμα 3.3.1. Θεωρούμε το παρακάτω σύνολο δεδομένων (Σχήμα 3.11). Όπως και στην μέθοδο της ενσάκισης, έτσι και στα τυχαία δάση, θα ασχοληθούμε πρώτα με τα δεδομένα εκπαίδευσης. Εφαρμόζοντας δειγματοληψία με επανάθεση, κατασκευάζονται, έστω $M = 6$, Bootstrapped σύνολα δεδομένων στα οποία προσαρμόζονται τα αντίστοιχα δένδρα αποφάσεων. Ωστόσο, μία θεμελιώδης διαφορά μεταξύ των δύο μεθόδων είναι, όπως αναφέρθηκε, ότι στα τυχαία δάση, πριν από κάθε τμήση, επιλέγεται ένα τυχαίο δείγμα m χαρακτηριστικών ως υποψήφιες μεταβλητές τμήσης. Για το συγκεκριμένο παράδειγμα ισχύει ότι $m \approx \sqrt{p} = \sqrt{4} = 2$.

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	Y
Δεδομένα εκπαίδευσης	110	5	Ναι	Όχι	22
	98	7	Ναι	Ναι	19
	125	12	Όχι	Ναι	25
	114	9	Ναι	Όχι	23
Δεδομένα ελέγχου	102	10	Ναι	Ναι	21
	109	14	Όχι	Ναι	15

Σχήμα 3.11: Δεδομένα του παραδείγματος 3.3.1.

Όπως δηλώνει το Σχήμα 3.12, για το πρώτο Bootstrapped υποσύνολο, επιλέγονται τυχαία οι μεταβλητές X_2 και X_3 από το σύνολο των 4 επεξηγηματικών μεταβλητών ως υποψήφιες μεταβλητές τμήσεις για την ρίζα του δένδρου. Τελικά, έστω ότι η μεταβλητή X_3 δίνει μικρότερο SSR από αυτό της X_2 (μέτρο νοθείας για την δενδρική παλινδρόμηση), οπότε θα είναι και η μεταβλητή που θα χρησιμοποιηθεί για την ρίζα του δένδρου για αυτό το Bootstrapped υποσύνολο. Προχωρώντας στην επόμενη τμήση, επαναλαμβάνεται η ίδια διαδικασία, επιλέγοντας τυχαία δύο μεταβλητές (Σχήμα 3.13), πέρα της X_3 και συγκρίνοντας τα αντίστοιχα μέτρα νοθείας τους.

X_1	X_2	X_3	X_4	Y
98	7	Ναι	Ναι	19
114	9	Ναι	Όχι	23
110	5	Ναι	Όχι	22
110	5	Ναι	Όχι	22

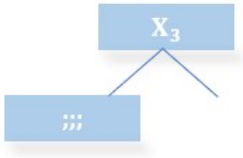


Σχήμα 3.12: Επιλογή των μεταβλητών X_2 και X_3 από το πρώτο Bootstrapped υποσύνολο και χρήση της X_3 στην ρίζα του δένδρου.

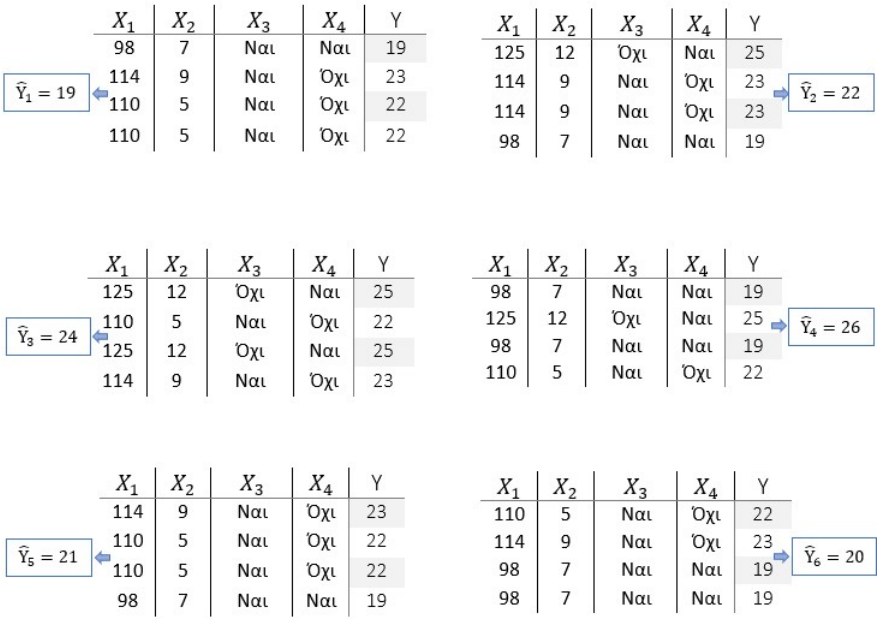
Εκείνη με το καλύτερο μέτρο νοθείας (σε επίπεδο ακρίβειας) επιλέγεται ως μεταβλητή τμήσης για την συγκεκριμένη τμήση.

Με αυτόν τον τρόπο, κατασκευάζονται διαδοχικά και τα 6 Bootstrapped υποσύνολα μαζί με τα αντίστοιχα δένδρα παλινδρόμησης. Με εφαρμογή της πρώτης παρατήρησης του συνόλου ελέγχου στο εν λόγω τυχαίο δάσος, προκύπτουν οι αντίστοιχες προβλέψεις, όπως φαίνεται στο Σχήμα 3.14.

X_1	X_2	X_3	X_4	Y
98	7	Ναι	Ναι	19
114	9	Ναι	Όχι	23
110	5	Ναι	Όχι	22
110	5	Ναι	Όχι	22



Σχήμα 3.13: Επιλογή των μεταβλητών X_1 και X_2 από το πρώτο Bootstrapped υποσύνολο για την απόφαση του επόμενου ορίου.



Σχήμα 3.14: Εκτιμήσεις των τιμών της εξαρτημένης μεταβλητής του προβλήματος, όπως αυτές προκύπτουν από το παραγόμενο μοντέλο των τυχαίων δασών.

Ως τελική εκτίμηση, όπως και στην ενσάκιση, βρίσκουμε τον μέσο όρο όλων των εκτιμήσεων που προέκυψαν για το εν λόγω δείγμα.

$$\hat{Y}_{rf} = \frac{1}{6}\sum_{i=1}^{6} \hat{Y}_i = \frac{1}{6}(19 + 22 + 24 + 26 + 21 + 20) = 22.$$

Ο βαθμός του σφάλματος, στα δεδομένα εκπαίδευσης, ενός τυχαίου δάσους υπολογίζεται με τον ίδιο τρόπο, όπως στην μέθοδο της ενσάκισης, κάνοντας χρήση των out-of-bag παρατηρήσεων, ώστε να βρεθεί το out-of-bag μέσο τετραγωνικό σφάλμα για ένα πρόβλημα παλινδρόμησης ή το ολικό out-of-bag σφάλμα ταξινόμησης για ένα πρόβλημα ταξινόμησης.

Αξίζει να σημειωθεί ότι στα τυχαία δάση δεν γίνεται κλάδεμα των δένδρων αποφάσεων. Το μέγεθος του δένδρου φθάνει στο μεγαλύτερο δυνατό. Επίσης, ισχύουν τα επόμενα πορίσματα [8]:

► Αύξηση στην συσχέτιση μεταξύ δύο οποιονδήποτε δένδρων στο δάσος οδηγεί σε αύξηση του βαθμού του σφάλματος του δάσους.

► Ένα δένδρο με χαμηλό βαθμό σφάλματος αποτελεί έναν ισχυρό ταξινομητή.

► Αύξηση της ισχύς των δένδρων στο δάσος προκαλεί μείωση στον βαθμό σφάλματος του δάσους.

► Μείωση του αριθμού m των στοιχείων (χαρακτηριστικών) που περιέχει ένα τυχαίο δείγμα επιφέρει μείωση στην συσχέτιση και την ισχύ.

► Αύξηση του αριθμού m των στοιχείων (χαρακτηριστικών) που περιέχει ένα τυχαίο δείγμα έχει ως αποτέλεσμα την αύξηση σε συσχέτιση και ισχύ.

3.4 Ενίσχυση (Boosting)

Η μέθοδος της **Ενίσχυσης (Boosting)**, όπως αυτή ονομάζεται, προτιμάται πολλές φορές από τα τυχαία δάση, διότι χρησιμοποιεί δένδρα μικρότερου μεγέθους, τα οποία επαρκούν για τις ανάγκες παραγωγής των απαραίτητων προβλέψεων. Το γεγονός αυτό καθιστά την ενίσχυση ιδανική μέθοδο για την δενδρική παλινδρόμηση, όπως και για την δενδρική ταξινόμηση, καθώς παρατηρείται σημαντική μείωση στην πολυπλοκότητα και εντυπωσιακή διευκόλυνση στην ερμηνευτική προσέγγιση.

Ο πρώτος αλγόριθμος που αναπτύχθηκε επάνω στην μέθοδο της ενίσχυσης είναι γνωστός ως αλγόριθμος "AdaBoost", ο οποίος σχεδιάστηκε από τους Yoav Freund και Robert Schapire και αποτέλεσε αξιέπαινο λόγο βράβευσης αυτών το 2003 [11]. Στην συνέχεια, βασιζόμενος σε κάποια χαρακτηριστικά του AdaBoost, ο περίφημος Jerome H. Friedman κατασκεύασε έναν ισχυρότερο και πιο αποτελεσματικό αλγόριθμο, το λεγόμενο "Gradient Boosting" [12]. Στους αλγορίθμους αυτούς, κάθε μοντέλο (δένδρο αποφάσεων) επιχειρεί να προβλέψει το σφάλμα του αμέσως προηγούμενου μοντέλου και να καταλήξει σε μία ακριβέστερη, πιο εύστοχη τελική εκτίμηση σε σχέση με το προηγούμενο.

3.4.1 AdaBoost

Σε ένα δάσος από δένδρα που δημιουργήθηκαν μέσω του AdaBoost, τα δένδρα αποτελούνται από έναν μοναδικό κόμβο με δύο φύλλα (με δύο τελικούς κόμβους). Αυτά τα είδη δένδρων αποφάσεων είναι γνωστά ως 'κούτσουρα' ή υιοθετώντας την αγγλική ορολογία "stumps". Τα τελευταία χαρακτηρίζονται και ως 'αδύναμοι ταξινομητές' ("weak learners"), διότι, σε αντίθεση με την περίπτωση των τυχαίων δασών, κάνουν χρήση μίας μόνο μεταβλητής (χαρακτηριστικού), προκειμένου να λάβουν μία απόφαση ταξινόμησης. Τα βασικά γνωρίσματα του AdaBoost αλγορίθμου είναι τα κάτωθι:

✓ Κάποιοι αδύναμοι ταξινομητές έχουν μεγαλύτερη επίδραση στην τελική ταξινόμηση από κάποιους άλλους. Όπως θα δούμε στην συνέχεια, αυτό εξαρ-

τάται από το πόσο καλή είναι η ταξινόμηση των στοιχείων (παρατηρήσεων), υπολογίζοντας το ολικό σφάλμα του κάθε ταξινομητή.

✓ Σε αντίθεση με τα τυχαία δάση, όπου η σειρά των δένδρων στο δάσος δεν έχει κάποια ιδιαίτερη σημασία, στο AdaBoost, η σειρά προτεραιότητας των δένδρων αποφάσεων (stumps) παίζει σημαντικό ρόλο και πρέπει να λαμβάνεται σοβαρά υπόψη. Τα σφάλματα που προκύπτουν από τον πρώτο (αδύναμο) ταξινομητή επηρεάζουν τον τρόπο κατασκευής του δεύτερου ταξινομητή, τα σφάλματα του οποίου επηρεάζουν τον τρόπο κατασκευής του τρίτου ταξινομητή, κλπ.

• Όσον αφορά τα στατιστικά προβλήματα ταξινόμησης, ο αλγόριθμος του AdaBoost προτιμάται και είναι αποτελεσματικότερος όταν εφαρμόζεται σε προβλήματα δυαδικής ταξινόμησης (binary classification), καθώς τέτοιου είδους δένδρα είναι σχετικά μικρά σε μέγεθος και περιέχουν μία μόνο απόφαση για ταξινόμηση. Για αυτόν ακριβώς τον λόγο, ο αλγόριθμος του AdaBoost χρησιμοποιεί "stumps" και όχι ολόκληρα δένδρα αποφάσεων.

Με ποιον τρόπο, λοιπόν, γίνεται η κατασκευή των αδύναμων ταξινομητών και πώς υπολογίζεται το ολικό σφάλμα αυτών;

Αρχικά, για κάθε διαθέσιμη παρατήρηση υπολογίζεται το αντίστοιχο βάρος (weight), το οποίο είναι ίσο με τη μονάδα διαιρεμένη προς τον συνολικό αριθμό των παρατηρήσεων. Έτσι, στο ξεκίνημα του αλγορίθμου, τα βάρη είναι τα ίδια για όλα τα στοιχεία (παρατηρήσεις) και πρέπει να αθροίζουν στην μονάδα. Στην συνέχεια, κατασκευάζονται όλοι οι δυνατοί ταξινομητές για το κάθε χαρακτηριστικό (επεξηγηματική μεταβλητή) του προβλήματος και, ύστερα, επιλέγεται ο καλύτερος από αυτούς κάνοντας χρήση κάποιου από τα μέτρα νοθείας ταξινόμησης (αναλύθηκαν εις βάθος στο Κεφάλαιο 2). Αφού βρεθεί ο κατάλληλος ταξινομητής για την πρώτη επανάληψη, υπολογίζεται το ολικό σφάλμα αυτού.

Το ολικό σφάλμα (Total Error) για έναν ταξινομητή είναι το άθροισμα των βαρών εκείνων των παρατηρήσεων από το σύνολο εκπαίδευσης που έχουν ταξινομηθεί λανθασμένα. Παίρνει τιμές μεταξύ του μηδενός και του ένα με μηδενικό ολικό σφάλμα να δηλώνει ότι ο αντίστοιχος ταξινομητής είναι ο ιδανικός, ενώ με ολικό σφάλμα ίσο με την μονάδα να αντιστοιχεί σε έναν αδύναμο, σχεδόν απορριπτέο ταξινομητή.

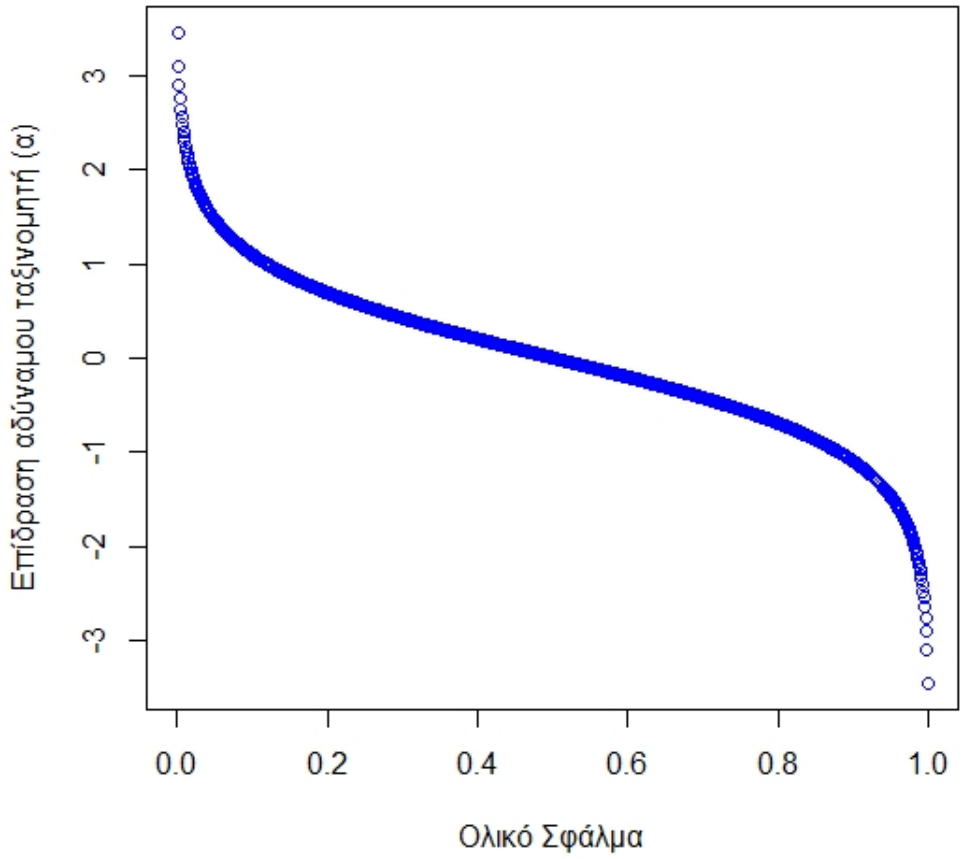
Έχοντας, λοιπόν, ως στόχο την εκτίμηση του ολικού σφάλματος της κάθε παρατήρησης, κατασκευάζεται ο αντίστοιχος ταξινομητής, ο οποίος περιλαμβάνει μία μόνο επεξηγηματική μεταβλητή, εκείνη που κρίθηκε ως η πιο κατάλληλη για την αντίστοιχη τιμήση.

Η επίδραση του κάθε ταξινομητή, συμβολίζεται με α, και περιγράφεται από την επόμενη σχέση, η οποία δίνει τον βαθμό σημαντικότητας του ταξινομητή, δηλαδή ένα μέτρο του κατά πόσο επηρεάζει τελικά την συνολική εκτίμηση ή

ταξινόμηση:

$$a \;=\; \frac{1}{2}\log\left(\frac{1-\text{Total Error}}{\text{Total Error}}\right). \tag{3.8}$$

Στο Σχήμα 3.15 που ακολουθεί φαίνεται αυτή η επίδραση του εκάστοτε ταξινομητή σαν συνάρτηση του ολικού σφάλματος αυτού που, όπως προκύπτει, είναι μία φθίνουσα συνάρτηση:



Σχήμα 3.15: Επίδραση αδύναμου ταξινομητή συναρτήσει του ολικού σφάλματος.

Παρατηρούμε ότι όταν το ολικό σφάλμα μηδενίζεται, δηλαδή ο ταξινομητής είναι άριστος, η επίδραση που έχει στην τελική ταξινόμηση είναι η μεγαλύτερη δυνατή. Στην αντίθετη περίπτωση που το ολικό σφάλμα είναι ίσο με την μονάδα,

ο ταξινομητής έχει πολύ κακή απόδοση και, άρα η επίδρασή του θα είναι πολύ μικρή (αρνητικός αριθμός).

Σε κάθε επανάληψη, υπολογίζεται το νέο βάρος της κάθε παρατήρησης από το σύνολο εκπαίδευσης, σύμφωνα με την σχέση:

$$\text{Weight} \;=\; \text{weight}(0)\; e^{\pm a}, \tag{3.9}$$

όπου η αρνητική επίδραση (-α) είναι για τις παρατηρήσεις που έχουν ταξινομηθεί σωστά, ενώ η θετική επίδραση (+α) είναι για τις παρατηρήσεις που έχουν ταξινομηθεί εσφαλμένα και με weight(0) συμβολίζεται το αρχικό βάρος με το οποίο ξεκινήσαμε που είναι το ίδιο για κάθε παρατήρηση.

Με άλλα λόγια, οι παρατηρήσεις που έχουν ταξινομηθεί σωστά λαμβάνουν λιγότερο βάρος στην επόμενη επανάληψη, ενώ οι παρατηρήσεις που έχουν ταξινομηθεί εσφαλμένα λαμβάνουν περισσότερο βάρος στην επόμενη επανάληψη, προκειμένου ο αλγόριθμος να καταφέρει να βρει έναν ταξινομητή για τις παρατηρήσεις που είναι πιο δύσκολο να ταξινομηθούν.

Αφού υπολογιστεί το νέο βάρος για την κάθε παρατήρηση του συνόλου εκπαίδευσης, καταγράφεται το κανονικοποιημένο βάρος (Normalized weight) της i-οστής παρατήρησης του συνόλου εκπαίδευσης που είναι ίσο με το βάρος της εν λόγω i-οστής παρατήρησης διαιρεμένο προς το συνολικό άθροισμα των βαρών όλων των διαθέσιμων παρατηρήσεων εκπαίδευσης, έστω N σε πλήθος ($i = 1, ..., N$). Τα κανονικοποιημένα βάρη αθροίζουν προσεγγιστικά στην μονάδα και αποτελούν, πλέον, τα νέα βάρη των παρατηρήσεων εκπαίδευσης βάσει των οποίων θα κατασκευαστούν οι επόμενοι ταξινομητές.

Είναι σημαντικό να αναφερθεί ότι σε κάθε επανάληψη τα βάρη πρέπει να ανανεώνονται, διαφορετικά η τελική πρόβλεψη θα είναι ίδια με αυτή του αρχικού μοντέλου.

Για τη δημιουργία του επόμενου αδύναμου ταξινομητή, διαμορφώνουμε ένα καινούργιο σύνολο δεδομένων, το οποίο θα περιέχει τουλάχιστον μία φορά τις παρατηρήσεις με τα μεγαλύτερα βάρη, δηλαδή τις παρατηρήσεις που είχαν ταξινομηθεί λανθασμένα από τον αμέσως προηγούμενο ταξινομητή. Τα σύνολα αυτά έχουν ίδιο μέγεθος με το αρχικό σύνολο δεδομένων και, γενικά, προκύπτουν ως εξής: Επιλέγοντας κάθε φορά τυχαία έναν αριθμό, έστω ν, μεταξύ του μηδενός και του ένα, ο αλγόριθμος 'τρέχει' N φορές, προκειμένου να διαλέξει N διαφορετικές παρατηρήσεις από το (αμέσως) προηγούμενο σύνολο δεδομένων. Με άλλα λόγια, το (αμέσως) προηγούμενο σύνολο δεδομένων χωρίζεται σε 'κουβάδες' ("buckets") και βλέποντας σε ποιον κουβά ανήκει ο ν, κατασκευάζουμε το νέο σύνολο δεδομένων. Συγκεκριμένα, για την διευκόλυνση του αναγνώστη ορίζουμε, χωρίς βλάβη της γενικότητας, ως $w_1, w_2, ..., w_N$ το βάρος της πρώτης, δεύτερης, ..., N-οστής παρατήρησης από το σύνολο εκπαίδευσης, αντίστοιχα, του αρχικού συνόλου δεδομένων. Αν ο αριθμός ν που διαλέξαμε βρίσκεται μεταξύ του 0 και του w_1, τότε προσθέτουμε την πρώτη παρατήρηση στο καινούργιο

σύνολο δεδομένων. Αν ο αριθμός ν βρίσκεται μεταξύ του w_1 και του $w_1 + w_2$, τότε προσθέτουμε την δεύτερη παρατήρηση στο καινούργιο σύνολο δεδομένων. Αν ο αριθμός ν βρίσκεται μεταξύ του $w_1 + w_2$ και του $w_1 + w_2 + w_3$, τότε προσθέτουμε την τρίτη παρατήρηση στο καινούργιο σύνολο δεδομένων, κλπ. Συνεχίζουμε να διαλέγουμε κάθε φορά έναν διαφορετικό αριθμό ν μεταξύ του μηδενός και του ένα έως ότου το καινούργιο σύνολο δεδομένων να φτάσει σε μέγεθος το αρχικό σύνολο δεδομένων. Το σύνολο δεδομένων που μόλις κατασκευάστηκε θα είναι και αυτό που θα χρησιμοποιηθεί για την δημιουργία του επόμενου αδύναμου ταξινομητή.

Επαναλαμβάνουμε την ίδια διαδικασία με πριν, αποδίδοντας, αρχικά, το ίδιο βάρος σε όλες τις παρατηρήσεις και βρίσκοντας τον κατάλληλο ταξινομητή για τα νέα, πλέον, δεδομένα. Ύστερα, εφαρμόζουμε τα προαναφέροντα βήματα.

Τέλος, ως τελική πρόβλεψη επιλέγουμε την κλάση που έχει την μεγαλύτερη επίδραση στο μοντέλο. Ο αλγόριθμος του AdaBoost (για ταξινόμηση) γίνεται καλύτερα κατανοητός με το παράδειγμα που ακολουθεί.

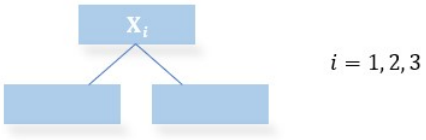
Παράδειγμα 3.4.1. Έστω ότι επιθυμούμε να προβλέψουμε την κατηγορική μεταβλητή Y με τη βοήθεια των επεξηγηματικών μεταβλητών X_1, X_2 και X_3, οι τιμές και κλάσεις των οποίων φαίνονται στον πίνακα του Σχήματος 3.16.

X_1	X_2	X_3	Y	Βάρος Δείγματος
Ναι	Όχι	93	Ναι	1/5
Ναι	Ναι	87	Όχι	1/5
Όχι	Ναι	90	Ναι	1/5
Ναι	Όχι	99	Όχι	1/5
Όχι	Όχι	84	Όχι	1/5

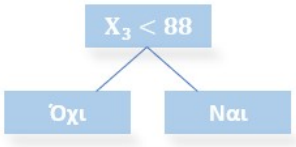
Σχήμα 3.16: Δεδομένα του παραδείγματος 3.4.1. μαζί με τα αντίστοιχα βάρη των παρατηρήσεων.

Ο πίνακας αυτός περιέχει και το αρχικό βάρος της κάθε παρατήρησης, το οποίο είναι ίσο με την μονάδα διαιρεμένη προς το συνολικό πλήθος των παρατηρήσεων. Έπειτα, κατασκευάζονται όλοι οι δυνατοί ταξινομητές (stumps) για το κάθε χαρακτηριστικό (επεξηγηματική μεταβλητή), οι οποίοι θα έχουν την γενική μορφή του Σχήματος 3.17. Κάνοντας χρήση κάποιου από τα μέτρα νοθείας για ταξινόμηση, έστω ότι ο καλύτερος ταξινομητής προκύπτει να είναι αυτός του Σχήματος 3.18.

Με τον συγκεκριμένο ταξινομητή έχουμε μία λάθος ταξινομημένη παρατήρηση, συνεπώς το ολικό σφάλμα (Total Error) του ταξινομητή θα είναι είναι ίσο



Σχήμα 3.17: Μορφή (αδύναμου) ταξινομητή για το χαρακτηριστικό $X_i, i = 1, 2, 3$.



Σχήμα 3.18: Καλύτερος ταξινομητής κατά την πρώτη επανάληψη του αλγορίθμου AdaBoost για το παράδειγμα 3.4.1.

με 1/5.

Η επίδραση του εν λόγω ταξινομητή δίνεται από τον τύπο (3.8) και για αυτό το παράδειγμα παίρνει την μορφή

$$a_1 \;=\; \frac{1}{2}\log\left(\frac{1-1/5}{1/5}\right) \;=\; \frac{1}{2}\log(4) \;=\; 0.693.$$

Η παρατήρηση που ταξινομήθηκε λανθασμένα θα λάβει περισσότερο βάρος στην συνέχεια, προκειμένου να δωθεί μεγαλύτερη έμφαση στην σωστή ταξινόμησή του, ενώ οι σωστά ταξινομημένες παρατηρήσεις θα αποκτήσουν μικρότερο βάρος στην συνέχεια (Σχήμα 3.19).

Το νέο βάρος της κάθε παρατήρησης (του συνόλου εκπαίδευσης) υπολογίζεται από τον τύπο (3.9). Για παράδειγμα, για την παρατήρηση που ταξινομήθηκε λανθασμένα από τον αδύναμο ταξινομητή που προέκυψε, το νέο βάρος θα είναι ίσο με

$$\text{Weight} \;=\; \text{weight}(0)\; e^{a_1} \;=\; \tfrac{1}{5}\, e^{0.693} \;=\; 0.4$$

Τα νέα βάρη των παρατηρήσεων φαίνονται στον πίνακα του Σχήματος 3.20.

Ωστόσο, τα τελευταία δεν αθροίζουν στην μονάδα, όπως τα αρχικά βάρη των παρατηρήσεων. Επομένως, υπολογίζεται στην συνέχεια το κανονικοποιημένο βάρος για κάθε μία από τις παρατηρήσεις. Για παράδειγμα, το κανονικοποιημένο

X_1	X_2	X_3	Y	Βάρος Δείγματος
Ναι	Όχι	93	Ναι	1/5
Ναι	Ναι	87	Όχι	1/5
Όχι	Ναι	90	Ναι	1/5
Ναι	Όχι	99	Όχι	1/5
Όχι	Όχι	84	Όχι	1/5

↑ Βάρος

Σχήμα 3.19: Αύξηση του βάρους της λάθος ταξινομημένης παρατήρησης.

X_1	X_2	X_3	Y	Βάρος Δείγματος	Νέο Βάρος Δείγματος
Ναι	Όχι	93	Ναι	1/5	0.1
Ναι	Ναι	87	Όχι	1/5	0.1
Όχι	Ναι	90	Ναι	1/5	0.1
Ναι	Όχι	99	Όχι	1/5	0.4
Όχι	Όχι	84	Όχι	1/5	0.1

Σχήμα 3.20: Νέα βάρη παρατηρήσεων για το παράδειγμα 3.4.1.

βάρος της πρώτης παρατήρησης θα είναι

$$N_1 = \frac{0.1}{0.8} = 0.125.$$

Έπειτα, για τη διαμόρφωση του καινούργιου συνόλου δεδομένων, διαχωρίζουμε το διαθέσιμο σύνολο δεδομένων σε buckets (μπλε διαστήματα στο επόμενο σύνολο δεδομένων), τα οποία είναι 5 στο πλήθος και φαίνονται στο Σχήμα 3.21. Ο αλγόριθμος θα τρέξει 5 φορές, με διαφορετικό αριθμό ν κάθε φορά (με τιμές μεταξύ του 0 και του 1), προκειμένου να διαλέξει αυθαίρετα 5 παρατηρήσεις από το διαθέσιμο σύνολο δεδομένων.

Για 5 αυθαίρετα ν προκύπτει το επόμενο σύνολο δεδομένων (Σχήμα 3.22).

X_1	X_2	X_3	Υ	Κανονικοποιημένο Βάρος	
Ναι	Όχι	93	Ναι	0.125	0 με 0.125
Ναι	Ναι	87	Όχι	0.125	0.125 με 0.25
Όχι	Ναι	90	Ναι	0.125	0.25 με 0.375
Ναι	Όχι	99	Όχι	0.5	0.375 με 0.875
Όχι	Όχι	84	Όχι	0.125	0.875 με 1

Σχήμα 3.21: Χωρισμός του συνόλου δεδομένων σε buckets.

	X_1	X_2	X_3	Υ	Κανονικοποιημένο Βάρος
ν = 0.64:	Ναι	Όχι	99	Όχι	0.5
ν = 0.42:	Ναι	Όχι	99	Όχι	0.5
ν = 0.13:	Ναι	Όχι	93	Ναι	0.125
ν = 0.27:	Όχι	Ναι	90	Ναι	0.125
ν = 0.80:	Ναι	Όχι	99	Όχι	0.5

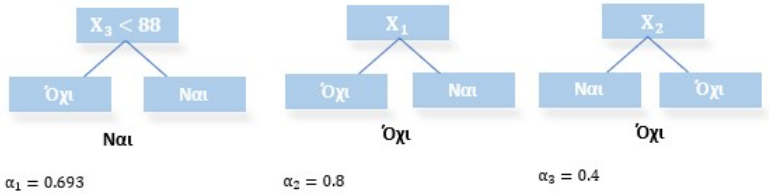
Σχήμα 3.22: Νέο σύνολο δεδομένων για 5 αυθαίρετα ν.

Παρατηρούμε ότι η παρατήρηση που είχε ταξινομηθεί λανθασμένα από τον πρώτο αδύναμο ταξινομητή έχει επιλεχθεί 3 φορές στο καινούργιο σύνολο δεδομένων, γεγονός που αποδεικνύει το κύριο γνώρισμα του AdaBoost ότι οι λανθασμένα ταξινομημένες παρατηρήσεις αποκτούν μεγαλύτερο βάρος στην επόμενη επανάληψη του αλγορίθμου. Με παρόμοια βήματα, δημιουργείται ο δεύτερος αδύναμος ταξινομητής με βάση το νέο σύνολο δεδομένων που μόλις κατασκευάστηκε. Έπειτα, βρίσκουμε το ολικό σφάλμα αυτού και την αντίστοιχη επίδραση a_2. Τα κανονικοποιημένα βάρη για το νέο σύνολο δεδομένων υπολογίζονται με τη βοήθεια των αντίστοιχων νέων βαρών των παρατηρήσεων και το νέο σύνολο δεδομένων χωρίζεται και πάλι σε buckets. Η διαδικασία αυτή επαναλαμβάνεται έως ότου να παρατηρηθεί κάποια μείωση στο σφάλμα ή μέχρι να επιτευχθεί ο επιθυμητός αριθμός αδύναμων ταξινομητών (καθορίζεται από τον ερευνητή και τις ανάγκες του προβλήματος).

Έστω ότι από τα δεδομένα ελέγχου έχουμε την παρατήρηση που φαίνεται στο Σχήμα 3.23 και έστω ότι από την παραπάνω διαδικασία προκύπτουν τα δένδρα (αδύναμοι ταξινομητές) μαζί με τις αντίστοιχες επιδράσεις και προβλέψεις τους, όπως αυτά φαίνονται στο Σχήμα 3.24. Συνεπώς, σύμφωνα με το Σχήμα

X_1	X_2	X_3	Y
Ναι	Όχι	91	Όχι

Σχήμα 3.23: Παρατήρηση από τα δεδομένα ελέγχου για το παράδειγμα 3.4.1.



Σχήμα 3.24: Τελικό μοντέλο ενίσχυσης για το παράδειγμα 3.4.1.

3.24, η μεταβλητή απόκρισης Y ταξινομείται στις κλάσεις 'Ναι', 'Όχι' και 'Όχι' αντίστοιχα. Συνεπώς, η ολική επίδραση για την κλάση 'Ναι' είναι 0.693, ενώ η ολική επίδραση για την κλάση 'Όχι' είναι $0.8 + 0.4 = 1.2$ και, επομένως διαλέγουμε την κλάση 'Όχι' ως τελική κλάση για την μεταβλητή Y της εν λόγω παρατήρησης.

• Ο αλγόριθμος του AdaBoost μπορεί να εφαρμοστεί τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης. Όταν πρόκειται για την τελευταία περίπτωση, οι συλλογισμοί και τα βήματα που ακολουθούνται είναι παρόμοια με αυτά της ταξινόμησης. Για την εύκολη διάκριση των δύο περιπτώσεων όταν εφαρμόζεται ο αλγόριθμος του AdaBoost, θα αναφερόμαστε στους αδύναμους ταξινομητές που είχαμε στην περίπτωση της ταξινόμησης ως αδύναμους εκτιμητές για το είδος των προβλημάτων, όπου έχουμε να εκτιμήσουμε κάποια ποσοτική μεταβλητή απόκρισης.

Ομοίως με πριν, αφού βρεθεί το αρχικό βάρος για κάθε παρατήρηση από το σύνολο εκπαίδευσης, κατασκευάζεται ο πρώτος αδύναμος εκτιμητής, λαμβάνοντας υπόψη το μέτρο νοθείας για την περίπτωση της παλινδρόμησης (SSR). Έπειτα, ακολουθούμε την ίδια διαδικασία με αυτήν που περιγράφτηκε προηγουμένως για το πρόβλημα της ταξινόμησης με την διαφοροποίηση ότι η τελική εκτίμηση για την μεταβλητή απόκρισης που μελετάμε βρίσκεται από τον σταθμισμένο μέσο όρο των εκτιμήσεων που προκύπτουν από τον κάθε αδύναμο εκτιμητή, αν εφαρμοστεί σε κάποιο σύνολο ελέγχου. Με την έννοια του σταθμισμένου μέσου όρου εννοούμε το άθροισμα των εκτιμήσεων που προκύπτουν από τους αδύναμους εκτιμητές, αν αυτοί εφαρμοστούν σε κάποιο σύνολο ελέγχου, πολλαπλασιασμένες με την αντίστοιχη επίδραση α των αντίστοιχων εκτιμητών, διαιρεμένο προς το άθροισμα αυτών των επιδράσεων.

Παράδειγμα 3.4.2. Θεωρούμε το σύνολο δεδομένων του παραδείγματος 3.4.1. με διαφορά στην μεταβλητή απόκρισης Y, η οποία εδώ είναι μία ποσοτική

τυχαία μεταβλητή (Σχήμα 3.25).

X_1	X_2	X_3	Y	Βάρος Δείγματος
Ναι	Όχι	93	14	1/5
Ναι	Ναι	87	21	1/5
Όχι	Ναι	90	17	1/5
Ναι	Όχι	99	12	1/5
Όχι	Όχι	84	23	1/5

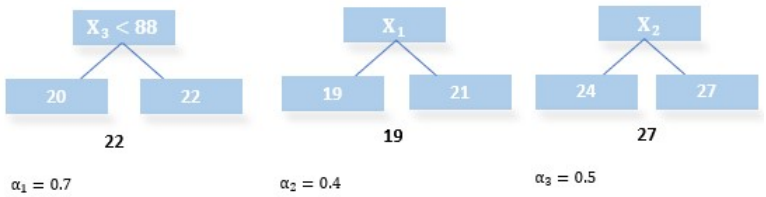
Σχήμα 3.25: Δεδομένα του παραδείγματος 3.4.2. μαζί με τα αντίστοιχα βάρη των παρατηρήσεων.

Η διαδικασία υλοποίησης του αλγορίθμου AdaBoost είναι παρόμοια με αυτή του παραδείγματος 3.4.1. Για το συγκεκριμένο πρόβλημα, το οποίο είναι ένα πρόβλημα παλινδρόμησης, οι αδύναμοι εκτιμητές κατασκευάζονται λαμβάνοντας υπόψη το μέτρο νοθείας για παλινδρόμηση (SSR). Έπειτα, ακολουθούμε την διαδικασία που περιγράφτηκε προηγουμένως και έστω ότι από τα δεδομένα ελέγχου έχουμε την παρατήρηση που φαίνεται στο Σχήμα 3.26

X_1	X_2	X_3	Y
Ναι	Όχι	91	25

Σχήμα 3.26: Παρατήρηση από τα δεδομένα ελέγχου για το παράδειγμα 3.4.2.

και έστω ότι από την παραπάνω διαδικασία προκύπτουν τα δένδρα (αδύναμοι εκτιμητές) μαζί με τις αντίστοιχες επιδράσεις και εκτιμήσεις τους, όπως φαίνονται στο Σχήμα 3.27. Επομένως, από το Σχήμα 3.27, προκύπτουν οι εκτιμήσεις



Σχήμα 3.27: Τελικό μοντέλο ενίσχυσης για το παράδειγμα 3.4.2.

22, 19 και 27 για την μεταβλητή απόκρισης Y με αντίστοιχες επιδράσεις 0.7, 0.4 και 0.5. Συνεπώς, η τελική εκτίμηση θα είναι ο σταθμισμένος μέσος όρος των εκτιμήσεων αυτών, δηλαδή

$$\hat{Y} = \frac{\sum_{i=1}^{3} a_i \hat{Y}_i}{\sum_{i=1}^{3} a_i} = \frac{0.7\times 22 + 0.4\times 19 + 0.5\times 27}{0.7 + 0.4 + 0.5} = 22.8.$$

3.4.2 Gradient Boosting

Ο αλγόριθμος του Gradient Boosting χρησιμοποιεί δένδρα μικρού μεγέθους, λιγότερο πολύπλοκα στην θέση των αδύναμων ταξινομητών του αλγόριθμου του AdaBoost. Με αυτό τον τρόπο, ενισχύεται η προβλεπτική ισχύς του μοντέλου, προσφέροντας πιο ακριβή αποτελέσματα σε προβλήματα ταξινόμησης, αλλά και παλινδρόμησης.

Το Gradient Boosting έχει τρεις (ρυθμιζόμενες) παραμέτρους:

1) Τον αριθμό των δένδρων αποφάσεων M. Η επιλογή του M γίνεται με τη μέθοδο της διασταυρωτικής επικύρωσης.

2) Την παράμετρο συρρίκνωσης λ, η οποία είναι ένας μικρός θετικός αριθμός που 'ρυθμίζει' την συνεισφορά του καινούργου δένδρου στο αμέσως επόμενο. Το λ παίρνει τιμές μεταξύ του μηδενός και του ένα. Πολύ μικρές τιμές του λ μπορεί να απαιτούν την χρήση μεγάλου M για μία καλύτερη απόδοση. Συνήθως, οι τιμές που δίνονται στο λ είναι 0.01 ή 0.001.

3) Τον αριθμό των τμήσεων d κάθε δένδρου. Είναι πιθανόν η τιμή $d = 1$ να αποδίδει καλά και σε αυτή την περίπτωση κάθε δένδρο είναι ένα κούτσουρο, όπως αυτά που συναντήσαμε στην μέθοδο του AdaBoost. Ισχύει ότι αν έχουμε d τμήσεις σε ένα δένδρο, τότε οι μεταβλητές (χαρακτηριστικά) που θα εμφανίζονται σε αυτό θα είναι το πολύ d. Το d ονομάζεται, γενικά, και 'βάθος αλληλεπίδρασης' (interaction depth), καθώς, όπως αναφέρθηκε, ελέγχει την αλληλεπίδραση ανάμεσα στα δένδρα που κατασκευάζονται. Όσο περισσότεροι είναι οι τελικοί κόμβοι (φύλλα) ενός δένδρου, τόσο μεγαλύτερο βάθος έχει το δένδρο και, συνεπώς, τόσο δυσκολότερη γίνεται η κατανόηση των κανόνων απόφασης του δένδρου σχετικά με τον διαχωρισμό των μεταβλητών.

• Ο αλγόριθμος του Gradient Boosting για προβλήματα παλινδρόμησης παρουσιάζεται στην συνέχεια:

Έστω ότι τα υπόλοιπα συμβολίζονται ως $r_i = y_i - \hat{y}_i =: y_i - \hat{f}(x)$ (παρατηρούμενη − προβλεπόμενη τιμή), με το $\hat{f}(x)$ να είναι ίσο με την προβλεπόμενη (εκτιμώμενη) τιμή της i-οστής παρατήρησης του συνόλου εκπαίδευσης της μεταβλητής απόκρισης ($i = 1, ..., N$). Για λόγους απλότητας, θέτουμε, αρχικά,

$\hat{f}(x) = 0$ και, άρα $r_i = y_i$ για όλα τα i στο δείγμα εκπαίδευσης. Για $j = 1, ..., M$ γίνεται επανάληψη των παρακάτω βημάτων:

(i) Προσαρμογή δένδρου $\hat{f}^j$ με d τμήσεις (άρα, $d + 1$ τελικούς κόμβους) στα δεδομένα εκπαίδευσης ($\boldsymbol{X}$, $\boldsymbol{r}$), όπου το τυχαίο διάνυσμα $\boldsymbol{X}$ δηλώνει τις τιμές των επεξηγηματικών μεταβλητών (στο σύνολο εκπαίδευσης). Το διάνυσμα $\boldsymbol{r}$ δηλώνει τις τιμές των υπολοίπων r_i με $i = 1, ..., N$ και το δένδρο προσαρμόζεται χρησιμοποιώντας αυτές ως αποκρίσεις, δηλαδή επιχειρεί να εκτιμήσει τις τιμές των υπολοίπων, αντί για τις τιμές της μεταβλητής απόκρισης, όπως συνηθίζεται σε άλλους αλγορίθμους.

(ii) Ανανέωση του $\hat{f}$, προσθέτοντας μία 'συρρικνωμένη' εκδοχή του νέου δένδρου

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^j(x). \tag{3.10}$$

(iii) Ανανέωση υπολοίπων

$$r_i \leftarrow r_i - \lambda \hat{f}^j(x_i). \tag{3.11}$$

Το τελικό μοντέλο θα είναι το εξής:

$$\hat{f}(x) \ = \ \sum_{j=1}^{M} \lambda \hat{f}^j(x). \tag{3.12}$$

Στην γενική μορφή της σχέσης (3.12) προστίθεται στο δεξί μέλος και ο όρος $\hat{f}(x)$, όμως, επειδή έχουμε κάνει την υπόθεση ότι $\hat{f}(x) = 0$, παραλείπτεται στην συγκεκριμένη περίπτωση. Τέλος, για μία παρατήρηση που προέρχεται από το σύνολο ελέγχου, η εκτίμηση της αντίστοιχης εξαρτημένης μεταβλητής υπολογίζεται από την σχέση (3.12), πολλαπλασιάζοντας την παράμετρο συρρίκνωσης με το άθροισμα των επιμέρους εκτιμήσεων, όπως αυτές προκύπτουν από τα δένδρα που έχουν ήδη κατασκευαστεί.

Επομένως, σε κάθε βήμα κατασκευάζεται ένα νέο δένδρο που προβλέπει τα καινούργια υπόλοιπα και, στην συνέχεια, οι προβλέψεις κάθε δένδρου προσθέτονται όλες μαζί διαδοχικά, αφού πολλαπλασιασθούν, πρώτα, με την παράμετρο συρρίκνωσης λ, όπως δηλώνει και η σχέση (3.10). Η ανάπτυξη αυτών των δένδρων συνεχίζεται έως ότου να επιτευχθεί ο μέγιστος προκαθορισμένος αριθμός δένδρων ή μέχρι να καταλήξουμε στο συμπέρασμα ότι προσθέτοντας επιπλέον δένδρα δεν μειώνει σημαντικά τα υπόλοιπα.

Ο αλγόριθμος του Gradient Boosting βασίζεται στην ιδέα ότι πραγματοποιώντας πολλά μικρά βήματα προς την σωστή κατεύθυνση προσφέρει καλύτερες προβλέψεις με ένα δείγμα ελέγχου [9]. Τα νέα υπόλοιπα που προκύπτουν μετά

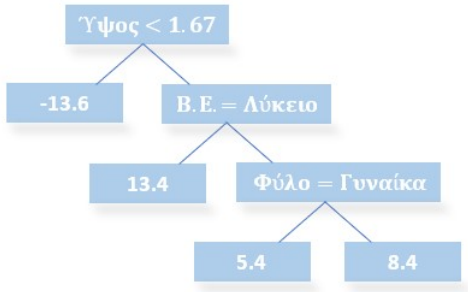
από κάθε επανάληψη, έχουν μικρότερες τιμές από τα προηγούμενα, γεγονός που αποδεικνύει αυτόν τον ισχυρισμό.

Παράδειγμα 3.4.3. Για το σύνολο δεδομένων (εκπαίδευσης) του Σχήματος 3.28 θα εφαρμόσουμε τον αλγόριθμο του Gradient Boosting της μεθόδου της ενίσχυσης.

Φύλο (X_1)	**Ύψος (m)** (X_2)	**Βαθμίδα εκπαίδευσης (Β.Ε.)** (X_3)	**Βάρος (kg)** (Υ)	Υπόλοιπο ($r_i, i = 1, \ldots, 5$)
Γυναίκα	1.5	Γυμνάσιο	52	-14.6
Γυναίκα	1.65	Γυμνάσιο	54	-12.6
Γυναίκα	1.69	Α.Ε.Ι.	72	5.4
Άντρας	1.73	Α.Ε.Ι.	75	8.4
Άντρας	1.78	Λύκειο	80	13.4

Σχήμα 3.28: Δεδομένα του παραδείγματος 3.4.3.

Αρχικά, υπολογίζουμε το υπόλοιπο για την κάθε παρατήρηση. Ως (αρχική) προβλεπόμενη τιμή παίρνουμε τον μέσο όρο των τιμών της εξαρτημένης μεταβλητής Y, δηλαδή $\hat{Y} = 66.6$ και, επομένως, για την πρώτη παρατήρηση, θα ισχύει ότι το αντίστοιχο υπόλοιπο θα είναι ίσο με $r_1 = Y_1 - \hat{Y} = 52 - 66.6 = -14.6$. Για την δεύτερη παρατήρηση το υπόλοιπο θα είναι $r_2 = Y_2 - \hat{Y} = 54 - 66.6 = -12.6$. Με παρόμοιο τρόπο υπολογίζονται και τα υπόλοιπα $r_i,\ i = 3, 4, 5$. Στην συνέχεια, κατασκευάζουμε ένα δένδρο (παλινδρόμησης) στα δεδομένα εκπαίδευσης, έστω $\hat{f}^1$, τέτοιο ώστε να προβλέπει τις τιμές των υπολοίπων που μόλις βρήκαμε. Αυτό το δένδρο περιγράφεται στο Σχήμα 3.29.



Σχήμα 3.29: Δένδρο $\hat{f}^1$ για τα δεδομένα (εκπαίδευσης) του παραδείγματος 3.4.3.

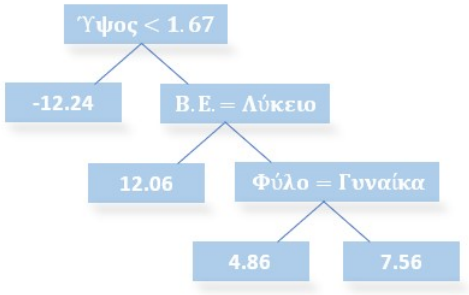
Έστω ότι η παράμετρος συρρίκνωσης είναι ίση με 0.1. Τότε, η προβλεπόμενη τιμή για την πρώτη παρατήρηση του συνόλου εκπαίδευσης θα είναι $\hat{Y}_1 = 66.6 + 0.1 \times (-13.6) = 65.24$. Το αντίστοιχο νέο υπόλοιπο θα είναι $r_1 = Y_1 - \hat{Y}_1 =$

$52 - 65.24 = -13.24$. Με παρόμοιο τρόπο υπολογίζονται όλα τα νέα υπόλοιπα και δηλώνονται στο Σχήμα 3.30.

Φύλο (X_1)	**Ύψος (m)** (X_2)	**Βαθμίδα εκπαίδευσης (Β.Ε.)** (X_3)	**Βάρος (kg)** (Y)	Υπόλοιπο ($r_i, i = 1, \ldots, 5$)
Γυναίκα	1.5	Γυμνάσιο	52	-13.24
Γυναίκα	1.65	Γυμνάσιο	54	-11.24
Γυναίκα	1.69	Α.Ε.Ι.	72	4.86
Άντρας	1.73	Α.Ε.Ι.	75	7.56
Άντρας	1.78	Λύκειο	80	12.06

Σχήμα 3.30: Νέα υπόλοιπα για τα δεδομένα του παραδείγματος 3.4.3.

Όπως αναμέναμε, τα υπόλοιπα μειώθηκαν ήδη από την δεύτερη επανάληψη του αλγορίθμου. Όπως και πριν, προσαρμόζουμε ένα νέο δένδρο ($\hat{f}^2$), το οποίο φαίνεται στο Σχήμα 3.31, για την πρόβλεψη των νέω υπολοίπων.



Σχήμα 3.31: Δένδρο $\hat{f}^2$ για την πρόβλεψη των νέων υπολοίπων.

Παρατηρούμε ότι τα δύο δένδρα έχουν παρόμοια μορφή. Παρ᾽ όλα αυτά, κάτι τέτοιο, συνήθως, δεν συμβαίνει σε περιπτώσεις όπου έχουμε να διαχειριστούμε μεγάλα σύνολα δεδομένων. Έπειτα, επαναλαμβάνουμε την ίδια διαδικασία, κατασκευάζοντας νέα υπόλοιπα και νέα δένδρα παλινδρόμησης έως ότου να επιτύχουμε τον επιθυμητό αριθμό δένδρων παλινδρόμησης (καθορίζεται από τον ερευνητή) ή μέχρι να παρατηρήσουμε σημαντική μείωση στα υπόλοιπα.

Έστω, για λόγους απλότητας και μόνο, ότι το τελικό μοντέλο αποτελείται από τα δένδρα παλινδρόμησης των Σχημάτων 3.29 και 3.31. Τότε, για μία παρατήρηση που προέρχεται από το σύνολο ελέγχου (Σχήμα 3.32), η εκτίμηση για το βάρος θα είναι ίση με

$$\hat{Y} = 66.6 + [(0.1 \times 8.4) + (0.1 \times 7.56)] = 68.2.$$

Φύλο (X_1)	Ύψος (m) (X_2)	Βαθμίδα εκπαίδευσης (Β.Ε.) (X_3)	Βάρος (kg) (Υ)
Άντρας	1.82	Α.Ε.Ι.	69

Σχήμα 3.32: Παρατήρηση από το σύνολο ελέγχου για το παράδειγμα 3.4.3.

• Ο αλγόριθμος του Gradient Boosting μπορεί να εφαρμοστεί, εξίσου καλά, και σε προβλήματα ταξινόμησης. Θα ασχοληθούμε, κυρίως, με την περίπτωση της δυαδικής ταξινόμησης, όπου οι δυνατές κλάσεις της μεταβλητής απόκρισης είναι δύο και θα αναφερόμαστε σε αυτές ως 'κλάση 1' και 'κλάση 0'.

Έστω ότι η αρχική πρόβλεψη για την κατηγορική μεταβλητή απόκρισης είναι ίση με τον λογάριθμο του λόγου συμπληρωματικών πιθανοτήτων που έχει η i-οστή παρατήρηση του συνόλου εκπαίδευσης της μεταβλητής απόκρισης ($i = 1, ..., N$) να ανήκει στην κλάση 1 και θα συμβολίζεται ως $\log(odds)$. Τότε ισχύει το εξής:

$$\log(odds) \;=\; \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right), \tag{3.13}$$

όπου $\hat{p}_i$ με $i = 1, ..., N$ εκφράζει την πιθανότητα η i-οστή παρατήρηση της μεταβλητής απόκρισης του συνόλου εκπαίδευσης να ανήκει στην κλάση 1. Η πιθανότητα αυτή ονομάζεται **πιθανότητα πρόβλεψης**. Συνεπώς, η συμπληρωματική πιθανότητα, $1 - \hat{p}_i$, είναι η πιθανότητα η i-οστή παρατήρηση της μεταβλητής απόκρισης του συνόλου εκπαίδευσης να μην ανήκει στην κλάση 1, δηλαδή να ανήκει στην κλάση 0 (δυαδική ταξινόμηση). Η βάση του λογαρίθμου είναι αυθαίρετη και αφήνεται στην κρίση του ερευνητή, ωστόσο, συνηθίζεται στην στατιστική να χρησιμοποιείται ο φυσικός λογάριθμος, δηλαδή ο λογάριθμος βάσης e.

Ένας εύκολος τρόπος να χρησιμοποιήσουμε τον $\log(odds)$ σε ένα πρόβλημα ταξινόμησης είναι να τον μετατρέψουμε σε πιθανότητα, λύνοντας την εξίσωση (3.13) ως προς $\hat{p}_i$ και, έτσι, προκύπτει ότι

$$\hat{p}_i \;=\; \frac{e^{\log(odds)}}{1+e^{\log(odds)}}. \tag{3.14}$$

Η (3.14) είναι γνωστή ως λογιστική συνάρτηση ή **συνάρτηση απώλειας**.

Έπειτα, ορίζουμε τα υπόλοιπα ως $r_i = y_i - \hat{p}_i$ (παρατηρούμενη − προβλεπόμενη τιμή), με το y_i ($i = 1, ..., N$) να παίρνει τις τιμές 1 ή 0 ανάλογα με το αν η i-οστή παρατήρηση της μεταβλητής απόκρισης του συνόλου εκπαίδευσης ανήκει ή όχι στην κλάση 1, αντίστοιχα. Για $j = 1, ..., M$ γίνεται επανάληψη των παρακάτω βημάτων:

(i) Προσαρμογή δένδρου $\hat{f}^j$ με d τμήσεις (άρα, $d+1$ τελικούς κόμβους) στα δεδομένα εκπαίδευσης $(\boldsymbol{X}, \boldsymbol{r})$, όπου, όπως και στην περίπτωση της παλινδρόμησης, το τυχαίο διάνυσμα $\boldsymbol{X}$ δηλώνει τις τιμές των επεξηγηματικών μεταβλητών και το διάνυσμα $\boldsymbol{r}$ δηλώνει τις τιμές των υπολοίπων r_i με $i = 1, ..., N$ και το δένδρο προσαρμόζεται χρησιμοποιώντας αυτές ως αποκρίσεις, δηλαδή αποσκοπεί στην εκτίμηση των τιμών των υπολοίπων. Έστω ότι το δένδρο καταλήγει να έχει K φύλλα (τελικούς κόμβους). Η τιμή στο k-οστό φύλλο με $k = 1, ..., K$ υπολογίζεται από τον τύπο:

$$(\text{leaf})_k \;=\; \frac{\sum_{i=1}^{N} r_{i_k}}{\sum_{i=1}^{N} [\, \hat{p}_i \times (1 - \hat{p}_i) \,]}. \tag{3.15}$$

Στον τύπο (3.15) ο αριθμητής δηλώνει όλα τα υπόλοιπα που ανήκουν στο φύλλο k και η πιθανότητα πρόβλεψης $\hat{p}_i$ που βρίσκεται στον παρονομαστή αναφέρεται στο i-οστό υπόλοιπο που ανήκει στο φύλλο k.

(ii) Ανανέωση του $\hat{f}$, προσθέτοντας μία 'συρρικνωμένη' εκδοχή του νέου δένδρου

$$\hat{f}(x) \leftarrow \hat{f}(x) + \log(odds) + \lambda \hat{f}^j(x). \tag{3.16}$$

Η πρόβλεψη που προκύπτει από την (3.16) καλείται '$\log(odds)$ πρόβλεψη'.

(iii)

$$\hat{p}_i \;=\; \frac{e^{\hat{f}(x)}}{1 + e^{\hat{f}(x)}}. \tag{3.17}$$

(iv) Ανανέωση υπολοίπων

$$r_i \;=\; y_i - \hat{p}_i. \tag{3.18}$$

Το τελικό μοντέλο θα είναι το εξής:

$$\hat{f}(x) \;=\; \log(odds) + \sum_{j=1}^{M} \lambda \hat{f}^j(x). \tag{3.19}$$

Προφανώς, αν θέσουμε εκ των προτέρων την αρχική πρόβλεψη να είναι μηδέν, δηλαδή $\log(odds) = 0$, τότε θα μηδενίζεται και στην σχέση (3.19). Η ανάπτυξη των δένδρων συνεχίζεται έως ότου να επιτευχθεί ο μέγιστος προκαθορισμένος αριθμός δένδρων ή μέχρι να υπάρξει μεγάλη μείωση στα υπόλοιπα.

Σαν τελευταίο βήμα δεν πρέπει να ξεχάσουμε να μετατρέψουμε την 'log(*odds*) πρόβλεψη' της (3.19) σε πιθανότητα, έτσι ώστε να μπορούμε να την χρησιμοποιούμε για να κάνουμε ταξινομήσεις (προβλέψεις) σε διαφορετικά σύνολα ελέγχου. Συνεπώς, η επιθυμητή πιθανότητα πρόβλεψης θα δίνεται από τον τύπο:

$$\hat{P} = \frac{e^{\hat{f}(x)}}{1 + e^{\hat{f}(x)}}. \tag{3.20}$$

Αν χρησιμοποιήσουμε σαν όριο το 0.5 για να πραγματοποιήσουμε τις ταξινομήσεις, δηλαδή αν $\hat{P} > 0.5$ για ένα σύνολο ελέγχου, τότε η μεταβλητή απόκρισης για το εν λόγω σύνολο ελέγχου θα ανήκει στην κλάση 1, διαφορετικά θα ανήκει στην κλάση 0.

Παράδειγμα 3.4.4. Το σύνολο δεδομένων που ακολουθεί στο Σχήμα 3.33 προέρχεται από ένα δημοφιλές πρόβλημα δυαδικής ταξινόμησης στην μηχανική μάθηση, το πρόβλημα της πρόβλεψης της μοίρας των επιβατών του Τιτανικού με βάση κάποια χαρακτηριστικά [12]. Η μεταβλητή 'Θέση' (X_1) δηλώνει την θέση του επιβάτη και είναι κατηγορική (με κλάσεις 1, 2 και 3), η μεταβλητή 'Ηλικία' (X_2) δηλώνει την ηλικία του επιβάτη όταν αυτός επέβαινε στον Τιτανικό, η μεταβλητή 'Κόστος' (X_3) εκφράζει το κόστος που έπρεπε να πληρώσει ο επιβάτης, προκειμένου να ταξιδέψει με τον Τιτανικό (μετρημένο σε £) και η μεταβλητή 'Φύλο' (X_4) δηλώνει το φύλο του επιβάτη. Με βάση, λοιπόν, τα συγκεκριμένα χαρακτηριστικά επιχειρούμε να προβλέψουμε εάν ο εν λόγω επιβάτης επιβίωσε ή όχι μετά από την τραγωδία του Τιτανικού (Y). Η κλάση '1'

Θέση (X_1)	Ηλικία (X_2)	Κόστος (£) (X_3)	Φύλο (X_4)	Επιβίωσε (Y)
3	22	7.25	Άντρας	0
1	38	71.28	Γυναίκα	1
2	26	7.93	Γυναίκα	1
1	35	53.10	Γυναίκα	1
3	8	21.07	Άντρας	0
3	27	11.13	Γυναίκα	1

Σχήμα 3.33: Δεδομένα του παραδείγματος 3.4.4.

της εξαρτημένης μεταβλητής Y δηλώνει ότι ο επιβάτης επιβίωσε, ενώ η κλάση '0' δηλώνει ότι ο επιβάτης δεν κατάφερε να επιβιώσει. Αρχικά, υπολογίζουμε το log(*odds*) που σύμφωνα με τον τύπο (3.13) θα ισούται με

$$\log(odds) = \log\left(\frac{4/6}{2/6}\right) = \log\left(\frac{4}{2}\right) = 0.7.$$

Στην συνέχεια, από τον τύπο (3.14) ισχύει ότι

$$P(\text{επιβίωσης}) = \hat{p}_i = \frac{e^{\log(odds)}}{1+e^{\log(odds)}} = \frac{e^{0.7}}{1+e^{0.7}} = 0.7,$$

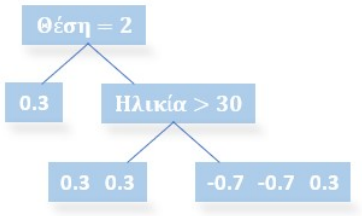
η οποία για το συγκεκριμένο παράδειγμα ταυτίζεται με το $\log(odds)$, λόγω των στρογγυλοποιήσεων.

Έπειτα, καταγράφουμε τα υπόλοιπα $r_i,\ i = 1, ..., 6$, για την κάθε παρατήρηση αντίστοιχα. Για παράδειγμα, τα υπόλοιπα για τις δύο πρώτες παρατηρήσεις είναι ίσα με $r_1 = Y_1 - \hat{p}_i = 0 - 0.7 = -0.7$ και $r_2 = Y_2 - \hat{p}_i = 1 - 0.7 = 0.3$ αντίστοιχα. Τα υπόλοιπα φαίνονται στον πίνακα του Σχήματος 3.34.

Θέση (X_1)	Ηλικία (X_2)	Κόστος (£) (X_3)	Φύλο (X_4)	Επιβίωσε (Y)	Υπόλοιπο ($\eta, i = 1, ..., 6$)
3	22	7.25	Άντρας	0	-0.7
1	38	71.28	Γυναίκα	1	0.3
2	26	7.93	Γυναίκα	1	0.3
1	35	53.10	Γυναίκα	1	0.3
3	8	21.07	Άντρας	0	-0.7
3	27	11.13	Γυναίκα	1	0.3

Σχήμα 3.34: Υπόλοιπα για τις παρατηρήσεις του παραδείγματος 3.4.4.

Με βάση το εν λόγω σύνολο δεδομένων κατασκευάζεται δένδρο $\hat{f}^1$, προκειμένου να προβλέψει τις τιμές των υπολοίπων $r_i,\ i = 1, ..., 6$ (Σχήμα 3.35). Η



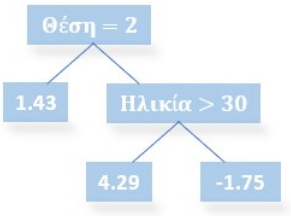
Σχήμα 3.35: Δένδρο $\hat{f}^1$ για την πρόβλεψη των υπολοίπων $r_i,\ i = 1, ..., 6$.

τιμή στο k-οστό φύλλο ($k = 1, 2, 3$) υπολογίζεται από τον τύπο (3.15). Έτσι, ισχύει ότι:

$$(\text{leaf})_1 = \frac{\sum_{i=1}^{6} r_{i_k}}{\sum_{i=1}^{6} [\,\hat{p}_i \times (1-\hat{p}_i)\,]} = \frac{0.3}{0.7(1-0.7)} = 1.43,$$

$$(\text{leaf})_2 = \frac{\sum_{i=1}^{6} r_{i_k}}{\sum_{i=1}^{6} [\,\hat{p}_i \times (1-\hat{p}_i)\,]} = \frac{0.3+0.3}{[0.7(1-0.7)]+[0.7(1-0.7)]} = 4.29,$$

$$(\text{leaf})_3 = \frac{\sum_{i=1}^{6} r_{i_k}}{\sum_{i=1}^{6} [\,\hat{p}_i \times (1-\hat{p}_i)\,]} = \frac{-0.7-0.7+0.3}{[0.7(1-0.7)]+[0.7(1-0.7)]+[0.7(1-0.7)]} = -1.75.$$



Σχήμα 3.36: Δένδρο $\hat{f}^1$, μετά τον προσδιορισμό των τιμών στα φύλλα.

Συνεπώς, το δένδρο $\hat{f}^1$ λαμβάνει την μορφή του Σχήματος 3.36 και έστω ότι η παράμετρος συρρίκνωσης είναι ίση με λ= 0.1. Τότε, για τον πρώτο επιβάτη στο σύνολο δεδομένων (εκπαίδευσης), η $\log(odds)$ πρόβλεψη του αν επιβίωσε ή όχι θα είναι ίση με $0.7 + [0.1 \times (-1.75)] = 0.525$, από την σχέση (3.16), όπου $\hat{f}^1(x) = -1.75$ και την οποία μετατρέπουμε σε πιθανότητα πρόβλεψης

$$\hat{p}_1 = \frac{e^{0.525}}{1 + e^{0.525}} = 0.63.$$

Με παρόμοιο τρόπο προκύπτουν οι πιθανότητες πρόβλεψης και για τις υπόλοιπες παρατηρήσεις.

$$\hat{f}^2(x) = 0.7 + [0.1 \times 4.29] = 1.129 \text{ και } \hat{p}_2 = 0.76,$$

$$\hat{f}^3(x) = 0.7 + [0.1 \times 1.43] = 0.843 \text{ και } \hat{p}_3 = 0.70,$$

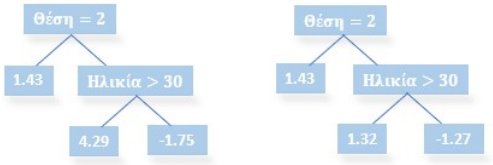
$$\hat{f}^4(x) = 0.7 + [0.1 \times 4.29] = 1.129 \text{ και } \hat{p}_4 = 0.76,$$

$$\hat{f}^5(x) = 0.7 + [0.1 \times (-1.75)] = 0.525 \text{ και } \hat{p}_5 = 0.63,$$

$$\hat{f}^6(x) = 0.7 + [0.1 \times (-1.75)] = 0.525 \text{ και } \hat{p}_6 = 0.63.$$

Χρησιμοποιούμε σαν όριο το 0.5 για να πραγματοποιήσουμε τις ταξινομήσεις, δηλαδή αν $\hat{p} > 0.5$, τότε η μεταβλητή απόκρισης θα ανήκει στην κλάση 1, διαφορετικά θα ανήκει στην κλάση 0. Επομένως, παρατηρούμε ότι η πιθανότητα $\hat{p}_6$ είναι μικρότερη από την αρχική πιθανότητα πρόβλεψης, δηλαδή από την $\log(odds) = 0.7$ και, παράλληλα, ισχύει ότι ο έκτος επιβάτης στο σύνολο δεδομένων εκπαίδευσης επιβίωσε (ανήκει στην κλάση 1). Αυτός είναι και ο λόγος που κατασκευάζουμε πολλά δένδρα αποφάσεων στην μέθοδο του Gradient Boosting. Έστω, για λόγους απλότητας, ότι το τελικό μοντέλο αποτελείται από τα δένδρα παλινδρόμησης που φαίνονται στο Σχήμα 3.37.

Έστω η παρατήρηση του Σχήματος 3.38 που προέρχεται από ένα ανεξάρτητο σύνολο δεδομένων (σύνολο ελέγχου).



Σχήμα 3.37: Τελικό μοντέλο για το παράδειγμα 3.4.4.

Θέση (X_1)	Ηλικία (X_2)	Κόστος (£) (X_3)	Φύλο (X_4)	Επιβίωσε (Y)
2	30	8.21	Άντρας	1

Σχήμα 3.38: Παρατήρηση από το σύνολο ελέγχου για το παράδειγμα 3.4.4.

Τότε η $\log(odds)$ πρόβλεψη του αν επιβίωσε ή όχι ο συγκεκριμένος επιβάτης υπολογίζεται από την σχέση (3.19), δηλαδή

$$\hat{f}(x) = \log(odds) + \sum_{j=1}^{M} \lambda \hat{f}^j(x) = 0.7 + [(0.1 \times 1.43) + (0.1 \times 1.43)]$$

Επομένως, θα ισχύει ότι $\hat{f}(x) = 0.986$. Τότε

$$\hat{p} = \frac{e^{0.986}}{1 + e^{0.986}} = 0.73 > 0.5$$

και, επομένως, μπορούμε να συμπεράνουμε ότι ο επιβάτης της εν λόγω παρατήρησης επιβίωσε, δηλαδή ανήκει στην κλάση '1'.

Ο αλγόριθμος του Gradient Boosting μπορεί να επεκταθεί και στην περίπτωση της πολλαπλής ταξινόμησης, όπου η μεταβλητή απόκρισης αποτελείται από πολλαπλές κλάσεις (τουλάχιστον τρεις). Συνήθως, όταν έχουμε τέτοιου είδους προβλήματα, επεκτείνουμε την δυαδική ταξινόμηση σε πολλαπλών κλάσεων ταξινόμηση (multi-class classification). Η επέκταση αυτή μπορεί να γίνει μέσω ευρετικών μεθόδων, οι οποίες διασπούν το πρόβλημα των πολλαπλών κλάσεων ταξινόμησης σε πολλαπλά προβλήματα δυαδικής ταξινόμησης και κατασκευάζουν ένα μοντέλο για κάθε ένα από αυτά. Δύο γνωστές μέθοδοι που χρησιμοποιούνται συχνά για αυτόν τον σκοπό είναι η λεγόμενη 'Ένα έναντι των Υπολοίπων' μέθοδος ή, αλλιώς, OvR (One-vs-Rest) και η 'Ένα έναντι ενός' μέθοδος ή, αλλιώς, OvO (One-vs-One) [10].

Η OvR μέθοδος περιλαμβάνει τον διαχωρισμό του πολλαπλών κλάσεων συνόλου δεδομένων σε πολλαπλά προβλήματα δυαδικής ταξινόμησης. Ένας δυαδικός ταξινομητής προσαρμόζεται σε κάθε πρόβλημα δυαδικής ταξινόμησης (δένδρο αποφάσεων σαν αυτό που κατασκευάζεται κατά την τελευταία επανάληψη του βήματος (i) της παραπάνω μορφής του αλγορίθμου Gradient Boosting) και οι προβλέψεις γίνονται με βάση το πιο αξιόπιστο μοντέλο, δηλαδή το μοντέλο με

το καλύτερο μέτρο νοθείας. Για παράδειγμα, θεωρούμε ένα πρόβλημα ταξινόμησης στο οποίο η μεταβλητή απόκρισης έχει τρεις πιθανές κλάσεις. Έστω ότι οι κλάσεις αυτές είναι οι εξής: 'πράσινο', 'μπλε' και 'κόκκινο'. Τότε η εν λόγω μέθοδος χωρίζει το πρόβλημα σε 3 σύνολα δεδομένων δυαδικής ταξινόμησης, όπως φαίνεται παρακάτω [10]:

- 1ο Πρόβλημα Δυαδικής Ταξινόμησης: πράσινο έναντι (μπλε, κόκκινο)
- 2ο Πρόβλημα Δυαδικής Ταξινόμησης: μπλε έναντι (πράσινο, κόκκινο)
- 3ο Πρόβλημα Δυαδικής Ταξινόμησης: κόκκινο έναντι (πράσινο, μπλε)

Παρόμοια, η OvO μέθοδος περιλαμβάνει και αυτή τον διαχωρισμό του πολλαπλών κλάσεων συνόλου δεδομένων σε πολλαπλά προβλήματα δυαδικής ταξινόμησης. Στην συνέχεια, χωρίζει το σύνολο δεδομένων σε ένα σύνολο δεδομένων για κάθε κλάση έναντι κάθε άλλης κλάσης, σε αντίθεση με την OvR μέθοδο, η οποία το χωρίζει σε ένα δυαδικό σύνολο δεδομένων για κάθε κλάση. Για παράδειγμα, θεωρούμε ένα πρόβλημα ταξινόμησης, όπου η μεταβλητή απόκρισης έχει τέσσερις πιθανές κλάσεις, 'πράσινο', 'μπλε', 'κόκκινο' και 'κίτρινο'. Τότε η εν λόγω μέθοδος χωρίζει το πρόβλημα σε 6 σύνολα δεδομένων δυαδικής ταξινόμησης, όπως φαίνεται παρακάτω [10]:

- 1ο Πρόβλημα Δυαδικής Ταξινόμησης: πράσινο έναντι μπλε
- 2ο Πρόβλημα Δυαδικής Ταξινόμησης: πράσινο έναντι κόκκινο
- 3ο Πρόβλημα Δυαδικής Ταξινόμησης: πράσινο έναντι κίτρινο
- 4ο Πρόβλημα Δυαδικής Ταξινόμησης: μπλε έναντι κόκκινο
- 5ο Πρόβλημα Δυαδικής Ταξινόμησης: μπλε έναντι κίτρινο
- 6ο Πρόβλημα Δυαδικής Ταξινόμησης: κόκκινο έναντι κίτρινο

Ο γενικός τύπος υπολογισμού των συνόλων δυαδικής ταξινόμησης στην εν λόγω μέθοδο είναι ⟨Αριθμός Κλάσεων × (Αριθμός Κλάσεων − 1)⟩ / 2.

Επομένως, μπορούμε, πλέον, να εφαρμόσουμε τον αλγόριθμο του Gradient Boosting για κάθε ένα από αυτά τα προβλήματα.

Ανακεφαλαιώνοντας, παρ᾽ όλο που ο αλγόριθμος του Gradient Boosting αποδίδει καλύτερα με ένα δείγμα ελέγχου, ταυτόχρονα, τείνει πιο εύκολα στην υπερπροσαρμογή των δεδομένων σε σχέση με τον αλγόριθμο του AdaBoost. Επίσης, απαιτείται ιδιαίτερη προσοχή όσον αφορά την επιλογή των αντίστοιχων παραμέτρων, έτσι ώστε ο αλγόριθμος να δώσει το καλύτερο δυνατό μοντέλο με τις βέλτιστες προβλέψεις. Μία σημαντική διαφορά μεταξύ των τυχαίων δασών και της ενίσχυσης είναι ότι στην τελευταία μέθοδο, κάθε νέο δένδρο λαμβάνει υπόψιν τα δένδρα που έχουν ήδη δημιουργηθεί, επομένως, δένδρα μικρού μεγέθους μπορούν να θεωρηθούν επαρκή.

Κεφάλαιο 4

Εφαρμογή σε πρόβλημα παλινδρόμησης

Η θεωρία της δενδρικής παλινδρόμησης που αναπτύχθηκε στα προηγούμενα κεφάλαια εφαρμόζεται σε ένα πρόβλημα παλινδρόμησης, προκειμένου ο αναγνώστης να κατανοήσει πως αυτή προσαρμόζεται και χρησιμοποιείται σε πρακτικό επίπεδο, επεξεργάζοντας πραγματικά δεδομένα και υλοποιώντας τους αλγορίθμους που προαναφέρθηκαν με την βοήθεια υπολογιστή. Στην συνέχεια, το ίδιο πρόβλημα παλινδρόμησης αναλύεται με την μέθοδο της πολλαπλής γραμμικής παλινδρόμησης, με απώτερο σκοπό την σύγκρισή της με την μέθοδο της δενδρικής παλινδρόμησης. Οι δύο μέθοδοι μελετούνται και αξιολογούνται, έτσι ώστε, τελικά, να αποφασιστεί ποια μέθοδος δίνει καλύτερες, ακριβέστερες προβλέψεις για το υπό εξέταση στατιστικό πρόβλημα.

4.1 Ανάλυση του προβλήματος

Το πρόβλημα παλινδρόμησης πάνω στο οποίο επιλέχθηκε να αναπτυχθεί και να εφαρμοστεί η μέθοδος της δενδρικής παλινδρόμησης είναι το πρόβλημα της 'Υπόθεσης του Κύκλου Ζωής' ("Life-Cycle Hypothesis").

Σύμφωνα με τη θεωρία της οικονομικής επιστήμης, η υπόθεση του κύκλου ζωής ή, εναλλακτικά, το μοντέλο του κύκλου ζωής είναι ένα μοντέλο που προσπαθεί να εξηγήσει τα πρότυπα κατανάλωσης των ατόμων. Υποδηλώνει ότι τα άτομα σχεδιάζουν τη συμπεριφορά κατανάλωσης και εξοικονόμησης κατά την διάρκεια του κύκλου ζωής τους. Με άλλα λόγια, περιγράφει τις καταναλωτικές και αποταμιευτικές συνήθειες των ανθρώπων σε μακροχρόνιο ορίζοντα.

Η θεωρία αυτή υποστηρίζει ότι ο κάθε άνθρωπος αναζητά τρόπους, ώστε να καταφέρει να 'ισορροπήσει' και να σταθεροποιήσει τις καταναλωτικές του ανάγκες και συνήθειες, καταφεύγοντας στον δανεισμό, όταν το εισόδημά του

είναι χαμηλό και στην αποταμίευση, όταν το εισόδημά του είναι υψηλό. Επιπροσθέτως, η υπόθεση του κύκλου ζωής υπονοεί ότι ο απώτερος στόχος του κάθε ανθρώπου είναι να πετύχει την βέλτιστη κατανομή της κατανάλωσής του για τα αντίστοιχα χρόνια ζωής του.

Πολλαπλές στατιστικές μελέτες έχουν υποδείξει ότι η διαχρονική εξέλιξη των εισοδημάτων των ανθρώπων παρουσιάζεται σαν μία ομοιόμορφη καμπύλη [6]: Στην αρχή βρίσκεται σε χαμηλά επίπεδα, έπειτα αυξάνεται σε μεγάλο βαθμό και, τέλος, με τη σύνταξη και, καθώς πλησιάζει το κλείσιμο του κύκλου ζωής του ατόμου, το εισόδημα σταδιακά μειώνεται. Αναλυτικότερα, όταν το άτομο βρίσκεται σε μικρή ηλικία δανείζεται από συγγενικά ή φιλικά πρόσωπα, καθώς δεν διαθέτει κάποιο σπουδαίο εισόδημα. Έπειτα, κατά τα μεσαία χρόνια της ζωής του και στο απόγειο της επαγγελματικής του καριέρας, το άτομο αρχίζει να αποταμιεύει μέρος από το σχετικά υψηλό εισόδημα που λαμβάνει και, τέλος, καταλήγει να ζει και να καλύπτει όλες τις δαπάνες του από τη σύνταξη και τον πλούτο που έχει ήδη συγκεντρώσει.

Ο Franco Modigliani, σε συνεργασία με τον Richard Brumberg, δημοσίευσε την θεωρία της υπόθεσης του κύκλου ζωής το 1954 και το 1986 κέρδισε το βραβείο Nobel στα οικονομικά [13].

Θα χρησιμοποιήσουμε δεδομένα που περιλαμβάνονται στην στατιστική γλώσσα προγραμματισμού R για την πρόβλεψη του 'λόγου αποταμίευσης' ("savings ratio") μίας χώρας. Ο λόγος αποταμίευσης είναι ίσος με την συνολική προσωπική αποταμίευση διαιρεμένη προς το διαθέσιμο εισόδημα. Το σύνολο δεδομένων που αντιστοιχεί στο συγκεκριμένο πρόβλημα ονομάζεται LifeCycleSavings και ανήκει στο πακέτο "datasets" της R. Περιλαμβάνει δεδομένα σχετικά με τον λόγο αποταμίευσης διάφορων χωρών κατά το χρονικό διάστημα 1960-1970 και αποτελείται από 50 παρατηρήσεις πάνω σε 5 ποσοτικές μεταβλητές. Η μεταβλητή "sr" δηλώνει τον λόγο αποταμίευσης και θα είναι η μεταβλητή απόκρισης για το εν λόγω πρόβλημα. Με "pop15" συμβολίζεται το ποσοστό του πληθυσμού που είναι κάτω των 15 ετών και με "pop75" το ποσοστό του πληθυσμού που είναι άνω των 75 ετών, με "dpi" χαρακτηρίζεται το κατά κεφαλήν διαθέσιμο εισόδημα (μετρημένο σε $) και, τέλος, το "ddpi" είναι το ποσοστό μεταβολής του κατά κεφαλήν διαθέσιμου εισοδήματος. Επομένως, οι τέσσερις τελευταίες μεταβλητές θα είναι οι (ποσοτικές) επεξηγηματικές μεταβλητές του προβλήματος, βάσει των οποίων θα γίνουν οι εκτιμήσεις (προβλέψεις) της μεταβλητής απόκρισης.

Η έννοια του 'κατά κεφαλήν' εισοδήματος μπορεί, επίσης, να δηλωθεί με τον λατινικό όρο "per capita" και αναφέρεται στο μέσο εισόδημα ανά άτομο σε μία συγκεκριμένη περιοχή, σε έναν συγκεκριμένο χρόνο.

Τα δεδομένα, λοιπόν, αυτών των πέντε μεταβλητών θα χρησιμοποιηθούν στην συνέχεια για το πρόβλημα της υπόθεσης του κύκλου ζωής και θα αξιοποιηθούν μέσω της μεθόδου της δενδρικής παλινδρόμησης, ώστε να προκύψουν

οι ανάλογες προβλέψεις. Ύστερα, θα γίνει σύγκριση αυτών με τις αντίστοιχες προβλέψεις της γραμμικής παλινδρόμησης.

4.1.1 Δενδρική Παλινδρόμηση

Η μεταβλητή απόκρισης sr είναι ποσοτική, επομένως το πρόβλημα της υπόθεσης του κύκλου ζωής κατατάσσεται σε πρόβλημα παλινδρόμησης.

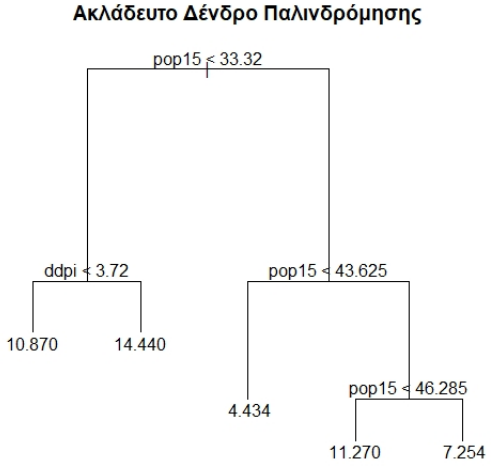
Αρχικά, φορτώνουμε στην R την βιβλιοθήκη του πακέτου "datasets" μέσα στο οποίο περιλαμβάνεται το σύνολο δεδομένων που θα χρησιμοποιήσουμε (LifeCycleSavings). Παράλληλα, φορτώνουμε και την βιβλιοθήκη "tree" για την δυνατότητα κατασκευής δένδρων παλινδρόμησης.

```
> library(tree)
> library(datasets)
> set.seed(1)
> train<-sample(1:nrow(LifeCycleSavings),0.8*nrow(
   LifeCycleSavings))
> tree_LifeCycleSavings<-tree(sr~.,LifeCycleSavings,subset=
   train)
> summary(tree_LifeCycleSavings)

Regression tree:
tree(formula = sr ~ ., data = LifeCycleSavings, subset =
   train)
Variables actually used in tree construction:
[1] "pop15" "ddpi"
Number of terminal nodes:  5
Residual mean deviance:  10.31 = 360.7 / 35
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.4630 -2.1260  0.1392  0.0000  1.5510  7.2870
> plot(tree_LifeCycleSavings)
> text(tree_LifeCycleSavings,pretty=0)
> title(main="Ακλάδευτο Δένδρο Παλινδρόμησης")
```

Κώδικας 4.1: Κατασκευή δένδρου παλινδρόμησης στα δεδομένα εκπαίδευσης του συνόλου LifeCycleSavings.

Στον Κώδικα 4.1, η συνάρτηση set.seed() ενεργοποιεί την γεννήτρια τυχαίων αριθμών και χρησιμοποιείται για την δημιουργία τυχαίων συνόλων, τα οποία μπορούμε να επικαλεστούμε και να αναπαράξουμε πολλές φορές. Αφού ορισθεί στην νέα μεταβλητή train το σύνολο εκπαίδευσης με τη βοήθεια της συνάρτησης sample(), γίνεται προσαρμογή του δένδρου παλινδρόμησης στα δεδομένα εκπαίδευσης, έχοντας ως μεταβλητή απόκρισης τον λόγο αποταμίευσης sr. Με την συνάρτηση sample(), η R επιλέγει τυχαίο δείγμα από τις 50 παρατηρήσεις του συνόλου δεδομένων LifeCycleSavings με μέγεθος όσο και το πλήθος των

γραμμών του συνόλου δεδομένων LifeCycleSavings πολλαπλασιασμένο επί 0.8, δηλαδή το 80% των συνολικών δεδομένων. Η εντολή summary() παρουσιάζει διάφορα αποτελέσματα που αφορούν το μοντέλο (δένδρο παλινδρόμησης), τα οποία προέκυψαν από την διαδικασία κατασκευής του δένδρου, χρησιμοποιώντας ως μέτρο νοθείας το άθροισμα των τετραγώνων των υπολοίπων, όπως περιγράφτηκε στην υποενότητα 2.2.1 του Κεφαλαίου 2. Άμεσα παρατηρούμε ότι μόνο οι μεταβλητές pop15 και ddpi χρησιμοποιήθηκαν στην κατασκευή του δένδρου, το οποίο καταλήγει σε πέντε τελικούς κόμβους (φύλλα). Στην δενδρική παλινδρόμηση, η υπολειπόμενη μέση απόκλιση (residual mean deviance) είναι, αλλιώς, το μέσο τετραγωνικό σφάλμα, MSE, το οποίο προκύπτει να είναι ίσο με 10.31 με το άθροισμα των τετραγώνων των υπολοίπων, SSR, να είναι ίσο με 360.7. Η συνάρτηση summary() δίνει και κάποιους επιπλέον περιγραφικούς δείκτες (ελάχιστη τιμή, 1ο τεταρτημόριο, διάμεσος, μέση τιμή, 3ο τεταρτημόριο και μέγιστη τιμή) των υπολοίπων. Το αρχικό, ακλάδευτο δένδρο παλινδρόμησης που κατασκευάστηκε για το εν λόγω πρόβλημα φαίνεται στο Σχήμα 4.1. Το όρισμα pretty=0 της εντολής text() δηλώνει ότι για την δημιουργία του



Σχήμα 4.1: Ακλάδευτο (ολικό) δένδρο παλινδρόμησης με βάση τα δεδομένα εκπαίδευσης του συνόλου LifeCycleSavings.

δενδρικού διαγράμματος χρησιμοποιούνται τα ονόματα των χαρακτηριστικών (επεξηγηματικών μεταβλητών), όπως αυτά δίνονται στα δεδομένα.

Η μεταβλητή ddpi, όπως αναφέρθηκε, δηλώνει το ποσοστό μεταβολής του κατά κεφαλήν διαθέσιμου εισοδήματος. Επομένως, το δένδρο υποδεικνύει ότι μεγάλες σχετικά αλλαγές στο διαθέσιμο εισόδημα (μεγαλύτερες ή ίσες του

3.72%) αντιστοιχούν σε μεγαλύτερα ποσοστά αποταμίευσης του εισοδήματος από το άτομο. Συγκεκριμένα, προβλέπεται ότι στις χώρες, όπου το ποσοστό των ατόμων του πληθυσμού που είναι κάτω των 15 ετών είναι μικρότερο του 33.32% και υπάρχει μεταβολή στο διαθέσιμο κατά κεφαλήν εισόδημα μεγαλύτερη ή ίση από 3.72%, τα άτομα αποταμιεύουν κατά 14.44%.

Εν συνεχεία, θα προχωρήσουμε στο κλάδεμα του δένδρου, προκειμένου να αποφύγουμε τυχόν προβλήματα υπερπροσαρμογής. Ακολουθώντας την μέθοδο του κλαδέματος της πολυπλοκότητας του κόστους που αναλύθηκε λεπτομερώς στο Κεφάλαιο 3, πρέπει να βρεθεί η βέλτιστη παράμετρος πολυπλοκότητας, δηλαδή εκείνο το α που ελαχιστοποιεί το σφάλμα (SSR) των υποδένδρων, όπως αυτά προέκυψαν από την διαδικασία της διασταυρωτικής επικύρωσης. Για να το επιτύχουμε αυτό, θα χρησιμοποιήσουμε το πακέτο "rpart" της R, καθώς και την συνάρτηση "prune()" του εν λόγω πακέτου. Η τελευταία παίρνει ως όρισμα ένα μοντέλο (δένδρο αποφάσεων), το οποίο έχει κατασκευαστεί με την βοήθεια της συνάρτησης "rpart()", επομένως, δημιουργούμε το προηγούμενο (ακλάδευτο) δένδρο παλινδρόμησης μέσω της συγκεκριμένης συνάρτησης, αφού πρώτα φορτώσουμε τα απαραίτητα πακέτα.

```
> library(rpart)
> library(rpart.plot)
> tree_LCS<-rpart(sr~.,LifeCycleSavings,subset=train,control=
rpart.control(xval=10))
> printcp(tree_LCS)

Regression tree:
rpart(formula = sr ~ ., data = LifeCycleSavings, subset =
   train, control = rpart.control(xval = 10))

Variables actually used in tree construction:
[1] ddpi  pop15

Root node error: 808.28/40 = 20.207

n= 40

        CP nsplit rel error  xerror    xstd
1 0.278965      0   1.00000 1.10339 0.24097
2 0.153833      1   0.72103 1.01623 0.27160
3 0.066198      2   0.56720 0.91159 0.24310
4 0.010000      3   0.50100 0.87524 0.22761
```

Κώδικας 4.2: Αναζήτηση του καλύτερου α (complexity parameter-cp).

Με το όρισμα rpart.control(xval=10) η R πραγματοποιεί διασταυρωτική επικύρωση 10 ομάδων (10-fold cross validation). Η διαδικασία της διασταυρωτικής επικύρωσης εμπεριέχεται στην συνάρτηση rpart(). Η τελευταία εντολή

του Κώδικα 4.2, δίνει έναν πίνακα που αφορά την παράμετρο πολυπλοκότητας α και παρέχει τα βέλτιστα κλαδέματα (εκείνα που οδηγούν σε ακριβέστερες προβλέψεις), ανάλογα με την τιμή του α. Το πλήθος των παρατηρήσεων εκπαίδευσης που χρησιμοποιήθηκαν για την κατασκευή του δένδρου είναι 40 (n=40) και το α που προτιμάται είναι αυτό που προσφέρει το χαμηλότερο σφάλμα, όπως αυτό προέκυψε από την διασταυρωτική επικύρωση (στήλη "xerror"). Στην πρώτη στήλη δίνονται οι διάφορες τιμές για το α, η δεύτερη στήλη δηλώνει τον αριθμό των τμήσεων, η τρίτη στήλη είναι το σχετικό σφάλμα του (εκάστοτε) μοντέλου και μπορούμε να θεωρήσουμε ότι είναι η μεταβλητότητα της μεταβλητής απόκρισης sr που εξηγείται από το μοντέλο (variance explained), ενώ η πέμπτη στήλη δηλώνει την αντίστοιχη τυπική απόκλιση των σφαλμάτων που προκύπτουν από την μέθοδο της διασταυρωτικής επικύρωσης 10 ομάδων.

```
> best<-tree_LCS$cptable[which.min(tree_LCS$cptable[,"xerror
   "]),"CP"]
> pruned_LCS<-prune(tree_LCS,cp=best)
> prp(pruned_LCS,main="Κλαδεμένο Δένδρο Παλινδρόμησης")
```

Κώδικας 4.3: Επιλογή του καλύτερου α και κλάδεμα του δένδρου παλινδρόμησης.

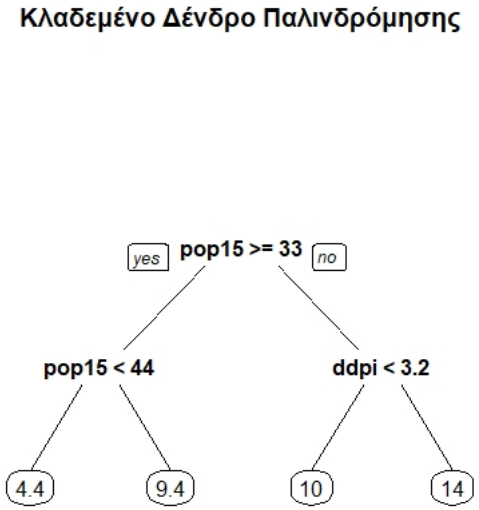
Έπειτα, όπως φαίνεται στον Κώδικα 4.3, αποθηκεύουμε το βέλτιστο α (= 0.01) σε μία νέα μεταβλητή ("best"), επιλέγοντας το χαμηλότερο σφάλμα από την στήλη "xerror". Τέλος, για το εν λόγω α, κλαδεύουμε το αρχικό δένδρο και το αποτέλεσμα φαίνεται στο Σχήμα 4.2, όπου οι τελικοί κόμβοι του δένδρου είναι 4.

Μπορούμε να πραγματοποιήσουμε προβλέψεις με βάση το κλαδεμένο δένδρο παλινδρόμησης στο σύνολο ελέγχου ("testdata"), όπως φαίνεται στον Κώδικα 4.4.

```
> testdata<-LifeCycleSavings[-train,]
> yhat<-predict(pruned_LCS,newdata=testdata)
> LifeCycleSavings_test<-testdata$sr
> plot(yhat,LifeCycleSavings_test,xlab="Προβλεπόμενη Τιμή",ylab=
"Παρατηρούμενη Τιμή",main="Διάγραμμα Προβλέψεων στο Σύνολο Ελέγχου")
> abline(0,1)
> mean((LifeCycleSavings_test-yhat)^2)
[1] 17.72186
```

Κώδικας 4.4: Προβλέψεις στο σύνολο ελέγχου με χρήση του κλαδεμένου δένδρου παλινδρόμησης για τα δεδομένα LifeCycleSavings.

Σύμφωνα με το Σχήμα 4.3, οι προβλέψεις που προκύπτουν με το κλαδεμένο δένδρο για το σύνολο ελέγχου είναι αρκετά ικανοποιητικές, καθώς οι μεγαλύτερες προβλεπόμενες τιμές τείνουν να πλησιάσουν τις αντίστοιχες παρατηρούμενες (πραγματικές) τιμές. Από τον κώδικα 4.4, το μέσο τετραγωνικό σφάλμα MSE



Σχήμα 4.2: Κλαδεμένο δένδρο παλινδρόμησης (με 4 φύλλα) για τα δεδομένα εκπαίδευσης του συνόλου LifeCycleSavings.

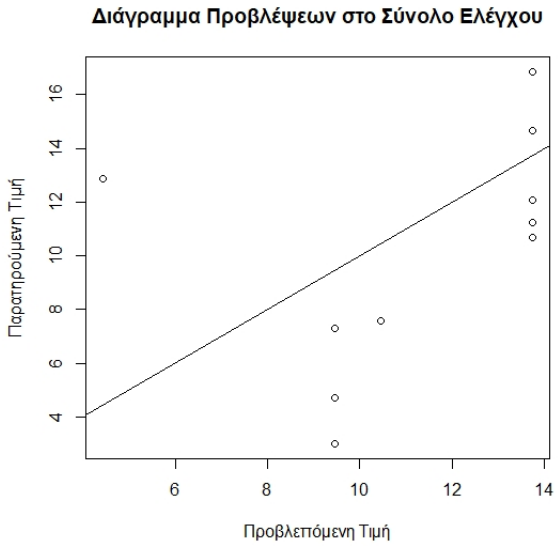
για το σύνολο ελέγχου είναι 17.72186. Η παρατηρούμενη και η προβλεπόμενη τιμή για την μεταβλητή απόκρισης sr δηλώνονται ως "LifeCycleSavings_test" και "yhat" αντίστοιχα. Η τετραγωνική ρίζα του MSE είναι περίπου 4.2097 και, επομένως, το συγκεκριμένο μοντέλο οδηγεί σε προβλέψεις που είναι ίσες, περίπου, με το 4.2097 του πραγματικού λόγου αποταμίευσης.

Οι προβλέψεις στο ίδιο σύνολο ελέγχου ("testdata"), αν χρησιμοποιούσαμε το αρχικό, ακλάδευτο δένδρο παλινδρόμησης φαίνονται στον Κώδικα 4.5.

```
> testdata<-LifeCycleSavings[-train,]
> yhat0<-predict(tree_LifeCycleSavings,newdata=testdata)
> LifeCycleSavings_test<-testdata$sr
> mean((LifeCycleSavings_test-yhat0)^2)
[1] 20.8751
```

Κώδικας 4.5: Προβλέψεις στο σύνολο ελέγχου με χρήση του ακλάδευτου δένδρου παλινδρόμησης για τα δεδομένα LifeCycleSavings.

Παρατηρούμε ότι το MSE για το σύνολο ελέγχου είναι 20.8751, ενώ το αντίστοιχο MSE με χρήση του κλαδεμένου δένδρου ήταν 17.72186, γεγονός που επιβεβαιώνει ότι το κλάδεμα βελτιώνει σημαντικά την ακρίβεια των τελικών



Σχήμα 4.3: Διάγραμμα προβλέψεων στο σύνολο ελέγχου με χρήση του κλαδεμένου δένδρου παλινδρόμησης για τα δεδομένα LifeCycleSavings.

προβλέψεων.

Έπειτα, εφαρμόζουμε τη μέθοδο της ενσάκισης και των τυχαίων δασών, οι οποίες επεξηγήθηκαν στο Κεφάλαιο 3, στα LifeCycleSavings δεδομένα. Θα χρησιμοποιήσουμε το πακέτο "randomForest" της R, επικαλούμενοι το γεγονός ότι τα τυχαία δάση αποτελούν ειδική περίπτωση της ενσάκισης για $m \approx \sqrt{p}$.

```
> library(randomForest)
> set.seed(1)
> bag_LifeCycleSavings<-randomForest(sr~.,data=
   LifeCycleSavings,subset=train,
mtry=4,importance=TRUE)
> bag_LifeCycleSavings

Call:
 randomForest(formula = sr ~ ., data = LifeCycleSavings, mtry
    = 4,      importance = TRUE, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 17.45832
                    % Var explained: 13.6
```

Κώδικας 4.6: Μέθοδος της ενσάκισης στα δεδομένα εκπαίδευσης του συνόλου LifeCycleSavings.

Έτσι, με την βοήθεια της συνάρτησης randomForest() κατασκευάζεται το μοντέλο (3.7), το οποίο θα αποτελείται από 500 δένδρα παλινδρόμησης, σύμφωνα με τα αποτελέσματα του Κώδικα 4.6. Η προσαρμογή των δένδρων αυτών έγινε λαμβάνοντας υπόψη και τις 4 επεξηγηματικές μεταβλητές για την κάθε τμήση του κάθε δένδρου παλινδρόμησης, όπως δηλώνει το όρισμα mtry=4 της εντολής randomForest(). Το MSE για τα δεδομένα εκπαίδευσης προκύπτει ότι είναι περίπου ίσο με 17.46, ενώ η μεταβλητότητα της μεταβλητής απόκρισης sr που εξηγείται από το εν λόγω μοντέλο (variance explained) εκφράζει πόσο καλά οι out-of-bag προβλέψεις εξηγούν την μεταβλητή απόκρισης του συνόλου εκπαίδευσης και λαμβάνει την τιμή 13.6 (αρκετά χαμηλή). Αυτό σημαίνει ότι το μοντέλο είναι 'φτωχό', δηλαδή έχει κακή απόδοση με ανακριβείς προβλέψεις. Το γεγονός αυτό πιθανόν να οφείλεται στο πρόβλημα υπερπροσαρμογής της μεθόδου της ενσάκισης για το οποίο έγινε αναφορά στο προηγούμενο κεφάλαιο και, ταυτόχρονα, υποδηλώνει ότι κάποια άλλη μέθοδος παλινδρόμησης θα ήταν ίσως προτιμότερη για το εν λόγω πρόβλημα.

Με βάση το μοντέλο που προκύπτει από τη μέθοδο της ενσάκισης πραγματοποιούνται προβλέψεις στο σύνολο ελέγχου ("testdata") για την μεταβλητή sr, όπως φαίνεται στον Κώδικα 4.7.

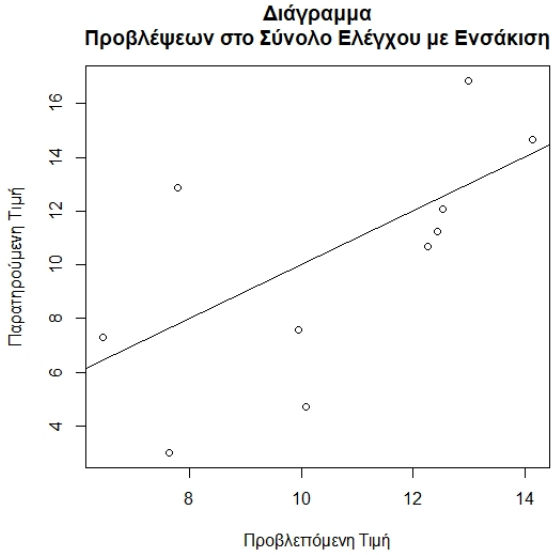
```
> yhat_bag<-predict(bag_LifeCycleSavings,newdata=testdata)
> plot(yhat_bag,LifeCycleSavings_test,xlab="Προβλεπόμενη Τιμή",
ylab="Παρατηρούμενη Τιμή",
main="Διάγραμμα Προβλέψεων στο Σύνολο Ελέγχου με Ενσάκιση")
> abline(0,1)
> mean((LifeCycleSavings_test-yhat_bag)^2)
[1] 10.20393
```

Κώδικας 4.7: Προβλέψεις στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου της ενσάκισης για τα δεδομένα LifeCycleSavings.

Σύμφωνα με το Σχήμα 4.4, παρατηρούμε ότι υπάρχουν κάποιες αποκλίσεις από τις παρατηρούμενες τιμές της εξαρτημένης μεταβλητής sr, πιθανόν, λόγω υπερπροσαρμογής του εν λόγω μοντέλου. Το MSE για το σύνολο ελέγχου στην περίπτωση της ενσάκισης είναι 10.20393, δηλαδή αρκετά μικρότερο από αυτό που προέκυψε όταν χρησιμοποιήσαμε το κλαδεμένο δένδρο (MSE=17.72186).

Στον Κώδικα 4.8, το πλήθος των δένδρων που αναπτύσσονται κατά την εκτέλεση της ενσάκισης έχει ορισθεί να είναι 30, μέσω του επιχειρήματος ntree:

```
> bag_LifeCycleSavings<-randomForest(sr~.,data=
   LifeCycleSavings,subset=train,
mtry=4,ntree=30)
> yhat_bag<-predict(bag_LifeCycleSavings,newdata=testdata)
> mean((LifeCycleSavings_test-yhat_bag)^2)
```



Σχήμα 4.4: Διάγραμμα προβλέψεων στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου της ενσάκισης για τα δεδομένα LifeCycleSavings.

```
[1] 11.88487
```

Κώδικας 4.8: Εφαρμογή του ορίσματος ntree στην μέθοδο της ενσάκισης για τα δεδομένα LifeCycleSavings.

Η εφαρμογή της μεθόδου των τυχαίων δασών στο σύνολο των δεδομένων εκπαίδευσης γίνεται με παρόμοιο τρόπο, όπως στην ενσάκιση. Χρησιμοποιείται η συνάρτηση randomForest() και ως τιμή στο όρισμα mtry ορίζουμε την τιμή 2 ή, με άλλα λόγια, λαμβάνεται τυχαίο δείγμα μεγέθους $m = \sqrt{p} = \sqrt{4} = 2$ χαρακτηριστικών (ανεξάρτητων μεταβλητών) ως υποψήφιες μεταβλητές τμήσης από το ολικό σύνολο των p χαρακτηριστικών.

```
> set.seed(1)
> rf_LifeCycleSavings<-randomForest(sr~.,data=
   LifeCycleSavings,subset=train,
mtry=2,importance=TRUE)
> rf_LifeCycleSavings

Call:
 randomForest(formula = sr ~ ., data = LifeCycleSavings, mtry
    = 2,      importance = TRUE, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2
```

```
          Mean of squared residuals: 17.02946
                    % Var explained: 15.73
```

Κώδικας 4.9: Μέθοδος των τυχαίων δασών στα δεδομένα εκπαίδευσης του συνόλου LifeCycleSavings.

Σύμφωνα με τον Κώδικα 4.9, το MSE για τα δεδομένα εκπαίδευσης προκύπτει ότι είναι περίπου ίσο με 17.03 (μικρότερο από το αντίστοιχο MSE της μεθόδου της ενσάκισης που ήταν, κατά προσέγγιση, ίσο με 17.46), ενώ η μεταβλητότητα της μεταβλητής απόκρισης sr που εξηγείται από το εν λόγω μοντέλο λαμβάνει στην περίπτωση των τυχαίων δασών τιμή ίση με 15.73%. Ομοίως με πριν, αυτό σημαίνει ότι το μοντέλο δεν είναι αρκετά αποδοτικό και χρειάζεται βελτίωση.

Στην συνέχεια, με βάση τον Κώδικα 4.10, πραγματοποιούνται οι αντίστοιχες προβλέψεις:

```
> yhat_rf<-predict(rf_LifeCycleSavings,newdata=testdata)
> plot(yhat_rf,LifeCycleSavings_test,xlab="Προβλεπόμενη Τιμή",
ylab="Παρατηρούμενη Τιμή",
main="Διάγραμμα Προβλέψεων στο Σύνολο Ελέγχου με Τυχαία Δάση")
> abline(0,1)
> mean((LifeCycleSavings_test-yhat_rf)^2)
[1] 10.61687
```

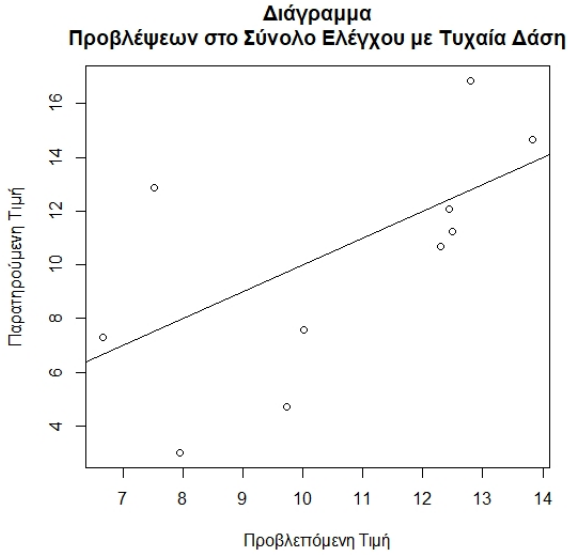
Κώδικας 4.10: Προβλέψεις στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου των τυχαίων δασών για τα δεδομένα LifeCycleSavings.

Από το Σχήμα 4.5 παρατηρούμε ότι, ομοίως με την περίπτωση της ενσάκισης, υπάρχουν κάποιες αποκλίσεις των προβλεπόμενων από τις παρατηρούμενες τιμές. Επιπροσθέτως, το MSE για το σύνολο ελέγχου στην περίπτωση των τυχαίων δασών (MSE=10.61687) δεν διαφέρει σημαντικά από αυτό της ενσάκισης (MSE=10.20393).

Για να αποκτήσουμε μία εικόνα σχετικά με την σημαντικότητα και, συνεπώς, την επίδραση της κάθε μεταβλητής στην κατασκευή του τυχαίου δάσους, επικαλούμαστε την συνάρτηση importance() της R. Ακόμα, με την συνάρτηση varImpPlot() δημιουργούνται διαγράμματα σχετικά με τα μέτρα σημαντικότητας των μεταβλητών (Κώδικας 4.11).

```
> importance(rf_LifeCycleSavings)
      %IncMSE  IncNodePurity
pop15 18.051983      252.9888
pop75  6.403249      161.9993
dpi    3.548197      115.9361
ddpi   4.738600      193.3919
> varImpPlot(rf_LifeCycleSavings)
```

Κώδικας 4.11: Σημαντικότητα της κάθε μεταβλητής στο τυχαίο δάσος για τα δεδομένα LifeCycleSavings.



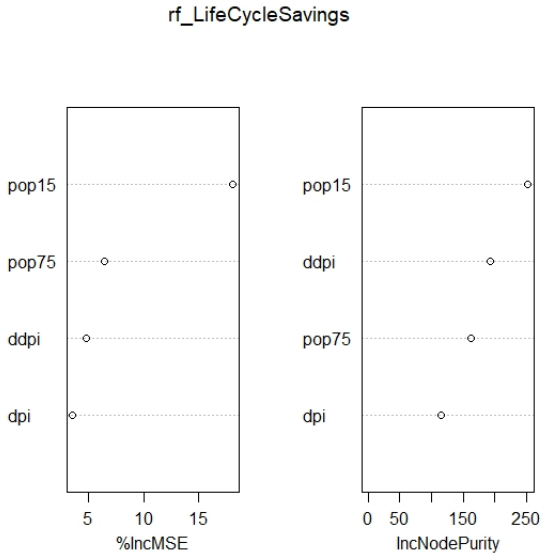
Σχήμα 4.5: Διάγραμμα προβλέψεων στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου των τυχαίων δασών για τα δεδομένα LifeCycleSavings.

Στον Κώδικα 4.11, το πρώτο μέτρο σημαντικότητας αναφέρεται στην μέση μείωση της αποδοτικότητας, άρα στην αύξηση του μέσου τετραγωνικού σφάλματος των προβλέψεων, βασιζόμενες στα out-of-bag δείγματα, όταν μία δοσμένη μεταβλητή αφαιρείται από το μοντέλο. Οι μεταβλητές που έχουν μεγάλες τιμές κρίνονται ως οι πιο σημαντικές. Αυτό, διότι αν κατασκευαστεί το μοντέλο, χωρίς μία μεταβλητή και δώσει χειρότερες προβλέψεις (μεγαλύτερο σφάλμα) από το αρχικό μοντέλο (με όλες τις μεταβλητές), τότε σημαίνει ότι η εν λόγω μεταβλητή είναι σημαντική.

Το δεύτερο μέτρο σημαντικότητας εκφράζει την ολική μείωση της ακαθαρσίας ενός κόμβου που οφείλεται σε τμήσεις, βάσει μίας συγκεκριμένης μεταβλητής. Η ολική αυτή μείωση προκύπτει από την μέση τιμή της εκάστοτε μείωσης κάθε δένδρου και στην περίπτωση της παλινδρόμησης η ακαθαρσία ενός κόμβου μετρείται μέσω του αθροίσματος των τετραγώνων των υπολοίπων (SSR) των δεδομένων εκπαίδευσης. Γενικά, προτιμάται εκείνος ο (τελικός) κόμβος που έχει την μεγαλύτερη καθαρότητα, ο κόμβος, δηλαδή, όπου όλες οι παρατηρήσεις του συνόλου εκπαίδευσης δίνουν τις ίδιες εκτιμήσεις-προβλέψεις.

Παρατηρώντας, λοιπόν, το Σχήμα 4.6 και με βάση τον Κώδικα 4.11, η σημαντικότερη μεταβλητή για την μέθοδο των τυχαίων δασών είναι η pop15, καθώς σε αυτήν οφείλεται η μεγαλύτερη αύξηση στο σφάλμα και η μεγαλύτερη αύξηση της καθαρότητας του κόμβου.

Τέλος, πραγματοποιείται η μέθοδος της ενίσχυσης (Gradient Boosting)



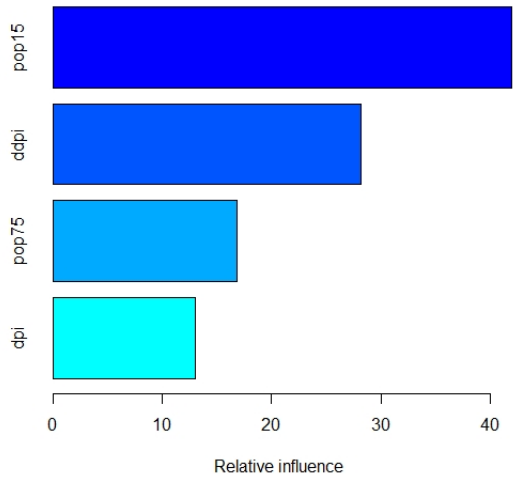
Σχήμα 4.6: Διάγραμμα μέτρων σημαντικότητας των μεταβλητών στην μέθοδο των τυχαίων δασών για τα δεδομένα LifeCycleSavings.

στο εν λόγω σύνολο δεδομένων, LifeCycleSavings, χρησιμοποιώντας το πακέτο "gbm" της R (Κώδικας 4.12).

```
> library(gbm)
> set.seed(1)
> boost_LifeCycleSavings<-gbm(sr~.,data=LifeCycleSavings[
   train,],distribution=
 "gaussian",n.trees=1000,interaction.depth=4,bag.fraction
   =0.85)
> summary(boost_LifeCycleSavings)
        var  rel.inf
pop15 pop15 41.94150
ddpi   ddpi 28.17941
pop75 pop75 16.84967
dpi     dpi 13.02943
```

Κώδικας 4.12: Μέθοδος της ενίσχυσης στα δεδομένα εκπαίδευσης του συνόλου LifeCycleSavings.

Στον Κώδικα 4.12 έχουμε εφαρμόσει τη μέθοδο της ενίσχυσης στα δεδομένα εκπαίδευσης (training data) με τη βοήθεια της συνάρτησης gbm() στην οποία έχουμε θέσει κάποια ορίσματα. Αρχικά, επειδή το πρόβλημα που μελετάμε είναι ένα πρόβλημα παλινδρόμησης, έχουμε θέσει ως κατανομή την γκαουσιανή μέσω του ορίσματος distribution="gaussian". Ακόμα, έχουμε δηλώσει τον αριθμό των παραγόμενων δένδρων κατά την ενίσχυση να είναι ίσος με 1000 δένδρα



Σχήμα 4.7: Σχετική επίδραση των μεταβλητών στην μέθοδο της ενίσχυσης για τα δεδομένα LifeCycleSavings.

(n.trees=1000), όπως και το βάθος αλληλεπίδρασης ή, αλλιώς, τον αριθμό των τμήσεων d κάθε δένδρου της διαδικασίας να είναι 4 (interaction.depth=4). Το όρισμα bag.fraction της συνάρτησης gbm() εκφράζει το κλάσμα των παρατηρήσεων εκπαίδευσης που διαλέγονται τυχαία για την κατασκευή του επόμενου δένδρου σε κάθε επανάληψη. Η R έχει ως προκαθορισμένη (default) τιμή το 0.5, δηλώνοντας έτσι ότι σε κάθε επανάληψη χρησιμοποιείται το μισό από το δείγμα εκπαίδευσης για την δημιουργία του επόμενου δένδρου. Ειδικότερα, εάν το δείγμα εκπαίδευσης είναι μικρό, τότε μπορεί να πάρει τιμές μεγαλύτερες του 0.5. Στο εν λόγω πρόβλημα, το δείγμα εκπαίδευσης είναι αρκετά μικρό, άρα μπορούμε να θέσουμε το bag.fraction=0.85.

Σύμφωνα με το Σχήμα 4.7, οι μεταβλητές pop15 και ddpi έχουν μεγάλη επίδραση κατά την διαδικασία της ενίσχυσης για την συγκεκριμένη επιλογή των παραμέτρων, ενώ η μεταβλητή dpi έχει την μικρότερη επίδραση κατά την κατασκευή του μοντέλου. Οι προβλέψεις της μεταβλητής απόκρισης sr από το σύνολο ελέγχου (testdata) με το μοντέλο ενίσχυσης φαίνονται στον Κώδικα 4.13:

```
> yhat_boost<-predict(boost_LifeCycleSavings,newdata=testdata
  ,n.trees=1000)
> mean((LifeCycleSavings_test-yhat_boost)^2)
```

```
[1] 27.38079
```

Κώδικας 4.13: Προβλέψεις στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου της ενίσχυσης για τα δεδομένα LifeCycleSavings.

Όσον αφορά την παράμετρο συρρίκνωσης λ, η R περιέχει ως προκαθορισμένη (default) τιμή το 0.001. Θέτοντας μία μεγαλύτερη τιμή για το λ, έστω λ=0.05, με το όρισμα shrinkage=0.05, παρατηρείται μία μείωση στο MSE για το σύνολο ελέγχου (Κώδικας 4.14):

```
> boost_LifeCycleSavings<-gbm(sr~.,data=LifeCycleSavings[
   train,],distribution=
 "gaussian",n.trees=1000,interaction.depth=4,shrinkage=0.05,
   bag.fraction=0.85)
> yhat_boost<-predict(boost_LifeCycleSavings,newdata=testdata
   ,n.trees=1000)
> mean((LifeCycleSavings_test-yhat_boost)^2)
[1] 23.05059
```

Κώδικας 4.14: Προβλέψεις στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου της ενίσχυσης για τα δεδομένα LifeCycleSavings και για λ=0.05.

Πίνακας 4.1: Σύνοψη των αποτελεσμάτων που προέκυψαν από την δενδρική παλινδρόμηση στο σύνολο ελέγχου των δεδομένων LifeCycleSavings, με βάση το Μέσο Τετραγωνικό Σφάλμα (MSE).

Μοντέλο (Δένδρο/Δένδρα Παλινδρόμησης)	MSE
Αρχικό, ακλάδευτο δένδρο παλινδρόμησης	20.8751
Κλαδεμένο δένδρο παλινδρόμησης	17.72186
Μοντέλο από την μέθοδο της ενσάκισης	10.20393
Μοντέλο από την μέθοδο των τυχαίων δασών	10.61687
Μοντέλο από την μέθοδο της ενίσχυσης (για λ= 0.001)	27.38079
Μοντέλο από την μέθοδο της ενίσχυσης (για λ= 0.05)	23.05059

Συνεπώς, σύμφωνα με τον Πίνακα 4.1, οδηγούμαστε στο συμπέρασμα ότι το καλύτερο μοντέλο (με την έννοια της ακρίβειας) για τα δεδομένα LifeCycleSavings είναι αυτό που προκύπτει μέσω της μεθόδου της ενσάκισης, καθώς δίνει το χαμηλότερο μέσο τετραγωνικό σφάλμα αν αυτό εφαρμοστεί στο σύνολο ελέγχου ("testdata") των εν λόγω δεδομένων.

4.1.2 Πολλαπλή Γραμμική Παλινδρόμηση

Έχοντας αναλύσει και επεξεργαστεί το πρόβλημα της υπόθεσης του κύκλου ζωής με τη μέθοδο της δενδρικής παλινδρόμησης, θα δούμε σε αυτή την ενότητα πώς μπορεί να εφαρμοστεί η κλασική μέθοδος της γραμμικής παλινδρόμησης

στα ίδια δεδομένα. Συγκεκριμένα, θα ασχοληθούμε με την πολλαπλή γραμμική παλινδρόμηση, καθώς το πρόβλημα περιλαμβάνει τέσσερις επεξηγηματικές μεταβλητές. Ωστόσο, θα εφαρμόσουμε πολλαπλή γραμμική παλινδρόμηση χρησιμοποιώντας μόνο τις μεταβλητές "pop15" και "ddpi", προκειμένου το μοντέλο της γραμμικής παλινδρόμησης να έχει την ίδια πολυπλοκότητα με τα μοντέλα που προέκυψαν από την μέθοδο της δενδρικής παλινδρόμησης.

Αρχικά, το μοντέλο που θα προσαρμόσουμε θα είναι ένα μοντέλο της μορφής

$$Y_i \ = \ a + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \ , i = 1, ..., 50. \tag{4.1}$$

Η μεταβλητή Y_i εκφράζει την μεταβλητή απόκρισης "sr" του εν λόγω προβλήματος. Η μεταβλητή X_{i1} αντιστοιχεί στην επεξηγηματική μεταβλητή "pop15" του προβλήματος και η μεταβλητή X_{i2} αντιστοιχεί στην μεταβλητή "ddpi" του προβλήματος. Ο δείκτης i δηλώνει την i-οστή παρατήρηση της εκάστοτε μεταβλητής και παίρνει τιμές από 1 έως 50, αφού διαθέτουμε 50 συνολικά παρατηρήσεις. Η σταθερά a εκφράζει την μέση τιμή της μεταβλητής απόκρισης Y όταν όλα τα $X_j, j = 1, ..., p$, με $p = 2$ είναι μηδέν. Ωστόσο, για το εν λόγω πρόβλημα αυτό δεν είναι εφικτό, σύμφωνα με τον ορισμό της κάθε επεξηγηματικής μεταβλητής, συνεπώς η τιμή αυτή δεν έχει ερμηνεία στην προκειμένη περίπτωση. Η παράμετρος $\beta_j, j = 1, 2$, εκφράζει την αναμενόμενη μεταβολή (αύξηση ή μείωση) της τιμής της τυχαίας μεταβλητής Y όταν η X_j αυξηθεί κατά μία μονάδα, δεδομένου ότι η $X_k, k \neq j$ παραμένει σταθερή. Τα ε_i ονομάζονται τυχαία σφάλματα και υποθέτουμε ότι είναι ανεξάρτητα μεταξύ τους και ισόνομα με $\varepsilon_i \sim N(0, \sigma^2)$ με την διασπορά των σφαλμάτων σ^2 να είναι άγνωστη.

Ομοίως με την περίπτωση της δενδρικής παλινδρόμησης, φορτώνουμε στην R την βιβλιοθήκη του πακέτου "datasets", το οποίο περιέχει το σύνολο δεδομένων "LifeCycleSavings" που θα χρησιμοποιήσουμε. Με τη βοήθεια της συνάρτησης lm() προσαρμόζουμε το γραμμικό μοντέλο, ορίζοντας ως μεταβλητή απόκρισης την μεταβλητή sr και ως επεξηγηματικές τις δύο μεταβλητές pop15 και ddpi (Κώδικας 4.15).

```
> library(datasets)
> sr<-LifeCycleSavings$sr
> pop15<-LifeCycleSavings$pop15
> pop75<-LifeCycleSavings$pop75
> dpi<-LifeCycleSavings$dpi
> ddpi<-LifeCycleSavings$ddpi
> mod_LifeCycleSavings<-lm(sr~pop15+ddpi)
```

Κώδικας 4.15: Ορισμός των μεταβλητών του συνόλου δεδομένων LifeCycleSavings και προσαρμογή του μοντέλου.

Ωστόσο, πρέπει να σημειωθεί σε αυτό το σημείο ότι η εφαρμογή της (πολλαπλής) γραμμικής παλινδρόμησης προαπαιτεί το μοντέλο (4.1) να πληρεί κάποιες

προϋποθέσεις, οι οποίες επεξηγούνται αναλυτικά στο Παράρτημα και θα αναφερθούν στην συνέχεια.

Αρχικά, γίνεται έλεγχος συσχέτισης των επεξηγηματικών μεταβλητών:

```
> cor(pop15,ddpi)
[1] -0.04782569
```

Κώδικας 4.16: Συσχέτιση μεταξύ των μεταβλητών του συνόλου δεδομένων LifeCycleSavings.

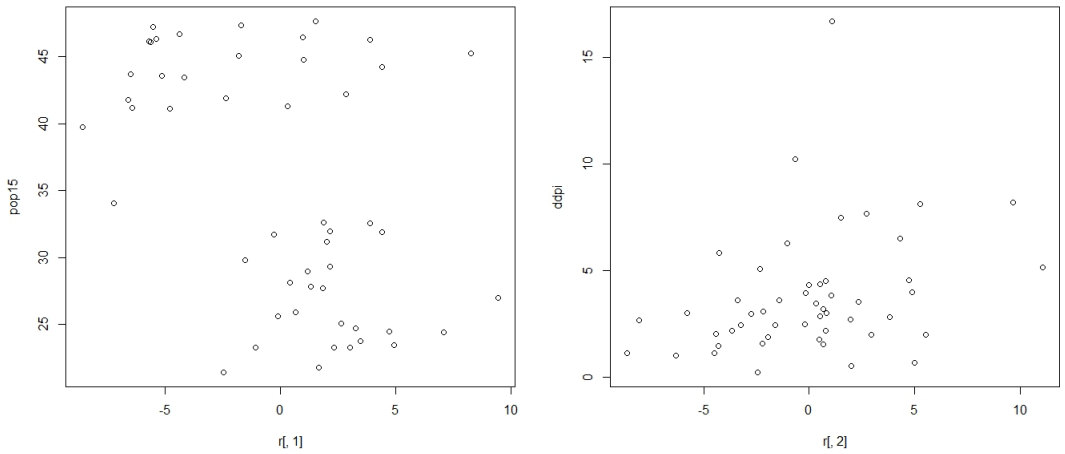
Όπως φαίνεται από τον Κώδικα 4.16, οι μεταβλητές συσχετίζονται μεταξύ τους.

Στην συνέχεια, ελέγχεται αν ικανοποιούνται οι προϋποθέσεις του γραμμικού μοντέλου.

- Γραμμικότητα:

```
> r<-residuals(mod_LifeCycleSavings,"partial")
> par(mfrow=c(1,2))
> plot(r[,1],pop15)
> plot(r[,2],ddpi)
```

Κώδικας 4.17: Εξέταση της υπόθεσης της γραμμικότητας για το μοντέλο mod_LifeCycleSavings.



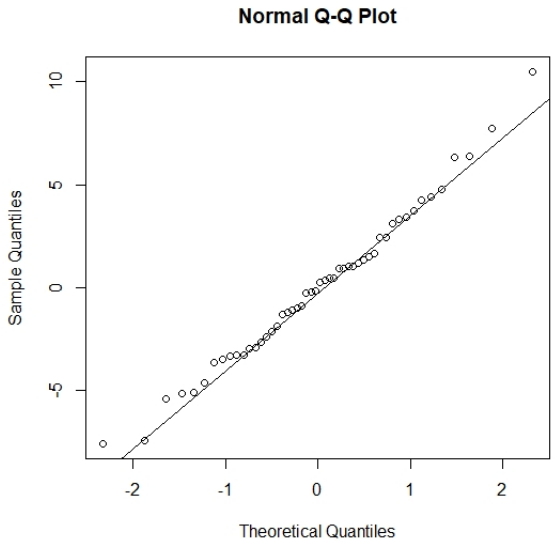
Σχήμα 4.8: Γραμμικότητα για το μοντέλο mod_LifeCycleSavings.

Η πρώτη εντολή στον Κώδικα 4.17 δίνει τα 'μερικά υπόλοιπα' για τα οποία γίνεται λόγος στο Παράρτημα που αναφέρθηκε. Με βάση το Σχήμα 4.8 παρατηρούμε ότι η υπόθεση της γραμμικότητας δεν είναι ξεκάθαρη, αλλά ούτε και απίθανη για τα εν λόγω δεδομένα.

- Κανονικότητα των Σφαλμάτων:

```
> qqnorm(residuals(mod_LifeCycleSavings))
> qqline(residuals(mod_LifeCycleSavings))
```

Κώδικας 4.18: Εξέταση της υπόθεσης της κανονικότητας των σφαλμάτων για το μοντέλο mod_LifeCycleSavings.



Σχήμα 4.9: Κανονικότητα σφαλμάτων για το μοντέλο mod_LifeCycleSavings.

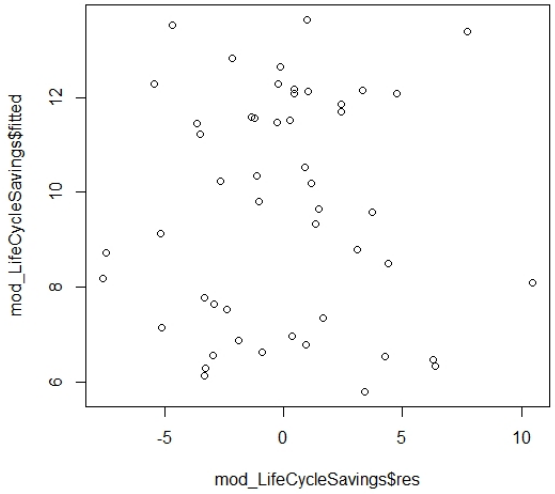
Σύμφωνα με το Σχήμα 4.9, το οποίο προκύπτει από τον Κώδικα 4.18, δεν υπάρχουν σημαντικές αποκλίσεις των ποσοστιαίων σημείων των υπολοίπων από τα ποσοστιαία σημεία της κανονικής κατανομής, καθώς τείνουν να πλησιάσουν την ευθεία. Επομένως, υπάρχουν σημαντικές ενδείξεις υπέρ της υπόθεσης της κανονικότητας των σφαλμάτων.

- Ομοσκεδαστικότητα:

```
> plot(mod_LifeCycleSavings$res,mod_LifeCycleSavings$fitted)
```

Κώδικας 4.19: Εξέταση της υπόθεσης της ομοσκεδαστικότητας για το μοντέλο mod_LifeCycleSavings.

Από το Σχήμα 4.10, όπως αυτό προκύπτει από την εφαρμογή του Κώδικα 4.19, παρατηρούμε ότι τα σημεία είναι τυχαία διασκορπισμένα, χωρίς να παρουσιάζουν κάποιο συστηματικό τρόπο συμπεριφοράς, επομένως ικανοποιείται η υπόθεση της ομοσκεδαστικότητας.



Σχήμα 4.10: Ομοσκεδαστικότητα για το μοντέλο mod_LifeCycleSavings.

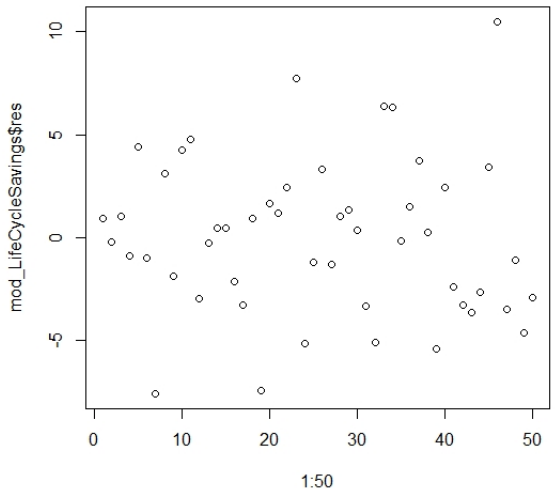
- Ανεξαρτησία των Σφαλμάτων:

```
> plot(1:50,mod_LifeCycleSavings$res)
```

Κώδικας 4.20: Εξέταση της υπόθεσης της ανεξαρτησίας των σφαλμάτων για το μοντέλο mod_LifeCycleSavings.

Ο Κώδικας 4.20 κατασκευάζει το Σχήμα 4.11, τα σημεία του οποίου δείχνουν να είναι τυχαία κατανεμημένα, δίχως να ακολουθούν κάποιο μοτίβο (pattern) και, άρα μπορούμε να συμπεράνουμε ότι ισχύει η υπόθεση της ανεξαρτησίας των σφαλμάτων.

Συνεπώς, εφόσον ικανοποιούνται οι τέσσερις προϋποθέσεις για την εφαρμογή της μεθόδου της πολλαπλής γραμμικής παλινδρόμησης, μπορούμε να εμπιστευτούμε τις προβλέψεις που δίνει το μοντέλο (4.1) για τις τιμές της μεταβλητής απόκρισης sr, το οποίο ορίσαμε στην R ως "mod_LifeCycleSavings". Εκτελώντας την εντολή summary() της R, λαμβάνουμε πληροφορίες για το εν λόγω γραμμικό μοντέλο και κάποιες εκτιμήσεις για τους συντελεστές β_j, $j = 1, 2, 3, 4$ και για την σταθερά a:



Σχήμα 4.11: Ανεξαρτησία σφαλμάτων για το μοντέλο mod_LifeCycleSavings.

```
> summary(mod_LifeCycleSavings)

Call:
lm(formula = sr ~ pop15 + ddpi)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5831 -2.8632  0.0453  2.2273 10.4753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.59958    2.33439   6.682 2.48e-08 ***
pop15       -0.21638    0.06033  -3.586 0.000796 ***
ddpi         0.44283    0.19240   2.302 0.025837 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
   1

Residual standard error: 3.861 on 47 degrees of freedom
Multiple R-squared:  0.2878,    Adjusted R-squared:  0.2575
F-statistic: 9.496 on 2 and 47 DF,  p-value: 0.0003438
```

Κώδικας 4.21: Η εντολή summary() για το πλήρες μοντέλο mod_LifeCycleSavings.

Σύμφωνα με τον Κώδικα 4.21, ο συντελεστής προσδιορισμού $R^2 = 0.2878$ και ο προσαρμοσμένος συντελεστής προσδιορισμού $\tilde{R}^2 = 0.2575$ απέχουν ση-

μαντικά από τη μονάδα, επομένως η προσαρμογή του μοντέλου μας δεν είναι πολύ ικανοποιητική. Επίσης, το στατιστικό ελέγχου F του F−test έχει την τιμή 9.496, ενώ η P−τιμή για τον εν λόγω έλεγχο είναι πολύ κοντά στο μηδέν, οπότε έχουμε ισχυρές ενδείξεις για να απορρίψουμε τη μηδενική υπόθεση $\beta_1 = \beta_2 = 0$ (λεπτομέρειες στο Παράρτημα). Ακόμα, παρατηρούμε ότι και οι δύο μεταβλητές είναι στατιστικά σημαντικές (σε επίπεδο σημαντικότητας 5%), όπως είχε προκύψει και στην μέθοδο της δενδρικής παλινδρόμησης.

Συνεπώς, κάνοντας χρήση του εν λόγω μοντέλου, πραγματοποιούνται προβλέψεις για την μεταβλητή απόκρισης sr στο ίδιο σύνολο ελέγχου με αυτό που είχαμε ορίσει στην μέθοδο της δενδρικής παλινδρόμησης ως "testdata" (Κώδικας 4.22).

```
> set.seed(1)
> train<-sample(1:nrow(LifeCycleSavings),0.8*nrow(
   LifeCycleSavings))
> testdata<-LifeCycleSavings[-train,]
> yhat<-predict(mod_LifeCycleSavings,testdata)
> LifeCycleSavings_test<-testdata$sr
> mean((LifeCycleSavings_test-yhat)^2)
[1] 7.441831
```

Κώδικας 4.22: Προβλέψεις με χρήση του μοντέλου mod_LifeCycleSavings σε δεδομένα από το σύνολο ελέγχου.

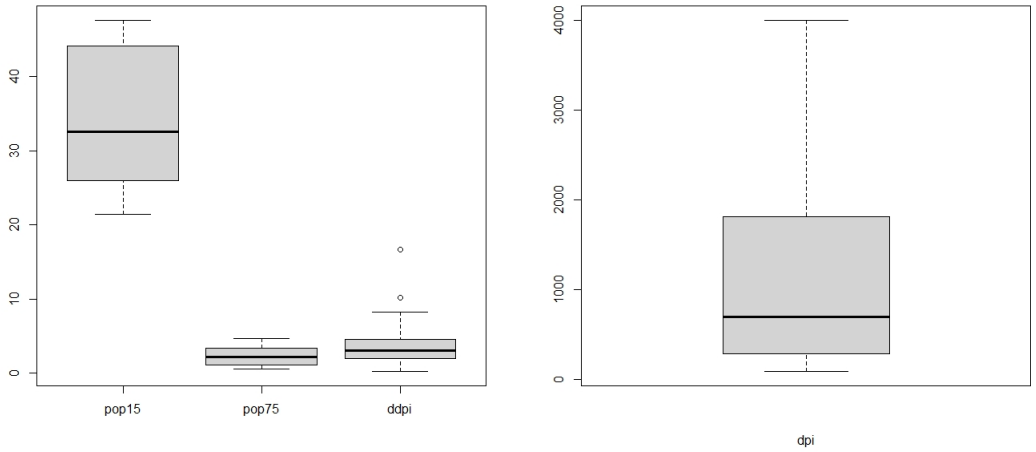
4.2 Συμπεράσματα

Συγκρίνοντας τη μέθοδο της δενδρικής παλινδρόμησης με αυτή της γραμμικής παλινδρόμησης στο πρόβλημα της υπόθεσης του κύκλου ζωής, προκύπτει από τους παραπάνω υπολογισμούς ότι η τελευταία δίνει μικρότερο μέσο τετραγωνικό σφάλμα για το σύνολο ελέγχου όταν πραγματοποιούνται προβλέψεις του λόγου αποταμίευσης (sr) σε αυτό. Συγκεκριμένα, το MSE στην μέθοδο της γραμμικής παλινδρόμησης είναι ίσο με 7.441831, επομένως η χρήση αυτής οδηγεί σε προβλέψεις που είναι ίσες, περίπου, με το 2.73 του πραγματικού λόγου αποταμίευσης, ενώ το MSE του βέλτιστου μοντέλου που προέκυψε από την μέθοδο της δενδρικής παλινδρόμησης (μοντέλο από την μέθοδο της ενσάκισης) είναι ίσο με 10.20393 και, άρα η χρήση αυτής οδηγεί σε προβλέψεις που είναι ίσες, περίπου, με το 3.19 του πραγματικού λόγου αποταμίευσης. Επιπροσθέτως, γνωρίζουμε ότι η μέθοδος της ενίσχυσης στην δενδρική παλινδρόμηση βελτιώνει σημαντικά την απόδοση και την ακρίβεια του μοντέλου έναντι των μεθόδων της ενσάκισης και των τυχαίων δασών. Παρ΄ όλα αυτά, οι δύο τελευταίες τεχνικές έδωσαν μικρότερα σφάλματα (για το σύνολο ελέγχου) από αυτήν της ενίσχυσης, καθώς τα μοντέλα της ενσάκισης και των τυχαίων δασών έδωσαν, αντίστοιχα, σφάλμα

10.20393 και 10.61687, ενώ κάνοντας χρήση του μοντέλου της ενίσχυσης στα δεδομένα ελέγχου προέκυψε σφάλμα ίσο με 27.38079. Το γεγονός αυτό, όταν συμβαίνει, οφείλεται, κυρίως, στην ύπαρξη ακραίων ή έκτροπων τιμών (outliers) στα δεδομένα, δηλαδή τιμών που είναι απομακρυσμένες από τις υπόλοιπες τιμές της μεταβλητής και οι οποίες επηρεάζουν σε μεγάλο βαθμό την διαδικασία της ενίσχυσης. Μία τακτική ανίχνευσης τυχόν ακραίων τιμών είναι η κατασκευή 'θηκοδιαγραμμάτων' ("boxplots") για κάθε μία από τις ανεξάρτητες μεταβλητές που περιλαμβάνονται στο σύνολο δεδομένων (Κώδικας 4.23).

```
> par(mfrow=c(1,2))
> boxplot(pop15,pop75,ddpi,names=c("pop15","pop75","ddpi"))
> boxplot(dpi,xlab=c("dpi"))
```

Κώδικας 4.23: Θηκοδιαγράμματα του συνόλου δεδομένων LifeCycleSavings για ανίχνευση ακραίων τιμών.

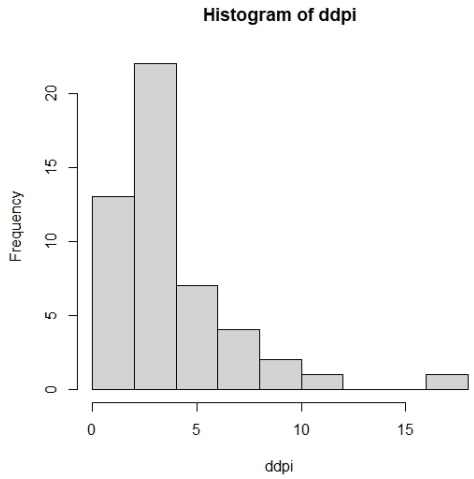


Σχήμα 4.12: Θηκοδιαγράμματα για το σύνολο δεδομένων LifeCycleSavings.

Σύμφωνα με το Σχήμα 4.12, η ύπαρξη μεμονωμένων σημείων στο θηκοδιάγραμμα της μεταβλητής ddpi δηλώνει την παρουσία έκτροπων τιμών στα δεδομένα της εν λόγω μεταβλητής. Το αντίστοιχο ιστόγραμμα συχνοτήτων (frequency histogram) που απεικονίζεται στο Σχήμα 4.13 μάς το επιβεβαιώνει (Κώδικας 4.24):

```
> hist(ddpi)
```

Κώδικας 4.24: Ιστόγραμμα της μεταβλητής ddpi του συνόλου δεδομένων LifeCycleSavings.



Σχήμα 4.13: Ιστόγραμμα της μεταβλητής ddpi του συνόλου δεδομένων LifeCycleSavings.

Η μέθοδος της ενίσχυσης κατασκευάζει κάθε νέο δένδρο με βάση τα υπόλοιπα (σφάλματα) των προηγούμενων δένδρων. Οι ακραίες τιμές μίας μεταβλητής δίνουν μεγαλύτερες τιμές υπολοίπων από μη-ακραίες τιμές, επομένως, κατά την διαδικασία της ενίσχυσης, θα δωθεί μεγαλύτερη έμφαση σε αυτά τα σημεία, συμβάλλοντας έτσι στην κακή απόδοση του μοντέλου.

Τέλος, η ύπαρξη γραμμικότητας μεταξύ των χαρακτηριστικών (προϋπόθεση του γραμμικού μοντέλου) έχει ως αποτέλεσμα η γραμμική παλινδρόμηση να μοντελοποιεί καλύτερα αυτή την (γραμμική) εξάρτηση απ΄ ότι τα τυχαία δάση.

Συνεπώς, όσον αφορά το πρόβλημα της υπόθεσης του κύκλου ζωής, η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης φαίνεται να υπερτερεί έναντι της μεθόδου της δενδρικής παλινδρόμησης.

Εν κατακλείδι, πρέπει να σημειωθεί ότι το σύνολο δεδομένων που χρησιμοποιήθηκε για το πρόβλημα ήταν αρκετά μικρό σε μέγεθος (50 παρατηρήσεις). Η δενδρική παλινδρόμηση και οι επεκτάσεις αυτής δουλεύουν και αποδίδουν καλύτερα σε περισσότερα δεδομένα, επομένως το μέγεθος του αρχικού συνόλου δεδομένων μπορεί να αποτελέσει άλλον έναν λόγο της προβλεπτικής 'αδυναμίας' της δενδρικής παλινδρόμησης.

Κεφάλαιο 5

Εφαρμογή σε πρόβλημα ταξινόμησης

Σε αυτό το κεφάλαιο γίνεται εφαρμογή της μεθόδου της δενδρικής ταξινόμησης, η οποία έχει ορισθεί από τα προηγούμενα κεφάλαια. Το πρόβλημα που θα αναλυθεί αφορά την δυαδική ταξινόμηση, όπου η εξαρτημένη μεταβλητή που μας ενδιαφέρει περιέχει μόνο δύο κατηγορίες ή κλάσεις. Η δενδρική ταξινόμηση μπορεί να επεκταθεί και σε προβλήματα πολλαπλής ταξινόμησης για τα οποία έγινε λόγος στο τέλος του τρίτου κεφαλαίου, ωστόσο μία τέτοια περίπτωση δεν αφορά το προς μελέτη στατιστικό πρόβλημα. Όπως και στο προηγούμενο κεφάλαιο, τα αποτελέσματα που θα προκύψουν από τη μέθοδο της δενδρικής ταξινόμησης θα συγκριθούν με τα αποτελέσματα που δίνει κάποια άλλη στατιστική μέθοδος, ώστε να διεξαχθούν τα ανάλογα συμπεράσματα, σχετικά με την καταλληλότητα της πρώτης μεθόδου (για το συγκεκριμένο πρόβλημα). Η δεύτερη μέθοδος που θα εξεταστεί είναι η μέθοδος της λογιστικής παλινδρόμησης, η οποία αναλύεται λεπτομερώς στο Παράρτημα που ακολουθεί, καθώς δεν αφορά το κύριο μέρος αυτής της εργασίας.

5.1 Ανάλυση του προβλήματος

Το πρόβλημα ταξινόμησης πάνω στο οποίο θα βασιστεί η μέθοδος της δενδρικής ταξινόμησης αφορά τα 'Αποτελέσματα Αναζήτησης Εξωπλανητών', όπως αυτά προέκυψαν από το Διαστημικό Παρατηρητήριο Kepler, ύστερα από εξέταση 10000 υποψήφιων εξωπλανητών ("Kepler Exoplanet Search Results").

Εξωπλανήτης καλείται ένας πλανήτης που δεν ανήκει στο δικό μας ηλιακό σύστημα. Το Διαστημικό Παρατηρητήριο Kepler είναι ένας διαστημικός δορυφόρος που κατασκεύασε η NASA, ο οποίος εκτοξεύτηκε και τέθηκε σε λειτουργία το 2009. Ο Kepler αποτελεί ένα διαστημικό τηλεσκόπιο, αφιερωμένο

στην αναζήτηση εξωπλανητών σε αστρικά συστήματα πέρα από το δικό μας, με απώτερο στόχο την ανακάλυψη νέων, κατοικήσιμων πλανητών στο μέγεθος της Γης που να βρίσκονται σε τροχιά γύρω από άλλα αστέρια.

Η αρχική αποστολή έληξε το 2013 λόγω μηχανικών βλαβών, αλλά το τηλεσκόπιο, παρ᾽ όλα αυτά, είναι λειτουργικό από το 2014 σε μία εκτεταμένη αποστολή "K2". Ο Kepler είχε επαληθεύσει 1284 νέους εξωπλανήτες από τον Μάιο του 2016. Από τον Οκτώβριο του 2017 υπάρχουν συνολικά πάνω από 3000 επιβεβαιωμένοι εξωπλανήτες (κάνοντας χρήση όλων των μεθόδων ανίχνευσης, συμπεριλαμβανομένων των επίγειων). Το τηλεσκόπιο είναι ακόμα ενεργό και συνεχίζει να συλλέγει νέα δεδομένα στην εκτεταμένη αποστολή του, βρισκόμενο σε ασφαλή τροχιά μακριά από τη Γη.

Τα προς επεξεργασία δεδομένα έχουν ληφθεί από τον ιστότοπο "Kaggle" (διαδικτυακή κοινότητα στατιστικής που προσφέρει στους χρήστες πληθώρα συνόλων δεδομένων που αφορούν ποικίλα θέματα). Το σύνολο δεδομένων που θα χρησιμοποιήσουμε ονομάζεται "Kepler Exoplanet Search Results" και αφορά το πρόβλημα της αναζήτησης εξωπλανητών, σύμφωνα με παρατηρήσεις και καταγραφές που πραγματοποιήθηκαν από το Kepler. Πιο αναλυτικά, είναι μία καταγραφή όλων των παρατηρηθέντων 'αντικειμένων ενδιαφέροντος' του Kepler, ή, για συντομία, όλων των "KOI" (Kepler's Object of Interest) που, κατά προσέγγιση, περιλαμβάνουν όλους τους 10000 υποψήφιους εξωπλανήτες στους οποίους έχει κάνει παρατηρήσεις ο Kepler. Τα εν λόγω δεδομένα αποτελούνται από 9564 γραμμές και 50 στήλες ή, αλλιώς, από μία εξαρτημένη μεταβλητή και 49 χαρακτηριστικά (ανεξάρτητες μεταβλητές), από 9564 παρατηρήσεις το καθένα. Για το συγκεκριμένο πρόβλημα θα θεωρήσουμε ως εξαρτημένη μεταβλητή (μεταβλητή απόκρισης) την μεταβλητή "koi_pdisposition" που περιλαμβάνεται στο εν λόγω σύνολο δεδομένων. Η μεταβλητή αυτή περιγράφει την κατάταξη που έχει η ανάλυση δεδομένων Kepler για τον υποψήφιο εξωπλανήτη και διακρίνεται σε "FALSE POSITIVE" και "CANDIDATE", δηλαδή σε 'Λανθασμένα Θετικός' και 'Υποψήφιος' αντίστοιχα. Επομένως, η μεταβλητή koi_pdisposition είναι μία κατηγορική μεταβλητή με δύο κατηγορίες που αφορούν το εκάστοτε αντικείμενο ενδιαφέροντος του Kepler με την πρώτη να δηλώνει ότι λανθασμένα θεωρήθηκε ως πλανήτης και την δεύτερη να δηλώνει ότι, πράγματι, το αντίστοιχο KOI είναι υποψήφιος εξωπλανήτης.

Για προφανείς λόγους, δεν θα αναφερθούν όλες οι ανεξάρτητες μεταβλητές που περιέχονται στο σύνολο δεδομένων. Αντί αυτού, περιγράφονται στην συνέχεια κάποιες από αυτές ενδεικτικά [15]:

• **"kepler_name"**

Τα ονόματα αυτά προορίζονται για να υποδεικνύουν με σαφήνεια μία κατηγορία αντικειμένων που έχουν επιβεβαιωθεί ή επικυρωθεί ως πλανήτες. Παρ᾽ όλα αυτά, η εν λόγω μεταβλητή δεν θα χρησιμοποιηθεί στην στατιστική ανάλυση που θα ακολουθήσει.

• **"koi_disposition"**
Η κατάταξη που είναι καταχωρημένη στην βιβλιογραφία προς αυτόν τον πιθανόν εξωπλανήτη ως 'Υποψήφιος', 'Λανθασμένα Θετικός' και 'Επιβεβαιωμένος' ("CANDIDATE", "FALSE POSITIVE", "CONFIRMED").
• **"koi_score"**
Μία τιμή μεταξύ του 0 και του 1 που υποδεικνύει την εμπιστοσύνη στην κατάταξη KOI. Για τους υποψήφιους εξωπλανήτες (CANDIDATES), μία υψηλότερη τιμή υποδηλώνει μεγαλύτερη εμπιστοσύνη στην κατάταξή του, ενώ για τους λανθασμένα θετικούς (FALSE POSITIVE), μία υψηλότερη τιμή υποδηλώνει λιγότερη εμπιστοσύνη σε αυτήν την κατάταξη.
• **"koi_fpflag_nt"**
Ένα KOI του οποίου η καμπύλη φωτός δεν είναι συνεπής μ᾽ αυτή ενός διερχόμενου πλανήτη. Αυτό περιλαμβάνει, αλλά δεν περιορίζεται σε, τεχνουργήματα οργάνων, μεταβλητά αστέρια που δεν εκλείπουν και ψευδείς ανιχνεύσεις.
• **"koi_fpflag_ss"**
Ένα KOI που παρατηρείται ότι έχει ένα σημαντικό δευτερεύον συμβάν, σχήμα διέλευσης ή μεταβλητότητα εκτός έκλειψης, γεγονός που υποδεικνύει ότι το συμβάν παρόμοιο με τη διέλευση πιθανότατα προκαλείται από ένα δυαδικό σύστημα έκλειψης.
• **"koi_fpflag_co"**
Η πηγή του σήματος προέρχεται από ένα κοντινό αστέρι, όπως συνάγεται από τη μέτρηση της κεντροειδούς θέσης της εικόνας τόσο εντός όσο και εκτός μετάδοσης ή από την ισχύ του σήματος διέλευσης στα εξωτερικά εικονοστοιχεία ("halo") του στόχου σε σύγκριση με το σήμα διέλευσης από τα pixel στο βέλτιστο διάφραγμα (ή πυρήνα).

Η ανάλυση και επεξεργασία του εν λόγω συνόλου δεδομένων πραγματοποιείται, κυρίως, με την μέθοδο της δενδρικής ταξινόμησης και, ύστερα, εφαρμόζεται λογιστική παλινδρόμηση, προκειμένου να διαπιστωθεί ποια από τις δύο μεθόδους είναι πιο ακριβής και ταξινομεί καλύτερα την μεταβλητή απόκρισης του συγκεκριμένου προβλήματος [14]. Και για αυτό το πρόβλημα ταξινόμησης θα γίνει χρήση της στατιστικής γλώσσας προγραμματισμού R.

5.1.1 Δενδρική Ταξινόμηση

Η κατασκευή των δένδρων ταξινόμησης θα γίνει με τη βοήθεια της βιβλιοθήκης "tree" της R. Αρχικά, φορτώνουμε τα δεδομένα που θα χρησιμοποιήσουμε για το πρόβλημα της αναζήτησης Kepler εξωπλανητών χρησιμοποιώντας το πακέτο "data.table" που υπάρχει στην R, καθώς αυτά είναι αποθηκευμένα σε ένα αρχείο τύπου .csv. Στην συνέχεια, αφαιρείται η στήλη "rowid", η οποία περιέχει την απαρίθμηση των γραμμών στο σύνολο δεδομένων, όπως και οι στήλες που περιέχουν ονομασίες πιθανών εξωπλανητών, οι στήλες των οποίων όλες οι πα-

ρατηρήσεις είναι ελλειπείς τιμές (NA's) και η στήλη "koi_disposition", καθώς, όπως προαναφέρθηκε, ως μεταβλητή απόκρισης θεωρείται η δυαδική μεταβλητή "koi_pdisposition". Τα βήματα που περιγράφηκαν υλοποιούνται στον Κώδικα 5.1:

```
> library(tree)
> library(data.table)
> exoplanets<-fread("exoplanet.csv")
> exoplanets$rowid<-NULL
> exoplanets$kepoi_name<-NULL
> exoplanets$kepler_name<-NULL
> exoplanets$koi_teq_err1<-NULL
> exoplanets$koi_teq_err2<-NULL
> exoplanets$koi_tce_delivname<-NULL
> exoplanets$koi_disposition<-NULL
> exoplanets<-as.data.frame(exoplanets)
```

Κώδικας 5.1: Προσαρμογή των δεδομένων για το πρόβλημα της αναζήτησης Kepler εξωπλανητών.

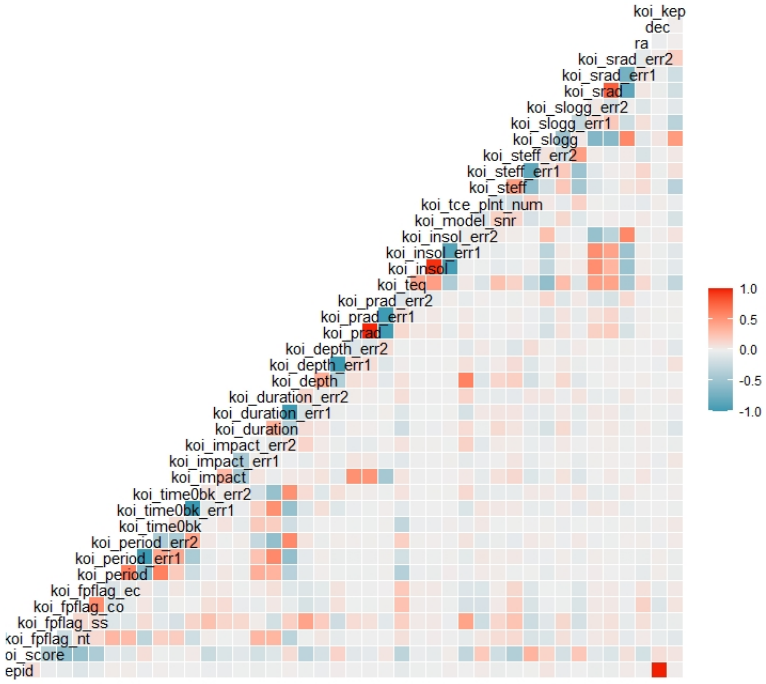
Η τελευταία εντολή του Κώδικα 5.1 μετατρέπει το σύνολο δεδομένων που μας ενδιαφέρει ("exoplanets") σε πλαίσιο δεδομένων (data frame).

Είναι προφανές ότι σε αυτό το πρόβλημα ταξινόμησης ερχόμαστε αντιμέτωποι με έναν μεγάλο όγκο δεδομένων. Από αυτά, κάποια ενδεχομένως να μην χρειάζονται ή να μην συμβάλλουν ιδιαίτερα στην τελική πρόβλεψη του μοντέλου, δηλαδή να μην προσφέρουν κάποια επιπλέον πληροφορία. Για αυτόν τον λόγο, σε περιπτώσεις όπου καλούμαστε να διαχειριστούμε και να επεξεργαστούμε πολλή πληροφορία, είναι σημαντικό να γίνεται ένας 'καθαρισμός' των δεδομένων ("data cleaning"), προκειμένου να διακρίνουμε τις στατιστικά σημαντικές μεταβλητές και να καταλήξουμε σε μία ακριβέστερη πρόβλεψη ή ταξινόμηση. Ο καθαρισμός αυτός περιλαμβάνει την αφαίρεση εκείνων των μεταβλητών που δεν περιέχουν κάποια πληροφορία ή που κρίνονται εξ' αρχής μη σημαντικές για την ανάλυση του προβλήματος ή των μεταβλητών που σχετίζονται σε μεγάλο βαθμό με κάποια άλλη μεταβλητή ή μεταβλητές και ενδέχεται αυτή η συσχέτιση να επηρεάσει την τελική πρόβλεψη (ύπαρξη μεροληψίας). Μία λύση για τον έλεγχο του τελευταίου ενδεχομένου είναι η κατασκευή ενός πίνακα συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών των δεδομένων, εφαρμόζοντας τον έλεγχο του Pearson (Pearson correlation coefficient test) στο σύνολο δεδομένων "exoplanets", όπως φαίνεται στον Κώδικα 5.2 (λεπτομέρειες για τον σχετικό έλεγχο αναφέρονται στο Παράρτημα).

```
> library(GGally)
> ggcorr(exoplanets, method = c("complete.obs", "pearson"))
Warning message:
In ggcorr(exoplanets, method = c("complete.obs", "pearson"))
   :
```

```
  data in column(s) 'koi_pdisposition' are not numeric and
   were ignored
```

Κώδικας 5.2: Δημιουργία πίνακα συσχέτισης μεταξύ των δεδομένων "exoplanets".



Σχήμα 5.1: Γράφημα του πίνακα συσχέτισης για τα δεδομένα "exoplanets".

Με βάση το Σχήμα 5.1, αφαιρούνται οι μεταβλητές που παρουσιάζουν τέλεια ή υψηλή θετική συσχέτιση με κάποια άλλη (≈ 1), καθώς και οι μεταβλητές που παρουσιάζουν τέλεια ή υψηλή αρνητική συσχέτιση με κάποια άλλη (≈ -1), όπως φαίνεται στον Κώδικα 5.3.

```
> exoplanets$dec<-NULL
> exoplanets$koi_period_err1<-NULL
> exoplanets$koi_time0bk_err1<-NULL
> exoplanets$koi_duration_err1<-NULL
> exoplanets$koi_depth_err1<-NULL
> exoplanets$koi_prad_err1<-NULL
> exoplanets$koi_prad_err2<-NULL
> exoplanets$koi_insol_err1<-NULL
> exoplanets$koi_insol_err2<-NULL
> exoplanets$koi_srad_err1<-NULL
> exoplanets$koi_srad_err2<-NULL
> exoplanets$koi_steff_err2<-NULL
```

```
> exoplanets$koi_impact_err2<-NULL
```

Κώδικας 5.3: Αφαίρεση χαρακτηριστικών μεγάλου βαθμού συσχέτισης με τα υπόλοιπα χαρακτηριστικά για το πρόβλημα της αναζήτησης Kepler εξωπλανητών.

Επιπλέον, εξακολουθούν να υπάρχουν χαρακτηριστικά των οποίων οι παρατηρήσεις περιέχουν κάποιες ελλιπείς τιμές. Μία συνηθισμένη τακτική είναι η αφαίρεση όλων των NA τιμών από τα δεδομένα. Ωστόσο, μ΄ αυτόν τον τρόπο χάνεται ένα μέρος της πληροφορίας που μπορεί τελικά να αποδειχθεί σημαντικό για την ευστοχία και την ακρίβεια του μοντέλου. Μία άλλη μέθοδος που χρησιμοποιείται όταν επιθυμούμε να αξιοποιήσουμε όσον το δυνατόν περισσότερα δεδομένα με όσον το δυνατόν μεγαλύτερη αμεροληψία, είναι ο υπολογισμός ή, αλλιώς, η 'επαναφορά' των ελλιπών τιμών ("data imputation"). Στον Κώδικα 5.4 φαίνεται ακριβώς αυτή η διαδικασία.

```
> Y_column<-which(names(exoplanets)=="koi_pdisposition")
> X<-exoplanets[,-Y_column]
> for(i in 1:ncol(X)){
X[,i][is.na(X[,i])]<-mean(X[,i],na.rm=TRUE)
}
> koi_pdisposition<-exoplanets$koi_pdisposition
> exoplanets<-data.frame(koi_pdisposition,X)
```

Κώδικας 5.4: Αντικατάσταση των ελλιπών τιμών στα δεδομένα "exoplanets" με τον αριθμητικό μέσο των παρατηρήσεων.

Αρχικά, με την πρώτη εντολή αποθηκεύουμε σε μία νέα μεταβλητή "Y_column" τον αριθμό της στήλης "koi_pdisposition", όπως αυτή εμφανίζεται στα δεδομένα exoplanets (2, καθώς αποτελεί την δεύτερη στήλη στα δεδομένα exoplanets). Η εν λόγω στήλη περιέχει κατηγορικές τιμές και, όπως αναφέρθηκε προηγουμένως, θα χρησιμοποιηθεί ως μεταβλητή απόκρισης για το πρόβλημα της αναζήτησης Kepler εξωπλανητών. Έπειτα, δημιουργούμε ένα καινούργιο πλαίσιο δεδομένων, το οποίο καλούμε "X" και αποτελείται από τις στήλες των δεδομένων exoplanets, πέρα της στήλης koi_pdisposition. Στην συνέχεια, για κάθε στήλη του X αντικαθιστούμε όλες τις NA τιμές με τον αριθμητικό μέσο των παρατηρήσεων της εκάστοτε μεταβλητής/στήλης. Τέλος, η τελευταία εντολή του Κώδικα 5.4 δημιουργεί εκ νέου ένα πλαίσιο δεδομένων με την ονομασία "exoplanets", το οποίο περιέχει ως στήλες την μεταβλητή απόκρισης (koi_pdisposition) και όλες τις επεξηγηματικές μεταβλητές που ήδη διαθέταμε, χωρίς την παρουσία πλέον ελλιπών τιμών.

Τελευταίο βήμα 'προετοιμασίας' των δεδομένων πριν την στατιστική ανάλυση αποτελεί η μετατροπή των κατηγοριών της μεταβλητής απόκρισης σε '1', αν το KOI είναι υποψήφιος εξωπλανήτης και σε '0', αν το KOI δεν μπορεί

να χαρακτηριστεί ως πλανήτης (λανθασμένα θετικός), καθώς και η αποθήκευσή τους σε μία νέα κατηγορική μεταβλητή Y, η οποία θα αντικαταστήσει την "koi_pdisposition" για λόγους ευκολίας, κάνοντας χρήση των νέων πλέον δεδομένων exoplanets (Κώδικας 5.5).

```
> Y<-ifelse(exoplanets$koi_pdisposition %in% c("CANDIDATE")
   ,"1","0")
> Y<-as.factor(Y)
> exoplanets$koi_pdisposition<-NULL
> exoplanets<-data.frame(exoplanets,Y)
```

Κώδικας 5.5: Αντικατάσταση της "koi_pdisposition" με μία νέα μεταβλητή Y.

Είμαστε έτοιμοι τώρα να ορίσουμε το σύνολο εκπαίδευσης, βάσει του οποίου θα κατασκευαστούν τα δένδρα ταξινόμησης και το σύνολο ελέγχου στο οποίο θα βασιστούν οι προβλέψεις των παραγόμενων μοντέλων (δένδρα αποφάσεων). Ο διαχωρισμός του ολικού συνόλου δεδομένων γίνεται ως εξής: το 80% των παρατηρήσεων θα αποτελέσουν το σύνολο εκπαίδευσης και το υπόλοιπο 20% θα αποτελέσει το σύνολο ελέγχου (Κώδικας 5.6).

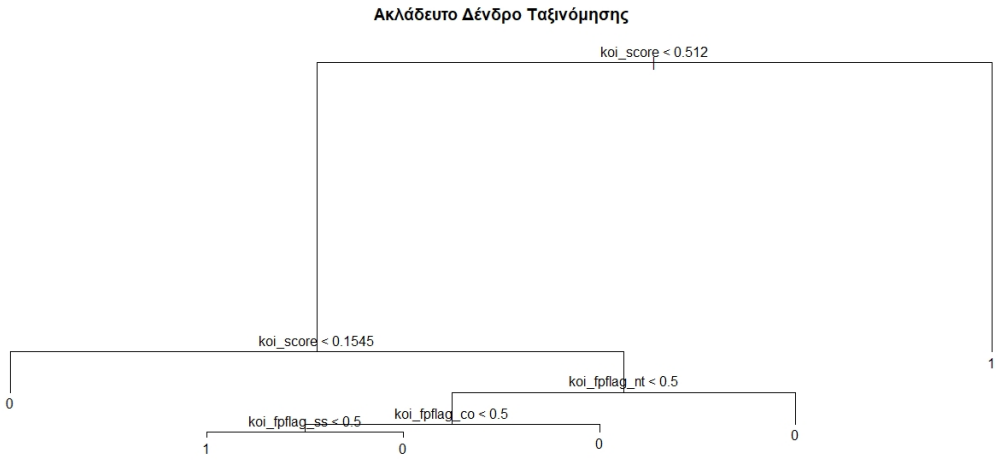
```
> set.seed(1)
> alpha<- 0.8
> train<- sample(1:nrow(exoplanets), alpha*nrow(exoplanets))
> testdata<- exoplanets[-train,]
```

Κώδικας 5.6: Δημιουργία συνόλου εκπαίδευσης και συνόλου ελέγχου για το πρόβλημα της αναζήτησης Kepler εξωπλανητών.

Στην συνέχεια, γίνεται προσαρμογή του δένδρου ταξινόμησης στα δεδομένα εκπαίδευσης, όπως φαίνεται στον Κώδικα 5.7, πραγματοποιώντας Αναδρομικό Δυαδικό Διαχωρισμό (Κεφάλαιο 3), έχοντας ως μεταβλητή απόκρισης την μεταβλητή Y που ορίσαμε προηγουμένως.

```
> tree_exoplanets<-tree(Y~.,exoplanets,subset=train)
> summary(tree_exoplanets)

Classification tree:
tree(formula = Y ~ ., data = exoplanets, subset = train)
Variables actually used in tree construction:
[1] "koi_score"     "koi_fpflag_nt" "koi_fpflag_co" "
   koi_fpflag_ss"
Number of terminal nodes:  6
Residual mean deviance:  0.08603 = 657.7 / 7645
Misclassification error rate: 0.01477 = 113 / 7651
> plot(tree_exoplanets)
> text(tree_exoplanets,pretty=0)
> title(" Ακλάδευτο Δένδρο Ταξινόμησης")
```

Κώδικας 5.7: Κατασκευή δένδρου ταξινόμησης στα δεδομένα εκπαίδευσης του συνόλου exoplanets.



Σχήμα 5.2: Ακλάδευτο (ολικό) δένδρο ταξινόμησης σύμφωνα με τα δεδομένα εκπαίδευσης του συνόλου exoplanets.

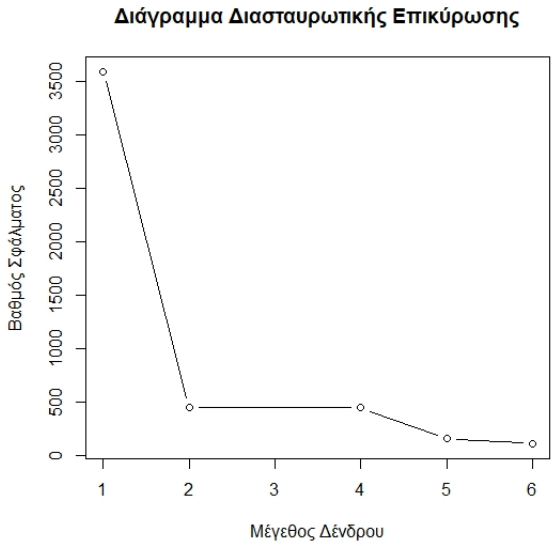
Σύμφωνα με το Σχήμα 5.2 παρατηρούμε ότι μόνο τέσσερις μεταβλητές χρησιμοποιούνται στην κατασκευή του δένδρου, το οποίο διαθέτει έξι τελικούς κόμβους. Σύμφωνα με αυτό, ένα KOI αποτελεί υποψήφιο εξωπλανήτη αν το αντίστοιχο koi_score είναι μεγαλύτερο ή ίσο του 0.512 ή αν κυμαίνεται μεταξύ του 0.1545 και του 0.512 και επιπλέον οι τιμές των koi_fpflag_nt, koi_fpflag_co koi_fpflag_ss είναι μικρότερες από 0.5.

Έπειτα, όπως και στην μέθοδο της δενδρικής παλινδρόμησης, προβαίνουμε σε κλάδεμα του δένδρου που μόλις κατασκευάσαμε, προκειμένου να αποφύγουμε προβλήματα υπερπροσαρμογής. Ελέγχουμε, μέσω της συνάρτησης cv.tree(), αν το κλάδεμα του δένδρου θα βελτιώσει την απόδοσή του (Κώδικας 5.8).

```
> set.seed(2)
> cv_exoplanets<-cv.tree(tree_exoplanets,FUN=prune.misclass)
> cv_exoplanets
$size
[1] 6 5 4 2 1

$dev
[1]  116  159  449  449 3592

$k
[1] -Inf   49   92   97 3144

$method
[1] "misclass"
```

```
attr(,"class")
[1] "prune"         "tree.sequence"
> plot(cv_exoplanets$size,cv_exoplanets$dev,type="b",
xlab="Μέγεθος Δένδρου",
ylab="Βαθμός Σφάλματος",main="Διάγραμμα Διασταυρωτικής Επικύρωσης")
```

Κώδικας 5.8: Εφαρμογή της cv.tree() συνάρτησης για τα δεδομένα exoplanets.



Σχήμα 5.3: Διάγραμμα Διασταυρωτικής Επικύρωσης για τα δεδομένα exoplanets.

Το όρισμα FUN=prune.misclass δηλώνει ότι ως κριτήριο διαμέρισης χρησιμοποιείται ο βαθμός του σφάλματος ταξινόμησης ("dev") στην διαδικασία της διασταυρωτικής επικύρωσης και του κλαδέματος, αντί για την προκαθορισμένη επιλογή της συνάρτησης cv.tree() που είναι η διασπορά. Η παράμετρος πολυπλοκότητας δηλώνεται με "k" στον Κώδικα 5.8 και από τον οποίο, σε συνδυασμό με το Σχήμα 5.3, φαίνεται ότι το δένδρο με 5 τελικούς κόμβους αποτελεί το αμέσως επόμενο δένδρο με τον χαμηλότερο βαθμό σφάλματος, όπως προκύπτει από τη μέθοδο της διασταυρωτικής επικύρωσης (αμέσως μετά το αρχικό δένδρο). Αν επιθυμούμε, παρ΄ όλα αυτά να προχωρήσουμε σε κλάδεμα του δένδρου, θα χρησιμοποιήσουμε την συνάρτηση prune.misclass(), έτσι ώστε να μετατρέψουμε το αρχικό δένδρο ταξινόμησης σε δένδρο 5 τελικών κόμβων, όπως φαίνεται στον Κώδικα 5.9.

```
> prune_exoplanets<-prune.misclass(tree_exoplanets,best=5)
> plot(prune_exoplanets)
> text(prune_exoplanets,pretty=0)
```

```
> title("Κλαδεμένο Δένδρο Ταξινόμησης")
```

Κώδικας 5.9: Κλάδεμα του δένδρου ταξινόμησης για τα δεδομένα exoplanets.

Σχήμα 5.4: Κλαδεμένο δένδρο ταξινόμησης (με 5 φύλλα) για τα δεδομένα εκπαίδευσης του συνόλου exoplanets.

Όπως φαίνεται στο Σχήμα 5.4, το δένδρο με 5 τελικούς κόμβους κάνει χρήση μόνο τριών μεταβλητών. Ο Κώδικας 5.10 περιγράφει την διαδικασία πραγματοποίησης προβλέψεων στο σύνολο ελέγχου χρησιμοποιώντας το κλαδεμένο δένδρο ταξινόμησης που προέκυψε.

```
> tree_pred<-predict(prune_exoplanets,testdata,type="class")
> table(tree_pred,testdata$Y)

tree_pred   0   1
        0 967   5
        1  42 899
> accuracy_Test<-(967+899)/1913
> print(paste('Ακρίβεια για τα δεδομένα ελέγχου:', accuracy_Test))
[1] "Ακρίβεια για τα δεδομένα ελέγχου: 0.975431259801359"
```

Κώδικας 5.10: Προβλέψεις στο σύνολο ελέγχου με χρήση του κλαδεμένου δένδρου ταξινόμησης για τα δεδομένα exoplanets.

Η ακρίβεια ταξινόμησης για τα δεδομένα ελέγχου προκύπτει να είναι περίπου ίση με 97.5%, δηλαδή περίπου το 97.5% των παρατηρήσεων από το σύνολο ελέγχου

ταξινομήθηκαν σωστά, σύμφωνα με το κλαδεμένο δένδρο ταξινόμησης με τους 5 τελικούς κόμβους.

Εφόσον η μέθοδος της διασταυρωτικής επικύρωσης έδειξε ότι καλύτερες προβλεψεις θα έχουμε με το αρχικό, ακλάδευτο δένδρο ταξινόμησης, πραγματοποιούμε προβλέψεις στο σύνολο ελέγχου κάνοντας χρήση του αρχικού δένδρου και, στην συνέχεια, κατασκευάζουμε τον πίνακα συνάφειας, όπου απεικονίζει τις προβλεπόμενες (από το δένδρο) κατηγορίες της μεταβλητής Y συναρτήσει των παρατηρούμενων κατηγοριών της ίδιας μεταβλητής για το σύνολο ελέγχου (testdata), όπως φαίνεται στον Κώδικα 5.11.

```
> tree_pred<-predict(tree_exoplanets,testdata,type="class")
> table(tree_pred,testdata$Y)

tree_pred   0   1
        0 983   7
        1  26 897
> accuracy_Test<-(983+897)/1913
> print(paste('Ακρίβεια για τα δεδομένα ελέγχου:', accuracy_Test))
[1] "Ακρίβεια για τα δεδομένα ελέγχου: 0.982749607945635"
```

Κώδικας 5.11: Προβλέψεις στο σύνολο ελέγχου με χρήση του αρχικού δένδρου ταξινόμησης για τα δεδομένα exoplanets.

Το όρισμα type="class" της συνάρτησης predict() δηλώνει ότι ασχολούμαστε με ένα πρόβλημα ταξινόμησης.

Επομένως, περίπου το 98.3% των παρατηρήσεων από το σύνολο ελέγχου ταξινομήθηκαν σωστά, σύμφωνα με το αρχικό δένδρο ταξινόμησης με τους 6 τελικούς κόμβους.

Η μέθοδος της ενσάκισης για το σύνολο δεδομένων exoplanets φαίνεται στον Κώδικα 5.12, όπου το όρισμα mtry=29 δηλώνει ότι λαμβάνονται υπόψιν όλες οι (επεξηγηματικές) μεταβλητές για την κάθε τμήση (ακολουθώντας τους συμβολισμούς της αντίστοιχης θεωρίας που αναφέρεται στο Κεφάλαιο 3, θα ισχύει ότι m=p=29).

```
> library(randomForest)
> set.seed(1)
> bag_exoplanets<-randomForest(Y~.,exoplanets,
subset=train,mtry=29,importance=TRUE)
> bag_exoplanets

Call:
 randomForest(formula = Y ~ ., data = exoplanets, mtry = 29,
        importance = TRUE, subset = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 29
```

```
        OOB estimate of  error rate: 0.89%
Confusion matrix:
     0    1 class.error
0 4031   28 0.006898251
1   40 3552 0.011135857
```

Κώδικας 5.12: Μέθοδος της ενσάκισης στα δεδομένα εκπαίδευσης των δεδομένων exoplanets.

Το μοντέλο της ενσάκισης για το εν λόγω πρόβλημα αποτελείται από 500 δένδρα αποφάσεων. Ακόμα, το OOB σφάλμα ταξινόμησης είναι 0.89% και η ακρίβεια της μεθόδου, δηλαδή το ποσοστό των ορθά ταξινομημένων παρατηρήσεων, όταν εφαρμόζεται στο σύνολο εκπαίδευσης, είναι περίπου ίσο με 99,4%, όπως προκύπτει από τον πίνακα συνάφειας (confusion matrix). Έπειτα, γίνονται προβλέψεις στο σύνολο ελέγχου, βάσει του μοντέλου ενσάκισης (Κώδικας 5.13).

```
> set.seed(2)
> yhat_bag<-predict(bag_exoplanets,newdata=testdata,type="
   class")
> table(yhat_bag,testdata$Y)

yhat_bag    0    1
       0 1005   15
       1    4  889
> accuracy_Test<-(1005+889)/1913
> print(paste('Ακρίβεια για τα δεδομένα ελέγχου:', accuracy_Test))
[1] "Ακρίβεια για τα δεδομένα ελέγχου: 0.990067956089911"
```

Κώδικας 5.13: Προβλέψεις στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου της ενσάκισης για τα δεδομένα exoplanets.

Παρατηρούμε ότι η μέθοδος της ενσάκισης βελτίωσε αρκετά την απόδοση του μοντέλου. Το 99% των παρατηρήσεων της μεταβλητής Y που ανήκουν στο σύνολο ελέγχου ταξινομούνται σωστά ως υποψήφιοι εξωπλανήτες ή, αντίστοιχα, ως μη πλανήτες.

Μπορούμε, στην συνέχεια, να εφαρμόσουμε και τη μέθοδο των τυχαίων δασών στο εν λόγω σύνολο δεδομένων για πιθανόν ακόμα μία περαιτέρω βελτίωση, σύμφωνα με τον Κώδικα 5.14.

```
> set.seed(1)
> rf_exoplanets<-randomForest(Y~.,exoplanets,subset=train,
mtry=5,importance=TRUE)
> yhat_rf<-predict(rf_exoplanets,newdata=testdata)
> table(yhat_rf,testdata$Y)

yhat_rf    0    1
```

```
      0 1001    9
      1    8  895
> accuracy_Test<-(1001+895)/1913
> print(paste('Ακρίβεια για τα δεδομένα ελέγχου:', accuracy_Test))
[1] "Ακρίβεια για τα δεδομένα ελέγχου: 0.991113434396236"
```

Κώδικας 5.14: Προβλέψεις στο σύνολο ελέγχου με χρήση του μοντέλου της μεθόδου των τυχαίων δασών για τα δεδομένα exoplanets.

Όπως εξηγήθηκε και στην θεωρία του Κεφαλαίου 3, πριν από κάθε τμήση επιλέγεται τυχαίο δείγμα m χαρακτηριστικών (ανεξάρτητων μεταβλητών) ως υποψήφιες μεταβλητές τμήσης. Το τυχαίο αυτό δείγμα των m χαρακτηριστικών επιλέγεται από το ολικό σύνολο των p χαρακτηριστικών. Συνήθως, για την μέθοδο των τυχαίων δασών ισχύει ότι $m \approx \sqrt{p}$. Για το εν λόγω πρόβλημα έχουμε ότι $m \approx \sqrt{29} \approx 5$. Πράγματι, υπάρχει μία σχετικά μικρή βελτίωση στην ακρίβεια από τη μέθοδο των τυχαίων δασών, καθώς είναι ακριβής κατά 99.11% στα δεδομένα ελέγχου, ενώ η αντίστοιχη ακρίβεια της μεθόδου της ενσάκισης για τα ίδια δεδομένα ήταν 99%.

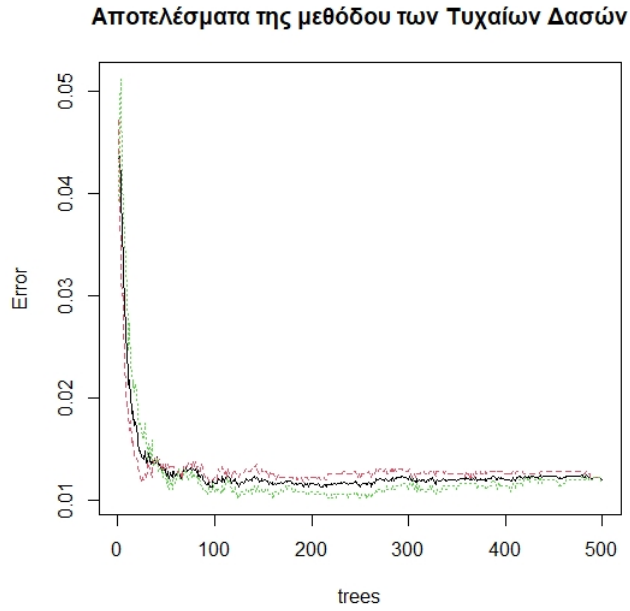
```
> plot(rf_exoplanets,
main="Αποτελέσματα της μεθόδου των Τυχαίων Δασών")
```

Κώδικας 5.15: Κατασκευή διαγράμματος του σφάλματος συναρτήσει του αριθμού δένδρων για το μοντέλο της μεθόδου των τυχαίων δασών για τα δεδομένα exoplanets.

Στο Σχήμα 5.5 φαίνεται το διάγραμμα ανάμεσα στο σφάλμα ταξινόμησης ("error") και τον αριθμό των δένδρων ("trees") για το μοντέλο που προέκυψε μέσα από τη μέθοδο των τυχαίων δασών (Κώδικας 5.15). Η μαύρη καμπύλη αντιπροσωπεύει το ολικό OOB σφάλμα, ενώ οι υπόλοιπες δύο δηλώνουν το σφάλματα των κλάσεων, μία καμπύλη για κάθε κλάση (στο εν λόγω πρόβλημα έχουμε 2 κλάσεις). Παρατηρούμε ότι όταν υπάρχουν πολλά δένδρα στο μοντέλο, το σφάλμα ταξινόμησης λαμβάνει σχετικά μικρές τιμές.

Αντίστοιχα, η μέθοδος της ενίσχυσης δίνει παρόμοια αποτελέσματα με αυτή των τυχαίων δασών, συνεπώς θα μείνουμε στην τελευταία την οποία και θα θεωρήσουμε ως τη μέθοδο που προσφέρει την βέλτιστη δυνατή απόδοση στις προβλέψεις για το πρόβλημα της αναζήτησης Kepler εξωπλανητών.

Τέλος, για την επιβεβαίωση των συμπερασμάτων μας μπορούμε να επικαλεστούμε την λεγόμενη 'Καμπύλη Λειτουργικών Χαρακτηριστών', η οποία είναι γνωστή και ως Καμπύλη ROC (Receiver Operating Characteristic Curve). Η καμπύλη αυτή είναι μια μέτρηση απόδοσης για τα προβλήματα ταξινόμησης ή, αλλιώς, μία καμπύλη πιθανότητας. Το εμβαδόν κάτω από την καμπύλη ROC, γνωστό ως "AUC" (Area Under the Curve), αντιπροσωπεύει το βαθμό ή το μέτρο της διαχωρισιμότητας. Δείχνει την διακριτική ικανότητα του μοντέλου



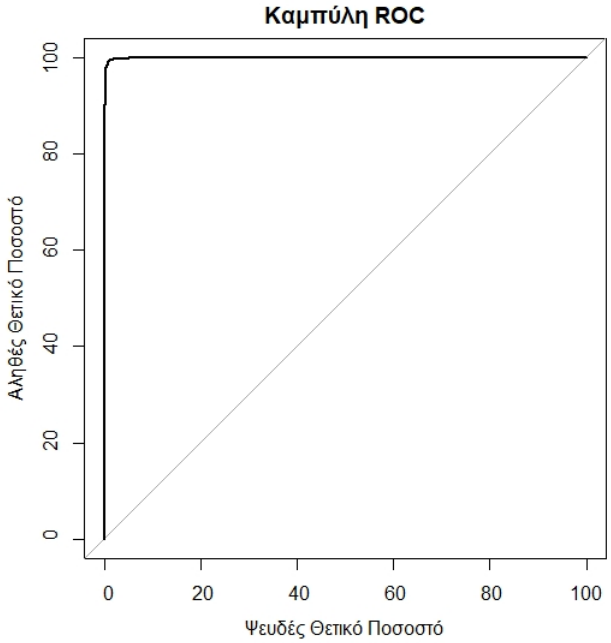
Σχήμα 5.5: Διάγραμμα σφάλματος συναρτήσει του πλήθους των δένδρων αποφάσεων για το μοντέλο των τυχαίων δασών για τα δεδομένα exoplanets.

μεταξύ των τάξεων. Όσο υψηλότερη είναι η AUC, τόσο καλύτερο είναι το μοντέλο στην πρόβλεψη '0' τάξεων ως '0' και '1' τάξεων ως '1'. Ο αναγνώστης μπορεί να καταφύγει στο Παράρτημα για περισσότερες πληροφορίες σχετικά με την καμπύλη ROC. Η κατασκευή της πραγματοποιείται με βάση τον Κώδικα 5.16 και η μορφή της φαίνεται στο Σχήμα 5.6.

```
> library(pROC)
> yhat_rf<-predict(rf_exoplanets,newdata=testdata,type="prob
    ")
> roc(testdata$Y,yhat_rf[,2],plot=TRUE,legacy.axes=TRUE,
    percent=TRUE,
xlab="Ψευδές Θετικό Ποσοστό",ylab="Αληθές Θετικό Ποσοστό",
main="Καμπύλη ROC")
Setting levels: control = 0, case = 1
Setting direction: controls < cases

Call:
roc.default(response = testdata$Y, predictor = yhat_rf[, 2],
       percent = TRUE, plot = TRUE, legacy.axes = TRUE, xlab =
     "False Positive Percentage",      ylab = "True Positive
    Percentage")

Data: yhat_rf[, 2] in 1009 controls (testdata$Y 0) < 904 cases
```

```
    (testdata$Y 1).
Area under the curve: 99.95%
```

Κώδικας 5.16: Δημιουργία καμπύλης ROC για το μοντέλο της μεθόδου των τυχαίων δασών για τα δεδομένα exoplanets.



Σχήμα 5.6: Καμπύλη ROC για το μοντέλο της μεθόδου των τυχαίων δασών για τα δεδομένα exoplanets.

Για τον σχεδιασμό της καμπύλης ROC επιλέγεται η δεύτερη στήλη της μεταβλητής "yhat_rf", η οποία αναφέρεται στην πιθανότητα που προβλέπεται ένα KOI να είναι υποψήφιος εξωπλανήτης (να ανήκει στην κλάση '1'). Ο οριζόντιος άξονας αντιπροσωπεύει το ποσοστό των ψευδώς θετικών αποτελεσμάτων, δηλαδή το ποσοστό λανθασμένης πρόβλεψης της κατάστασης $Y = 0$ (το KOI να μην αποτελεί πλανήτη). Αντίστοιχα, ο κατακόρυφος άξονας αντιπροσωπεύει το ποσοστό των αληθώς θετικών αποτελεσμάτων, δηλαδή το ποσοστό ορθής πρόβλεψης της κατάστασης $Y = 1$ (το KOI να αποτελεί υποψήφιο εξωπλανήτη). Το εμβαδόν κάτω από την καμπύλη είναι 0.9995, δηλαδή τείνει να πλησιάσει την μονάδα. Αυτό υποδεικνύει ότι το μοντέλο που προκύπτει από την μέθοδο των τυχαίων δασών έχει σχεδόν άριστη απόδοση στην διάκριση μεταξύ των θετικών και αρνητικών κλάσεων που για το συγκεκριμένο πρόβλημα αυτό σημαίνει ότι η ικανότητα του εν λόγω μοντέλου να διαχωρίζει τους υποψήφιους εξωπλανήτες από τα αντικείμενα που λανθασμένα ταξινομούνται ως πλανήτες είναι πάρα πολύ καλή (σχεδόν βέλτιστη).

Πίνακας 5.1: Σύνοψη των αποτελεσμάτων που προέκυψαν από την δενδρική ταξινόμηση στο σύνολο ελέγχου των δεδομένων exoplanets, με βάση την ακρίβεια ταξινόμησης.

Μοντέλο (Δένδρο/Δένδρα Ταξινόμησης)	Ακρίβεια (σε %)
Αρχικό, ακλάδευτο δένδρο ταξινόμησης	98.3
Κλαδεμένο δένδρο ταξινόμησης	97.5
Μοντέλο από την μέθοδο της ενσάκισης	99.0
Μοντέλο από την μέθοδο των τυχαίων δασών	99.1

Επομένως, από τον Πίνακα 5.1, οδηγούμαστε στο συμπέρασμα ότι το καλύτερο μοντέλο (με την έννοια της ακρίβειας) για τα δεδομένα exoplanets είναι αυτό που προκύπτει μέσω της μεθόδου των τυχαίων δασών, καθώς προσφέρει την μεγαλύτερη ακρίβεια στις προβλέψεις, αν αυτό εφαρμοστεί στο σύνολο ελέγχου ("testdata") των εν λόγω δεδομένων.

5.1.2 Λογιστική Παλινδρόμηση

Η ανάλυση του προβλήματος ταξινόμησης της αναζήτησης Kepler εξωπλανητών μπορεί να πραγματοποιηθεί χρησιμοποιώντας και τη μέθοδο της λογιστικής παλινδρόμησης. Η μεταβλητή απόκρισης του αντίστοιχου συνόλου δεδομένων θεωρούμε και σε αυτήν την περίπτωση ότι είναι η μεταβλητή "koi_pdisposition" την οποία, για λόγους συντομίας, καλούμε Y. Ομοίως με πριν, η Y παίρνει μόνο δύο τιμές που αντιστοιχούν σε 2 ενδεχόμενα: Στο ενδεχόμενο "CANDIDATE" και στο ενδεχόμενο "FALSE POSITIVE", τα οποία κωδικοποιούμε στην συνέχεια σε '1' και '0' αντίστοιχα.

Θα εστιάσουμε την προσοχή μας σε ένα από τα δύο ενδεχόμενα, την 'επιτυχία', δηλαδή, χρησιμοποιώντας τους ανάλογους συμβολισμούς, στο ενδεχόμενο $Y = 1$ με πιθανότητα $p = P(\text{επιτυχία})$. Επομένως, η Y είναι μία τυχαία μεταβλητή που ακολουθεί την κατανομή Bernoulli, δηλαδή $Y \sim B(p)$ με $\mathbb{E}(Y) = p$ και $\text{Var}(Y) = p(1-p)$. Η συμπληρωματική πιθανότητα της p δηλώνει την πιθανότητα αποτυχίας, $1-p$, δηλαδή την πιθανότητα το αντικείμενο ενδιαφέροντος του δορυφόρου Kepler να μην αποτελεί κάποιο είδος πλανήτη.

Ακολουθώντας την ίδια διαδικασία με πριν, φορτώνουμε στην R τα δεδομένα που θα χρησιμοποιήσουμε για το εν λόγω πρόβλημα και αφαιρούμε τις ίδιες στήλες με αυτές που αφαιρέθηκαν στην περίπτωση της δενδρικής ταξινόμησης, κάνοντας τις ίδιες επιλογές μεταβλητών. Αφού προετοιμάσουμε τα δεδομένα, ορίσουμε την ίδια μεταβλητή απόκρισης Y και χωρίσουμε το σύνολο δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου σε αντιστοιχία 80% και 20%, προσαρμόζουμε το μοντέλο της λογιστικής παλινδρόμησης με χρήση της συνάρτησης glm() του πακέτου "glmnet". Ωστόσο, ως επεξηγηματικές με-

ταβλητές θεωρούμε μόνο τις μεταβλητές που χρησιμοποιήθηκαν τελικά στην δενδρική ταξινόμηση και, συγκεκριμένα στο ακλάδευτο, αρχικό δένδρο, προκειμένου τα δύο μοντέλα να έχουν την ίδια πολυπλοκότητα και να έχει νόημα η σύγκριση αυτών. Οι μεταβλητές αυτές είναι οι koi_score, koi_fpflag_nt, koi_fpflag_co και koi_fpflag_ss. Η προσαρμογή του μοντέλου φαίνεται στον Κώδικα 5.17.

```
> library(glmnet)
> log_exoplanets<-glm(Y ~ koi_score + koi_fpflag_nt +
   koi_fpflag_co +
koi_fpflag_ss, exoplanets[train,], family="binomial")
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(log_exoplanets)

Call:
glm(formula = Y ~ koi_score + koi_fpflag_nt + koi_fpflag_co +
    koi_fpflag_ss, family = "binomial", data = exoplanets[
   train,])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.00305  -0.00005   0.00000   0.04870   2.35479

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.6710     0.2533 -10.544   <2e-16 ***
koi_score       9.4266     0.5061  18.624   <2e-16 ***
koi_fpflag_nt -22.2347   649.7812  -0.034    0.973
koi_fpflag_co -19.7616   580.7821  -0.034    0.973
koi_fpflag_ss  -3.3458     0.2871 -11.654   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10578.02  on 7650  degrees of freedom
Residual deviance:   719.79  on 7646  degrees of freedom
AIC: 729.79

Number of Fisher Scoring iterations: 20
```

Κώδικας 5.17: Προσαρμογή του μοντέλου λογιστικής παλινδρόμησης για τα δεδομένα exoplanets.

Το όρισμα family="binomial" χρησιμοποιείται για τον καθορισμό του μοντέλου παλινδρόμησης ως μοντέλο δυαδικής λογιστικής παλινδρόμησης. Σύμφωνα με

τον Κώδικα 5.17, μόνο 2 μεταβλητές είναι στασιστικά σημαντικές σε επίπεδο σημαντικότητας 5% (έχουν P-τιμή<0.05). Παρ΄ όλα αυτά, το εν λόγω μοντέλο αποδεικνύεται ότι είναι αποδοτικότερο και με χαμηλότερο AIC από το μοντέλο, το οποίο θα περιείχε μόνο τις σημαντικές μεταβλητές. Επιλέγοντας τις στατιστικά σημαντικές μεταβλητές μέσω ανάλογων διαστημάτων εμπιστοσύνης θα καταλήγαμε, επίσης, σε ένα κατώτερο μοντέλο από αυτό του Κώδικα 5.17. Προκειμένου να εξετάσουμε την ακρίβεια του μοντέλου που μόλις κατασκευάσαμε, δημιουργούμε τα σύνολα του Κώδικα 5.18.

```
> train_prob<-predict(log_exoplanets, exoplanets[train,],
   type="response")
> train_pred<-ifelse(train_prob <0.5, "0", "1")
> train_ac<-ifelse(train_pred == exoplanets[train,]$Y,"1","0")
> train_ac<-as.numeric(train_ac)
> train_Ac<-mean(train_ac)
> test_prob<-predict(log_exoplanets, testdata, type="response
   ")
> test_pred<-ifelse(test_prob < 0.5, "0", "1")
> test_ac<-ifelse(test_pred == testdata$Y,"1","0")
> test_ac<-as.numeric(test_ac)
> test_Ac<-mean(test_ac)
```

Κώδικας 5.18: Κατασκευή κατάλληλων συνόλων προβλέψεων για την εύρεση της ακρίβειας του μοντέλου της λογιστικής παλινδρόμησης για τα δεδομένα exoplanets.

Οι τρεις πρώτες εντολές του Κώδικα 5.18 αποθηκεύουν τις προβλέψεις που πραγματοποιούνται με βάση το μοντέλο λογιστικής παλινδρόμησης στο σύνολο εκπαίδευσης για την εξαρτημένη μεταβλητή Y σε μία μεταβλητή "train_prob". Στην συνέχεια, γίνεται ταξινόμηση αυτών των προβλέψεων στις αντίστοιχες κλάσεις 0 και 1, χρησιμοποιώντας ως όριο την τιμή 0.5 (πιο συνηθέστερη επιλογή) και, ύστερα, αν αυτές ταυτίζονται με τις αντίστοιχες κατηγορίες της Y από το σύνολο εκπαίδευσης, τότε η μεταβλητή "train_ac" λαμβάνει τον χαρακτήρα '1' (επιτυχία), διαφορετικά λαμβάνει τον χαρακτήρα '0'. Τέλος, η τελευταία μεταβλητή μετατρέπεται σε αριθμητική μεταβλητή (numeric) και υπολογίζεται η μέση τιμή αυτής, η οποία θα αποτελεί και την μέση ακρίβεια του εν λόγω μοντέλου, όταν αυτό εφαρμόζεται στα δεδομένα εκπαίδευσης.

Με παρόμοιο τρόπο, κατασκευάζεται και η μεταβλητή "test_Ac" που αντιπροσωπεύει την μέση ακρίβεια του εν λόγω μοντέλου, όταν αυτό εφαρμόζεται στα δεδομένα ελέγχου.

```
> cat("Ακρίβεια στα δεδομένα εκπαίδευσης: ", train_Ac, "\n",
    "Ακρίβεια στα δεδομένα ελέγχου: ", test_Ac, sep="")
Ακρίβεια στα δεδομένα εκπαίδευσης:   0.9817017
```

```
Ακρίβεια στα δεδομένα ελέγχου: 0.9785677
```

Κώδικας 5.19: Ακρίβεια του μοντέλου της λογιστικής παλινδρόμησης

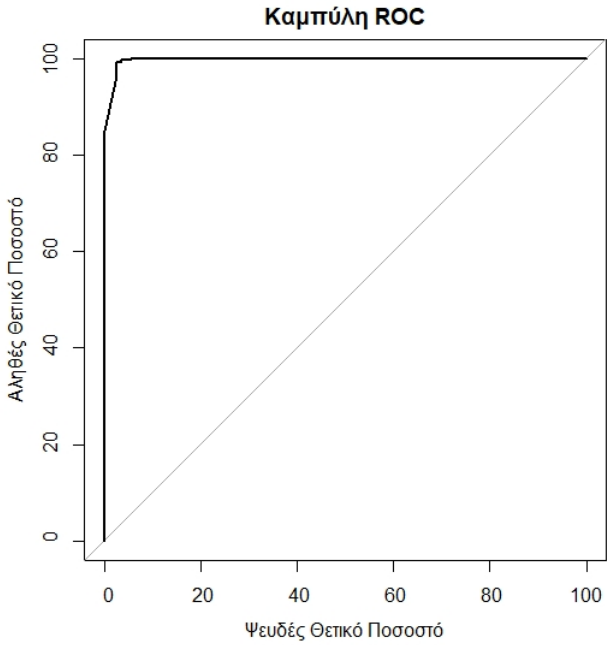
Από τον Κώδικα 5.19 παρατηρούμε ότι στο δείγμα ελέγχου ταξινομήθηκαν σωστά περίπου 97.9% παρατηρήσεις, χαμηλότερο ποσοστό από αυτό που προέκυψε από την χρήση του ολικού, ακλάδευτου δένδρου ταξινόμησης της προηγούμενης παραγράφου (98.3%).

Τέλος, κατασκευάζεται η καμπύλη ROC (Σχήμα 5.7), προκειμένου να ελεγχθεί η ικανότητα διάκρισης του μοντέλου που προέκυψε από τη μέθοδο της λογιστικής παλινδρόμησης μεταξύ των κλάσεων (Κώδικας 5.20).

```
> library(pROC)
> roc(testdata$Y,test_prob
,plot=TRUE,legacy.axes=TRUE,percent=TRUE,
xlab="Ψευδές Θετικό Ποσοστό",ylab="Αληθές Θετικό Ποσοστό",
main="Καμπύλη ROC")
Setting levels: control = 0, case = 1
Setting direction: controls < cases

Call:
roc.default(response = testdata$Y, predictor = test_prob,
   percent = TRUE,     plot = TRUE, legacy.axes = TRUE,
xlab = "Ψευδές Θετικό Ποσοστό",ylab = "Αληθές Θετικό Ποσοστό",
main = "Καμπύλη ROC")

Data: test_prob in 1009 controls (testdata$Y 0) < 904 cases (
   testdata$Y 1).
Area under the curve: 99.74%
```

Κώδικας 5.20: Δημιουργία καμπύλης ROC για το μοντέλο της μεθόδου της λογιστικής παλινδρόμησης για τα δεδομένα exoplanets.

Το εμβαδόν κάτω από την καμπύλη (AUC) είναι, επίσης, πολύ κοντά στην μονάδα και ίσο με 0.9974. Ο βαθμός διαχωρισιμότητας, δηλαδή, του εν λόγω μοντέλου είναι 99.74%.



Σχήμα 5.7: Καμπύλη ROC για το μοντέλο της μεθόδου της λογιστικής παλινδρόμησης για τα δεδομένα exoplanets.

5.2 Συμπεράσματα

Το πρόβλημα της αναζήτησης Kepler εξωπλανητών αναλύθηκε μέσω των μεθόδων της δενδρικής ταξινόμησης και της λογιστικής παλινδρόμησης. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιείχε χαρακτηριστικά που είχε εντοπίσει ο δορυφόρος Kepler. Όσον αφορά τον αλγόριθμο των δένδρων αποφάσεων, την καλύτερη απόδοση έδωσε το μοντέλο της μεθόδου των τυχαίων δασών, όπου, περίπου, 99.1% παρατηρήσεις ταξινομήθηκαν σωστά στο σύνολο ελέγχου και η αντίστοιχη καμπύλη ROC είχε σχεδόν 100% ικανότητα να διακρίνει αν το εκάστοτε KOI είναι υποψήφιος εξωπλανήτης ή όχι. Το γεγονός αυτό δείχνει ότι οι παρατηρήσεις της εξαρτημένης μεταβλητής προβλέπονται σχεδόν κατά 100% σωστά και οι παρατηρήσεις (KOI's) που προβλέπεται ότι ανήκουν στην κλάση 1 (είναι υποψήφιοι εξωπλανήτες) έχουν εξαιρετικά χαμηλή πιθανότητα να ανήκουν στην κλάση 0 (να μην θεωρούνται πλανήτες) και το αντίστροφο.

Από την άλλη πλευρά, το μοντέλο που προσαρμόστηκε με τη μέθοδο της λογιστικής παλινδρόμησης ήταν κατώτερο από αυτό των τυχαίων δασών υπό την έννοια της ακρίβειας. Συγκεκριμένα, το 97.9% των παρατηρήσεων του συνόλου ελέγχου ταξινομήθηκαν σωστά και, επιπλέον, το εν λόγω μοντέλο σημείωσε ποσοστό 99.74%, όσον αφορά την ικανότητα διαχωρισιμότητας μεταξύ των δύο

κλάσεων του προβλήματος, το οποίο, επίσης, πλησιάζει την πληρότητα (100%), όχι, όμως, τόσο, όσο το μοντέλο των τυχαίων δασών.

Συνεπώς, από την ανάλυση και την σύγκριση των δύο στατιστικών μεθόδων που προηγήθηκαν, μπορούμε να καταλήξουμε στο συμπέρασμα ότι η μέθοδος της δενδρικής ταξινόμησης θα προτιμηθεί έναντι της λογιστικής παλινδρόμησης, καθώς φαίνεται να είναι αποτελεσματικότερη και πιο ακριβής, παράγοντας τα βέλτιστα μοντέλα για το πρόβλημα της αναζήτησης εξωπλανητών Kepler. Τα μοντέλα αυτά μπορούν να συμβάλλουν στον εντοπισμό εκείνων των χαρακτηριστικών που εξηγούν τις διαφορές ανάμεσα σε έναν υποψήφιο εξωπλανήτη και σε έναν λανθασμένα θετικό πλανήτη, καθώς και στην πραγματοποίηση κατάλληλων ρυθμίσεων, προκειμένου να κατασκευαστούν ακόμα καλύτερα μοντέλα, τα οποία θα διευκολύνουν σημαντικά την διαδικασία αναζήτησης εξωπλανητών στο μέλλον.

Κεφάλαιο 6

Επίλογος

Σκοπός της παρούσας διπλωματικής εργασίας ήταν η επεξήγηση και η εμβάθυνση της μεθόδου της δενδρικής παλινδρόμησης και ταξινόμησης (αλγόριθμος CART), μία εξαιρετικά χρήσιμη μέθοδος στην στατιστική. Η σύγχρονη ανάγκη για ανάλυση, επεξεργασία και επίλυση προβλημάτων που συνοδεύονται από μεγάλους όγκους δεδομένων καθιστά ιδανική την εν λόγω μέθοδο, καθώς ερευνητές και στατιστικοί καταφεύγουν σε αυτή, λόγω της ευκολίας και της ερμηνευτικής απλότητας που αυτή διαθέτει. Οι διαθέσιμες προεκτάσεις της μεθόδου της δενδρικής παλινδρόμησης και ταξινόμησης (μέθοδος της ενσάκισης, μέθοδος των τυχαίων δασών, μέθοδος της ενίσχυσης) βελτιώνουν σημαντικά την ακρίβεια των προβλέψεων και, ταυτόχρονα, συμβάλλουν στο να παραμένουν αυτές αναλλοίωτες σε τυχόν μεταβολές στα δεδομένα.

Η τεχνική της δενδρικής παλινδρόμησης και ταξινόμησης χρησιμοποιεί δένδρα αποφάσεων, τα οποία αποτελούν μοντέλα πρόβλεψης, όπου σε κάθε κλαδί του δένδρου πραγματοποιούμε αποφάσεις που αφορούν τα χαρακτηριστικά (επεξηγηματικές μεταβλητές) που έχουμε στην διάθεση μας, έτσι ώστε να βελτιωθεί η ποιότητα της πρόβλεψης ή της ταξινόμησης (πιο ακριβής με μικρότερο σφάλμα). Η συγκεκριμένη τεχνική είναι εύκολη στην ερμηνεία και μπορεί να χειριστεί ποσοτικές, αλλά και κατηγορικές μεταβλητές με αποτελεσματικότητα. Επίσης, η απόδοση των δένδρων αποφάσεων δεν επηρεάζεται από το ενδεχόμενο μη γραμμικής συσχέτισης μεταξύ των δεδομένων. Ωστόσο, η χρήση δένδρων αποφάσεων για την πραγματοποίηση προβλέψεων των τιμών μίας εξαρτημένης μεταβλητής μπορεί να οδηγήσει σε προβλήματα υπερπροσαρμογής (overfitting) με αποτέλεσμα αυτά (δένδρα αποφάσεων) να μην αποδίδουν τόσο καλά σε καινούργια δεδομένα (testing data), τα οποία είναι ανεξάρτητα από τα δεδομένα που χρησιμοποιήθηκαν για την κατασκευή τους (training data). Επιπρόσθετα, ο αλγόριθμος CART δεν επιστρέφει πάντα τα βέλτιστα δένδρα αποφάσεων, καθώς σε κάθε βήμα κατασκευής αυτών επηρεάζεται από την ύπαρξη μεταβλητών που έχουν ισχυρότερη επίδραση από κάποιες άλλες (greedy approach).

Στην εν λόγω εργασία αναλύθηκαν λεπτομερώς όλες οι μέθοδοι που προαναφέρθηκαν συνοδευόμενες από αντίστοιχα παραδείγματα για επιπλέον κατανόηση και ευκολία του αναγνώστη. Έπειτα, η μέθοδος της δενδρικής παλινδρόμησης και της δενδρικής ταξινόμησης μελετήθηκε μέσω κάποιων εφαρμογών, κάνοντας χρήση πραγματικών δεδομένων με τη βοήθεια της στατιστικής γλώσσας προγραμματισμού R. Στην πρώτη περίπτωση της δενδρικής παλινδρόμησης, αυτή εφαρμόστηκε στο οικονομικό πρόβλημα της Υπόθεσης του Κύκλου Ζωής (Life-Cycle Hypothesis) και, στην συνέχεια, τα αποτελέσματα που προέκυψαν συγκρίθηκαν με αυτά που θα είχαμε αν χρησιμοποιούσαμε πολλαπλή γραμμική παλινδρόμηση για το ίδιο πρόβλημα. Αποδείχθηκε ότι η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης έδωσε καλύτερα, πιο ακριβή αποτελέσματα από την μέθοδο της δενδρικής παλινδρόμησης με χαμηλότερο σφάλμα. Τα ευρήματα αυτά αποδόθηκαν, κυρίως, στο γεγονός της ύπαρξης ακραίων τιμών στα δεδομένα, οι οποίες ενισχύουν την μεροληψία όταν εφαρμόζονται οι επεκτάσεις της δενδρικής παλινδρόμησης και στο γεγονός ότι το αρχικό σύνολο δεδομένων ήταν αρκετά μικρό σε μέγεθος. Ακόμα, η ένδειξη γραμμικής συσχέτισης μεταξύ των μεταβλητών συνέβαλε, επίσης, στην υπεροχή της πολλαπλής γραμμικής παλινδρόμησης για το συγκεκριμένο πρόβλημα.

Στην περίπτωση της δενδρικής ταξινόμησης, αυτή εφαρμόστηκε στο πρόβλημα της αναζήτησης εξωπλανητών, μέσω του δορυφόρου Kepler. Αφού υπολογίστηκε η ακρίβεια ταξινόμησης ενός αντικειμένου ως πιθανό εξωπλανήτη ή όχι με την μέθοδο και τις προεκτάσεις της δενδρικής ταξινόμησης, συγκρίθηκε με την αντίστοιχη ακρίβεια που έδωσε η μέθοδος της λογιστικής παλινδρόμησης για το ίδιο πρόβλημα ταξινόμησης. Τελικά, αποδείχθηκε ότι η πρώτη μέθοδος και, συγκεκριμένα, η μέθοδος των τυχαίων δασών είχε καλύτερη απόδοση από το λογιστικό μοντέλο που κατασκευάστηκε.

Συνεπώς, όπως φάνηκε και από τις εφαρμογές, η μεθοδολογία των δένδρων αποφάσεων προτιμάται, συνήθως, από ερευνητές και στατιστικούς για περιπτώσεις προβλημάτων, όπου το πλήθος των διαθέσιμων δεδομένων είναι μεγάλο και αυτά δεν χαρακτηρίζονται από την γραμμικότητα, δηλαδή όταν οι τιμές κάποιου ή κάποιων χαρακτηριστικών δεν εξαρτώνται απαραίτητα από τις τιμές κάποιου άλλου χαρακτηριστικού.

Κεφάλαιο 7

Παράρτημα

7.1 Πολλαπλή Γραμμική Παλινδρόμηση

Έστω ότι ασχολούμαστε με κάποιο πρόβλημα για το οποίο υπάρχουν υποψίες ή ενδείξεις ότι οι επεξηγηματικές μεταβλητές που έχουμε στην διάθεσή μας, έστω $\boldsymbol{X}= (X_1, ..., X_p)$ με $p \geq 2$, συνδέονται γραμμικά με την μεταβλητή απόκρισης Y. Τέτοια είδη προβλημάτων αποτελούν μία επέκταση του απλού γραμμικού μοντέλου στο οποίο οι τιμές της μεταβλητής απόκρισης επηρεάζονται από την μοναδική επεξηγηματική ή ανεξάρτητη μεταβλητή του προβλήματος ($p = 1$).

Λαμβάνοντας υπόψη τη σχέση που περιγράφει το απλό γραμμικό μοντέλο

$$Y_i \ = \ a + \beta X_i + \varepsilon_i, \ i = 1, 2, ..., n, \tag{7.1}$$

μπορούμε κατ' αναλογία να συμπεράνουμε ότι το **γενικό ή πολλαπλό γραμμικό μοντέλο** θα δίνεται από τη σχέση

$$Y_i \ = \ a + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i, \ i = 1, 2, ..., n, \tag{7.2}$$

όπου

• $Y_i, \ i = 1, 2, ..., n$ εκφράζει τις τιμές των παρατηρήσεων της μεταβλητής απόκρισης Y,

• $X_{ij}, \ i = 1, 2, ..., n, \ j = 1, ..., p$ αντιστοιχεί στην i-οστή παρατήρηση της επεξηγηματικής μεταβλητής X_j,

• $a, \ \beta_1, \ \beta_2, \ ..., \ \beta_p$ οι άγνωστες παράμετροι του μοντέλου και

• $\varepsilon_i, \ i = 1, 2, ..., n$ τα τυχαία σφάλματα, για τα οποία υποθέτουμε τα εξής:

✓ $\mathbb{E}(\varepsilon_i) = 0$, για κάθε i,

✓ $V(\varepsilon_i) = \sigma^2$, για κάθε i, δηλαδή τα τυχαία σφάλματα ικανοποιούν την υπόθεση της ομοσκεδαστικότητας, την οποία θα αναφέρουμε στην συνέχεια,

✓ $cov(\varepsilon_i, \varepsilon_j) = 0$, για κάθε $i \neq j$, δηλαδή τα ε_i είναι ασυσχέτιστα μεταξύ τους.

Η διαφορά

$$e_i \ = \ Y_i - \hat{Y}_i = \ Y_i - (\hat{a} + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip}), \tag{7.3}$$

όπου $\hat{Y}_i$ είναι η εκτίμηση της i-οστής τιμής της τυχαίας μεταβλητής Y, σύμφωνα με το πολλαπλό γραμμικό μοντέλο (7.2) και $\hat{a}, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ οι εκτιμήσεις των παραμέτρων του μοντέλου, ονομάζεται **υπόλοιπο ή κατάλοιπο** (residual). Τα υπόλοιπα e_i μπορούν να θεωρηθούν ως οι εκτιμήσεις των τυχαίων σφαλμάτων ε_i.

Η εκτίμηση των παραμέτρων του γραμμικού μοντέλου (με τη μέθοδο των ελαχίστων τετραγώνων) απαιτεί την ικανοποίηση κάποιων σημαντικών προϋποθέσεων. Για την περίπτωση του πολλαπλού γραμμικού μοντέλου, αυτές περιγράφονται παρακάτω [1]:

1) **Γραμμικότητα**
Στην περίπτωση που οι p επεξηγηματικές μεταβλητές είναι ασυσχέτιστες μεταξύ τους, ο έλεγχος της γραμμικότητας μπορεί να γίνει με την δημιουργία p διαφορετικών διαγραμμάτων διασποράς, ένα για κάθε μεταβλητή. Στην πιο συνηθισμένη περίπτωση, όπου οι επεξηγηματικές μεταβλητές συσχετίζονται, η παραπάνω διαδικασία δεν μπορεί να εφαρμοστεί. Εναλλακτικά, πρέπει να ελέγξουμε αν η τιμή της επεξηγηματικής μεταβλητής $X_j, j = 1, ..., p$ συνδέεται γραμμικά με τη δεσμευμένη μέση τιμή της μεταβλητής απόκρισης Y, δοθέντος του $\boldsymbol{X}$=$\boldsymbol{x}$,

$$\mathbb{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right] \ = \ a + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{7.4}$$

δεδομένου ότι οι άλλες τιμές των υπόλοιπων επεξηγηματικών μεταβλητών συνδέονται γραμμικά με τη δεσμευμένη μέση τιμή της Y. Παίρνοντας τις τιμές των εκτιμητών των συντελεστών που προκύπτουν χρησιμοποιώντας το 'πλήρες' μοντέλο πολλαπλής γραμμικής παλινδρόμησης και ύστερα από πράξεις, καταλήγουμε στην σχέση

$$p_j(x_{ij}) \approx \hat{b}_j x_{ij} + \hat{\varepsilon}_i \ \equiv \ P_{ij}, \ i = 1, ..., n \tag{7.5}$$

Οι όροι P_{ij} ονομάζονται **j-μερικά υπόλοιπα** (partial residuals). Συνεπώς, η υπόθεση της γραμμικότητας στο πολλαπλό γραμμικό μοντέλο ελέγχεται εξετάζοντας την γραμμικότητα της συνάρτησης $p_j(\cdot)$ με το αντίστοιχο διάγραμμα διασποράς των σημείων (x_{ij}, P_{ij}), για κάθε $j = 1, ..., p$. Αν υπάρχουν αποκλίσεις από την γραμμικότητα, μπορούμε να μετασχηματίσουμε κατάλληλα την προβληματική ή προβληματικές επεξηγηματικές μεταβλητές, προκειμένου να εξαλείψουμε το πρόβλημα.

2) **Κανονικότητα των Σφαλμάτων**
Για τον έλεγχο της υπόθεσης ότι τα τυχαία σφάλματα $\varepsilon_i, i = 1, ..., n$ ακολουθούν την κανονική κατανομή θεωρούμε τις εκτιμήσεις αυτών, δηλαδή τα υπόλοιπα e_i της σχέσης (7.3). Η κανονικότητα των υπολοίπων εξετάζεται είτε γραφικά μέσω του διαγράμματος της κανονικής κατανομής, είτε με τη βοήθεια κάποιου ελέγχου. Συνήθως, σε περίπτωση που διαπιστωθεί ότι υπάρχουν αποκλίσεις από την κανονική κατανομή, μετασχηματίζουμε με κατάλληλο τρόπο την μεταβλητή απόκρισης Y (χρησιμοποιώντας τον λογάριθμό της για παράδειγμα).

3) **Ομοσκεδαστικότητα**
Με την εξέταση της ομοσκεδαστικότητας ελέγχεται αν η δεσμευμένη κατανομή της τυχαίας μεταβλητής Y, δοθέντος $\boldsymbol{X}=\boldsymbol{x}$, έχει σταθερή διασπορά ανεξάρτητα της τιμής $\boldsymbol{x}$. Με άλλα λόγια, ελέγχουμε αν η διασπορά των τυχαίων σφαλμάτων παραμένει αναλλοίωτη για τις διάφορες τιμές $\boldsymbol{x}$ του τυχαίου διανύσματος των επεξηγηματικών μεταβλητών $\boldsymbol{X}$. Ως εκτιμήτρια των τυχαίων σφαλμάτων ε_i έχουμε ορίσει τα υπόλοιπα e_i $(i = 1, ..., n)$, δηλαδή ισχύει ότι $\hat{\varepsilon}_i = e_i$ $(i = 1, ..., n)$. Η εγκυρότητα της υπόθεσης της ομοσκεδαστικότητας ελέγχεται με το διάγραμμα διασποράς μεταξύ των υπολοίπων e_i και των προβλεπόμενων τιμών $\hat{Y}_i$. Εάν στο διάγραμμα εμφανίζεται κάποιο μοτίβο (pattern) ή ακολουθία, δηλαδή εάν τα σημεία παρουσιάζουν κάποιο συγκεκριμένο τρόπο συμπεριφοράς, τότε υπάρχουν ισχυρές ενδείξεις εναντίον της υπόθεσης της ομοσκεδαστικότητας (ύπαρξη 'ετεροσκεδαστικότητας'). Σε αυτήν την περίπτωση χρησιμοποιούμε κάποιον μετασχηματισμό της Y. Αντίθετα, αν τα σημεία φαίνεται να είναι τυχαία κατανεμημένα στο διάγραμμα, τότε μπορούμε να συμπεράνουμε ότι ισχύει η απαίτηση της ομοσκεδαστικότητας.

4) **Ανεξαρτησία των Σφαλμάτων**
Η ανεξαρτησία των τυχαίων σφαλμάτων μπορεί να ελεγχθεί με την κατασκευή ενός διαγράμματος διασποράς μεταξύ των υπολοίπων e_i $(i = 1, ..., n)$ και της σειράς των δεδομένων. Όπως και με την συνθήκη της ομοσκεδαστικότητας, αν τα σημεία του διαγράμματος κατανέμονται τυχαία, χωρίς να παρουσιάζουν κάποιο συστηματικό τρόπο συμπεριφοράς,

τότε λέμε ότι ικανοποιείται η υπόθεση της ανεξαρτησίας των σφαλμάτων. Διαφορετικά, έχουμε πρόβλημα 'αυτοσυσχέτισης' (autocorrelation), το οποίο μπορεί να ξεπεραστεί προσθέτοντας κάποια ανάλογη επεξηγηματική μεταβλητή στο μοντέλο.

Οι ίδιες προϋποθέσεις ισχύουν και για το απλό γραμμικό μοντέλο με την μόνη διαφορά ότι ο έλεγχος της γραμμικότητας πραγματοποιείται κατευθείαν από το διάγραμμα διασποράς των σημείων (x_i, y_i) με $i = 1, ..., n$, όπου x_i οι τιμές των παρατηρήσεων της επεξηγηματικής ή ανεξάρτητης μεταβλητής X και y_i οι τιμές των παρατηρήσεων της εξαρτημένης μεταβλητής ή μεταβλητής απόκρισης Y.

7.2 Μέτρα Αξιολόγησης ενός Μοντέλου

7.2.1 Συντελεστής Προσδιορισμού

Ο **συντελεστής προσδιορισμού** (coefficient of determination) συμβολίζεται με R^2 και δίνεται από την ποσότητα

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{7.6}$$

Με $\hat{y}_i$ $(i = 1, ..., n)$ δηλώνονται οι προβλεπόμενες τιμές της εξαρτημένης μεταβλητής και η ποσότητα $\bar{y}$ $(i = 1, ..., n)$ εκφράζει την μέση τιμή των παρατηρήσεων της εξαρτημένης μεταβλητής. Ο συντελεστής προσδιορισμού μετρά την επιτυχία του μοντέλου στην εξήγηση της συμπεριφοράς της εξαρτημένης μεταβλητής Y ή, αλλιώς, εκφράζει το ποσοστό της διασποράς (μεταβλητότητας) της τυχαίας μεταβλητής Y που εξηγείται με βάση το μοντέλο. Παίρνει τιμές στο διάστημα [0,1] και όσο πλησιάζει την μονάδα (100%), τόσο καλύτερη είναι η προσαρμογή, δηλαδή τόσο περισσότερο επιβεβαιώνονται οι υποψίες μας ότι το μοντέλο που χρησιμοποιούμε είναι το κατάλληλο για τα δεδομένα που έχουμε στην διάθεσή μας. Αν, για παράδειγμα, χρησιμοποιούμε πολλαπλή γραμμική παλινδρόμηση με p επεξηγηματικές μεταβλητές και έναν υψηλό συντελεστή προσδιορισμού, τότε, αυτό σημαίνει ότι υπάρχει ισχυρή γραμμική συσχέτιση ανάμεσα στις τυχαίες μεταβλητές Y και $X_1, ..., X_p$.

7.2.2 Διορθωμένος ή Προσαρμοσμένος Συντελεστής Προσδιορισμού

Αν στο μοντέλο που χρησιμοποιούμε προσθέσουμε επιπλέον μεταβλητές, τότε ο συντελεστής προσδιορισμού θα αυξηθεί, καθώς το άθροισμα των τετραγώνων των υπολοίπων (SSR) στον τύπο (7.6) μειώνεται ακόμα και στην περίπτωση που οι μεταβλητές που προστίθενται δεν προσφέρουν σημαντική πληροφορία για την τιμή της μεταβλητής απόκρισης, δηλαδή δεν είναι στατιστικά σημαντικές. Επομένως, λόγω ύπαρξης τέτοιων περιπτώσεων, προτιμάται να χρησιμοποιείται ο **διορθωμένος συντελεστής προσδιορισμού** $\tilde{R}^2$ (adjusted coefficient of determination), ο οποίος ορίζεται ως

$$\tilde{R}^2 \;=\; 1 - \frac{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1}}{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}. \tag{7.7}$$

Σε αντίθεση με τον συντελεστή προσδιορισμού, ο διορθωμένος συντελεστής προσδιορισμού δεν παρουσιάζει αύξηση όταν προστίθενται μία οποιαδήποτε επεξηγηματική μεταβλητή στο μοντέλο, αλλά μόνο όταν η μεταβλητή αυτή προσφέρει κάποια βελτίωση στο μοντέλο. Με άλλα λόγια, λαμβάνει υπόψη την πολυπλοκότητα του μοντέλου, δηλαδή τον αριθμό των επεξηγηματικών μεταβλητών p που περιλαμβάνονται σε αυτό και το γεγονός αυτό τον καθιστά ιδανικό μέτρο καταλληλότητας του μοντέλου, αλλά και αποτελεσματικό μέτρο σύγκρισης δύο μοντέλων.

Ο διορθωμένος συντελεστής προσδιορισμού συνδέεται με τον R^2 σύμφωνα με την σχέση (7.8):

$$\tilde{R}^2 \;=\; R^2 - (1-R^2)\frac{p}{n-p-1}. \tag{7.8}$$

7.2.3 Έλεγχος Pearson

Ένας άλλος τρόπος ελέγχου της σχέσης μεταξύ δύο συνεχών μεταβλητών είναι μέσω του **συντελεστή συσχέτισης Pearson** (Pearson' s correlation coefficient), ο οποίος συμβολίζεται με r και για δοσμένο δείγμα $(x_i, y_i),\ i = 1, ..., n$ δύο μεταβλητών με n παρατηρήσεις η καθεμία είναι ίσος με

$$r \;=\; \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2}} \tag{7.9}$$

με n το μέγεθος του δείγματος και $\bar{x}$, $\bar{y}$ οι δειγματικοί μέσοι των $\boldsymbol{x}$ και $\boldsymbol{y}$ αντίστοιχα. Δείχνει την κατεύθυνση της σχέσης μεταξύ των μεταβλητών (θετική ή αρνητική) και πόσο ισχυρή είναι. Παίρνει τιμές μεταξύ του -1 (τέλεια αρνητική συσχέτιση) και του 1 (τέλεια θετική συσχέτιση) με την μηδενική τιμή να ισοδυναμεί με την απουσία κάποιας σχέσης μεταξύ των μεταβλητών. Εκτός από την συνέχεια των μεταβλητών, η χρήση του συντελεστή Pearson προϋποθέτει, επίσης, αυτές να ακολουθούν την κανονική κατανομή και να σχετίζονται γραμμικά μεταξύ τους.

Ο **έλεγχος κατά Pearson** (Pearson' s correlation coefficient test) ελέγχει την μηδενική υπόθεση $H_0 : r = 0$ (δεν υπάρχει συσχέτιση ανάμεσα στις μεταβλητές) έναντι της εναλλακτικής υπόθεσης $H_1 : r \neq 0$ (υπάρχει συσχέτιση ανάμεσα στις μεταβλητές). Το στατιστικό ελέγχου κάτω από τη μηδενική υπόθεση είναι

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim St(n-2) \tag{7.10}$$

και ακολουθεί την κατανομή Student με n−2 βαθμούς ελευθερίας.

7.3 Λογιστική Παλινδρόμηση

Η μέθοδος της λογιστικής παλινδρόμησης χρησιμεύει στην ταξινόμηση των τιμών μίας μεταβλητής απόκρισης Y, χρησιμοποιώντας βασικά στοιχεία και κατανομές από τη θεωρία των πιθανοτήτων. Συνήθως, η Y είναι μία δίτιμη τυχαία μεταβλητή, δηλαδή λαμβάνει μόνο δύο τιμές, οι οποίες αντιστοιχούν σε δύο ενδεχόμενα. Μία αυθαίρετη και βολική κωδικοποίηση αυτών των ενδεχομένων αποτελούν οι τιμές 0 και 1. Εστιάζουμε την προσοχή μας σε ένα από τα δύο ενδεχόμενα, την 'επιτυχία', δηλαδή, χρησιμοποιώντας τους ανάλογους συμβολισμούς, στο ενδεχόμενο $Y = 1$ με πιθανότητα $p = P($επιτυχία$)$. Συνεπώς, η Y είναι μία τυχαία μεταβλητή που ακολουθεί την κατανομή Bernoulli, δηλαδή $Y \sim B(p)$ με $\mathbb{E}(Y) = p$ και $V(Y) = p(1-p)$.

Υπάρχουν περιπτώσεις στην λογιστική παλινδρόμηση όπου η εξαρτημένη μεταβλητή Y εκφράζει τον αριθμό των επιτυχιών σε n δοκιμές και, επομένως οι πιθανές τιμές της κυμαίνονται μεταξύ του μηδενός και του n. Τότε, αν υποθέσουμε ότι όλες οι δοκιμές είναι ανεξάρτητες μεταξύ τους και ότι η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή, μπορούμε να επεκτείνουμε την πρώτη περίπτωση (η Y παίρνει μόνο δύο τιμές) και να συμπεράνουμε ότι η Y θα ακολουθεί την Διωνυμική (binomial) κατανομή με παραμέτρους $Y \sim b(n, p)$. Ωστόσο, στο πρόβλημα αναζήτησης Kepler εξωπλανητών του Κεφαλαίου 5, η μεταβλητή απόκρισης Y έπαιρνε μόνο δύο τιμές ('1' για το ενδεχόμενο "CANDIDATE" και '0' για το ενδεχόμενο "FALSE POSITIVE"). Συνεπώς, θα δωθούν τα βασικά χαρακτηριστικά της λογιστικής παλινδρόμησης για αυτήν την περίπτωση,

τα οποία μπορούν με παρόμοιο τρόπο να επεκταθούν και στην περίπτωση της Διωνυμικής κατανομής.

Αν $\boldsymbol{X}$ είναι το τυχαίο διάνυσμα των επεξηγηματικών μεταβλητών από τις οποίες εξαρτάται η Y, τότε το **μοντέλο της λογιστικής παλινδρόμησης** που προσαρμόζεται [3] είναι ένα γενικευμένο γραμμικό μοντέλο της μορφής

$$\eta_x \;=\; g(\mathbb{E}(Y_{\boldsymbol{X}})) \;=\; g(\mu_{\boldsymbol{X}}) \;=\; \boldsymbol{X}'\boldsymbol{\beta} \tag{7.11}$$

με τα ακόλουθα χαρακτηριστικά:

1) $Y_{\boldsymbol{X}} \sim B(\mu_{\boldsymbol{X}})$,

2) $\eta_x \;=\; g(\mu_{\boldsymbol{X}}) \;=\; \ln\frac{\mu_{\boldsymbol{X}}}{n_{\boldsymbol{X}}-\mu_{\boldsymbol{X}}} \;=\; \ln\frac{p_{\boldsymbol{X}}}{1-p_{\boldsymbol{X}}} \;=\; \mathrm{logit}(p_{\boldsymbol{X}}) \;=\; \boldsymbol{X}'\boldsymbol{\beta}$ (συνάρτηση σύνδεσης),

3) οι παρατηρήσεις $Y_{\boldsymbol{X}}$ είναι ανεξάρτητες μεταξύ τους,

όπου $n_{\boldsymbol{X}}$ δηλώνει τον αριθμό των επαναλήψεων της τιμής του διανύσματος $\boldsymbol{X}$ των ανεξάρτητων μεταβλητών και για την κατανομή Bernoulli ισχύει ότι $n_{\boldsymbol{X}} = 1$. Με $\boldsymbol{\beta}$ συμβολίζεται το διάνυσμα των συντελεστών παλινδρόμησης (παραμέτρων) του μοντέλου συμπεριλαμβανομένου του σταθερού όρου a. Από την συνάρτηση σύνδεσης, η πιθανότητα επιτυχίας p μπορεί να υπολογιστεί από τον τύπο

$$p_{\boldsymbol{X}} \;=\; \frac{e^{\eta_x}}{1+e^{\eta_x}}. \tag{7.12}$$

Τα υπόλοιπα που χρησιμοποιούνται πιο συχνά στα γενικευμένα γραμμικά μοντέλα και, ειδικότερα στην λογιστική παλινδρόμηση, είναι τα λεγόμενα **υπόλοιπα deviance** (deviance residuals) που δίνονται από τον τύπο

$$r_i^D \;=\; \mathrm{sgn}(y_i-\hat{\mu}_i)\sqrt{d_i(y_i,\hat{\mu}_i)}, i = 1, ..., n, \tag{7.13}$$

με $d_i(y_i,\hat{\mu}_i) = 2\left[y_i\ln(y_i/\hat{\mu}_i) - (y_i-\hat{\mu}_i)\right]$ να εκφράζει την συμβολή της i-οστής παρατήρησης στην 'απόκλιση' του μοντέλου ή, αλλιώς, στην ελεγχοσυνάρτηση deviance

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) \;=\; D(\hat{\boldsymbol{\beta}}) \;=\; 2\sum_{i=1}^{n}\left[y_i\ln(y_i/\hat{\mu}_i) - (y_i-\hat{\mu}_i)\right] \;=\; \sum_{i=1}^{n} d_i(\hat{\boldsymbol{\beta}}) \;=\; \sum_{i=1}^{n}(r_i^D)^2. \tag{7.14}$$

Η συνάρτηση $\mathrm{sgn}(y_i-\hat{\mu}_i)$ δίνει θετικό πρόσημο, εάν $y_i \geq \hat{\mu}_i$ και αρνητικό πρόσημο, εάν $y_i < \hat{\mu}_i$.

Στον κώδικα 5.17 του Κεφαλαίου 5 φαίνονται τα αποτελέσματα που έδωσε η R με βάση το μοντέλο λογιστικής παλινδρόμησης που προσαρμόστηκε στα δεδομένα "exoplanets". Στα αποτελέσματα αυτά αναφέρονται κάποια αριθμητικά μέτρα των υπολοίπων deviance, τα οποία φαίνεται να είναι κεντραρισμένα στο 0, οι εκτιμήσεις των συντελεστών που υπάρχουν στο μοντέλο, η απόκλιση του μοντέλου που περιέχει μόνο τον σταθερό όρο ("Null deviance") και αυτή των υπολοίπων ("Residual deviance") με τους αντίστοιχους βαθμούς ελευθερίας, καθώς και η τιμή του κριτηρίου AIC. Από τις P-τιμές και την τιμή της στατιστικής συνάρτησης του ελέγχου Wald ("z value") της κάθε επεξηγηματικής μεταβλητής, μπορούμε να συμπεράνουμε αν η εν λόγω επεξηγηματική μεταβλητή είναι στατιστικά σημαντική ή όχι (σε επίπεδο σημαντικότητας 5%), πραγματοποιώντας έλεγχο της μηδενικής υπόθεσης $H_0 : \beta_j = 0$ έναντι της εναλλακτικής υπόθεσης $H_1 : \beta_j \neq 0$. Τέλος, η R δίνει τον αριθμό των Fisher επαναλήψεων (Fisher Scoring iterations) που εκφράζει πόσο γρήγορα συγκλίνει η συνάρτηση που προσαρμόστηκε με την εντολή "glm()" στους εκτιμητές μέγιστης πιθανοφάνειας των συντελεστών.

7.3.1 Καμπύλη ROC

Η 'Καμπύλη Λειτουργικών Χαρακτηριστών', η οποία, συνήθως, αναφέρεται ως **Καμπύλη ROC** (Receiver Operating Characteristic Curve) αποτελεί μια μέτρηση απόδοσης για τα προβλήματα ταξινόμησης ή, αλλιώς, έναν διαφορετικό δείκτη καλής προσαρμογής. Έστω μία δυαδική τυχαία μεταβλητή Y, η οποία λαμβάνει τις τιμές $Y = 1$ και $Y = 0$. Έστω, ακόμα, μία ποσότητα $\hat{p}$ που εκφράζει την εκτίμηση της πιθανότητας επιτυχίας, δηλαδή της πιθανότητας του ενδεχομένου $Y = 1$, καθώς και ένα όριο p_0, τέτοιο ώστε [3]:
▶ αν $\hat{p} > p_0$, τότε εκτιμάται ότι $Y = 1$, ενώ
▶ αν $\hat{p} \leq p_0$, τότε εκτιμάται ότι $Y = 0$.

Ορίζονται, στην συνέχεια, δύο απαραίτητες ποσότητες για την κατασκευή μίας καμπύλης ROC:

• **Ευαισθησία** (Sensitivity): Είναι 'το ποσοστό των αληθώς θετικών αποτελεσμάτων', δηλαδή το ποσοστό ορθής πρόβλεψης (ταξινόμησης) της κατάστασης $Y = 1$ και συμβολίζεται ως **TPR** (True Positive Rate).

• **Ειδικότητα** (Specificity): Είναι 'το ποσοστό των αληθώς αρνητικών αποτελεσμάτων', δηλαδή το ποσοστό ορθής πρόβλεψης (ταξινόμησης) της κατάστασης $Y = 0$ και συμβολίζεται ως **1−FPR**, όπου FPR είναι το ποσοστό των ψευδώς θετικών αποτελεσμάτων (False Positive Rate), δηλαδή το ποσοστό

λανθασμένης πρόβλεψης της κατάστασης $Y = 1$.

Τα παραπάνω ποσοστά υπολογίζονται με τη βοήθεια ενός πίνακα συνάφειας, ο οποίος κατασκευάζεται εκ νέου κάθε φορά, για κάθε όριο p_0. Συγκεκριμένα, για κάθε νέα τιμή του p_0, σημειώνουμε την τιμή που λαμβάνει η ποσότητα FPR στον οριζόντιο άξονα και την τιμή που λαμβάνει η ποσότητα TPR στον κατακόρυφο άξονα αντίστοιχα. Έπειτα, ενώνουμε τα σημεία με μία γραμμή, κατασκευάζοντας την επιθυμητή καμπύλη ROC.

Όοσν αφορά την στατιστική γλώσσα προγραμματισμού R, για λόγους ευκολίας και καλύτερης κατανόησης των αξόνων, στον κατακόρυφο άξονα μίας καμπύλης ROC απεικονίζεται το ποσοστό των αληθώς θετικών αποτελεσμάτων (Ευαισθησία), ενώ στον οριζόντιο άξονα απεικονίζεται το ποσοστό των ψευδώς θετικών αποτελεσμάτων (1−Ειδικότητα).

Επιπλέον, στο γράφημα απεικονίζεται και η ευθεία $y = x$ που παριστάνει τα σημεία στα οποία το ποσοστό των αληθώς θετικών αποτελεσμάτων ισούται με το ποσοστό των ψευδώς θετικών αποτελεσμάτων.

Επομένως, για κάθε νέο όριο p_0, υπολογίζονται οι ποσότητες 'Ευαισθησία' και 'Ειδικότητα' και τα αντίστοιχα σημεία αποτυπώνονται στο διάγραμμα.

Το εμβαδόν της επιφάνειας που βρίσκεται κάτω από την καμπύλη ROC, γνωστό ως "AUC" (Area Under the Curve), αντιπροσωπεύει το βαθμό ή το μέτρο της διαχωρισιμότητας. Δείχνει, δηλαδή, την διακριτική ικανότητα του μοντέλου μεταξύ των τάξεων. Όσο υψηλότερη είναι η AUC, τόσο καλύτερο είναι το μοντέλο στην πρόβλεψη '0' τάξεων ως '0' και '1' τάξεων ως '1'. Επίσης, με βάση το AUC μπορεί να γίνει σύγκριση δύο καμπυλών ROC. Εκείνη με το μεγαλύτερο AUC, δηλαδή εκείνη με την μεγαλύτερη ικανότητα να διακρίνει τις κλάσεις μεταξύ τους, είναι και η καλύτερη και αυτή που τελικά προτιμάται.

Βιβλιογραφικές Αναφορές

[1] Φουσκάκης, Δ. (2013). *Ανάλυση Δεδομένων με Χρήση της R*. Εκδόσεις Τσότρας. Αθήνα.

[2] Πετρίδης, Δ. (2015). *Ανάλυση Πολυμεταβλητών Τεχνικών, Εφαρμογές Περιπτώσεων*. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις.

[3] Καρώνη, Χ. και Οικονόμου, Π. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης: Με χρήση MINITAB και R*. Εκδόσεις Συμεών. $2^{η}$ έκδοση. Αθήνα.

[4] Κοκολάκης, Γ. και Φουσκάκης, Δ. (2009). *Στατιστική. Θεωρία & Εφαρμογές*. Εκδόσεις Συμεών. Αθήνα.

[5] Κουγιουμτζής, Δ. (2016-2017). Εφαρμοσμένη Στατιστική Ανάλυση, Μέρος Α. *Εφαρμοσμένη Στατιστική Ανάλυση*. Τεχνικές Ανάλυσης και Συλλογής Δεδομένων–Στοιχεία Επιχειρησιακής Έρευνας. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

[6] https://www.economicshelp.org/blog/27080/concepts/life-cycle-hypothesis/.

[7] James, G., Witten, D., Hastie, T. and Tibshiran, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. New York.

[8] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Routledge. New York.

[9] https://towardsdatascience.com/understanding-gradient-boosting-from-scratch-with-small-dataset-587592cc871f.

[10] https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/.

[11] Nikolaou, N. (2014). *Introduction to AdaBoost*. University of Manchester. United Kingdom.

[12] https://blog.paperspace.com/gradient-boosting-for-classification/.

[13] Deaton, A. (2005). *Franco Modigliani and the Life Cycle Theory of Consumption* . Research Program in Development Studies and Center for Health and Wellbeing. Princeton University. New Jersey.

[14] Saha, R. (2019). *Comparing Classification Models on Kepler Data*. University of Alberta. Canada.

[15] https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html.