ΕΘΝΙΚΟ ΜΕΤΣΟΒΕΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΠΜΣ: ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

Μεταπτυχιακή Διπλωματική Εργασία

Μοντελοποίηση Δεδομένων με χρήση Λογιστικής Παλινδρόμησης, Δέντρων Ταξινόμησης και Παλινδρόμησης, και Εφαρμογές στην ανάλυση των Καμπύλων ROC

Δημουλάς Δημήτριος

Επιβλέπων : Χρήστος Κουκουβίνος, Καθηγητής ΕΜΠ

Αθήνα, Οκτώβριος 2011

Περιεχόμενα

Περίληψη

Η πρόβλεψη της έκβασης ενός φαινόμενου, το να καταφέρουμε δηλαδή να εκτιμήσουμε με κάποιον τρόπο το αποτέλεσμά του, είναι ένα βασικό κίνητρο στην ανάπτυξη στατιστικών μεθόδων και εργαλείων. Είτε πρόκειται για προβλέψεις που αφορούν την υγεία ασθενών, είτε αφορούν φυσικά φαινόμενα, είτε οικονομικά, δεκάδες μοντέλα έχουν κατασκευαστεί, όπως και μέθοδοι για την αξιολόγηση της απόδοσης αυτών των μοντέλων.

Η θεωρητική βάση για την πρόβλεψη με μαθηματικές μεθόδους βασίζεται στη ταξινόμηση, όπου κάθε στοιχείο ή δεδομένο που έχουμε διαθέσιμο, το ταξινομούμε μέσω μιας διαδικασίας σε μια κατηγορία, μαζί με τα υπόλοιπα στοιχεία που μοιράζονται κοινές ιδιότητες. Δύο πολύ διαδεδομένες μέθοδοι για την πρόβλεψη αποτελούν τα δέντρα ταξινόμησης και παλινδρόμησης (Classification and Regression Tree -CART) και η λογιστική παλινδρόμηση. Για την αξιολόγηση της ικανότητας πρόβλεψης μπορεί να χρησιμοποιηθεί σαν δείκτης το εμβαδό κάτω από την Receiver Operating Characteristics (ROC) καμπύλη που σχεδιάζεται με βάση τα αποτελέσματα των μεθόδων που ακολουθήθηκαν.

Το πρώτο κεφάλαιο της παρούσας εργασίας, αποτελείται από θεωρητικά στοιχεία τα οποία σχετίζονται με την έννοια της ταξινόμησης, την κατασκευή και την αξιολόγηση των δέντρων CART, τη μέθοδο της λογιστικής παλινδρόμησης και τις καμπύλες ROC. Επίσης, παρουσιάζονται κάποια στοιχεία σχετικά με μεθόδους διαχείρισης των δεδομένων μας, όπως το cross-validation και το bootstrapping.

Το δεύτερο κεφάλαιο της εργασίας αφορά μελέτες που έχουν πραγματοποιηθεί (Austin (2007), Austin et al. (2010), Faltus et al. (2008), Tsien et. al. (1998) οι οποίες συγκρίνουν, με χρήση του εμβαδού κάτω από μια καμπύλη ROC, την προβλεπτική ικανότητα που εμφάνισε η χρήση των δέντρων CART και η λογιστική παλινδρόμηση πάνω σε ιατρικά δεδομένα.

Ευχαριστίες

Θα ήθελα κατ’ αρχήν να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Χρήστο Κουκουβίνο που δέχτηκε να μου αναθέσει την παρούσα εργασία.

Επίσης, ένα πολύ μεγάλο ευχαριστώ στην υποψήφια διδάκτορα Χριστίνα Παρπούλα. Η βοήθεια που μου προσέφερε κατά τη διάρκεια της συγγραφής αυτής της εργασίας ήταν ανεκτίμητη.

Τέλος, θέλω να ευχαριστήσω όλους τους συντρόφους και φίλους που μου στάθηκαν και με βοήθησαν κατά τη συγγραφή της εργασίας μέχρι και την ολοκλήρωσή της.

Κεφάλαιο 1:

1.1 Το πρόβλημα της πρόβλεψης ως πρόβλημα ταξινόμησης

Σαν ταξινόμηση ορίζουμε τη διαδικασία κατά την οποία εξετάζουμε τις ιδιότητες των αντικειμένων ενός συνόλου (π.χ. μιας βάσης δεδομένων) και στη συνέχεια αντιστοιχίζουμε αυτά τα αντικείμενα σε κατηγορίες, ανάλογα με τις κοινές τους ιδιότητες, με βάση κάποιο μοντέλο ταξινόμησης.

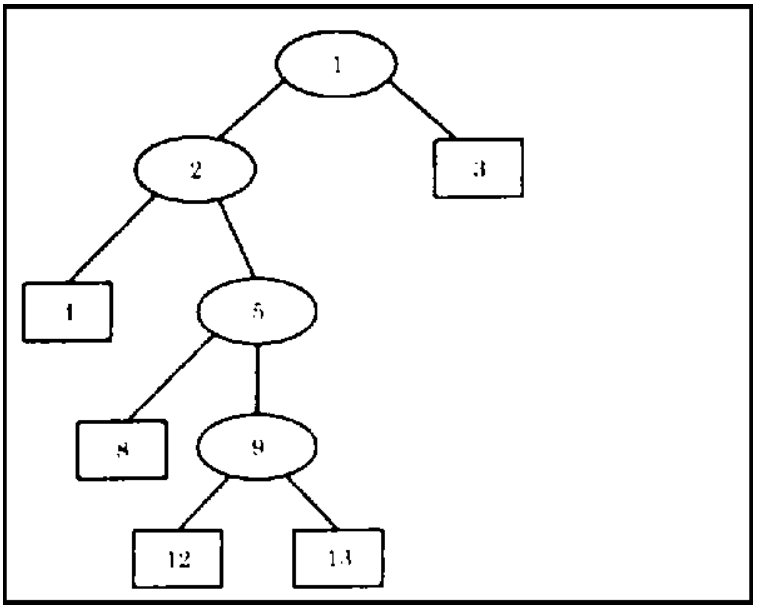
Με μαθηματικούς όρους, θεωρούμε την δειγματική βάση δεδομένων σαν το σύνολο εκπαίδευσης (training set) και τη συμβολίζουμε $\mathrm{E} = \{t_1, t_2, ..., t_n\}$. Κάθε εγγραφή στη δειγματική βάση αποτελείται από ένα σύνολο πολλαπλών χαρακτηριστικών και έχει μια γνωστή ετικέτα κλάσης (label), όπως οι εγγραφές σε μια βάση δεδομένων W. Το σύνολο των κλάσεων θα το συμβολίσουμε $\mathrm{C} = \{c_1, c_2, ..., c_m\}$. Σκοπός είναι να αναλύσουμε τα δεδομένα που έχουμε και στη συνέχεια να αναπτύξουμε μια ακριβή περιγραφή ή ένα μοντέλο που θα περιγράφει κάθε κλάση, χρησιμοποιώντας τα χαρακτηριστικά που έχουν τα δεδομένα. Να ορίσουμε, δηλαδή, μια απεικόνιση $\mathrm{f} : \mathrm{E} \to \mathrm{C}$ όπου κάθε εγγραφή t_i αντιστοιχίζεται σε μια κλάση c_j. Οι περιγραφές των κλάσεων που προκύπτουν από αυτή τη διαδικασία, μπορούμε να τις χρησιμοποιήσουμε για να ταξινομήσουμε μελλοντικά δεδομένα (test set) σε αυτή τη βάση W ή για να αναπτύξουμε καλύτερες περιγραφές για τις κλάσεις, με άλλα λόγια καλύτερους «κανόνες ταξινόμησης». Θεωρώντας την διαδικασία της ταξινόμησης σαν το διαμερισμό του συνόλου E σε κλάσεις ισοδυναμίας, τότε το πρόβλημα της πρόβλεψης ουσιαστικά είναι ένα πρόβλημα ταξινόμησης σε άπειρο αριθμό κλάσεων.

Η ταξινόμηση ανήκει στην κατηγορία της μάθησης με επίβλεψη (supervised learning), λόγω του ότι ξέρουμε εκ των προτέρων τις ομάδες ταξινόμησης όπως και το αποτέλεσμα του υποδείγματος. Αναλόγως του βαθμού αποδοχής της περιγραφής, μπορούμε να μετράμε το βαθμό αξιοπιστίας της και για δεδομένα τα οποία δεν χρησιμοποιήθηκαν στη διαδικασία διαμόρφωσής της. Οι διάφορες τεχνικές ταξινόμησης δημιουργούν ένα μοντέλο μέσω της αξιολόγησης των δεδομένων από το σύνολο εκπαίδευσης και στη συνέχεια το εφαρμόζουν σε νέα δεδομένα. Η ταξινόμηση αποτελεί αντικείμενο μελέτης πολλών επιστημών, όπως είναι η εξόρυξη δεδομένων (data mining), η μηχανική μάθηση (machine learning) και η στατιστική. Εφαρμόζεται σε πολλά διαφορετικά πεδία, όπως για διαγνώσεις στην ιατρική, στην βιολογία, στο marketing. Δυο από τις πιο διαδεδομένες τεχνικές ταξινόμησης είναι τα δέντρα αποφάσεων και η παλινδρόμηση τις οποίες θα εξετάσουμε αναλυτικά.

1.2 Πρόβλεψη μέσω δέντρων απόφασης

Τα δέντρα απόφασης ή δέντρα ταξινόμησης (decision trees ή classification trees) είναι μια από τις πιο συχνά χρησιμοποιούμενες τεχνικές ταξινόμησης, για το λόγο ότι προσφέρει σαφή και κατανοητά αποτελέσματα μέσα σε λίγο χρόνο. Η συγκεκριμένη τεχνική μπορεί να χρησιμοποιηθεί για την ταξινόμηση και την πρόβλεψη τόσο ονομαστικών, όσο και αριθμητικών ποσοτήτων.

Στην περίπτωση που εξετάζουμε ονομαστικές ποσότητες, τότε τα δέντρα λέγονται δέντρα ταξινόμησης. Στην περίπτωση που εξετάζουμε αριθμητικές ποσότητες, τότε ονομάζονται δέντρα παλινδρόμησης (regression trees). Τα δέντρα ταξινόμησης και τα δέντρα παλινδρόμησης βασίζονται στο ίδιο σκεπτικό, με κάποιες όμως διαφορές που θα δούμε παρακάτω.



Εικόνα 1.1*: Παράδειγμα ενός τυπικού δέντρου απόφασης*

1.2.1 Περιγραφή των δέντρων απόφασης

Η συνθήκη με βάση την οποία κινείται η τεχνική των δέντρων απόφασης παλινδρόμησης, είναι η διαίρεση των δεδομένων σε υποσύνολα με βάση κάποιο χαρακτηριστικό και με κριτήριο το κέρδος πληροφορίας, εξασφαλίζοντας ότι τα στοιχεία του κάθε υποσυνόλου θα έχουν κατά το δυνατόν την ίδια τιμή σε αυτό το χαρακτηριστικό. Αυτή η διαδικασία, χρησιμοποιώντας κόμβους και διακλαδώσεις, ξεκινώντας από ένα κόμβο-ρίζα, οδηγεί στην απεικόνιση που λέγεται δέντρο απόφασης.

Τα στοιχεία (υποδείγματα) ταξινομούνται με το να διατάσσονται, ξεκινώντας από τη ρίζα, σε κάποιο κόμβο-φύλλο. Ο κάθε κόμβος αναφέρεται στην εξέταση ενός συγκεκριμένου χαρακτηριστικού (attribute) του υποδείγματος και κάθε κλάδος που ακολουθεί, αναφέρεται σε μια από τις πιθανές τιμές για αυτό το υπόδειγμα. Κάθε διαδρομή από τον κόμβο-ρίζα σε κάποιο φύλλο, αποτελεί μια ένωση χαρακτηριστικών. Σε κάθε κόμβο, η εξέταση του συγκεκριμένου χαρακτηριστικού γίνεται με βάση μια σταθερά, ή με βάση τις τιμές δυο χαρακτηριστικών, ή μέσω μιας συνάρτησης ενός ή περισσότερων χαρακτηριστικών. Κάθε φύλλο δίνει στα στοιχεία που καταλήγουν σε αυτό μια ταξινόμηση ή ένα σύνολο ταξινομήσεων ή μια κατανομή πιθανότητας.

Εάν τα χαρακτηριστικά τα οποία εξετάζουμε είναι ονομαστικά, τότε ο αριθμός των κλάδων που εξέρχονται από κάθε κόμβο ισούται συνήθως με τον αριθμό των διακριτών τιμών που μπορεί να πάρει το χαρακτηριστικό που εξετάζεται, άρα το συγκεκριμένο χαρακτηριστικό εξετάζεται μια μόνο φορά. Μπορούμε εναλλακτικά να διαιρέσουμε τις πιθανές τιμές των χαρακτηριστικών σε δύο, συνήθως, υποσύνολα οπότε τα κλαδιά θα ελέγχουν σε ποιο από αυτά τα σύνολα ανήκει η συγκεκριμένη τιμή του χαρακτηριστικού. Σε αυτήν την περίπτωση, το κάθε χαρακτηριστικό μπορεί να ελέγχεται περισσότερες από μια φορές. Εάν τα χαρακτηριστικά που εξετάζουμε είναι αριθμητικά, τότε η τιμή του χαρακτηριστικού συγκρίνεται με μια σταθερά, αν πρόκειται για ακέραιο αριθμό, ή συγκρίνεται με ένα σύνολο τιμών, αν πρόκειται για πραγματικό αριθμό. Επομένως ένα χαρακτηριστικό μπορεί να ελεγχθεί πάνω από μια φορά. Εναλλακτικά, μπορούμε να διαιρέσουμε τις πιθανές τιμές σε τρία ή περισσότερα υποσύνολα και να συγκρίνουμε την τιμή του χαρακτηριστικού με αυτά τα υποσύνολα.

Για να ταξινομήσουμε ένα άγνωστο στοιχείο, ακολουθούμε την πορεία από τη ρίζα προς τους κόμβους, εξετάζοντας τις τιμές των χαρακτηριστικών του σε κάθε έναν από αυτούς, μέχρι να καταλήξουμε σε ένα φύλλο, οπότε και το στοιχείο ταξινομείται σύμφωνα με την τάξη αυτού του φύλλου. Σε περίπτωση που το υπό έλεγχο χαρακτηριστικό παρουσιάζει ελλείπουσες τιμές (missing values), τότε έχουμε τις εξής επιλογές: Μπορούμε να καταχωρήσουμε την άγνωστη τιμή σαν μια ξεχωριστή τιμή του χαρακτηριστικού και αντίστοιχα να δημιουργήσουμε έναν επιπλέον κλάδο στον κόμβο. Μια άλλη επιλογή είναι να εκχωρήσουμε το υπόδειγμα στον κλάδο με την μεγαλύτερη συχνότητα. Εναλλακτικά μπορούμε να διαμελίσουμε το υπόδειγμα και να εκχωρήσουμε τα κομμάτια σε κάθε κλάδο, με στάθμιση ανάλογη της συχνότητας του κλάδου. Στη συνέχεια αθροίζουμε τις καταληκτικές υποδείξεις από κάθε κλάδο με την ίδια στάθμιση.

Το δέντρο που κατασκευάζουμε θέλουμε να είναι όσο το δυνατόν μικρότερο, να έχουμε δηλαδή το μέγιστο κέρδος πληροφορίας από κάθε βήμα. Το κέρδος πληροφορίας αυξάνεται με την αύξηση της μέσης ομοιογένειας των υποσυνόλων. Επομένως, για να διαλέξουμε το χαρακτηριστικό-ρίζα του δέντρου, θα πρέπει να βρούμε ποιο είναι το χαρακτηριστικό εκείνο το οποίο δημιουργεί τα περισσότερο ομοιογενή υποσύνολα κλάδων και άρα μας εξασφαλίζει το μέγιστο κέρδος

πληροφορίας. Για να ποσοτικοποιήσουμε το κέρδος πληροφορίας, χρησιμοποιούμε την έννοια της εντροπίας. Υπάρχουν και άλλες συναρτήσεις που μπορούμε να χρησιμοποιήσουμε σαν κριτήρια διάσπασης, όπως το λόγο κέρδους πληροφορίας (gain ratio) ή το δείκτη Gini.

Για την κατασκευή δέντρων απόφασης κινούμαστε από πάνω προς τα κάτω (top down) και χρησιμοποιούμε επαναληπτικά τη μέθοδο του «διαίρει και βασίλευε». Συνοπτικά η διαδικασία κατασκευής είναι η εξής:

- Επιλέγουμε το χαρακτηριστικό για τον αρχικό κόμβο – ρίζα του δέντρου και δημιουργούμε κλάδους για κάθε πιθανή τιμή αυτού του χαρακτηριστικού.

- Σπάμε τα υποδείγματα σε υποσύνολα - ένα για κάθε κλάδο που εξέρχεται από τη ρίζα.

- Επαναλαμβάνουμε τα παραπάνω βήματα σε κάθε κλάδο ξεχωριστά, χρησιμοποιώντας μόνο τα υποδείγματα από το υποσύνολο του κλάδου στον οποίο δουλεύουμε κάθε φορά.

- Όταν όλα τα υποδείγματα ενός κόμβου ανήκουν στην ίδια τάξη, τότε σταματάμε τη διαδικασία.

1.2.2 Αλγόριθμοι C&RT (Classification and Regression Tree)

Ένας C&RT αλγόριθμος αποτελεί μια επαναληπτική διαδικασία που διαχωρίζει τα δεδομένα σε δυο υποσύνολα ώστε τα δεδομένα μέσα σε κάθε ένα από τα υποσύνολα να είναι περισσότερο ομοιογενή από όσο ήταν στο αρχικό σύνολο. Στη συνέχεια η διαδικασία επαναλαμβάνεται σε κάθε ένα από τα υποσύνολα έως ότου το κριτήριο ομοιογένειας ή κάποιο άλλο κριτήριο διακοπής επιτευχθεί.

Οι C&RT αλγόριθμοι είναι βασισμένοι στη θεωρία των δέντρων ταξινόμησης/παλινδρόμησης και προσφέρουν αρκετή ευελιξία για τους εξής λόγους: Το ίδιο πεδίο πρόβλεψης μπορεί να χρησιμοποιηθεί σε διαφορετικά επίπεδα στο ίδιο δέντρο. Χρησιμοποιούν υποκατάστατο διαχωρισμό για την καλύτερη χρήση των δεδομένων με ελλείπουσες τιμές. Τα άνισα κόστη λανθασμένης ταξινόμησης μπορούν να θεωρηθούν μέσα στη διαδικασία κατασκευής του δέντρου. Επιτρέπουν τον καθορισμό priors (εκ των προτέρων πιθανότητες) σε προβλήματα ταξινόμησης. Στατιστικά πακέτα δίνουν τη δυνατότητα στο χρήστη να εφαρμόσει αυτόματο κλάδεμα για να παράγει ένα πιο γενικευμένο δέντρο.

1.2.3 Κατασκευή C&RT δέντρων

1.2.3.1 Αρχικοί υπολογισμοί – πεδία συχνότητας και συντελεστές βάρους:

Σαν πεδίο συχνότητας ορίζουμε το πλήθος των παρατηρήσεων που αντιπροσωπεύονται από κάθε καταχώρηση. Μέσω των πεδίων συχνότητας μπορούμε να κατηγοριοποιήσουμε τα δεδομένα μας και έτσι να κατασκευάσουμε πίνακες δεδομένων με λιγότερες παρατηρήσεις. Το τελικό αποτέλεσμα της ανάλυσής μας θα πρέπει να παραμένει το ίδιο είτε χρησιμοποιήσουμε πεδία συχνότητας είτε όλα τα δεδομένα case by case. Το άθροισμα των τιμών για ένα πεδίο συχνότητας πρέπει να είναι ίσο με τον συνολικό αριθμό των παρατηρήσεων στο δείγμα. Για ευκολία στους υπολογισμούς, μπορούμε να ορίσουμε συντελεστές βάρους. Σαν συντελεστή βάρους ονομάζουμε την τιμή μέσω της οποίας σταθμίζουμε την συνεισφορά μιας καταχώρησης στην ανάλυσή μας σε αναλογία με το πλήθος των στοιχείων του δείγματος που η καταχώρηση αυτή αναπαριστά.

Για παράδειγμα: Σε μια προώθηση Marketing, 10.000 νοικοκυριά ανταποκρίνονται και 1.000.000 νοικοκυριά όχι. Για να μειώσουμε το μέγεθος των δεδομένων μπορούμε να συμπεριλάβουμε όλα τα δείγματα που ανταποκρίνονται (10.000 δείγματα) με συντελεστή βάρους 1 και μόνο το 1% των δειγμάτων που δεν ανταποκρίνονται (10.000 δείγματα) με συντελεστή βάρους 100.

1.2.3.2 Αλγόριθμος κατασκευής

Ο αλγόριθμος C&RT διαχωρίζει κάθε κόμβο με τέτοιο τρόπο ώστε ο θυγατρικός κόμβος που προκύπτει κάθε φορά να είναι πιο καθαρός (με βάση ένα μέτρο καθαρότητας) σε σχέση με τον μητρικό κόμβο. Σε έναν καθαρό κόμβο, όλες οι καταχωρήσεις έχουν την ίδια τιμή στο πεδίο-στόχο.

Ξεκινώντας από τον αρχικό κόμβο-ρίζα, τα βήματα είναι τα ακόλουθα:

0. Για κάθε πεδίο πρόβλεψης, βρες τον καλύτερο διαχωρισμό ανάλογα το είδος του.
    - Για αριθμητικά πεδία: Ταξινόμησε τις τιμές των πεδίων στον κόμβο από την μικρότερη στη μεγαλύτερη. Διάλεξε κάθε σημείο με τη σειρά σαν σημείο διαχωρισμού και υπολόγισε το στατιστικό μη-καθαρότητας για τους θυγατρικού κόμβους που προκύπτουν από τον διαχωρισμό. Επέλεξε σαν καλύτερο σημείο διαχωρισμού για τον κόμβο αυτό που μειώνει περισσότερο την μη-καθαρότητα σε σχέση με τον προηγούμενο κόμβο από τον οποίο προέκυψε.
    - Για κατηγορικά πεδία: Εξέτασε κάθε πιθανό διαχωρισμό των τιμών σε δυο υποσύνολα. Για κάθε πιθανό διαχωρισμό, υπολόγισε το στατιστικό μη-καθαρότητας των θυγατρικών κόμβων που προκύπτουν. Επέλεξε σαν καλύτερο σημείο διαχωρισμού για το πεδίο, αυτό το οποίο μειώνει

περισσότερο τη μη-καθαρότητα σε σύγκριση με τον κόμβο από τον οποίο προήλθε.

1. Βρες τον καλύτερο διαχωρισμό για τον κόμβο. Προσδιόρισε το πεδίο του οποίου ο διαχωρισμός δίνει την μεγαλύτερη μείωση στη μη-καθαρότητα για τον κόμβο και επέλεξε αυτόν τον διαχωρισμό συνολικά για τον κόμβο.

2. Έλεγξε αν ικανοποιείται κάποιο κριτήριο διακοπής. Εάν ο αρχικός κόμβος ή ο διαχωρισμός δεν ικανοποιούν κάποιο κριτήριο διακοπής, τότε επανέλαβε τον διαχωρισμό για να δημιουργηθούν δυο θυγατρικοί κόμβοι και επανέλαβε τον αλγόριθμο σε κάθε έναν από αυτούς.

1.2.4 Κενά – ελλείπουσες τιμές

Σε περίπτωση που το πεδίο πρόβλεψης που χρησιμοποιείται για διαχωρισμό παρουσιάζει κάποιο κενό ή μια ελλείπουσα τιμή σε κάποιο συγκεκριμένο κόμβο, τότε ένα άλλο πεδίο το οποίο αποδίδει διαχωρισμό παρόμοιο στα πλαίσια του συγκεκριμένου κόμβου χρησιμοποιείται στη θέση του πεδίου με το κενό ή την ελλείπουσα τιμή. Η τιμή του πεδίου-αντικαταστάτη χρησιμοποιείται για να εκχωρήσει την εγγραφή σε έναν από τους θυγατρικούς κόμβους.

Για λόγους ταχύτητας και εξοικονόμηση μνήμης, μόνο ένας περιορισμένος αριθμός από πεδία-αντικαταστάτες ορίζεται για κάθε διαχωρισμό που πραγματοποιείται στο δέντρο. Εάν μια καταχώρηση παρουσιάζει ελλείπουσες τιμές στο πεδίο διαχωρισμού και σε όλα τα πεδία-αντικαταστάτες, τότε χρησιμοποιείται ο θυγατρικός κόμβος με την υψηλότερη σταθμισμένη πιθανότητα. Η σταθμισμένη πιθανότητα υπολογίζεται από τον τύπο:

$$\frac{N_{f,j}(t)}{N_f(t)}$$

Το $N_{f,j}(t)$ είναι το άθροισμα από τα βάρη συχνότητας (frequency weights) για τις καταχωρήσεις της κατηγορίας j για τον κόμβο *t.*

Το $N_f(t)$ είναι το άθροισμα από τα βάρη συχνότητας για όλες τις καταχωρήσεις στον κόμβο *t.*

Σε περίπτωση που το μοντέλο κατασκευάστηκε χρησιμοποιώντας ίσα ή καθορισμένα από τον χρήστη priors, τότε αυτά ενσωματώνονται μέσα στον υπολογισμό ως εξής:

$$\frac{\pi(j)}{p_f(t)} \times \frac{N_{f,j}(t)}{N_f(t)},$$

όπου $\pi(j)$ είναι η prior πιθανότητα για την κατηγορία j και $p_f(t)$ είναι η σταθμισμένη πιθανότητα μιας εγγραφής που εκχωρείται στον κόμβο με

$$p_f(t) = \sum_j \frac{\pi(j) N_{f,j}(t)}{N_{f,j}}.$$

$N_{f,j}(t)$ είναι το άθροισμα από τα βάρη συχνότητας στον κόμβο t που ανήκουν στην κατηγορία j. Σε περίπτωση που δεν ορίζονται βάρη συχνότητας είναι ο αριθμός των καταχωρήσεων.

$N_{f,j}$ είναι το άθροισμα από τα βάρη συχνότητας των καταχωρήσεων που ανήκουν σε κατηγορία μέσα σε ολόκληρο το δείγμα εκπαίδευσης.

Οι καταχωρήσεις που παρουσιάζουν ελλείπουσες τιμές στο πεδίο-στόχο αγνοούνται κατά την κατασκευή του δέντρου. Κατά την ταξινόμηση νέων καταχωρήσεων, τα κενά διαχειρίζονται με τον ίδιο τρόπο όπως και κατά την κατασκευή του δέντρου. Με χρήση, δηλαδή, πεδίων-αντικαταστατών (όπου είναι εφικτό) και διαχωρισμό με βάση τις σταθμισμένες πιθανότητες, αν χρειάζεται.

1.2.5 Μη-καθαρότητα

Ανάλογα με το είδος του πεδίου-στόχου, μπορούμε να χρησιμοποιήσουμε τρία διαφορετικά μέτρα για τη μη-καθαρότητα στα C&RT μοντέλα. Στην περίπτωση που έχουμε συμβολικά πεδία-στόχους, μπορούμε να χρησιμοποιήσουμε τον δείκτη Gini ή τον δείκτη Twoing. Για συνεχή πεδία-στόχους, χρησιμοποιούμε τη μέθοδο Απόκλισης Ελάχιστων Τετραγώνων - LSD (Least Squared Deviation).

1.2.5.1 Ο δείκτης Gini

Ο δείκτης Gini $g(t)$ σε έναν κόμβο t ενός δέντρου C&RT, ορίζεται ως

$$g(t) = \sum_{j \neq i} p(j/t) p(i/t)$$

ή

$$g(t) = 1 - \sum_j p^2(j/t)$$

όπου i και j είναι κατηγορίες στο πεδίο-στόχο και

$$p(j/t) = \frac{p(j,t)}{p(t)}$$

$$p(j,t) = \frac{\pi(j) N_j(t)}{N_j}$$

$$P(t) = \sum_j p(j,t)$$

$\pi(j)$ είναι η τιμή της πιθανότητας prior για την κατηγορία j, $N_j(t)$ είναι ο αριθμός των καταχωρήσεων στην κατηγορία j του κόμβου t και N_j είναι ο αριθμός των καταχωρήσεων της κατηγορίας j στον αρχικό κόμβο-ρίζα.

Σε περίπτωση που χρησιμοποιούμε τον δείκτη Gini για να βρούμε τη βελτίωση ενός διαχωρισμού κατά την ανάπτυξη ενός δέντρου, χρησιμοποιούμε μόνο τις καταχωρήσεις στον κόμβο t και στον αρχικό κόμβο-ρίζα που έχουν έγκυρες τιμές για το πεδίο διαχωρισμού για να υπολογίσουμε το $N_j(t)$ και το N_j αντίστοιχα.

Όταν οι καταχωρίσεις σε έναν κόμβο διανέμονται ομαλά δια μέσου των κατηγοριών, ο δείκτης Gini παίρνει την μεγαλύτερη τιμή του $1 - \frac{1}{k}$, με k να είναι ο αριθμός των κατηγοριών για το πεδίο-στόχο. Αν όλες οι καταχωρήσεις σε έναν κόμβο ανήκουν στην ίδια κατηγορία, τότε ο δείκτης Gini ισούται με $g(t) = 0$.

Η συνάρτηση του κριτηρίου Gini για τον διαχωρισμό s στον κόμβο t, ορίζεται ως:

$$\Phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R),$$

με p_L να είναι η σύνολο των καταχωρήσεων στον κόμβο t που στέλνονται στον αριστερό θυγατρικό κόμβο και αντίστοιχα p_R το σύνολο των καταχωρήσεων που στέλνονται στον δεξιό θυγατρικό κόμβο.

Ορίζουμε

$$p_L = \frac{p(t_L)}{p(t)}$$

και

$$p_R = \frac{p(t_R)}{p(t)}$$

και επιλέγουμε τον διαχωρισμό s για να μεγιστοποιήσουμε την τιμή της $\Phi(s,t)$.

1.2.5.2 Ο δείκτης Twoing

Ο δείκτης Twoing βασίζεται στο διαχωρισμό των κατηγοριών-στόχους σε δυο υπερκλάσεις και στην εύρεση του καλύτερου διαχωρισμού στο πεδίο πρόβλεψης με βάση αυτές τις υπερκλάσεις.

Ορίζουμε τις υπερκλάσεις C_1, C_2 ως:

$$C_1 = \{ j : p(j / t_L) \geq p(j / t_R \}$$

$$C_2 = C - C_1$$

με C να είναι το σύνολο των κατηγοριών του πεδίου-στόχου και $p(j / t_L)$, $p(j / t_R)$ είναι όμοιο με το $p(j / t)$ που ορίζεται για τον δείκτη Gini, για τους αριστερούς και τους δεξιούς θυγατρικούς κόμβους αντίστοιχα.

Ορίζεται η συνάρτηση του κριτήριου Twoing για τον διαχωρισμό s στον κόμβο t ως εξής:

$$\Phi(s,t) = p_L p_R \left[\sum_j \left| p(j / t_L) - p(j / t_R) \right| \right]^2$$

με t_L, t_R να είναι οι κόμβοι που δημιουργούνται από τον διαχωρισμό s ο οποίος επιλέγεται για να μεγιστοποιεί το παραπάνω κριτήριο.

1.2.5.3 Απόκλιση ελαχίστων τετραγώνων (LSD)

Το μέτρο LSD για τη μη-καθαρότητα χρησιμοποιείται όταν έχουμε συνεχή πεδία-στόχους. Συμβολίζεται με $R(t)$ και αποτελεί την σταθμισμένη διακύμανση μέσα στον κόμβο t:

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i \left(y_i - \overline{y(t)} \right)^2$$

όπου το $N_w(t)$ αποτελεί τον σταθμισμένο αριθμό των καταχωρήσεων στον κόμβο t, το w_i είναι η τιμή του σταθμισμένου πεδίου για μια οποιαδήποτε καταχώρηση i, το f_i είναι η τιμή του πεδίου συχνότητας, y_i είναι η τιμή του πεδίου-στόχου και $y(t)$ είναι ο σταθμισμένος μέσος για τον κόμβο t.

Η συνάρτηση του κριτηρίου LSD για τον διαχωρισμό στον κόμβο t είναι:

$\Phi(s,t) = R(t) - p_L R(t_L) - p_R R(t_R)$. Ο διαχωρισμός s επιλέγεται έτσι ώστε να μεγιστοποιείται η συνάρτηση $\Phi(s,t)$.

1.2.6 Κανόνες τερματισμού διαδικασίας

Η διαδικασία διαχωρισμού των κόμβων στο δέντρο σταματάει με βάση κάποιους κανόνες. Όταν ικανοποιηθεί τουλάχιστον ένας από τους παρακάτω κανόνες τερματισμού για κάθε κόμβο του δέντρου, τότε ο αλγόριθμός θα σταματήσει:

- Όλες οι καταχωρήσεις έχουν την ίδια τιμή για το πεδίο-στόχο (ο κόμβος είναι καθαρός).

- Όλες οι καταχωρήσεις στον κόμβο έχουν την ίδια τιμή για όλα τα πεδία πρόβλεψης τα οποία χρησιμοποιούνται από το μοντέλο.

- Το βάθος του δέντρου για τον τρέχον κόμβο είναι το μέγιστο βάθος που έχει προκαθοριστεί . Ο τρέχον κόμβος καθορίζεται από τον αριθμό των διαδοχικών διαχωρισμών κόμβων.

- Ο αριθμός των καταχωρήσεων στον κόμβο είναι μικρότερος από το ελάχιστο μέγεθος του μητρικού κόμβου, όπως αυτό έχει προκαθοριστεί.

- Ο αριθμός των καταχωρήσεων σε οποιονδήποτε από τους θυγατρικούς κόμβους, που προκύπτουν από τον καλύτερο διαχωρισμό κόμβου, είναι μικρότερος από το ελάχιστο μέγεθος του θυγατρικού κόμβου, όπως έχει προκαθοριστεί.

- Ο καλύτερος διαχωρισμός για τον κόμβο αποδίδει μια μείωση στην μη-καθαρότητα η οποία είναι μικρότερη από την ελάχιστη μεταβολή μη-καθαρότητας όπως αυτή έχει προκαθοριστεί.

1.2.7 Κέρδη – κόστη

Τα κέρδη (profits) αποτελούν αριθμητικές τιμές, σχετικές με τις κατηγορίες ενός συμβολικού πεδίου-στόχου. Χρησιμοποιούνται για να εκτιμήσουν το κέρδος ή την απώλεια που σχετίζεται με ένα τμήμα και καθορίζουν τη σχετική τιμή κάθε τιμής του πεδίου-στόχου. Οι τιμές χρησιμοποιούνται για να υπολογιστούν τα κέρδη, αλλά δεν χρησιμοποιούνται κατά τη διαδικασία κατασκευής του δέντρου.

Υπολογίζουμε το κέρδος ενός κόμβου στο δέντρο από τον τύπο:

$$\sum_j f_j(t)P_j$$

όπου j είναι η κατηγορία του πεδίου-στόχου, $f_j(t)$ το άθροισμα των τιμών των πεδίων συχνότητας για όλες τις καταχωρήσεις στον κόμβο t με κατηγορία j για το πεδίο-στόχο και P_j είναι η προκαθορισμένη τιμή κέρδους για την κατηγορία j.

Σε περίπτωση που καθορίζονται τα κόστη, ισχύουν τα παρακάτω για τα μέτρα μη-καθαρότητας:

- Για τον δείκτη Gini, αν καθορίζονται κόστη, αυτός υπολογίζεται ως εξής:
$$g(t) = \sum_{j \neq i} C(i/j)\, p(j/t)\, p(i/t),$$
με το $C(i/j)$ να προσδιορίζει το κόστος της λανθασμένης ταξινόμησης μιας κατηγορίας j σαν κατηγορία i.

- Για το κριτήριο Twoing, αν καθορίζονται κόστη, αυτά δεν χρησιμοποιούνται στον διαχωρισμό των κόμβων. Τα κόστη θα ενσωματωθούν στην εκχώρηση κόμβου και στην εκτίμηση του ρίσκου (όπως αυτό περιγράφεται στη συνέχεια).

- Τα κόστη δεν εφαρμόζονται στα δέντρα παλινδρόμησης, άρα δεν χρησιμοποιούνται στην περίπτωση που εφαρμόζουμε την LSD μέθοδο.

1.2.8 Εκ των πρότερων πιθανότητες (priors)

Οι εκ των προτέρων πιθανότητες (priors), είναι αριθμητικές τιμές που επηρεάζουν τα ποσοστά λανθασμένης ταξινόμησης για τις κατηγορίες του πεδίου-στόχου. Καθορίζουν την αναλογία των καταχωρήσεων που αναμένονται να ανήκουν σε κάθε κατηγορία του πεδίου-στόχου πριν από την ανάλυση. Οι priors εμπλέκονται στη διαδικασία κατασκευής του δέντρου και στην εκτίμηση του ρίσκου. Υπάρχουν τρεις είδη priors που χρησιμοποιούνται συνήθως:

1. **Εμπειρικές priors**

Οι priors υπολογίζονται με βάση τα δεδομένα εκπαίδευσης που διαθέτουμε. Σε κάθε κατηγορία αντιστοιχεί μια prior πιθανότητα η οποία είναι η σταθμισμένη αναλογία των εγγραφών στα δεδομένα εκπαίδευσης τα οποία ανήκουν σε αυτή την κατηγορία. Υπολογίζονται από τον τύπο:

$$\pi(j) = \frac{N_{w,j}}{N_w}.$$

Κατά τη διαδικασία κατασκευής του δέντρου και της ανάθεσης κατηγοριών, το N_s λαμβάνει υπ' όψη τα βάρη συχνότητας, σε περίπτωση που αυτά καθορίζονται. Στην

εκτίμηση του ρίσκου, συμπεριλαμβάνονται μόνο τα βάρη συχνότητας για τον υπολογισμό των εμπειρικών priors.

2. **Ίσες priors**

Μπορούμε να ορίσουμε την prior πιθανότητα για κάθε μια από τις J κατηγορίες στην ίδια τιμή:

$$\pi(j) = \frac{1}{J}.$$

3. **Προκαθορισμένες priors**

Μπορούμε να χρησιμοποιήσουμε προκαθορισμένες priors για κάθε μια κατηγορία. Θα πρέπει ωστόσο αυτές να συμμορφώνονται με τον περιορισμό ότι το άθροισμα όλων των priors πρέπει να ισούται με την μονάδα. Σε περίπτωση που δεν ικανοποιούν τον παραπάνω περιορισμό, υπάρχει τρόπος να παράγουμε προσαρμοσμένες priors οι οποίες να διατηρούν της αναλογίες των προκαθορισμένων priors και επιπλέον ικανοποιούν τον περιορισμό. Αυτό επιτυγχάνεται χρησιμοποιώντας τον τύπο:

$$\pi'(j) = \frac{\pi(j)}{\sum_j \pi(j)}$$

όπου $\pi'(j)$ είναι η προσαρμοσμένη prior για την κατηγορία j, η βασισμένη στην προκαθορισμένη prior $\pi(j)$ για την κατηγορία j.

1.2.9 Η διαδικασία κλαδέματος (pruning)

Σαν κλάδεμα (pruning) αναφέρεται η διαδικασία κατά την οποία ένα αναπτυσσόμενο δέντρο εξετάζεται για να αφαιρεθούν οι διαχωρισμοί των κάτω επιπέδων οι οποίοι δεν συνεισφέρουν σημαντικά στην ακρίβεια του δέντρου.

Διάφορα λογισμικά (όπως για παράδειγμα το πακέτο Clementine για το SPSS) προσπαθούν να δημιουργήσουν το μικρότερο δυνατό δέντρο του οποίου το ρίσκο λανθασμένης ταξινόμησης δεν είναι πολύ μεγαλύτερο από το ρίσκο λανθασμένης ταξινόμησης του μεγαλύτερου πιθανού δέντρου. Ένα κλαδί του δέντρου αφαιρείται εάν το κόστος το οποίο σχετίζεται με την μεγαλύτερη πολυπλοκότητα του δέντρου υπερβαίνει το κέρδος το οποίο σχετίζεται με τον να έχουμε ένα άλλο επίπεδο κόμβων (ένα κλαδί). Χρησιμοποιεί ένα δείκτη οποίος μετράει το ρίσκο λανθασμένης

ταξινόμησης και την πολυπλοκότητα του δέντρου, εφόσον θέλουμε να τα ελαχιστοποιήσουμε και τα δύο. Το μέτρο πολυπλοκότητας (cost-complexity) ορίζεται ως:

$$R_a(T) = R(T) + a\left|T\right|$$

όπου $R(T)$ είναι το ρίσκο λανθασμένης ταξινόμησης του δέντρου Τ και $\left|T\right|$ είναι ο αριθμός των τερματικών κόμβων για το δέντρο Τ. Ο όρος a αντιπροσωπεύει το κόστος πολυπλοκότητας ανά τερματικό κόμβο για το δέντρο. Η τιμή του όρου a υπολογίζεται από τον αλγόριθμο κατά τη διάρκεια του κλαδέματος.

Για οποιοδήποτε δέντρο μπορούμε να παράγουμε, υπάρχει ένα μέγιστο μέγεθος $(T_{\max})$ στο οποίο κάθε τερματικός κόμβος περιέχει μόνο μια καταχώρηση. Χωρίς κόστος πολυπλοκότητας, για $a = 0$, το μέγιστο δέντρο έχει το χαμηλότερο ρίσκο, από τη στιγμή που κάθε εγγραφή προβλέπεται τέλεια. Συνεπώς, όσο μεγαλύτερη η τιμή του a, τόσο μικρότερος ο αριθμός των τερματικών κόμβων στο $T(a)$, με $T(a)$ να είναι το δέντρο με το μικρότερο κόστος πολυπλοκότητας για το δοσμένο a. Όσο η ποσότητα a αυξάνεται από το μηδέν, παράγει μια πεπερασμένη ακολουθία από υπό-δέντρα $T_1, T_2, \ldots$ το καθένα με λιγότερους τερματικούς κόμβους διαδοχικά. Το κλάδεμα του κόστους-πολυπλοκότητας δουλεύει αφαιρώντας τον πιο αδύναμο διαχωρισμό.

Συμβολίζοντας με $\{t\}$ έναν οποιονδήποτε ξεχωριστό κόμβο και με T_t τον υπό-κλάδο του $\{t\}$, μπορούμε να αναπαραστήσουμε το κόστος πολυπλοκότητας για το $\{t\}$ με τις παρακάτω εξισώσεις:

$$R_a(\{t\}) = R(t) + a$$

και

$$R_a(T_t) = R(T_t) + a\left|T_t\right|.$$

Σε περίπτωση που το $R_a(T_t)$ είναι μικρότερο από το $R_a(\{t\})$, τότε το κλαδί T_t έχει μικρότερο κόστος πολυπλοκότητας από αυτό του ξεχωριστού κόμβο $\{t\}$. Η διαδικασία ανάπτυξης του δέντρου εξασφαλίζει ότι $R_a(\{t\}) \geq R_a(T_t)$ για $a = 0$. Καθώς το a αυξάνει από το μηδέν, το $R_a(\{t\})$ και το $R_a(T_t)$ αυξάνονται γραμμικά, με το τελευταίο να αυξάνει με ταχύτερο ρυθμό. Τελικά, θα πετύχουμε ένα όριο a' τέτοιο ώστε $R_a(\{t\}) < R_a(T_t)$ για όλα τα $a > a'$. Επομένως, όταν το a γίνεται μεγαλύτερο του a', το κόστος πολυπλοκότητας του δέντρου μπορεί να μειωθεί εάν κόψουμε τον υπό-κλάδο T_t κάτω από το $\{t\}$.

Για να υπολογίσουμε το όριο a' μπορούμε να λύσουμε την ανισότητα $R_a(\{t\}) \geq R_a(T_t)$ για να βρούμε την μεγαλύτερη τιμή του a για την οποία ισχύει η ανισότητα, την οποία μπορούμε να απεικονίσουμε ως $g(t)$. Η ανισότητα είναι η:

$$a \leq g(t) = \frac{R(t) - R(T_t)}{|T_t| - 1}.$$

Σαν τον πιο αδύναμο σύνδεσμο (t) στο δέντρο Τα, μπορούμε να ορίσουμε τον κόμβο ο οποίος παίρνει την μικρότερη τιμή του $g(t)$:

$$g(\bar{t}) = \min_{t \in T} g(t).$$

Επομένως, όσο αυξάνει το a, ο κόμβος $\bar{t}$ είναι ο πρώτος κόμβος για τον οποίο ισχύει $R_a(\{t\}) = R_a(T_t)$. Σε αυτό το σημείο, προτιμάται το $\{t\}$ από το $T_{\bar{t}}$ και ο υπό-κλάδος κλαδεύεται.

Συνολικά, ο αλγόριθμος του κλαδέματος μπορεί να συνοψιστεί στα ακόλουθα βήματα:

- Ορίζουμε $a_1 = 0$ και ξεκινάμε από το πλήρως αναπτυγμένο δέντρο $T_1 = T(0)$.
- Αυξάνουμε το a μέχρι το κλάδεμα ενός κλαδιού. Υπολογίζουμε την εκτίμηση ρίσκου για το κλαδεμένο δέντρο.
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να απομείνει μόνο ο αρχικός κόμβος-ρίζα. Λαμβάνουμε έτσι μια σειρά από δέντρα $T_1, T_2 \ldots, T_k$.
- Στην περίπτωση που θέλουμε να ισχύει ο κανόνας του τυπικού σφάλματος, διαλέγουμε το μικρότερο δέντρο T_{opt} για το οποίο ισχύει $R(T_{opt}) \leq \min_k R(T_k) + m \times SE(R(T))$.
- Στην περίπτωση που δεν θέλουμε να ισχύει ο κανόνας του τυπικού σφάλματος, τότε επιλέγουμε το δέντρο με την μικρότερη εκτίμηση ρίσκου $R(T)$.

1.2.10 Δευτερεύοντες υπολογισμοί

Μπορούμε να προχωρήσουμε σε δευτερεύοντες υπολογισμούς οι οποίοι δεν σχετίζονται άμεσα με την κατασκευή του δέντρου, αλλά προσφέρουν πληροφορίες για το μοντέλο και για την αποδοτικότητά του.

1.2.10.1 Εκτίμηση ρίσκου

Η εκτίμηση ρίσκου περιγράφει το ρίσκο να υπάρξει σφάλμα στις προβλεπόμενες τιμές σε συγκεκριμένους κόμβους του δέντρου ή σε ολόκληρο το δέντρο.

α) Στα δέντρα ταξινόμησης με συμβολικά πεδία-στόχους, η εκτίμηση ρίσκου $r(t)$ του κόμβου t υπολογίζεται ως:

$$r(t) = \frac{1}{N_f} \sum_j N_{f,j}(t) C(j^*(t) / j)$$

,με $C(j^*(t) / j)$ να είναι το κόστος του να ταξινομηθεί μια καταχώρησης που έχει τιμή στόχο j σαν $j^*(t)$. Το $N_{f,j}(t)$ είναι το άθροισμα των βαρών συχνότητας (frequency weights) για τις καταχωρήσεις στον κόμβο t στην κατηγορία j, ή ο αριθμός των εγγραφών αν τα βάρη συχνότητας δεν προσδιορίζονται. Το N_f είναι το άθροισμα των βαρών συχνότητας για όλες τις καταχωρήσεις στα δεδομένα εκπαίδευσης.

β) Εάν το μοντέλο χρησιμοποιεί προκαθορισμένες priors, η εκτίμηση του ρίσκου γίνεται μέσω του τύπου:

$$r(t) = \sum_j \frac{\pi(j) N_{f,j}(t)}{N_{f,j}} C(j^*(t) / j).$$

Τα βάρη δεν χρησιμοποιούνται στον υπολογισμό των εκτιμήσεων ρίσκου.

Για δέντρα παλινδρόμησης με αριθμητικά πεδία στόχους, η εκτίμηση του ρίσκου $r(t)$ του κόμβου t υπολογίζεται ως:

$$r(t) = \frac{1}{N_f(t)} \sum_{i \in t} \left(f_i y_i - \bar{y}(t) \right)^2$$

,όπου f είναι το βάρος συχνότητας για την εγγραφή i, την ανατεθειμένη στον κόμβο t, το y_i είναι η τιμή του πεδίου-στόχου για την καταχώρηση i και $\bar{y}(t)$ είναι ο σταθμισμένος μέσος του πεδίου-στόχου για όλες τις καταχωρήσεις στον κόμβο t.

γ) Η εκτίμηση του ρίσκου $R(t)$ για ένα δέντρο T (ταξινόμησης ή παλινδρόμησης), υπολογίζεται παίρνοντας το άθροισμα των εκτιμήσεων ρίσκου για τους τερματικούς κόμβους $r(t)$. Δηλαδή:

$$R(t) = \sum_{t \in T'} r(t)$$

,με T' να είναι το σύνολο των τερματικών κόμβων στο δέντρο.

1.2.10.2 Συνόψιση κέρδους (gain)

Μέσω της συνόψισης κέρδους (gain) παρέχονται περιγραφικά στατιστικά για τους τερματικούς κόμβους ενός δέντρου.

Εάν το πεδίο-στόχος είναι συνεχές, η συνόψιση κέρδους δείχνει τον σταθμισμένο μέσο της τιμής-στόχου για κάθε τερματικό κόμβο:

$$g(t) = \sum_{i \in t} w_i f_i x_i \ .$$

Εάν το πεδίο-στόχος είναι συμβολικό, η συνόψιση κέρδους δείχνει το σταθμισμένο ποσοστό των εγγραφών σε μια επιλεγμένη κατηγορία στόχου:

$$g(t) = \frac{\sum_{i \in t} f_i x_i(j)}{\sum_{i \in t} f_i} \ .$$

όπου $x_i(j) = 1$ στην περίπτωση που η εγγραφή x_i είναι στην κατηγορία j, και $x_i(j) = 0$ σε άλλη περίπτωση. Εάν τα κέρδη (profits) προσδιορίζονται για το δέντρο, η συνόψιση κέρδους (gain) είναι η μέση τιμή κέρδους για κάθε τερματικό κόμβο:

$$g(t) = \sum_{i \in t} f_i P(x_i)$$

με $P(x_i)$ να είναι η τιμή κέρδους που εκχωρείται στην τιμή στόχο που παρατηρείται στην εγγραφή x_i.

1.2.11 Παραγόμενο μοντέλο/scoring

Παρακάτω, παρουσιάζονται οι υπολογισμοί που πραγματοποιούνται από το παραγόμενο C&RT μοντέλο:

1.2.11.1 Προβλεπόμενες τιμές

Με την εισαγωγή στο γενικευμένο μοντέλο των εγγραφών, το μοντέλο δημιουργεί μια πρόβλεψη για κάθε εγγραφή. Κάθε τερματικός κόμβος ενός δέντρου έχει μια συγκεκριμένη προβλεπόμενη τιμή που σχετίζεται με το δέντρο.

Για τα δέντρα ταξινόμησης με συμβολικά πεδία-στόχους, κάθε προβλεπόμενη κατηγορία τερματικού κόμβου είναι η κατηγορία με το χαμηλότερο σταθμισμένο κόστος για τον κόμβο. Το σταθμισμένο κόστος υπολογίζεται από τον τύπο

$$\min_i \sum_j C(i/j)\, p(i/j)$$

με $C(i/j)$ να είναι το προκαθορισμένο κόστος της λανθασμένης ταξινόμησης μιας κατηγορίας j σαν κατηγορία i, και $p(j/t)$ να είναι η δεσμευμένη σταθμισμένη πιθανότητα μιας εγγραφής να είναι στην κατηγορία j, δεδομένου του να είναι στον κόμβο t, η οποία ορίζεται σαν:

$$p(j/t) = \frac{p(j,t)}{\sum_j p(j,t)}$$

με το $p(j,t)$ να ορίζεται από τον τύπο

$$p(j,t) = \pi(j)\frac{N_{w,j}(t)}{N_{w,j}}$$

όπου $\pi(j)$ είναι η prior για την κατηγορία j, το $N_{w,j}(t)$ είναι ο σταθμισμένος αριθμός των εγγραφών στον κόμβο t στην κατηγορία j ή ο αριθμός των εγγραφών αν δεν ορίζονται συχνότητες ή βάρη:

$$N_{w,j}(t) = \sum_{i \in t} w_i f_i j(i) \quad \text{και}$$

Το $N_{w,j}$ είναι ο σταθμισμένος αριθμός των εγγραφών στην κατηγορία j για οποιονδήποτε κόμβο:

$$N_{w,j} = \sum_{i \in T} w_i f_i j(i)$$

Για τα δέντρα παλινδρόμησης με αριθμητικά πεδία-στόχους, κάθε προβλεπόμενη κατηγορία τερματικού κόμβου είναι ο σταθμισμένος μέσος των τιμών στόχου για τις εγγραφές μέσα στον κόμβο. Αυτός ο σταθμισμένος μέσος υπολογίζεται από τον τύπο:

$$\bar{y}(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i y_i$$

με το $N_w(t)$ να ορίζεται ως:

$$N_w(t) = \sum_{i \in t} w_i f_i$$

1.2.12 Εμπιστοσύνη

Για τα δέντρα ταξινόμησης, οι τιμές εμπιστοσύνης για τις ταξινομημένες εγγραφές που χρησιμοποιούνται από το παραγόμενο μοντέλο υπολογίζονται ως εξής:

Η τιμή εμπιστοσύνης για μια ταξινομημένη εγγραφή, είναι η τροποποιημένη από τη διόρθωση Laplace αναλογία των σταθμισμένων εγγραφών στα δεδομένα εκπαίδευσης στον καθορισμένο τερματικό κόμβο της συγκεκριμένης εγγραφής, οι οποίες ανήκουν στην προβλεπόμενη κατηγορία:

$$\frac{N_{f,j}(t)+1}{N_f(t)+k}$$

Για τα δέντρα παλινδρόμησης δεν ορίζεται τιμή εμπιστοσύνης.

1.3 Πρόβλεψη μέσω λογιστικής παλινδρόμησης

1.3.1 Το λογιστικό μοντέλο παλινδρόμησης (Logistic Regression model)

Το λογιστικό μοντέλο παλινδρόμησης άρχισε να χρησιμοποιείται ευρέως κατά τη δεκαετία του ’50, κυρίως με εφαρμογές στη βιοστατιστική. Ανήκει στην κατηγορία των γενικευμένων γραμμικών μοντέλων. Χαρακτηριστικό αυτών των μοντέλων είναι ότι η διακύμανση της απόκρισης, είναι συνάρτηση της μέσης αναμενόμενης τιμής της. Η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί στις περιπτώσεις όπου η μεταβλητή είναι δίτιμη, με άλλα λόγια η πρόβλεψη είναι το αποτέλεσμα μιας διαδικασίας Bernoulli, όπως επιτυχία/αποτυχία ή π.χ. σε ένα πείραμα, αν το φάρμακο ενεργεί ή όχι σε έναν ασθενή.

Ας υποθέσουμε ότι έχουμε n ανεξάρτητες πειράματικες εκτελέσεις με δίτιμη απόκριση y (0 ή 1), η οποία εξαρτάται από ένα σύνολο μεταβλητών παλινδρόμησης $x_1, x_2, ..., x_k$. Οι μεταβλητές αυτές μπορεί να είναι για παράδειγμα το ύψος ή η ηλικία ενός ασθενή και η απόκριση να είναι το αν ενεργεί ένα φάρμακο ή όχι σε αυτόν. Εάν ορίσουμε την τιμή $y=1$ σαν επιτυχία και την τιμή $y=0$ σαν αποτυχία, τότε μπορούμε να μοντελοποιήσουμε την μέση απόκριση $P(x_i)$, όπου $P(x_i)$ η πιθανότητα επιτυχίας του πειράματος, και x_i να είναι οι μεταβλητές παλινδρόμησης στο $i-$οστό σημείο δεδομένων.

Το λογιστικό μοντέλο για την πιθανότητα $P(x_i)$ θα είναι:

$$P(x_i)=\frac{1}{1+e^{-x_i'\beta}} \quad (1)$$

όπου ο όρος $x_i'\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$ είναι ο linear predictor . Επιπλέον ισχύει η σχέση $0 \le P(x_i) \le 1$. Πρέπει να σημειωθεί ότι οι τιμές των παραμέτρων επηρεάζουν σημαντικά την μορφή της λογιστικής καμπύλης.

1.3.2 Εκτίμηση παραμέτρων με τη μέθοδο της μέγιστης πιθανοφάνειας

Ας υποθέσουμε ότι τα δεδομένα μας είναι χωρισμένα σε κατηγορίες. Δηλαδή, έχουμε n_i στο πλήθος πειραματικές μονάδες στο $i-$οστό σημείο δεδομένων (για παράδειγμα, μπορούμε να θεωρήσουμε ότι το n_i είναι το πλήθος των πειραματόζωων στα οποία έχουμε παρέχει μια συγκεκριμένη δοσολογία φαρμάκου). Το μοντέλο μας, βάση της εξίσωσης (1), γράφεται στη μορφή:

$$E(y_i) = n_i P(x_i) = n_i \frac{1}{1+e^{-x_i'\beta}} \quad , \quad i = 1, 2, ...m .$$

με $y_1, y_2, ..., y_m$ να είναι οι παρατηρούμενες τιμές των ανεξάρτητων διωνυμικών τυχαίων μεταβλητών. Σε αυτήν την περίπτωση ισχύει

$$\text{var}(y_i) = n_i P(x_i)[1 - P(x_i)]$$

και το άθροισμα $\sum_{i=1}^{m} n_i = n$ είναι το συνολικό πλήθος του δείγματός μας.

Η συνάρτηση πιθανότητας μιας απλής διωνυμικής τυχαίας μεταβλητής y με παραμέτρους n, P δίνεται από τον τύπο:

$$\binom{n}{y} P^y (1-P)^{n-y} .$$

Ωστόσο, ο όρος $\binom{n}{y}$ δεν περιλαμβάνει το β, οπότε δεν μπορεί να χρησιμοποιηθεί. Επομένως, η log πιθανοφάνεια για το λογιστικό μοντέλο παλινδρόμησης δίνεται από τον τύπο:

$$\ln[L(P; y)] = \sum_{i=1}^{m} \left\{ y_i \ln\left[\frac{P(x_i)}{1-P(x_i)}\right] + n_i \ln[1 - P(x_i)] \right\} \qquad (2).$$

Μπορούμε τώρα να εισάγουμε τη μορφή του λογιστικού μοντέλου στην εξίσωση (1).

Ο όρος $\ln\left[\frac{P(x_i)}{1-P(x_i)}\right]$ ονομάζεται logit και μπορεί να γραφτεί στη μορφή:

$$\ln\left[\frac{P(x_i)}{1-P(x_i)}\right] = x_i'\beta = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j\,, \qquad i = 1,2,...,m\,, \qquad m \geq k+1$$

Επομένως, η log πιθανοφάνεια της εξίσωσης (2) μπορεί να πάρει τη μορφή:

$$\ln[L(\beta; y)] = \sum_{i=1}^{m}\sum_{j=1}^{k} y_i x_{ij}\beta_j - \sum_{i=1}^{m} n_i \ln\left(1+\exp\sum_{j=1}^{k} x_{ij}\beta_j\right) \qquad (3)$$

Η εξίσωση (3) πρέπει να μεγιστοποιηθεί ως προς τον όρο β_j. Σε μορφή πινάκων, η εξίσωση (3) μπορεί να γραφτεί ως εξής:

$$\ln[L(\beta; y)] = \beta' Xy - \sum_{i=1}^{m} n_i \ln[1+\exp(x_i'\beta)] \qquad (4)$$

με X να είναι ο κλασικός πίνακας του μοντέλου που συναντάμε και στην γραμμική παλινδρόμηση και y το διάνυσμα της απόκρισης.

Παραγωγίζουμε την εξίσωση (4) ως προς β:

$$\frac{\partial \ln L(\beta_i; y)}{\partial \beta} = X'y - \sum_{i=1}^{m}\left[\frac{n_i}{1+e^{x_i'\beta}}\right] e^{x_i'\beta} x_i$$

Από τη σχέση $\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} = \frac{1}{1+e^{-x_i'\beta}} = P(x_i)$, έχουμε ότι:

$$\frac{\partial \ln L(\beta; y)}{\partial \beta} = X'y - \sum_{i=1}^{m} n_i P(x_i) x_i\,.$$

Ο όρος $n_i P(x_i)$ αποτελεί τον μέσο της διωνυμικής τυχαίας μεταβλητής. Μπορούμε, λοιπόν, να εκφράσουμε το δεξί μέλος της παραπάνω σχέσης σε μορφή πινάκων ως $X'(y-\mu)$, όπου:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}$$

και $\mu_i = n_i P(x_i)$. Σαν αποτέλεσμα, ο εκτιμητής μέγιστης πιθανοφάνειας (Maximum Likelihood Estimator-MLE) είναι η λύση της εξίσωσης (score equation):

$$X'(y-\mu) = 0 \qquad (5)$$

Για την λύση της εξίσωσης (5) μπορούμε να χρησιμοποιήσουμε μια επαναληπτική διαδικασία για να παράγουμε τις εκτιμήσεις $b_0, b_1, ..., b_k$ των όρων $\beta_0, \beta_1, ..., \beta_k$ για τις $p = k+1$ παραμέτρους του μοντέλου. Μια τέτοια επαναληπτική μέθοδος είναι αυτή των σταθμισμένων ελάχιστων τετραγώνων (weighted least squares).

Ο τύπος για το σταθμισμένο άθροισμα ελαχίστων τετραγώνων των υπολοίπων είναι:

$$S = \sum_{i=1}^{m} \left[\frac{(y_i - m_i)^2}{\sigma_i^2} \right]$$

όπου $\mu_i = n_i P(x_i)$ και σ_i^2 είναι η διωνυμική διακύμανση στο $i-$οστό σημείο δεδομένων με

$$\sigma_i^2 = n_i P(x_i)[1 - P(x_i)] = n_i \frac{e^{-x'_i \beta}}{\left(1 + e^{-x'_i \beta}\right)^2}$$

Ελαχιστοποιούμε το S:

$$\min S = \min_{\beta} \sum_{i=1}^{m} \left[\frac{(y_i - m_i)^2}{\sigma_i^2} \right].$$

Η διακύμανση σ_i^2 είναι σταθερή, επομένως παραγωγίζουμε μόνο τον αριθμητή του S και παίρνουμε:

$$2 \left[\frac{\sum_{i=1}^{m} (y_i - \mu_i)}{\sigma_i^2} \right] \left(\frac{\partial \mu_i}{\partial \beta} \right).$$

Ισχύει ότι:

$$\frac{\partial \mu_i}{\partial \beta} = n_i P(x_i)[1 - P(x_i)] x_i = \sigma_i^2 x_i .$$

Επομένως, η λύση που παίρνουμε από την ελαχιστοποίηση του σταθμισμένου αθροίσματος τετραγώνων των υπολοίπων με σταθερό σ_i^2 είναι:

$$\sum_{i=1}^{m} (y_i - \mu_i) x_i = 0$$

η οποία είναι παρόμοια με την εξίσωση $X'(y - \mu) = 0$, εξίσωση (5). Άρα μια επαναληπτική μέθοδος όπως η παραπάνω μπορεί να χρησιμοποιηθεί για να

προσδιοριστούν οι αριθμητικές τιμές των $b_0, b_1, ..., b_k$, δηλαδή των εκτιμητών μέγιστης πιθανοφάνειας.

1.3.3 Παραδείγματα εφαρμογής της λογιστικής παλινδρόμησης

Ακολουθούν δύο παραδείγματα στα οποία γίνεται χρήση μοντέλων λογιστικής παλινδρόμησης για την λύση ενός προβλήματος.

1.3.3.1 1ο Παράδειγμα εφαρμογής της λογιστικής παλινδρόμησης

Με χρήση ενός απλουστευμένου μοντέλου θα εκτιμήσουμε τον κίνδυνο θανάτου σε διάστημα 10 χρόνων από ένα καρδιακό νόσημα. Το μοντέλο περιλαμβάνει μόνο τρείς παράγοντες ρίσκου, την ηλικία, το φύλλο και το επίπεδο της χοληστερόλης στο αίμα. Οι παράμετροι του μοντέλου είναι οι εξής:

$\beta_0 = -5.0$
$\beta_1 = +2.0$
$\beta_2 = -1.0$
$\beta_3 = +1.2$
x_1 = η ηλικία σε χρόνια, πάνω απο τα 50
x_2 = το φύλο, όπου 0 είναι αρσενικό και 1 θηλυκό
x_3 = το επίπεδο χοληστερόλης σε mmol/L πάνω απο το 5.0

Μπορούμε να εκφράσουμε το μοντέλο ως εξής:

$$\text{risk of death} = \frac{1}{1+e^{-z}}$$

όπου $z = -5.0 + 2.0x_1 - 1.0x_2 + 1.2x_3$

Σε αυτό το μοντέλο, ο κίνδυνος θανάτου αυξάνεται με την ηλικία (το z αυξάνεται κατά 2.0 για κάθε χρόνο πάνω από τα 50), μειώνεται εάν έχουμε θηλυκό ασθενή (το z μειώνεται κατά 1.0 στην περίπτωση θηλυκού ασθενή) και αυξάνεται με αυξημένο επίπεδο της χοληστερόλης (το z αυξάνεται κατά 1.2 για κάθε 1 mmol/L παραπάνω από τα 5 mmol/L).

Θα χρησιμοποιήσουμε το παραπάνω μοντέλο σε ένα ασθενή 50 χρονών με επίπεδο χοληστερόλης 7.0 mmol/L. Το ρίσκο θανάτου είναι:

$$\frac{1}{1+e^{-z}} \text{ , όπου } z = -5.0 + (+2.0)(50-50) + (-1.0)0 + (+1.2)(7.0-5.0)$$

Το αποτέλεσμα μας λέει ότι το ρίσκο να πεθάνει ο συγκεκριμένος ασθενής από καρδιακή ασθένεια μέσα σε 10 χρόνια είναι 0.07 ή αλλιώς 7%.

1.3.3.2 2^o Παράδειγμα εφαρμογής της λογιστικής παλινδρόμησης

Το παράδειγμα αυτό πραγματεύεται τη χρήση της λογιστικής παλινδρόμησης για την ανάλυση της επίδρασης μιας ουσίας σε ένα πείραμα τοξικότητας. Ο Πίνακας 1.2 δείχνει την επίδραση διαφορετικών δόσεων νικοτίνης στην κοινή μύγα των φρούτων:

Συγκέντρωση x (g/100cc)	Πλήθος εντόμων n	Αριθμός εντόμων που πέθαναν y	Ποσοστό
0.10	47	8	17.0
0.15	53	14	26.4
0.20	55	24	43.6
0.30	52	32	61.5
0.50	46	38	82.6
0.70	54	50	92.6
0.95	52	50	96.2

Πίνακας 1.2: *Δεδομένα πειράματος τοξικότητας για το 2^o παράδειγμα*

Με χρήση της λογιστικής παλινδρόμησης θα καταλήξουμε σε ένα κατάλληλο μοντέλο και θα εκτιμήσουμε τις αποτελεσματικές δόσεις (ED), τις τιμές δηλαδή της νικοτίνης που οδηγούν σε μια συγκεκριμένη τιμή πιθανότητας P. Τέτοιες ποσότητες χρησιμοποιούνται συχνά για να χαρακτηρίσουν τα αποτελέσματα μια πειραματικής δοκιμασίας. Θα εκτιμήσουμε την ED_{50}, όπου ED_P είναι η τιμή του x για την οποία η πιθανότητα ενός συμβάντος, στην προκειμένη περίπτωση ο θάνατος μια μύγας των φρούτων, παίρνει την τιμή P.

Η χρήση του στατιστικού πακέτου PROC LOGIST δίνει τα παρακάτω αποτελέσματα:

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>Chi-Square	Standardized Estimate
INTERCPT	1	-1.7361	0.2420	51.4482	0.0001	
X	1	6.2954	0.7422	71.9399	0.0001	1.024917

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr>Chi-Square	Standardized Estimate
INTERCPT	1	3.1236	0.3349	86.9818	0.0001	
LOGX	1	2.1279	0.2214	92.3628	0.0001	0.898802

Πίνακας 1.3*: Αποτελέσματα PROC LOGISTIC για τα δεδομένα του 2^{ου} παραδείγματος*

Χρησιμοποιήθηκαν δύο λογιστικά μοντέλα με διαφορετική μορφή το καθένα για τον Linear predictor. Αρχικά χρησιμοποιήθηκε το τυπικό μοντέλο της εξίσωσης (1) με τον τυπικό linear predictor $\beta_0 + \beta_1 x$. Επιπλέον, χρησιμοποιήθηκε το μοντέλο

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 \ln x)}}.$$

Συχνά αντικαθιστούμε το x με το $\ln x$ σε τέτοιου είδους πειράματα. Αυτή η πρακτική είναι ιδιαίτερα χρήσιμη όταν το x έχει μεγάλο εύρος τιμών. Οι *p-values* των παραμέτρων που παράχθηκαν από τα στατιστικά χ^2 του Wald είναι αρκετά σημαντικές και για τα δύο μοντέλα, άρα έχουμε δύο υποψήφια μοντέλα. Μια μέθοδος για να συγκρίνουμε τα δύο υποψήφια μοντέλα είναι να συγκρίνουμε τα εύρη των διαστημάτων εμπιστοσύνης γύρω από το $\hat{y}$ (Lewis, Montgomery, Myers 2001). Μια άλλη σχετική μέθοδος είναι να παρατηρήσουμε το τυπικό σφάλμα του εκτιμώμενου predictor $x'b$ για τα δύο μοντέλα. Στον Πίνακα 1.4 παρουσιάζονται τα τυπικά σφάλματα των linear predictors.

$b_0 + b_1 x$	$b_0' + b_i' \ln x$
0.1844	0.2440
0.1607	0.1763
0.1428	0.1439
0.1336	0.1408
0.2139	0.2041
0.3432	0.2646
0.5194	0.3246

Πίνακας 1.4: *Τυπικά σφάλματα των δύο μοντέλων για το 2^o παράδειγμα*

Στην περίπτωσή μας είναι δύσκολο να επιλέξουμε ανάμεσα στα δύο μοντέλα χρησιμοποιώντας τις παραπάνω πληροφορίες, παρ' όλο που τα τυπικά σφάλματα είναι αρκετά μικρότερα για το log μοντέλο στις υψηλές δόσεις. Η χρήση των υπολοίπων (residuals) για την εξέταση αυτών των μοντέλων με τον ίδιο τρόπο που χρησιμοποιούνται στα συνηθισμένα γραμμικά μοντέλα θέλει προσοχή, καθώς τα υπόλοιπα δεν έχουν κοινή διακύμανση.

Θα υπολογίσουμε τον ED_{50} χρησιμοποιώντας και τα δύο μοντέλα για τον linear predictor:

Για το μοντέλο $b_0 + b_1 x$, έχουμε το ED_{50} να δίνεται από την εξίσωση:

$$\mathrm{ED}_{50} = \frac{b_0}{b_1}$$

Στο παράδειγμά μας αυτό ισούται με 0.277g/100cc.

Για το μοντέλο $b_0' + b_i' \ln x$, το ED_{50} δίνεται από την εξίσωση:

$$\mathrm{ED}_{50} = e^{-1.42} = 0.242\mathrm{g}/100\mathrm{cc}$$

1.3.4 Άλλες μορφές στατιστικής συμπερασματολογίας για τη λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση χρησιμοποιείται σε πολλές διαφορετικές περιπτώσεις για την εξαγωγή συμπερασμάτων, όπως για παράδειγμα σε κλινικές δοκιμές όπου πρέπει

να συγκρίνουμε τα αποτελέσματα διαφορετικών θεραπειών των οποίων το αποτέλεσμα έχει δυαδική μορφή. Για την βελτίωση του μοντέλου δοκιμάζεται η σημασία της κάθε μεταβλητής.

Σε αρκετές περιπτώσεις, τα δεδομένα που έχουμε στη διάθεσή μας δεν είναι ομαδοποιημένα, δηλαδή, $n_i = 1$. Στη περίπτωση όμως που οι πειραματικές μονάδες του δείγματος είναι σχετικά ομοιογενείς, τότε η λογιστική παλινδρόμηση μπορεί να πάρει τη μορφή μιας καμπύλης «δόσης-απόκρισης», όπου μετράει την ανταπόκριση ενός ασθενούς ανάλογα με τη δοσολογία που του χορηγείται. Σε μια τέτοια περίπτωση, ισχύει ότι $k = 1$ και $p = 2$ και το μοντέλο παίρνει την μορφή:

$$P(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \quad (6)$$

Τα όρια εμπιστοσύνης για τα β_0 και β_1, όπως επίσης τα όρια εμπιστοσύνης για τους συντελεστές της $P(x_i)$ έχουν σημασία για τους ερευνητές. Το ενδιαφέρον για την μελέτη του κάθε συντελεστή παλινδρόμησης ξεχωριστά, έρχεται από την ανάγκη να προσδιορίσουμε τους λόγους πιθανοτήτων (odd ratios).

Η ιδέα του προσδιορισμού του odd ratio, είναι αποτέλεσμα της χρήσης του logit (P) που δίνεται από τον τύπο:

$$\log\left[\frac{P}{(1-P)}\right]$$

Στη γενική σχέση (1) του λογιστικού μοντέλου παλινδρόμησης , το logit $[P(x_i)]$ δίνεται από τη σχέση:

$$\ln\left[\frac{P(x_i)}{1-P(x_i)}\right] = x_i'\beta \quad (7)$$

και μέσω αυτού του μετασχηματισμού του P γραμμικοποιείται η λογιστική συνάρτηση.

Παράδειγμα:

Ας υποθέσουμε για παράδειγμα, ότι η μεταβλητή x είναι κατηγορική και ότι μια ομάδα από τα πειραματικά μας υποκείμενα χωρίζονται σε αυτά που τους χορηγήθηκε μια δόση από βιταμίνη C (θέτουμε $x = 0$) και σε αυτά που δεν τους χορηγήθηκε τίποτα ($x = 1$), οπότε και η απόκριση που μπορεί να είναι η μόλυνση του αναπνευστικού συστήματος ή όχι, παίρνει τις τιμές $y = 1$ και $y = 0$ αντίστοιχα. Αν τώρα χρησιμοποιήσουμε την εξίσωση (7) για το μοντέλο της εξίσωσης (6), τότε στο παράδειγμα μας έχουμε τη σχέση:

$$\ln\left[\frac{P(x_i)}{1-P(x_i)}\right] = \beta_0 + \beta_1 x_i$$

Αν τώρα θεωρήσουμε ένα υποκείμενο στο οποίο χορηγείται η βιταμίνη C, δηλαδή $x = 0$, τότε η ποσότητα $\exp(\beta_0)$ μπορεί να μεταφραστεί σαν το λόγο συχνοτήτων για τα υποκείμενα που μολύνθηκαν προς αυτά που δεν μολύνθηκαν, για όλο τον πληθυσμό που μελετάμε. Όσον αφορά την ομάδα υποκειμένων στα οποία δεν χορηγήθηκε βιταμίνη ($x = 1$), τότε έχουμε:

$$\ln\left[\frac{P(x_i)}{1-P(x_i)}\right] = \beta_0 + \beta_1$$

Μπορούμε να χρησιμοποιήσουμε την παραπάνω ερμηνεία του β_0 για να βρούμε την αντίστοιχη odd ratio ερμηνεία του β_1. Για την ομάδα υποκειμένων που δεν δέχτηκε θεραπεία ισχύει:

$$\ln\left[\frac{\Pr(Y=1|x=1)}{\Pr(Y=0|x=1)}\right] = \ln\left[\frac{\Pr(Y=1|x=0)}{\Pr(Y=0|x=0)}\right] + \beta_1$$

Άρα, η ποσότητα $\exp(\beta_1)$ μπορεί να ερμηνευτεί σαν την λόγο συχνοτήτων της ομάδας που δεν δέχτηκε θεραπεία, σε σχέση με αυτή που δέχτηκε. Προφανώς, ένας ερευνητής ερμηνεύει μια τιμή $\beta_0 \ll 0$, όπως επίσης και μια τιμή $\beta_1 \gg 0$, να είναι ευνοϊκή προς την θεραπεία.

1.3.4.1 Ιδιότητες της διασποράς των εκτιμητών μέγιστης πιθανοφάνειας στην λογιστική παλινδρόμηση

Οι εκτιμητές μέγιστης πιθανοφάνειας παρουσιάζουν ασυμπτωτικές ιδιότητες στη διακύμανση και την συνδιακύμανση, οι οποίες εκφράζονται σαν συνάρτηση του πίνακα πληροφορίας. Στην περίπτωση ενός γραμμικού μοντέλου με κανονικά, ανεξάρτητα και όμοια κατανεμημένα (iid) σφάλματα, ο πίνακας πληροφορίας για τους εκτιμώμενους συντελεστές παλινδρόμησης δίνεται από τον τύπο:

$$\mathrm{I}(b) = \frac{\mathrm{X'X}}{\sigma^2}$$

όπου σ^2 είναι η διακύμανση του σφάλματος. Σε αυτήν την περίπτωση, ο πίνακας διακύμανσης – συνδιακύμανσης (variance – covariance matrix) των εκτιμώμενων συντελεστών είναι:

$$\mathrm{I}^{-1}(b) = (\mathrm{X'X})^{-1}\sigma^2$$

Ο πίνακας πληροφορίας παρουσιάζει, κατά μια έννοια, την ποιότητα των πληροφοριών των παραμέτρων που διατίθενται από τα δεδομένα μας. Ένας σχετικά μεγάλος πίνακας πληροφορίας σημαίνει μικρότερες διακυμάνσεις στους εκτιμώμενους συντελεστές του μοντέλου. Ο υπολογισμός του πίνακα πληροφορίας γίνεται με διάφορες μεθόδους, όπως με την βοήθεια της εξίσωσης (5):

$$\mathrm{I}(b) = \mathrm{var}\left[\mathrm{X}'(y-\mu)\right]$$

Σε αυτή τη σχέση, με var συμβολίζουμε τον πίνακα διακύμανσης-συνδιακύμανσης (variance-covariance matrix).

Χρησιμοποιώντας τον τελεστή τυπικής διακύμανσης, η παραπάνω εξίσωση αποκτά την μορφή:

$$\mathrm{var}\left[\mathrm{X}'(y-\mu)\right] = \mathrm{X}'\,\mathrm{var}\left[(y-\mu)\right]\mathrm{X}$$

Για το μοντέλο της λογιστικής παλινδρόμησης, έχουμε υποθέσει για τις $y_1, y_2, ..., y_m$ ανεξάρτητες παρατηρήσεις, ότι κάθε y_i παρατήρηση είναι μια διωνυμική τυχαία μεταβλητή με μέσο $n_i P(x_i)$ και διασπορά $\sigma_i^{\,2} = n_i[P(x_i)][1-P(x_i)]$. Επομένως, ισχύει ότι:

$$\mathrm{V} = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_m^2\}$$

και

$$\mathrm{I}(b) = \mathrm{X}'\mathrm{V}\mathrm{X}$$

Ο ασυμπτωτικός πίνακας διακύμανσης -συνδιακύμανσης δίνεται, λοιπόν, από τον τύπο:

$$\mathrm{var}(b) =)(\mathrm{X}'\mathrm{V}\mathrm{X})^{-1}$$

Επομένως, τα εκτιμώμενα τυπικά σφάλματα βρίσκονται στα διαγώνια στοιχεία του V, τον οποίο αντικαθιστά ο V από τη στιγμή που τα β της $P(x_i)$ έχουν αντικατασταθεί από τα εκτιμώμενα b.

1.3.4.2 Συμπερασματολογία με χρήση της μεθόδου Wald στη λογιστική παλινδρόμηση

Η πρώτη εφαρμογή της μεθόδου Wald έχει να κάνει με έλεγχο υποθέσεων για κάθε ξεχωριστό συντελεστή του μοντέλου της λογιστικής παλινδρόμησης. Πιο συγκεκριμένα, θέλουμε να ελέγξουμε:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

με το β_j να εμφανίζεται στον linear predictor $x_i^{'}\beta$ του λογιστικού μοντέλου στην εξίσωση (1).

Για έναν εκτιμητή μέγιστης πιθανοφάνειας b_j ισχύει ότι:

$$z_j = \frac{b_j - \beta_j}{\sigma b_j}$$

ο οποίος ακολουθεί ασυμπτωτικά την τυπική κανονική κατανομή $N(0,1)$ και έτσι ισχύει ότι το

$$z_j^2 = \left(\frac{b_j}{\sigma b_j} \right)^2$$

ακολουθεί ασυμπτωτικά την x_1^2 κατανομή ,υπό την H_0 υπόθεση, όπου σb_j είναι το κατάλληλο διαγώνιο στοιχείο του ασυμπτωτικού πίνακα variance-covariance των b. Στην πράξη, αντικαθιστούμε τα σb_j με τα σb_j. Ο έλεγχος που διεξάγουμε, είναι ο συνηθισμένος μονομερής ή διμερής έλεγχος (one or two-sided test).

Μια δεύτερη μορφή της Wald συμπερασματολογίας έχει να κάνει με τον υπολογισμό του διαστήματος εμπιστοσύνης της διωνυμικής πιθανότητας για κάποια δοσμένα ή αυθαίρετα δεδομένα. Θα μπορούσε να χρησιμοποιηθεί η μέθοδος Δέλτα για το σκοπό αυτό αλλά λόγω της ύπαρξης του linear predictor $x_i^{'}\beta$ στο λογιστικό μοντέλο ακολουθείται μια εναλλακτική διαδικασία υπολογισμού των διαστημάτων εμπιστοσύνης.

Να θυμίσουμε ότι στη λογιστική παλινδρόμηση, η μέση απόκριση στο $x = x_i$ δίνεται από τον τύπο $\frac{1}{1+e^{-x_i^{'}\beta}}$ και άρα είναι πιθανότητα. Η σημειακή εκτίμηση της πιθανότητας δίνεται από το $y_i = P(x_i)$.

Στο λογιστικό μοντέλο

$$P = \frac{1}{1+e^{-x^{'}\beta}}$$

το P είναι μια μονότονη εξίσωση του $x'\beta$. Μπορούμε να ορίσουμε ένα $100(1-a)\%$ διάστημα εμπιστοσύνης στο P, χρησιμοποιώντας ένα διάστημα εμπιστοσύνης στο $x'\beta$. Ο linear predictor περιλαμβάνει προφανώς όρους που είναι γραμμικοί στο β

και μπορούμε να εκμεταλλευτούμε το γεγονός ότι ο b, ο εκτιμητής μέγιστης πιθανοφάνειας του β, είναι ασυμπτωτικά κανονικός. Άρα, ένα άνω διάστημα εμπιστοσύνης για το $x'\beta$, παράγει ένα άνω διάστημα εμπιστοσύνης για το P. Ασυμπτωτικά ισχύει ότι

$$x'b \sim N[x'\beta, x'(\mathrm{X'VX})^{-1}x],$$

το διάστημα εμπιστοσύνης για το $x'\beta$ δίνεται από το

$$x'b \pm z_{a/2}\sqrt{x'(\mathrm{X'VX})^{-1}x}.$$

Σε βιολογικές και χημικές εφαρμογές, όπου τα δεδομένα είναι ομαδοποιημένα και η i-οστή παρατηρούμενη απόκριση y_i είναι διωνυμική με παραμέτρους $P(x_i)$ και n_i, έχει ένα ενδιαφέρον να υπολογίσουμε το διάστημα πρόβλεψης για το y_i. Για το σκοπό αυτό, θα χρειαστούμε μια έκφραση για τη διακύμανση του

$$P(x_i) = \frac{1}{1+e^{-x_i'\beta}} \qquad i = 1, 2, ..., m.$$

Χρησιμοποιώντας τη μέθοδο Δέλτα, παίρνουμε τη σχέση:

$$\mathrm{var}\left[P(x_i)\right] = \left(\frac{\partial P(x_i)}{\partial b}\right)' (\mathrm{X'VX})^{-1} \left(\frac{\partial P(x_i)}{\partial b}\right)$$

Μια πολύ σημαντική ιδιότητα της λογιστικής παλινδρόμησης είναι η παρακάτω:

$$\frac{\partial P(x_i)}{\partial \beta} = n_i[P(x_i)][1-P(x_i)]x_i$$

ή πιο γενικά:

$$\frac{\partial \mu_i}{\partial \beta} = [\mathrm{var}(y_i)]x_i$$

Από αυτήν τη σχέση προκύπτει:

$$\mathrm{var}[P(x_i)] = [\mathrm{var}(y_i)]^2 x_i'(\mathrm{X'VX})^{-1}x_i$$

Άρα, το διάστημα πρόβλεψης μπορεί να βρεθεί όπως και σε όλα τα γραμμικά μοντέλα.

Κατ' αρχήν,

$$\frac{y_i - P(x_i)}{n_i[P(x_i)][1-P(x_i)]\sqrt{1+x_i'(\mathrm{X'VX})^{-1}x_i}} \sim N(0,1) \text{ ασυμπτωτικά.}$$

Άρα, ένα κατάλληλο $100(1-a)\%$ διάστημα εμπιστοσύνης για το y_i, μπορεί να βρεθεί από την σχέση:

$$P(x_i) \pm z_{a/2} \left\{ n_i [P(x_i)][1-P(x_i)] \right\} \sqrt{1 + x_i^{'} (\mathrm{X'VX})^{-1} x_i} \quad \text{για } i = 1, 2, ..., m$$

Στην πράξη, πρέπει να αντικαταστήσουμε το $P(x_i)$ στον πίνακα V.

1.3.4.3 Συμπερασματολογία με χρήση πιθανοφάνειας στη λογιστική παλινδρόμηση

Με την συμπερασματολογία πιθανοφάνειας, μπορούμε να ενισχύσουμε τον έλεγχο υποθέσεων, χρησιμοποιώντας την log likelihood. Η χρήση της μοιάζει αρκετά με τη χρήση της αρχής του επιπλέον αθροίσματος τετραγώνων (extra sum of squares principal) των γραμμικών μοντέλων. Στα γραμμικά μοντέλα, μπορούμε να χρησιμοποιήσουμε κάτω από την μηδενική υπόθεση ένα μοντέλο ελαττωμένο, δηλαδή, η μηδενική υπόθεση θέτει σε ένα υποσύνολο συντελεστών παλινδρόμησης την τιμή μηδέν. Ο έλεγχος χρησιμοποιεί τη διαφορά στο άθροισμα τετραγώνων του σφάλματος:

$$\mathrm{SS_E(reduced)} - \mathrm{SS_E(full)}$$

Στη λογιστική παλινδρόμηση, η διαφορά στο άθροισμα τετραγώνων του σφάλματος αντικαθιστάται από τη διαφορά της log πιθανοφάνειας.

Στην λογιστική παλινδρόμηση, ισχύει ότι ασυμπτωτικά

$$-2 \ln \left[\frac{L(\text{reduced})}{L(\text{full})} \right] \sim \chi^2_{\Delta}$$

όπου το $L(.)$ είναι η πιθανοφάνεια και στην περίπτωσή μας, θέλουμε την πιθανοφάνεια για το πλήρες και για το ελαττωμένο μοντέλο. Η παράμετρος Δ είναι η διαφορά στον αριθμό των παραμέτρων ανάμεσα στο πλήρες και το ελαττωμένο μοντέλο.

Ας υποθέσουμε ότι ο linear predictor είναι $\beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2 + \beta_3 \chi_3$ και ότι θέλουμε να εξετάσουμε την υπόθεση $H_0 : \beta_1 = \beta_2 = 0$

Το στατιστικό ελέγχου για το λόγο πιθανοφάνειας (likelihood ratio test statistic) δίνεται από το:

$$2 \left[\ln L[b_0, b_1, b_2, b_3] - \ln L[b_0^*, b_3^*] \right]$$

με $L(b_0^*, b_3^*)$ να είναι η πιθανοφάνεια για το λογιστικό μοντέλο στο οποίο έχουμε επικαλεστεί την μηδενική υπόθεση και τα β_1 και β_2 έχουν μηδενιστεί. Σαν αποτέλεσμα, η υπόθεση απορρίπτεται εάν η log-πιθανοφάνεια αυξηθεί σημαντικά εισάγοντας τα β_1, β_2 στο μοντέλο μαζί με τα β_0, β_3. Στην περίπτωσή μας, η κατανομή που χρησιμοποιείται για την άνω-ουρά (upper-tail) ενός μονόπλευρου ελέγχου είναι η χ_2^2. Άρα, η συμπερασματολογία με χρήση πιθανοφάνειας, μπορεί να χρησιμοποιηθεί για ελέγχους ενός σετ υποθέσεων.

1.3.5 Έλεγχος καλής προσαρμογής (Goodness-of-Fit tests)

1.3.5.1 Έλεγχος της προσαρμογής με χρήση της απόκλισης

Ένας από τους ελέγχους που μπορούμε να χρησιμοποιήσουμε για να ελέγξουμε την προσαρμογή του μοντέλου μας, χρησιμοποιώντας το κριτήριο του λόγου πιθανοφάνειας, είναι με τη χρήση της απόκλισης (deviance). Η απόκλιση μας βοηθάει να προσδιορίσουμε εάν το προσαρμοσμένο λογιστικό μοντέλο μας είναι σημαντικά χειρότερο από το κορεσμένο μοντέλο που δίνεται από το

$$E(y_i) = P_i \qquad , i = 1, 2, ..., m$$

Ο όρος P_i δεν είναι ίδιος με τον $P(x_i)$. Το κορεσμένο μοντέλο έχει την διωνυμική παρατήρηση y_i σαν τον εκτιμητή P_i. Επομένως, έχουμε μηδέν βαθμούς ελευθερίας μετά την εκτίμηση για τα υπόλοιπα. Σε περίπτωση που δεν έχουμε ομαδοποιημένα δεδομένα, ο εκτιμητής θα είναι απλά 0 ή 1. Το κορεσμένο μοντέλο θα παρουσιάζει πιθανοφάνεια η οποία δεν θα είναι μικρότερη από την πιθανοφάνεια του προσαρμοσμένου μοντέλου παλινδρόμησης.

Η απόκλιση για ένα προσαρμοσμένο μοντέλο λογιστικής παλινδρόμησης δίνεται από τον τύπο:

$$D(\beta) = -2 \ln \left[\frac{L(\beta)}{L(P)} \right],$$

όπου $L(\beta)$ είναι η πιθανοφάνεια του προσαρμοσμένου λογιστικού μοντέλου με το β να έχει αντικατασταθεί από τον εκτιμητή μέγιστης πιθανοφάνειας και $L(P)$ είναι η πιθανοφάνεια του κορεσμένου μοντέλου στο οποίο εκτιμάμε τις m παραμέτρους $P_1, P_2, ..., P_m$ με τον τρόπο που περιγράφηκε παραπάνω.

Για το κανονικό έλεγχο, χρησιμοποιούμε το γεγονός ότι:

$$D(\beta) \sim \chi^2_{m-p} \text{ ασυμπτωτικά}$$

Μια μη-σημαντική τιμή του $D(\beta)$ στην άνω-ουρά ενός μονόπλευρου ελέγχου, σημαίνει ότι η προσαρμογή του μοντέλου δεν είναι σημαντικά χειρότερη από αυτή του κορεσμένου μοντέλου. Έτσι, μια σχετικά μικρή τιμή της απόκλισης είναι ευνοϊκή για το προσαρμοσμένο μοντέλο. Γενικά, εάν η ποσότητα $\frac{D(\beta)}{m-p}$, όπου το $m-p$ είναι ο μέσος της χ^2_{m-p} κατανομής, δεν είναι σημαντικά μεγαλύτερη του 1, τότε η ποιότητα της προσαρμογής είναι λογική-καλή.

Στον έλεγχο ενός σετ υποθέσεων με τη χρήση του κριτηρίου του λόγου πιθανοφάνειας, μπορούμε εναλλακτικά να χρησιμοποιήσουμε τις διαφορές στην απόκλιση, όπως χρησιμοποιούμε τις διαφορές του αθροίσματος τετραγώνων των σφαλμάτων στα γραμμικά μοντέλα. Το στατιστικό υπόθεσης της αναλογίας πιθανοφάνειας είναι

$$-2\ln[L(\text{reduced}) / L(\text{full})]$$

και μπορεί να αντικατασταθεί με $D(\text{reduced}) - D(\text{full})$, αφού το $L(P)$ εξουδετερώνει τη διαφορά στις αποκλίσεις.

Στο προηγούμενο παράδειγμα με τον έλεγχο $H_0 : \beta_1 = \beta_2 = 0$ στη λογιστική παλινδρόμηση με τέσσερις παραμέτρους, μπορούμε να γράψουμε:

$$2\ln L(b_0, b_1, b_2, b_3) - 2\ln L(b_0^*, b_3^*) = D(\beta_0, \beta_3) - D(\beta_0, \beta_1, \beta_2, \beta_3)$$

Η χρήση της απόκλισης και άρα η log-πιθανοφάνεια στο σετ υποθέσεων, δεν είναι τόσο ευπαθής σε μικρά δείγματα όσο η χρήση της απόκλισης σε έναν έλεγχο προσαρμογής.

1.3.5.2 Το στατιστικό Pearson στη λογιστική παλινδρόμηση

Εκτός από την απόκλιση, ένα άλλο στατιστικό που χρησιμοποιείται συχνά για τον έλεγχο προσαρμογής είναι το στατιστικό χ^2 του Pearson το οποίο δίνεται από τον τύπο:

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(y_i - n_i P_i\right)^2}{n_i P_i \left(1 - P_i\right)}$$

και ακολουθεί την κατανομή χ^2_{m-p} ασυμπτωτικά. Το στατιστικό Pearson, όπως και η απόκλιση, έχουν την ίδια ασυμπτωτική κατανομή. Το στατιστικό χ^2 δεν φαίνεται να είναι καλύτερη μέθοδος για τον έλεγχο της προσαρμογής από την απόκλιση, και

αυτό γιατί η κατανομή του σε μικρά δείγματα είναι αβέβαιη. Ωστόσο, και οι δύο μέθοδοι δίνουν σε πολλές περιπτώσεις τα ίδια αποτελέσματα και πολύ σπάνια τα αποτελέσματά τους έρχονται σε αντίθεση.

1.3.5.3 Το στατιστικό των Hosmer και Lemeshaw

Οι Hosmer και Lemeshaw (1989), ανέπτυξαν έναν έλεγχο, εναλλακτικό στην απόκλιση και στο χ^2 του Pearson. Η διαδικασία είναι απλή και έχει ως εξής:

Για κάθε παρατήρηση, υπολογίζονται οι τιμές P που προκύπτουν από το μοντέλο. Στη συνέχεια, αυτές οι τιμές ταξινομούνται κατά μέγεθος και τοποθετούνται σε ομάδες που περιλαμβάνουν περίπου 10 διαστήματα. Σε κάθε διάστημα, υπολογίζονται οι αναμενόμενες τιμές (ή οι αναμενόμενες συχνότητες). Υπολογίζεται ένα στατιστικό χ^2 του τύπου:

$$\sum_{i=1}^{m}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

Η κατανομή αναφοράς (reference distribution) είναι η χ^2 με βαθμούς ελευθερίας το πλήθος των διαστημάτων μείον 2.

1.3.6 Διαστήματα εμπιστοσύνης για τους συντελεστές παλινδρόμησης και τα odd ratios

Τα διαστήματα εμπιστοσύνης για τους συντελεστές παλινδρόμησης, επακόλουθα και για τα odd ratios, μπορούν να υπολογιστούν χρησιμοποιώντας τη συμπερασματολογία του Wald ή μέσω πιθανοφανειών. Εδώ θα εξετάσουμε την Wald προσέγγιση, μια και παραλληλίζεται με την αντίστοιχη προσέγγιση που χρησιμοποιείται στα γραμμικά μοντέλα.

Οι εκτιμητές μέγιστης πιθανοφάνειας ακολουθούν ασυμπτωτικά κανονική κατανομή (asymptotically normal). Επομένως, τα $100(1-a)\%$ Wald διαστήματα εμπιστοσύνης για έναν συντελεστή παλινδρόμησης β_j στο linear predictor, δίνονται από το παρακάτω τύπο:

$$b_j \pm z_{a/2} \text{ (s.e. of } b_j)$$

όπου $z_{a/2}$ είναι το α/2 ποσοστιαίο σημείο της κανονικής κατανομής. Από τη στιγμή που το odd ratio κανονικά υπολογισμένο για τη μεταβλητή x_j είναι το e^{β_j}, τότε ένα $100(1-a)\%$ διάστημα εμπιστοσύνης για το odds ratio είναι $\left[e^{\text{LL}}, e^{\text{UL}}\right]$, όπου τα LL

και UL είναι το κάτω όριο και το άνω όριο εμπιστοσύνης αντίστοιχα για τον συντελεστή β_j.

1.3.7 Η έννοια της υπερβολικής διασποράς στη λογιστική παλινδρόμηση

Σε ένα μοντέλο λογιστικής παλινδρόμησης, η έλλειψη μιας καλής προσαρμογής προκαλείται συνήθως από έναν από τους παρακάτω λόγους:

- Η διωνυμική υπόθεση είναι λαθεμένη.
- Η επιλογή του logit μοντέλου είναι ακατάλληλη (ίσως κάποιο άλλο μοντέλο να είναι πιο κατάλληλο).
- Η δομή που χρησιμοποιείται στο linear predictor είναι λαθεμένη. Ενδεχομένως να υπάρχουν αλληλεπιδράσεις ή άλλη όροι ανώτερης τάξης που αγνοήθηκαν ή να έπρεπε να χρησιμοποιήσουμε τον λογάριθμο μιας μεταβλητής.
- Τα δεδομένα μας έχουν ακραίες τιμές.

Για τον έλεγχο της προσαρμογής μπορούμε να χρησιμοποιήσουμε μια από της προηγούμενες τεχνικές, όπως των Hosmer-Lemeshaw ή απλά να ελέγξουμε αν η μέση απόκλιση (δηλαδή η απόκλιση δια τους βαθμούς ελευθερίας) είναι κοντά στη μονάδα, η οποία χρησιμοποιείται συνήθως όταν τα δεδομένα είναι ομαδοποιημένα και υπάρχει ένα λογικό μέγεθος δείγματος σε κάθε ομάδα.

Υπάρχει φυσικά η περίπτωση η επιλογή της κατανομής και του μοντέλου να είναι οι κατάλληλες, καθώς και στο μοντέλο να μην υπάρχουν ακραίες τιμές αλλά παρ' όλα αυτά η μέση απόκλιση να είναι προβληματική. Σε μια τέτοια περίπτωση λέμε ότι παρουσιάζεται υπερβολική διασπορά (overdispersion).

Η υπερβολική διασπορά παρουσιάζεται όταν η μεταβλητότητα που αντιπροσωπεύεται από την διωνυμική υπόθεση $nP(1-P)$ δεν είναι επαρκής. Επομένως, υπάρχει μια επιπλέον εμπλεκόμενη παράμετρος κλίμακας, $\sigma^2 > 1$, για την οποία η διακύμανση μιας ξεχωριστής παρατήρησης γίνεται $nP(1-P)\sigma^2$ αντί για $nP(1-P)$. Στην περίπτωση που έχουμε $\sigma^2 < 1$, τότε μιλάμε για υπό-διασπορά (underdispersion).

1.3.7.1 Μεταβλητότητα ανάμεσα στις διωνυμικές παραμέτρους ή συσχέτιση ανάμεσα σε διωνυμικές παρατηρήσεις

Είναι πολύ πιθανή η υπερβολική διασπορά σε περιπτώσεις που έχουμε ανομοιογενείς πειραματικές μονάδες. Ένας τρόπος να παρουσιάσουμε αναλυτικά το φαινόμενο της υπερβολικής διασποράς είναι να υποθέσουμε ότι υπάρχει μεταβλητότητα στη διωνυμική παράμετρο p.

Μέσα σε ένα σύνολο όπου οι πειραματικές συνθήκες είναι σταθερές, η παράμετρος p έχει κατανομή με μέσο μ και διακύμανση $\varphi > 0$. Τότε, εάν Y είναι η διωνυμική τυχαία μεταβλητή, ισχύει

$$E(Y) = E\left[E(Y|p)\right] = nE(p) = p\,.$$

Όμως

$$\text{var}(Y) = \text{var}\left[E(Y|p)\right] + E\left[\text{var}(Y|p)\right].$$

Ισχύει ότι

$$\text{var}\left[E(Y|p)\right] = \text{var}[np] = n^2\varphi \quad \text{και}$$

$$E\left[\text{var}(Y|p)\right] = nE\left[p(1-p)\right].$$

όπου $nE\left[p(1-p)\right] = n\left[E(p) - E(p^2)\right] = n[\mu - (\varphi + \mu^2)]$.

Αυτό έχει σαν αποτέλεσμα ότι

$$\text{var}(Y) = n^2\varphi + n\mu - n\varphi - n\mu^2 = n\mu(1-\mu) + n\varphi(n-1) > n\mu(1-\mu)$$

Εάν θεωρήσουμε ότι η ανομοιογένεια στις πειραματικές μονάδες προκαλεί ένα φαινόμενο ανάλογο με το να μεταβάλλεις τυχαία τη παράμετρο p, τότε η διακύμανση της διωνυμικής τυχαίας μεταβλητής αυξάνεται παραπάνω από όσο δικαιολογεί η διωνυμική διακύμανση. Από την άλλη εάν $\varphi = 0$, τότε η $\text{var}(Y)$ ελαττώνεται στην διακύμανση της συνηθισμένης διωνυμικής τυχαίας μεταβλητής.

1.3.7.2 Επίδραση της υπερβολικής διασποράς στα αποτελέσματα

Στην περίπτωση που παρουσιάζεται υπερβολική διασπορά στην λογιστική παλινδρόμηση, η παράμετρος κλίμακας $\sigma^2 > 1$ εισέρχεται στον πίνακας διακύμανσης-συνδιακύμανσης (variance-covariance) με τη μορφή

$$\text{var}(b) = (\text{X'VX})^{-1}\sigma^2$$

και με αυτόν τον τρόπο τα τυπικά σφάλματα υποεκτιμούνται, από τη στιγμή που αγνοείται το $\sigma^2 > 1$.

Στην περίπτωση που παρουσιάζεται υπερβολική διασπορά σε ένα κατάλληλο κανονικό μοντέλο, οι εκτιμητές μέγιστης πιθανοφάνειας των β θα παραμείνουν ασυμπτωτικά αμερόληπτοι.

1.3.7.3 Προσαρμογές λόγω υπερβολικής διασποράς

Ένας πειραματιστής μπορεί να προβεί σε κατάλληλες προσαρμογές για να «διορθώσει» τα τυπικά σφάλματα των συντελεστών παλινδρόμησης. Στην ομαδοποιημένη λογιστική παλινδρόμηση, όπου το μέγεθος του δείγματος σε κάθε ομάδα είναι επαρκώς μεγάλο, είναι λογικό να εκτιμήσουμε την παράμετρο κλίμακας από την μέση απόκλιση, δηλαδή την απόκλιση δια τους βαθμούς ελευθερίας.

Διαισθητικά, μια άλλη εκτίμηση είναι να χρησιμοποιήσουμε το στατιστικό του Pearson διαιρεμένο με *n-p* βαθμούς ελευθερίας. Δηλαδή,

$$\frac{1}{n-p}\sum_{i=1}^{m}\left[\frac{\left(y_i - y_i\right)^2}{n_i\, p_i(1-p_i)}\right] = \frac{\chi^2}{n-p}$$

Η διαίρεση του τετραγωνισμένου υπολοίπου δια την διωνυμική διακύμανση τυποποιείται για την διωνυμική διακύμανση. Μια περίπτωση όμως υπερβολικής διασποράς έχει τον παράγοντα σ^2 ενταγμένο μέσα στο $\text{var}(y_i)$. Επομένως, η παραπάνω ποσότητα παράγει μια εκτίμηση 1 σε μια καθαρή διωνυμική περίπτωση, αλλά εκτιμά $\sigma^2 > 1$ σε μια περίπτωση υπερβολικής διασποράς όπου το μοντέλο περιλαμβάνει την μοναδική παράμετρο κλίμακας σ^2 στο $\text{var}(y_i)$. Άρα, η διόρθωση για το τυπικό σφάλμα είναι να πολλαπλασιάσουμε τα αρχικά τυπικά σφάλματα με τον παράγοντα $\sqrt{\frac{\text{dev}}{(n-p)}}$ ή με τον $\sqrt{\frac{\chi^2}{n-p}}$. Εάν αυτοί οι δύο παράγοντες δίνουν παρόμοια αποτελέσματα, τότε τα πράγματα είναι ικανοποιητικά για την ανάλυση. Είναι ξεκάθαρο ότι η υπερβολική διασπορά επιδρά σημαντικά και στο στατιστικό χ^2 του Wald.

1.4 Καμπύλες Λειτουργικού Χαρακτηριστικού Δέκτη (ROC)

1.4.1 Εισαγωγή

Η χρήση των καμπύλων ROC (Receiver Operating Characteristic) ξεκίνησε στις αρχές της δεκαετίας του ’50, σαν εφαρμογή της στατιστικής θεωρίας αποφάσεων σε προβλήματα της θεωρίας λήψης σήματος. Ξεκίνησαν σαν μια γραφική μέθοδο για τη μέτρηση της ποιότητας λήψης σήματος από έναν δέκτη σε ατελή διαγνωστικά συστήματα. Στη συνέχεια οι καμπύλες ROC βρήκαν εφαρμογή και σε άλλες επιστήμες όπως στην Ιατρική, την Ψυχολογία και την Πληροφορική.

Ας υποθέσουμε ότι έχουμε ένα δείγμα, όπως για παράδειγμα έναν πληθυσμό με ασθενείς και υγιείς ανθρώπους. Ας υποθέσουμε επίσης ότι μπορούμε με κάποια μέθοδο να χωρίσουμε το δείγμα σε δυο ομάδες, τα D+ και D-, όπως για παράδειγμα μπορούμε να χωρίσουμε τον πληθυσμό σε ασθενείς (D+) και υγιείς (D-). Ο μελετητής ορίζει ένα σημείο στη συνεχή κλίμακα με βάση το οποίο χωρίζει τα στοιχεία του δείγματος (σημείο απόφασης) και τα ταξινομεί στις δυο αυτές ομάδες D+ και D-, π.χ. με βάση αν οι κλινικές μετρήσεις είναι μεγαλύτερες ή μικρότερες αυτού του σημείου. Λόγω της ατέλειας του ελέγχου, μπορεί κάποια στοιχεία να έχουν ταξινομηθεί λάθος, όπως για παράδειγμα κάποιοι ασθενείς να ταξινομήθηκαν σαν υγιείς και το αντίθετο.

Ας θεωρήσουμε το παράδειγμα ενός προβλήματος διπλής κλάσης (δυαδικής ταξινόμησης), ένα πρόβλημα δηλαδή όπου ένα αποτέλεσμα χαρακτηρίζεται ως θετικό(p) ή ως αρνητικό (n). Υπάρχουν τέσσερις δυνατές εκβάσεις για ένα δυαδικό ταξινομητή. Αν το αποτέλεσμα της πρόβλεψης είναι p και η πραγματική τιμή είναι επίσης p, τότε αυτό ονομάζεται αληθώς θετικό (TP ή ΑΨ). Εάν η πραγματική τιμή είναι n, τότε ονομάζεται ψευδώς θετικό (FP ή ΨΘ). Στην περίπτωση που η πρόβλεψη είναι n και η πραγματική τιμή επίσης n, τότε ονομάζεται αληθώς αρνητικό (TN ή ΑΑ) και εάν η πρόβλεψη είναι n αλλά η πραγματική τιμή p, τότε ονομάζεται ψευδώς αρνητικό (FN ή ΨΑ).

Ορίζουμε την ευαισθησία ή ποσοστό αληθώς θετικών (TPR) ως εξής:

$$\text{TPR}=\frac{\text{TP}}{\text{P}}=\frac{\text{TP}}{\text{TP+FN}}$$

Επίσης, ορίζουμε την ειδικότητα ή ποσοστό αληθώς αρνητικών (TNR) ως εξής:

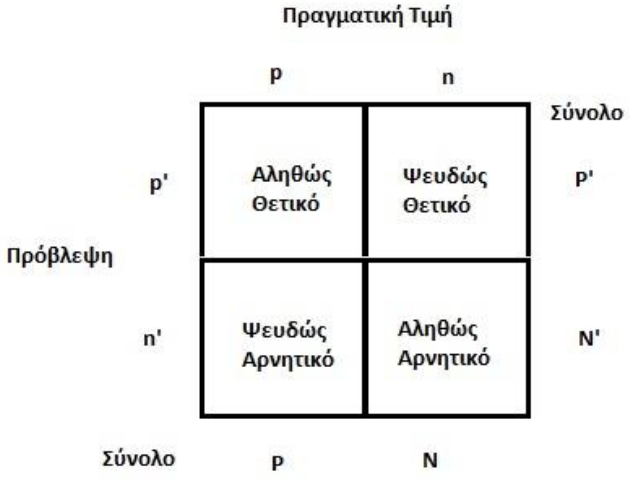
$$\text{TNR}=\frac{\text{TN}}{\text{N}}=\frac{\text{TN}}{\text{FP+TN}}=1\text{-FPR}$$

με το ποσοστό των ψευδώς θετικών (FPR) να ορίζεται ως εξής:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\left(\mathrm{FP} + \mathrm{TN}\right)} \ .$$

Μια καμπύλη ROC είναι μια γραφική παράσταση της ευαισθησίας των αληθώς θετικών ενδείξεων έναντι του 1 –της ειδικότητας ή των ψευδώς θετικών ενδείξεων σε έναν σύστημα δυαδικής ταξινόμησης καθώς το όριο ταξινόμησης ποικίλει. Το λειτουργικό χαρακτηριστικό ενός δέκτη, μπορεί επίσης να παρασταθεί από το γράφημα του ποσοστού των αληθώς θετικών (TPR – True Positive Rate) έναντι του ποσοστού των ψευδώς θετικών (FPR – False Positive Rate). Τέλος, η ROC είναι γνωστή και σαν καμπύλη σχετικού λειτουργικού χαρακτηριστικού, επειδή αποτελεί μια σύγκριση των δυο λειτουργικών χαρακτηριστικών (TPR και FPR) καθώς το κριτήριο αλλάζει.

Σε ένα πείραμα με P θετικές και N αρνητικές περιπτώσεις, τα τέσσερα αποτελέσματα μπορούν να παρουσιαστούν με ένα 2×2 πίνακα συνάφειας ως εξής:



Πίνακας 1.5: *Πίνακας συνάφειας για πείραμα με P θετικές και N αρνητικές περιπτώσεις*

Η καμπύλη ROC είναι το συνεχές γράφημα που ορίζουν τα σημεία FP και TP για όλα τα δυνατά σημεία απόφασης στο μοναδιαίο τετράγωνο $[0,1]\times[0,1]$ που ξεκινά από το σημείο (0,0) και καταλήγει στο σημείο (1,1). Για μια επαρκή κλίμακα σημείων απόφασης, χρειάζεται να κατασκευάσουμε αντίστοιχου πλήθους πίνακες σαν τον Πίνακα 1.5 με τα αντίστοιχα σημεία FP και TP και στη συνέχεια να τα ενώσουμε με ευθύγραμμα τμήματα τα οποία και ορίζουν την καμπύλη ROC.

Το εμβαδό που προκύπτει κάτω από την καμπύλη ROC χρησιμοποιείται ως δείκτης διαχωρισμού για την κατανομή των στοιχείων στις δύο ομάδες. Ο υπολογισμός του εμβαδού γίνεται μη-παραμετρικά, χρησιμοποιώντας τον κανόνα του

τραπεζίου με βάση τα σημεία FP,TP. Όταν οι δύο κατανομές συμπίπτουν απόλυτα, τότε το εμβαδό παίρνει την ελάχιστη τιμή που είναι 0.5. Αν οι κατανομές δεν συμπίπτουν πουθενά, τότε παίρνει τη μέγιστη τιμή που είναι 1.0.

Τα αποτελέσματα ενός ελέγχου, μπορεί να δίνονται σε διακριτή, διαβαθμισμένη κλίμακα, ανάλογα με την πεποίθηση για το πώς πρέπει να χωριστούν τα δείγματα στις διάφορες ομάδες.

Για παράδειγμα, σε ένα δείγμα ασθενών, κατασκευάζεται ένας πίνακας παρόμοιος με τον παρακάτω Πίνακα 1.6.

	Κατηγορία Ταξινόμησης					
GS	--	-	-/+	+	++	Σ
D-	33	6	6	11	2	58
D+	3	2	2	11	33	51

***Πίνακας 1.6**: Παράδειγμα ταξινόμησης 109 ακτινογραφιών σε κλίμακα 5 σημείων "- -" σίγουρα υγιής , "+ +" σίγουρα ασθενή*

Από έναν πίνακα διπλής εισόδου ($2\times n$) όπως αυτός του παραδείγματος, ορίζονται n+1 πίνακες διάστασης 2×2, μεταβάλλοντας σταδιακά το σημείο απόφασης από κατηγορία σε κατηγορία, περιλαμβάνοντας και τις ακραίες περιπτώσεις.

Ενώνοντας τα n+1 σημεία FP,TP που ορίζονται, κατασκευάζουμε την καμπύλη ROC στη διακριτή περίπτωση.

1.4.2 Σημεία και περιοχές της καμπύλης ROC με προβλεπτική ικανότητα

Στο επίπεδο που απεικονίζεται μια καμπύλη ROC, μπορούμε να ορίσουμε επιμέρους σημεία και περιοχές που έχουν σημαντική προβλεπτική ικανότητα, όπως είναι τα παρακάτω:

- Το σημείο που η καμπύλη ROC τέμνει την κάθετη στη διαγώνιο
- Η περιοχή μεταξύ της ROC καμπύλης και της διαγωνίου
- Η περιοχή κάτω από τη ROC καμπύλη
- Η απόσταση d' μεταξύ του μέσου της κατανομής στο σύστημα υπό θόρυβο μείον το μέσο της κατανομής στο σύστημα υπό σήματα, δια την τυπική

απόκλιση, με την προϋπόθεση ότι οι δυο αυτές κατανομές είναι Κανονικές με την ίδια τυπική απόκλιση. Βάσει των υποθέσεων αυτών, μπορεί να αποδειχθεί ότι το σχήμα της ROC εξαρτάται μόνο από την d'.

1.4.3 Καμπύλες ROC και κλασσική στατιστική θεωρία

Υπάρχει σχέση ανάμεσα στις καμπύλες ROC και στην κλασσική στατιστική θεωρία και πιο συγκεκριμένα, στη θεωρία του ελέγχου υποθέσεων. Στο πρόβλημα ελέγχου μιας απλής στατιστικής υπόθεσης $H_0 : f = f_0$ έναντι μιας εναλλακτικής $H_A : f = f_A$, μπορούμε να αντιστοιχήσουμε στην κατανομή της μηδενικής υπόθεσης την κατανομή των στοιχείων του ενός συνόλου (π.χ. των υγιών) και στην εναλλακτική, την κατανομή των στοιχείων του άλλου συνόλου (π.χ. των ασθενών). Τότε, το σημείο απόφασης θα αντιστοιχεί στο σημείο που ορίζει το επίπεδο σημαντικότητας (σφάλμα τύπου Ι), το οποίο είναι:

$$a = P(\text{απόρριψη } H_0 \,|\, H_0) = \int_C f_0(x)dx$$

όπου C είναι η κρίσιμη περιοχή ή αλλιώς, η περιοχή δεξιά του σημείου απόφασης.

Αντίστοιχα θα ισχύει:

$$FP = P(\text{αποτέλεσμα ελέγχου D+} \,|\, \text{D-}) = \int_C f_{D-}(x)dx$$

Με όμοιο τρόπο, για την εναλλακτική υπόθεση H_A, ορίζουμε το σφάλμα τύπου ΙΙ και την ισχύ σαν:

$$1-\beta = P(\text{μη-απόρριψη } H_A \,|\, H_A) = \int_C f_A(x)dx$$

Που αντιστοιχεί στο κλάσμα TP:

$$TP = P(\text{αποτέλεσμα ελέγχου D+} \,|\, \text{D+}) = \int_C f_{D+}(x)dx$$

Εάν μεταβάλουμε το επίπεδο σημαντικότητας και πάρουμε το γράφημα που ορίζεται από τα σημεία (α, 1-β), ορίζεται η αντίστοιχη καμπύλη ROC για τον έλεγχο της απλής μηδενικής υπόθεσης H_0 έναντι της απλής εναλλακτικής H_A. Η αντιστοιχία ανάμεσα στους ελέγχους υποθέσεων και τις καμπύλες ROC φαίνεται και στο παρακάτω Πίνακα 1.7:

	Απόφαση			Ιατρικός έλεγχος	
	Μη-απόρριψη	Απόρριψη			
	Ho	Ho		D-	D+
Ho			GS		
ΑΛΗΘΗΣ	1-α	α	D-	TN	FP
ΨΕΥΔΗΣ	β	1-β	D+	FN	TP

Πίνακας 1.7*: Αντιστοιχία ανάμεσα στους ελέγχους υποθέσεων και τις καμπύλες ROC*

1.4.4 Χρήση του εμβαδού κάτω από μια καμπύλη ROC

Η περιοχή κάτω από μια καμπύλη ROC (Area Under the ROC Curve - AUC) εκφράζει την πιθανότητα να ταξινομηθεί ένα τυχαία επιλεγμένο θετικό δείγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό.

Στο παράδειγμα με τους ασθενείς, η περιοχή κάτω από την καμπύλη εκφράζει την πιθανότητα η τιμή ενός test για έναν ασθενή (D+) να είναι μεγαλύτερη από την τιμή του test για ένα άτομο που δεν έχει ασθένεια (D-), δηλαδή εκφράζει την $P(D+ > D-)$ η οποία δίνεται από τον τύπο:

$$w = \frac{1}{d_+ d_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} I\left(D_i^+, D_j^-\right)$$

με d_+ και d_- να είναι ο αριθμός των ατόμων με ασθένεια ή όχι. Το D_i^+ είναι η τιμή του διαγνωστικού test για το i άτομο της ομάδας ασθενών και D_j^- είναι η τιμή του διαγνωστικού test για το j άτομο της ομάδας των μη-ασθενών. Το $I(D_i^+, D_j^-)$ παίρνει την τιμή 1 αν $D_i^+ > D_j^-$ και την τιμή ½ αν $D_i^+ = D_j^-$ και την τιμή 0 αν $D_i^+ < D_j^-$.

Στην περίπτωση που στο πείραμά μας, μικρές τιμές του test υποδεικνύουν μεγάλη πιθανότητα εμφάνισης της νόσου, δηλαδή το test συνδέεται αρνητικά με τη νόσο, τότε υπολογίζουμε την ποσότητα $w' = 1 - w$ ή μετασχηματίζουμε το test με τρόπο ώστε να συνδέεται θετικά με τη νόσο. Μια τιμή για το AUC, $w = 0.5$ αντιστοιχεί σε ένα test το οποίο μαντεύει τυχαία αν κάποιος είναι ασθενής ή όχι.

Το AUC μπορεί να συνδεθεί με μια σειρά από στατιστικά. Συνδέεται με το στατιστικό U του Mann-Whitney το οποίο ελέγχει αν τα θετικά κατατάσσονται υψηλότερα από τα αρνητικά. Είναι ισοδύναμη με το test Wilcoxon και σχετίζεται με το συντελεστή του Gini(G) μέσω της σχέσης $G + 1 = 2\text{AUC}$, με:

$$G = 1 - \sum_{k=1}^{n} \left(X_k - X_{k-1} \right) \left(Y_k - Y_{k-1} \right)$$

Χρησιμοποιώντας αυτόν τον τρόπο, η AUC μπορεί να υπολογιστεί χρησιμοποιώντας ένα αριθμό τραπεζοειδών προσεγγίσεων.

Υπάρχουν ωστόσο περιπτώσεις, όπως σπάνιες ασθένειες, που ο εκτιμητής αυτός δεν είναι αξιόπιστος και μπορεί να δώσει αρνητική τιμή. Κάποιες φορές είναι προτιμότερο, αντί να εξετάζουμε ολόκληρη την καμπύλη ROC, να εξετάζουμε μια συγκεκριμένη περιοχή. Υπάρχει η δυνατότητα να υπολογιστούν μερικές AUC (partial AUC), όπως για παράδειγμα, να επικεντρώσουμε στην περιοχή της καμπύλης με χαμηλό ποσοστό ψευδώς θετικών, η οποία έχει συχνά πρωταρχικό ενδιαφέρον στην οπτικοποίηση των δεδομένων του τεστ.

1.4.4.1 Υπολογισμός του εμβαδού κάτω από την ROC

Θα συμβολίζουμε το εμβαδό κάτω από τη ROC καμπύλη με θ. Έστω $X_1, X_2, ..., X_n$ ανεξάρτητες μεταξύ τους μετρήσεις υγιών από μια συνεχή συνάρτηση κατανομής F_0 και έστω $Y_1, Y_2, ..., Y_m$ οι ανεξάρτητες μεταξύ τους μετρήσεις των ασθενών από συνεχή συνάρτηση κατανομής F_1. Τότε ισχύει $\text{FP}(c) = 1 - F_0(c)$ και $\text{TP}(c) = 1 - F_1(c)$, $c \in (-\infty, +\infty)$ και η ROC ορίζεται από τα σημεία $\left(1 - F_0(c), 1 - F_1(c)\right)$. Το εμβαδό κάτω από τη ROC ορίζεται ως:

$$\theta = \int_0^1 F_0(c) dF_1(c) = P(X < Y)$$

οι κατανομές F_0, F_1 εκτιμώνται από τις εμπειρικές τους κατανομές.

Το εμβαδό μπορεί να υπολογιστεί από το συνολικό αριθμό των ζευγών $X < Y$ προς το συνολικό πλήθος των ζευγών (X, Y) που ισούνται με mn.

Ορίζουμε ως θ τον αμερόληπτο εκτιμητή του θ. Χρησιμοποιώντας τη δείκτρια συνάρτηση I, το θ θα ισούται με:

$$\theta = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} I\left(X_i < Y_j \right)}{mn} = \frac{U}{mn}$$

Εάν οι κατανομές F_0, F_1 είναι άγνωστες, τότε ο θ έχει ελάχιστη διασπορά μεταξύ των εκτιμητών του θ. Το U στον αριθμητή είναι το στατιστικό Wilcoxon-Mann-Whiney.

Μια εύχρηστη έκφραση για τη διακύμανση του θ είναι η:

$$\mathrm{Var}(\theta) = \frac{\theta(1-\theta) + (m-1)\left(Q_1 - \theta^2\right) + (n-1)\left(Q_2 - \theta^2\right)}{mn}$$

με $Q_1 = P\left(X_i < Y_j, X_i < Y_l\right),\ Q_2 = P\left(X_i < Y_j, X_k < Y_j\right)$

$i,k = 1,...,n$ και $j,l = 1,...,m$

Η έκφραση αυτή προκύπτει από την απόδειξη του Lehmann για τη διακύμανση του στατιστικού Mann-Whitney υπό την υπόθεση H_1.

Ισχύει επιπλέον ότι:

$$\frac{\theta - \theta}{\sqrt{\mathrm{Var}\left(\theta\right)}} \sim N(0,1)$$

καθώς $n,m \to \infty$ και ισχύει η ασυμπτωτική κανονικότητα για $n,m \geq 8$.

Ένα άνω φράγμα για τη διακύμανση του θ δίνεται από τη σχέση:

$$\mathrm{Var}_{\max}\left(\theta\right) = \frac{\theta(1-\theta)}{\max(n,m)}.$$

Το εμβαδό κάτω από την ROC εκφράζει την πιθανότητα ένα ζεύγος μετρήσεων, μιας ενός υγιούς και μιας ενός ασθενούς, να ταξινομηθεί με τη σωστή σειρά (η μέτρηση του ασθενούς να είναι υψηλότερη από του υγιούς). Η ποσότητα Q_1 στην διακύμανση εκφράζει την πιθανότητα δύο τυχαίες μετρήσεις ασθενών να είναι υψηλότερες από μια μέτρηση υγιούς και η ποσότητα Q_2 την πιθανότητα μια τυχαία επιλεγμένη μέτρηση ασθενούς να είναι υψηλότερη από δυο τυχαία επιλεγμένες μετρήσεις υγιών.

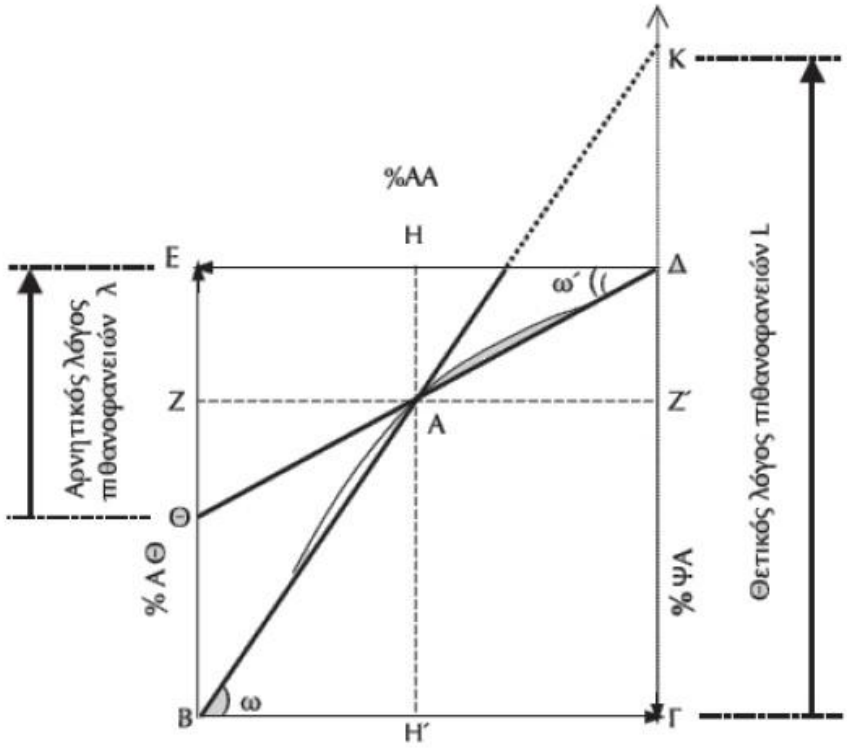
1.4.5 Λόγοι πιθανοφανειών και καμπύλη ROC

Σε κάθε σημείο πάνω στην καμπύλη ROC αντιστοιχεί και ένα ζεύγος ποσοστών για τα αληθώς θετικά (TP ή ΑΘ) και ψευδώς θετικά (FP ή ΨΘ) αποτελέσματα. Επομένως, αντιστοιχεί και ένα ζεύγος των συμπληρωματικών τους ποσοστών, για τα ψευδώς αρνητικά (FN ή ΨΑ) και τα αληθώς αρνητικά (TN ή ΑΑ) αποτελέσματα.

Εάν τα αποτελέσματα του πειράματος εκφράζονται διχοτομικά, τότε ο λόγος του ποσοστού TP προς το ποσοστό FP (δηλαδή το $\frac{\%TP}{\%FP}$), είναι ο θετικός λόγος

πιθανοφανειών L του πειράματος. Αντίστοιχα, ο λόγος του ποσοστού των FN προς το ποσοστό των TN (δηλαδή το $\frac{\%FN}{\%TN}$),είναι ο αρνητικός λόγος πιθανοφανειών λ του πειράματος.

Ένας τρόπος να υπολογίσουμε γραφικά τους λόγους πιθανοφανειών είναι ο φαίνεται στο παρακάτω σχήμα:



Εικόνα 1.8: *Γραφική απεικόνιση του θετικού και του αρνητικού λόγου πιθανοφανειών για το διαχωριστικό όριο Α, όταν τα αποτελέσματα εκφράζονται διχοτομικά.*

Παίρνουμε ένα σημείο Α πάνω στην καμπύλη ROC. Από το Α φέρνουμε τις κάθετες προς τους άξονες ευθείες ZZ' και HH'. Η απόσταση $BZ=H'A$ είναι ίση με το ποσοστό των TP (το %TP). Η απόσταση $BH'=AZ$ είναι ίση με το ποσοστό των FP (το %FP). Η εφαπτομένη της γωνίας ω του τριγώνου BAH', δηλαδή ο λόγος $\frac{H'A}{BH'}$, ισούται με τον θετικό λόγο πιθανοφανειών στο σημείο Α, τον L_A. Επειδή τα ορθογώνια τρίγωνα BAH' και $BK\Gamma$ είναι όμοια, ισχύει ότι:

$$L_A = \frac{\%TP}{\%FP} = \frac{BZ}{BH'} = \frac{H'A}{BH'} = \varepsilon\varphi\omega = \frac{\Gamma K}{B\Gamma} = \Gamma K .$$

Η απόσταση, δηλαδή, ΓΚ παριστάνει γραφικά το θετικό λόγο πιθανοφανειών. Αντίστοιχα, η απόσταση ΘΕ παριστάνει γραφικά τον αρνητικό λόγο πιθανοφανειών:

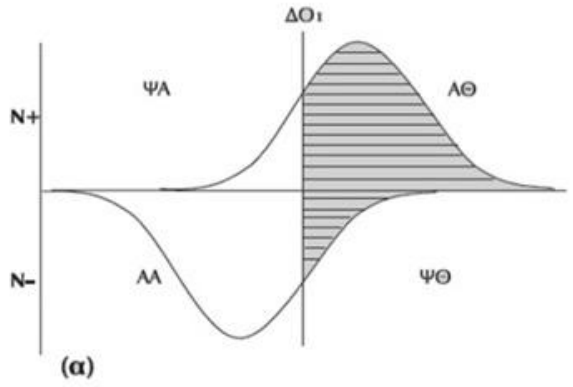
$$\lambda_A = \frac{\%FN}{\%TN} = \frac{\text{Z}'\Delta}{\Delta\text{H}} = \frac{\text{AH}}{\Delta\text{H}} = \varepsilon\varphi\omega' = \frac{\Theta\text{E}}{\Delta\text{E}} = \Theta\text{E}$$

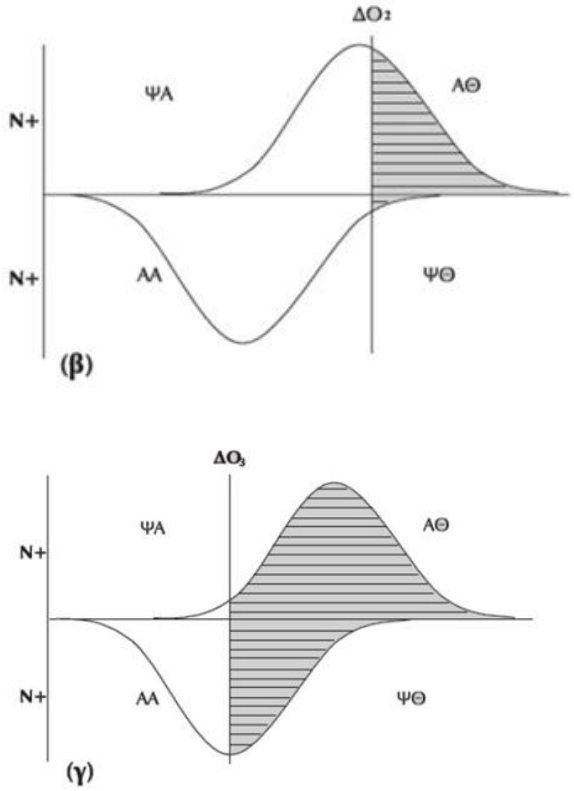
1.4.6 Το διαχωριστικό όριο

Για κάθε διαγνωστικό πείραμα, ορίζονται δυο χαρακτηριστικά. Η διαχωρίζουσα μεταβλητή (separator variable) και το διαχωριστικό όριο (cut off point). Η διαχωρίζουσα μεταβλητή είναι μια μετρήσιμη ιδιότητα σε ποιοτική, διατάξιμη ή σε μετρική (συνεχή ή διακριτή) κλίμακα. Το διαχωριστικό όριο είναι μια συγκεκριμένη τιμή στην κλίμακα μέτρησης της διαχωρίζουσας μεταβλητής, από την οποία τιμή και πέρα , όλες οι τιμές (δηλαδή οι τιμές για τα αποτελέσματα του πειράματος πέρα αυτής) θεωρούνται θετικές (π.χ. παθολογικές). Αντίστοιχα, οι τιμές κάτω αυτής θεωρούνται αρνητικές (π.χ. μη-παθολογικές).

Η επιλογή διαχωριστικού ορίου καθορίζει την ικανότητα να διακριθούν τα αποτελέσματα σωστά, όπως π.χ. να διακριθούν οι ασθενής από τους υγιείς.

Οι βιολογικές μεταβλητές εμφανίζουν μεγάλη διασπορά στις τιμές για τους νοσούντες και μη πληθυσμούς. Επιπλέον, υφίσταται μερική επικάλυψη των κατανομών των τιμών μιας μεταβλητής στις δύο ομάδες.





Εικόνα 1.9: *Κατανομές συχνότητας για τα αποτελέσματα μιας δοκιμασίας σε νοσούντα (N+) και υγιή (N-) άτομα.*

Στα παραπάνω σχήματα (Εικ.1.9 (α), (β), (γ)),οι κανονικές κατανομές εκφράζουν τα αποτελέσματα μιας δοκιμασίας σε δυο υποθετικούς πληθυσμούς νοσούντων (N+) και υγιών (N-) ατόμων. Η καμπύλη πάνω από τον οριζόντιο άξονα αντιστοιχεί στον πληθυσμό των νοσούντων και η καμπύλη κάτω από τον οριζόντιο άξονα αντιστοιχεί στον πληθυσμό των υγιών. Η μερική επικάλυψη των δυο κατανομών αποκλείει την τέλεια διακριτική ικανότητα της δοκιμασίας. Η ολική επιφάνεια που περιλαμβάνεται κάτω από καθεμία καμπύλη συχνοτήτων ισούται με 1 (ή 100%).

Για καθένα από τα πιθανά διαχωριστικά όρια (ΔΟ), το γραμμοσκιασμένο τμήμα της ολικής επιφάνειας δεξιά του διαχωριστικού ορίου και πάνω από τον οριζόντιο άξονα απεικονίζει το ποσοστό των αληθώς θετικών (%TP) και κάτω από τον οριζόντιο άξονα, το ποσοστό των ψευδώς θετικών (%FP) αποτελεσμάτων. Αντίστοιχα, το τμήμα της ολικής επιφάνειας αριστερά του διαχωριστικού ορίου και πάνω από τον οριζόντιο άξονα απεικονίζει το ποσοστό των ψευδώς αρνητικών (%FN) και κάτω από τον οριζόντιο άξονα, το ποσοστό των αληθώς αρνητικών (%TN) αποτελεσμάτων.

Το ΔO_1 εξασφαλίζει το μέγιστο άθροισμα ΑΘ και ΑΑ αποτελεσμάτων της δοκιμασίας. Όταν το ΔΟ μετατοπίζεται προς τα δεξιά στον οριζόντιο άξονα (ΔO_2),

δηλαδή προς τις «παθολογικές» τιμές, μειώνεται το %TP αποτελεσμάτων. Επομένως, αυξάνεται το %FN (1-%TP) αποτελεσμάτων. Η μετατόπιση του ΔΟ προς τα αριστερά (ΔO_3), δηλαδή προς τις «φυσιολογικές) τιμές, έχει αντίθετο αποτέλεσμα. Το %TP αυξάνεται και κατά συνέπεια μειώνεται το %FN αποτελεσμάτων. Συγχρόνως αυξάνεται και το %FP αποτελεσμάτων.

Καθώς το ΔΟ μετατοπίζεται σταδιακά προς τις «φυσιολογικές» τιμές, ισχύουν τα παρακάτω:

α) Το %TP και το %FP αποτελεσμάτων αυξάνονται ενώ το %TN και το %FN αποτελεσμάτων μειώνονται.

β) Ο θετικός λόγος πιθανοφανειών $L = \frac{\%TP}{\%FP}$ από τις πολύ υψηλές τιμές που έχει αρχικά, μειώνεται προοδευτικά μέχρι να πάρει την τιμή 1.

γ) Ο αρνητικός λόγος πιθανοφανειών $\lambda = \frac{\%FN}{\%TN}$ επίσης μειώνεται προοδευτικά και από τις πολύ υψηλές τιμές που έχει αρχικά παίρνει τελικά την τιμή 0.

Η διακύμανση της διακριτικής ικανότητας της δοκιμασίας σε συνάρτηση με τη μετατόπιση του Δ.Ο παριστάνεται γραφικά με την καμπύλη ROC.

1.4.6.1 Άλλοι παράγοντες που καθορίζουν την επιλογή του βέλτιστου διαχωριστικού ορίου

Το ΔΟ συνίσταται στο συνδυασμό του ποσοστού των TP και του ποσοστού των FP αποτελεσμάτων που μεγιστοποιεί την προσδοκώμενη χρησιμότητα (expected utility), όταν εφαρμόζεται σε ένα συγκεκριμένο πρόβλημα. Για τον καθορισμό αυτού του ΔΟ, λαμβάνονται υπ' όψη και επιπλέον παράγοντες όπως:

- Ο επιπολασμός του νοσήματος, η πιθανότητα δηλαδή του νοσήματος πριν από την εφαρμογή της διαδικασίας. Εάν, για παράδειγμα, ο επιπολασμός είναι μικρός και η θεραπεία δαπανηρή και με σοβαρές παρενέργειες, είναι προτιμότερη η μετακίνηση του ΔΟ προς τα κάτω-αριστερά, προς τις περισσότερο «παθολογικές» δηλαδή τιμές. Αντίθετα, αν ο επιπολασμός είναι μεγαλύτερος και οι συνέπειες από τη μη χορήγηση θεραπείας σοβαρές, είναι προτιμότερη η μετακίνηση του ΔΟ προς τα πάνω-δεξιά, προς τις πιο «φυσιολογικές» τιμές.

- Το αναμενόμενο καθαρό όφελος (net benefit) από τη θεραπεία των ατόμων, που σχετίζεται με τα FN αποτελέσματα και το αναμενόμενο καθαρό κόστος (net cost) από τη θεραπεία των μη νοσούντων, που σχετίζεται με τα FP αποτελέσματα της δοκιμασίας. Είναι φανερό ότι το όφελος από τη χορήγηση

της θεραπείας θα έχουν μόνο οι άρρωστοι που πάσχουν από το νόσημα. Η χορήγηση θεραπείας σε μη πάσχοντες ή η μη χορήγηση σε πάσχοντες, σημαίνει κόστος.

Το καθαρό όφελος, προκύπτει από τη διαφορά μεταξύ της χρησιμότητας της χορήγησης θεραπείας και της χρησιμότητας μη χορήγησης σε πάσχοντες:

$$\text{(καθαρό όφελος): } O = X\left(\Delta+, \frac{\Theta+}{N+}\right) - X\left(\Delta-, \frac{\Theta-}{N+}\right) = X_{A\Theta} - X_{\Psi A}$$

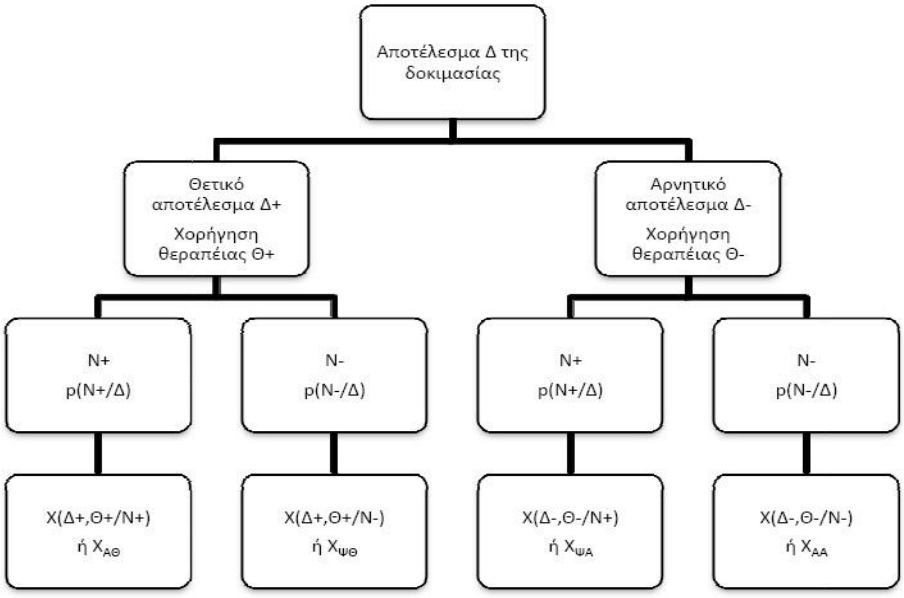
Το καθαρό κόστος, προκύπτει από τη διαφορά μεταξύ της χρησιμότητας μη χορήγησης θεραπείας σε μη πάσχοντες και της χρησιμότητας χορήγησης θεραπείας σε μη πάσχοντες:

$$\text{(καθαρό κόστος): } K = X\left(\Delta-, \frac{\Theta-}{N+}\right) - X\left(\Delta+, \frac{\Theta+}{N-}\right) = X_{AA} - X_{\Psi\Theta}$$

Το βέλτιστο ΔΟ της δοκιμασίας αντιστοιχεί στο σημείο που η κλίση της ROC (η κλίση της εφαπτόμενης στο σημείο αυτό ευθείας), ισούται με το γινόμενο του odds της απουσίας του νοσήματος επί το λόγο του καθαρού κόστους προς το καθαρό όφελος:

$$\varepsilon\varphi\gamma = \frac{1-p(N+)}{p(N+)} \times \frac{K}{O}.$$

Η επιλογή του βέλτιστου ΔΟ μπορεί να βρεθεί και με την τεχνική των δέντρων απόφασης όπως φαίνεται στην Εικόνα 1.10:



Εικόνα 1.10: *Δέντρο απόφασης για την επιλογή του βέλτιστου διαχωριστικού ορίου*

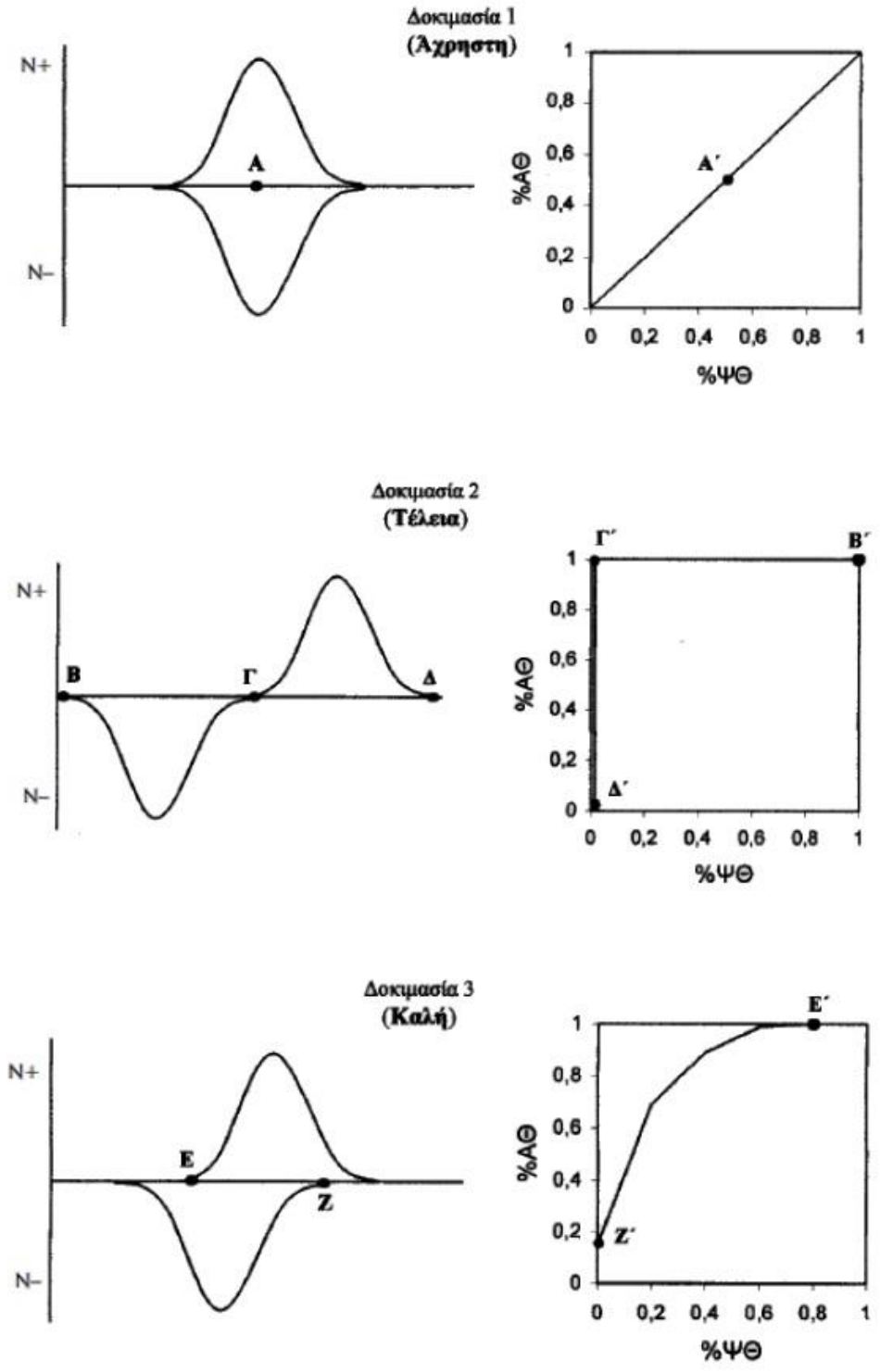
1.4.7 Εκτίμηση της διακριτικής ικανότητας ενός πειράματος

Στα παρακάτω σχήματα (Εικόνα 1.11) φαίνονται οι κατανομές συχνότητας των αποτελεσμάτων (συνεχής ποιοτική μεταβλητή) τριών διαγνωστικών δοκιμασιών με διαφορετική διακριτική ικανότητα και οι αντίστοιχες καμπύλες ROC.

Στην δοκιμασία 1, οι κατανομές συχνότητας των αποτελεσμάτων σε νοσούντα (N+) και μη νοσούντα (N-) άτομα, εμφανίζουν πλήρη επικάλυψη. Για κάθε τιμή του ΔΟ (π.χ. σημείο Α), το %TP συμπίπτει με το %FP και ο λόγος πιθανοφανειών ισούται με 1. Δηλαδή, το αποτέλεσμα της δοκιμασίας (θετικό ή αρνητικό) εμφανίζεται με την ίδια συχνότητα στους πάσχοντες και μη από το νόσημα, άρα δεν αποτελεί πληροφορία και η δοκιμασία στερείται οποιασδήποτε διαγνωστικής ικανότητας. Η ROC που περιγράφει τα λειτουργικά χαρακτηριστικά μιας τέτοιας δοκιμασίας, συμπίπτει με τη διαγώνιο του τετραγώνου.

Στην δοκιμασία 2, υπάρχει ένα εύρος τιμών του ΔΟ οι οποίες επιτυγχάνουν πλήρη και απόλυτη διάκριση των πασχόντων από τους μη πάσχοντες. Η αντίστοιχη καμπύλη ROC ανέρχεται κατακόρυφα από την κάτω αριστερή γωνία έως την άνω αριστερή και συνεχίζει παράλληλα προς τον οριζόντιο άξονα του %TN αποτελεσμάτων. Μια τέτοια κατάσταση είναι η ιδανική, σπανίως όμως συναντιέται στην πράξη.

Στη δοκιμασία 3, η καμπύλη ROC και η κατανομή, δίνουν έναν πιο συνηθισμένο, ρεαλιστικό χαρακτηρισμό για την αποτελεσματικότητα μιας διαγνωστικής διαδικασίας. Η καμπύλη ROC κείται μεταξύ των δυο προηγούμενων «ακραίων» περιπτώσεων.

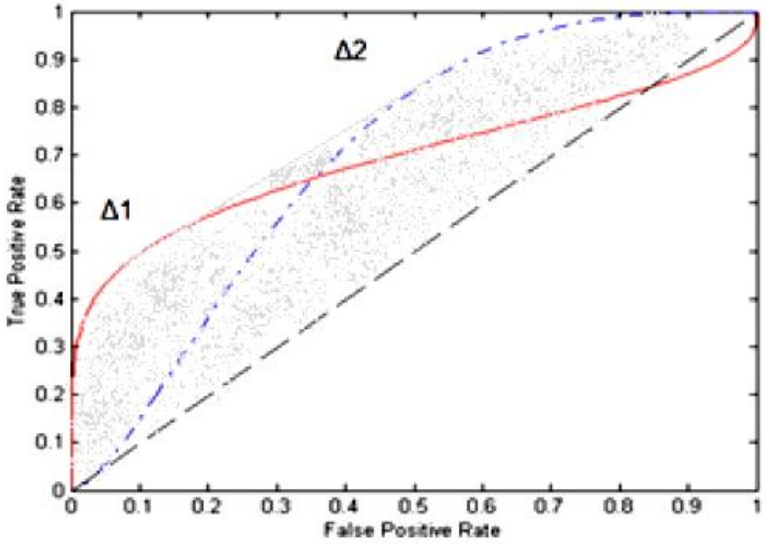


Εικόνα 1.11: *Κατανομές συχνότητας των αποτελεσμάτων και αντίστοιχες καμπύλες ROC για τρεις δοκιμασίες με διαφορετική διακριτική ικανότητα*

1.4.8 Σύγκριση διαγνωστικών δοκιμασιών

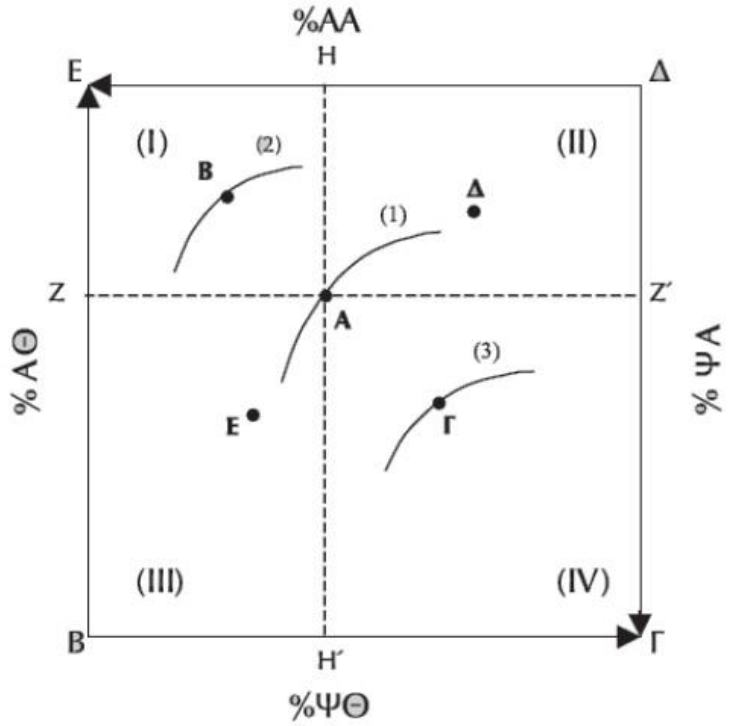
Μια καμπύλη ROC που εμφανίζει μεγαλύτερη κυρτότητα προς την άνω αριστερή γωνία (και επομένως το εμβαδό κάτω από αυτήν είναι μεγαλύτερο), εκφράζει μια δοκιμασία, ένα πείραμα, με μεγαλύτερη διακριτική ικανότητα. Καμία όμως δοκιμασία δεν είναι πάντα καλύτερη μιας άλλης.

Στο παρακάτω σχήμα (Εικόνα 1.12) η δοκιμασία Δ_1 είναι καλύτερη για μικρό ποσοστό FP και η δοκιμασία Δ_2 είναι καλύτερη για μεγάλο ποσοστό FP. Η σκιασμένη περιοχή αποτελεί την κυρτή θήκη της καμπύλης ROC (ROC convex hull) ενός συνόλου σημείων στο ROC επίπεδο. Ένας ταξινομητής επιλεγμένος μέσα από την κυρτή θήκη θεωρείται ιδανικός.



Εικόνα 1.12*: Καμπύλες ROC με την ίδια κυρτότητα αλλά με διαφορετική διακριτική ικανότητα*

Η σύγκριση διαφορετικών δοκιμασιών χρησιμοποιώντας τον θετικό και αρνητικό λόγο πιθανοφανειών είναι προτιμότερη έναντι του συνηθισμένου τρόπου σύγκρισης χρησιμοποιώντας της ευαισθησία και την ειδικότητα.



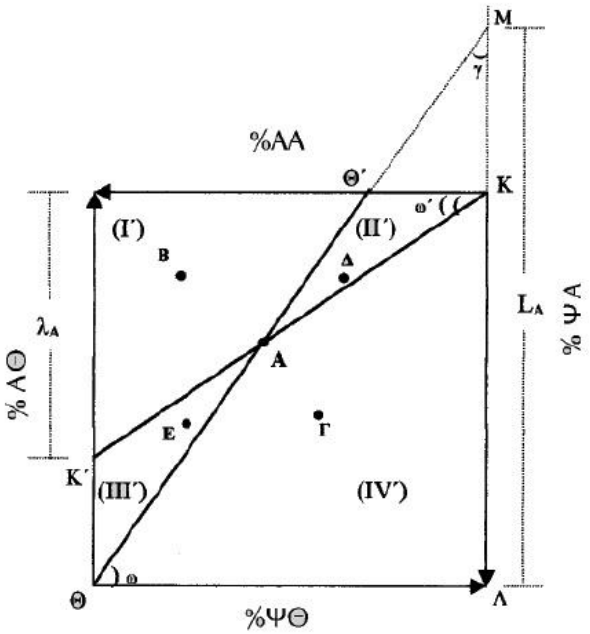
Εικόνα 1.13: *Σύγκριση σημείων καμπύλων ROC βάσει των αποτελεσμάτων %ΑΘ και %ΑΑ*

Στην Εικόνα 1.13, οι ευθείες ZZ' και HH' ορίζουν τέσσερις περιοχές:

- Κάθε σημείο Β που ανήκει στο τετράπλευρο Ι, ανήκει σε μια δεύτερη δοκιμασία με μεγαλύτερα %TP και %TN αποτελεσμάτων από τη δοκιμασία 1 (σημείο Α)., άρα και έχει μεγαλύτερη διακριτική ικανότητα.

- Κάθε σημείο Γ που ανήκει στο τετράπλευρο IV, ανήκει σε μια τρίτη δοκιμασία με μικρότερα %TP και %FN από τη δοκιμασία 1, άρα έχει μικρότερη διακριτική ικανότητα.

- Τα σημεία Δ και Ε που ανήκουν στα τετράπλευρα II και III αντίστοιχα, μπορεί να ανήκουν στην ίδια δοκιμασία με το σημείο Α, στην οποία έχει μετατοπιστεί το ΔΟ και έχει μεταβληθεί η διαγνωστική της ικανότητα, ή να ανήκει σε διαφορετική δοκιμασία. Σε μια τέτοια περίπτωση, βάση μόνο των %TP και %TN αποτελεσμάτων, δεν είναι ξεκάθαρο ποια δοκιμασία είναι προτιμότερη.



Εικόνα 1.14: *Σύγκριση σημείων καμπύλων ROC με βάση του θετικού (L) και αρνητικού (λ) λόγου πιθανοφανειών.*

Στην Εικόνα 1.14, οι ευθείες $\Theta\Theta'$ και $\Kappa\Kappa'$ ορίζουν τις τέσσερις περιοχές Ι',ΙΙ',ΙΙΙ,ΙV'. Από τη σύγκριση των λόγων πιθανοφάνειας στα σημεία Β, Γ, Δ, Ε που βρίσκονται σε αυτές τις περιοχές με τους λόγους πιθανοφάνειας στο Α, προκύπτουν τα εξής:

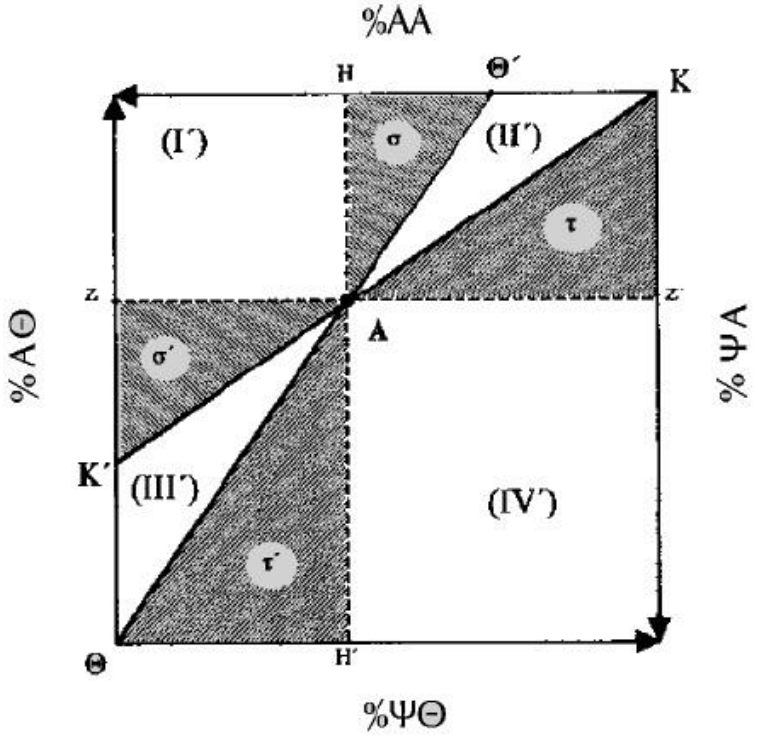
- Στην περιοχή Ι', οι λόγοι πιθανοφανειών είναι $L_{\Β} > L_{\Α}$ και $\lambda_{\Β} < \lambda_{\Α}$. Η διακριτική ικανότητα είναι γενικά μεγαλύτερη από της Α.

- Στην περιοχή ΙΙ', οι λόγοι πιθανοφάνειας είναι $L_{\Delta} < L_{\Α}$ και $\lambda_{\Delta} < \lambda_{\Α}$. Η διακριτική ικανότητα είναι μεγαλύτερη για την επιβεβαίωση της απουσίας του νοσήματος.

- Στην περιοχή ΙΙΙ', οι λόγοι πιθανοφάνειας είναι $L_{\Ε} > L_{\Α}$ και $\lambda_{\Ε} < \lambda_{\Α}$. Η διακριτική ικανότητα είναι γενικά μεγαλύτερη για την επιβεβαίωση της παρουσίας του νοσήματος.

- Στην περιοχή ΙV', οι λόγοι πιθανοφάνειας είναι $L_{\Γ} < L_{A}$ και $\lambda_{\Γ} > \lambda_{\Α}$. Η διακριτική ικανότητα είναι γενικά μικρότερη.

Υπάρχει η περίπτωση μια δοκιμασία να έχει μικρότερο %TN ή %TP σε σχέση με μια άλλη και παρ' όλα αυτά να έχει γενικά μεγαλύτερη διακριτική ικανότητα.



Εικόνα 1.15: *Περιοχές διαπραγμάτευσης του %ΑΘ και %ΑΑ αποτελεσμάτων*

Οι δοκιμασίες αυτές (σαν σημεία) βρίσκονται στις περιοχές σ και σ'. Αντίθετα στις περιοχές τ και τ', η απώλεια σε σχέση με το κέρδος από τη μεταβολή στη διαγνωστική ποιότητα της δοκιμασίας είναι αρκετά μεγάλη, έτσι ώστε η διακριτική ικανότητα αυτής να είναι σημαντικά μικρότερη.

1.4.8 Επιλογή βέλτιστου σημείου απόφασης με βάση την καμπύλη ROC

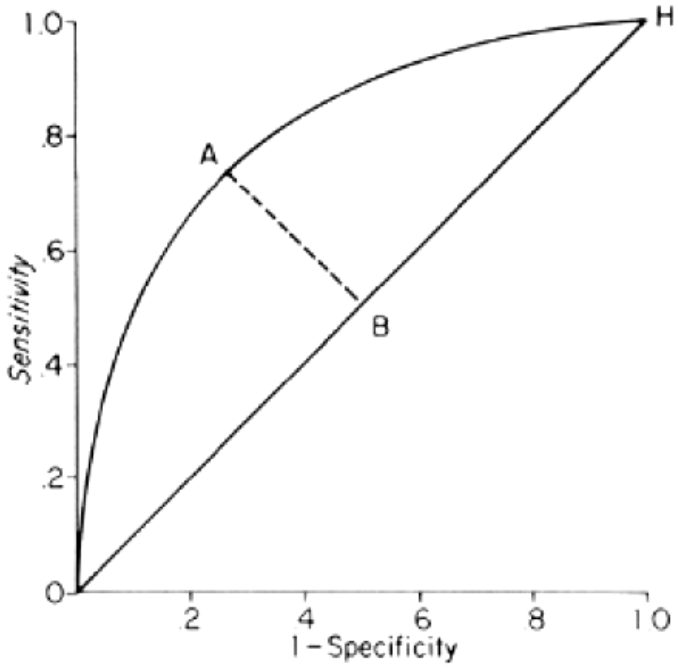
Μια συνεχής καμπύλη ROC είναι το γράφημα της συμμεταβολής των ζευγών των εσφαλμένων θετικών (FP) και των ορθώς θετικών αποφάσεων (TP), για όλα τα σημεία απόφασης. Σε μια ιατρική έρευνα, ενδεχομένως να χρειαστεί να επιλέξουμε ένα συγκεκριμένο σημείο απόφασης ως λειτουργικό σημείο, με βάση κάποιο κριτήριο. Με βάση αυτό το σημείο, θα κατατάσσουμε τους εξεταζόμενους σε υγιείς και ασθενείς με το μικρότερο δυνατό σφάλμα. Για την επιλογή του βέλτιστου σημείου απόφασης, μπορούμε να χρησιμοποιήσουμε την καμπύλη ROC.

Υπάρχουν τρία κριτήρια για την επιλογή, αναλόγως τη φύση του διαγνωστικού ελέγχου:

1. Το αυστηρό κριτήριο, που χρησιμοποιείται και στον βιομηχανικό ποιοτικό έλεγχο, είναι αυτό που επιλέγουμε σημείο απόφασης που αντιστοιχεί σε FP=0.01 ή 0.05. Επιλέγουμε δηλαδή το σφάλμα τύπου I σύμφωνα με τη μεθοδολογία των στατιστικών ελέγχων υποθέσεων. Το αυστηρό κριτήριο, θεωρεί μη αποδεκτές της μετρήσεις στα άκρα της κατανομής των φυσιολογικών μετρήσεων.

2. Το επιεικές κριτήριο επιλέγει το σημείο απόφασης έτσι ώστε το κλάσμα TP να είναι κοντά στη μονάδα. Αυτό το κριτήριο χρησιμοποιείται από τις συσκευές εντοπισμού βλαβών στα αεροσκάφη.

3. Το τρίτο (και πιο δημοφιλές) κριτήριο που χρησιμοποιείται, βρίσκεται κάπου ανάμεσα στα δυο προηγούμενα και βασίζεται στην μεγιστοποίηση των ορθών αποφάσεων, θετικών και αρνητικών. Το σημείο της καμπύλης όπου οι ορθές αποφάσεις μεγιστοποιούνται, δηλαδή το σημείο στο οποίο $TN+TP = \max$ και που αντίστοιχα ελαχιστοποιούνται οι εσφαλμένες αποφάσεις, δηλαδή $FN+FP = \min$, είναι αυτό με τη μέγιστη απόσταση από την κύρια διαγώνιο. Αυτό αντιστοιχεί στο σημείο απόφασης που ορίζεται από το σημείο τομής των συναρτήσεων πιθανότητας των κατανομών των υγειών και των ασθενών.



Εικόνα 1.16: *Βέλτιστο σημείο απόφασης με βάση την καμπύλη ROC*

Έστω ότι οι συντεταγμένες του σημείου Α είναι $(FP,TP)=(1\text{-}TN,TP)$. Το ευθύγραμμο τμήμα ΟΗ έχει κλίση 1 και ικανοποιεί την $y = x$. Το ΑΒ είναι κάθετο στο ΟΗ και ικανοποιεί την $TP=(1\text{-}TN)+\lambda$, όπου λ είναι ο σταθερός όρος. Λύνοντας ως προς το λ, έχουμε $\lambda = TP\text{-}TN+1$.

Το σημείο Β ανήκει και στις δύο ευθείες, άρα οι συντεταγμένες του θα είναι $\left(\frac{TP+1\text{-}TN}{2}, \frac{FP+1\text{-}TN}{2}\right)$ και το μήκος του ΑΒ που είναι

$$\sqrt{\left[(1-TN)-\frac{TP+1\text{-}TN}{2}\right]^2+\left[TP-\frac{TP+1\text{-}TN}{2}\right]^2} = \frac{\sqrt{2}}{2}(TP+TN-1)$$

θα γίνεται μέγιστο όταν η ποσότητα TN+TP γίνεται μέγιστη.

Στην περίπτωση που οι υγιείς και οι ασθενείς προέρχονται από κανονικές κατανομές $N(\mu_1,\sigma_1^2)$ και $N(\mu_2,\sigma_2^2)$ αντίστοιχα, το βέλτιστο σημείο απόφασης, έστω t, δίνεται από τη σχέση:

$$t = \frac{\left(\left(\mu_2\sigma_1^2-\mu_1\sigma_2^2\right)+\sqrt{\mu_2\sigma_1^2-\mu_1\sigma_2^2-\left(\sigma_2^2-\sigma_1^2\right)\left(\mu_1^2\sigma_2^2-\mu_2^2\sigma_1^2\right)-2\sigma_1^2\sigma_2^2\ln\frac{\sigma_2}{\sigma_1}}\right)}{\sigma_2^2-\sigma_1^2}$$

ενώ για $\sigma_1^2=\sigma_2^2$, ισχύει $t=\frac{\mu_1+\mu_2}{2}$.

Για ένα τυχαίο, απλό δείγμα από ένα πληθυσμό, το ποσοστό εσφαλμένων αποφάσεων συνολικά (misclassification rate – MR) δίνεται από τη σχέση: $MR=Prev(1\text{-}TP)+(1\text{-}Prev)(1\text{-}TN)$, με Prev να είναι ο επιπολασμός της ασθένειας στον πληθυσμό (το ποσοστό του πληθυσμού που έχει την ασθένεια τη δεδομένη στιγμή).

Το σημείο απόφασης που ελαχιστοποιεί την παραπάνω συνάρτηση βρίσκεται από την:

$\frac{dMR}{dFP} = -Prev\frac{dTP}{dFP}+(1-Prev)$ και εξισώνοντας με το μηδέν, έχουμε: $\frac{dTP}{dFP}=\frac{1\text{-}Prev}{Prev}$.

Δηλαδή, το βέλτιστο σημείο απόφασης πάνω στην καμπύλη ROC είναι αυτό το οποίο έχει κλίση

$$S=\frac{1\text{-}Prev}{Prev}.$$

Εάν στα ποσοστά των ορθών και εσφαλμένων ποσοστών υπεισέρχονται κόστη και κέρδη, ανάλογα με την απόφαση το σημείο στην ROC με τη βέλτιστη κλίση είναι:

$$S_{opt} = \frac{1-\text{Prev}}{\text{Prev}} \times \frac{B_{TN} - C_{FP}}{B_{TP} - C_{FN}}$$

με B_{TP} και B_{TN} να είναι τα κέρδη από τις ορθές αποφάσεις και C_{FP} και C_{FN} τα κόστη από τις εσφαλμένες αποφάσεις.

Η βέλτιστη κλίση είναι σχετικά μεγάλη και το κριτήριο σχετικά αυστηρό, στην περίπτωση που ο επιπολασμός είναι μικρός ή όταν ο αριθμητής της διαφοράς κέρδους-κόστους είναι μεγαλύτερος από τον παρονομαστή, δηλαδή, όταν η ορθή απόφαση είναι πιο σημαντική στις ορθώς αρνητικές αποφάσεις.

Η βέλτιστη κλίση είναι μικρή και το κριτήριο πιο επιεικές, στην περίπτωση που ο επιπολασμός είναι μεγάλος, ή όταν ο παρονομαστής της διαφοράς κόστους-κέρδους είναι μεγάλος σε σχέση με τον αριθμητή, δηλαδή είναι σημαντικό να εντοπιστεί ορθώς η ασθένεια.

Η επιλογή του σημείου απόφασης μπορεί να γίνει και με βάση το μέτρο πληροφορίας I του Shannon το οποίο ορίζεται ως εξής:

$$I = \text{TP} \cdot \text{Prev} \cdot \log_2 \frac{\text{TP}}{G} + \text{FP} \cdot (1-\text{Prev}) \cdot \log_2 \frac{\text{FP}}{G} + (1-\text{TP}) \cdot \text{Prev} \cdot \log_2 \frac{1-\text{TP}}{1-G}$$
$$+(1-\text{FP}) \cdot (1-\text{Prev}) \cdot \log_2 \frac{1-\text{FP}}{1-G}$$

με $G = \text{TP} \cdot \text{Prev} + \text{FP} \cdot (1-\text{Prev})$. Το σημείο απόφασης είναι αυτό που μεγιστοποιεί την πληροφορία I.

1.4.9 Το πρόβλημα διαχωρισμού σε τρεις κλάσεις – επιφάνεια ROC

Υπάρχουν προβλήματα στα οποία ο διαχωρισμός του δείγματος πρέπει να γίνει σε τρεις ομάδες αντί του συνηθισμένου διαχωρισμού σε δύο (όπως ασθενής – υγιής). Τέτοιο πρόβλημα είναι, για παράδειγμα, η εξέταση για καρκίνο η οποία κατηγοριοποιεί τα αποτελέσματα στις κατηγορίες: υγιείς, καλοήθης όγκος, κακοήθης όγκος. Στο πρόβλημα της τριχοτόμησης, μπορούμε να θεωρήσουμε κάθε μια από τις τρεις αποφάσεις, σαν τις κορυφές ενός ισόπλευρου τριγώνου. Το σημείο απόφασης μπορεί να βρίσκεται κάπου μέσα στο τρίγωνο αυτό και χωρίζει το επίπεδο σε τρία μέρη, σύμφωνα με τα οποία κατατάσσονται οι μετρήσεις σε τρείς ομάδες. Σε αυτήν την περίπτωση, χρειάζεται ένας διαγνωστικός κανόνας που αποτελείται από το συνδυασμό δύο ελέγχων, για τον ορισμό των τριών περιοχών του τριγώνου.

Η παρακάτω μέθοδος αναπαριστά γραφικά την ευαισθησία (TPF) ως προς την 1-ειδικότητα (FPF) και τη συνάρτηση της ασαφούς κατάστασης (IDF, Indeterminate fraction) συγχρόνως. Έστω X,Y,Z οι μετρήσεις του διαγνωστικού test για τις τρεις

κλάσεις και (c_0, s_0) τα κριτήρια απόφασης που μεγιστοποιούν την ακρίβεια της πρόβλεψης. Τότε:

$$TPR = P\left[Y > c_0 \text{ και } Z > s_0 \text{ / δείγμα απο ασθενείς}\right] = P\left[Y > c_0 / Z > s_0\right] P\left[Z > s_0\right]$$

$$FPR = P\left[X > c_0 \text{ και } Z > s_0 \text{ / δείγμα απο ασθενείς}\right] = P\left[X > c_0 / Z > s_0\right] P\left[Z > s_0\right]$$

$$IDF = P\left[Z \le s_0\right]$$

όπου $0 \le FPF \le P\left[Z > s_0\right] = 1 - IDF$

Για κάθε ορισμό της περιοχής απόφασης, υπάρχουν τρία ποσοστά ορθών αποφάσεων και έξι ποσοστά εσφαλμένων αποφάσεων. Εάν μεταβάλλουμε τον κανόνα ορισμού των περιοχών του τριγώνου, που είναι ανάλογος με το διαχωριστικό όριο στην περίπτωση της καμπύλης ROC, τα τρία ποσοστά ορθών αποφάσεων διαγράφουν τη λεγόμενη επιφάνεια ROC στο μοναδιαίο κύβο.

Ο όγκος κάτω από την επιφάνεια ROC είναι ένας δείκτης της ποιότητας του διαχωρισμού των τριών ομάδων, σύμφωνα με τον υπό μελέτη διαγνωστικό κανόνα και μεταβάλλεται από 1/6 (που είναι ο τυχαίος διαχωρισμός) έως και 1 (που είναι ο τέλειος διαχωρισμός).

1.4.10 Χρήση λογιστικής παλινδρόμησης για την εκτίμηση μιας ROC καμπύλης

Θεωρούμε τις $X_1, X_2, ..., X_n$ ανεξάρτητες μεταξύ τους μετρήσεις των υγειών από μια συνεχή συνάρτηση κατανομής F και αντίστοιχα, τις $Y_1, Y_2, ..., Y_n$ ανεξάρτητες μετρήσεις των ασθενών από μια συνεχή συνάρτηση κατανομής G. Τότε, η ROC καμπύλη είναι το γράφημα του:

$$R(s) = 1 - F[G^{-1}(1-s)] \text{ για } s \in [0,1]$$

Το εμβαδό κάτω από την καμπύλη ROC ορίζεται ως:

$$A = \int_0^1 R(s)ds = P(Y_i > X_i).$$

Μια εναλλακτική, ημιπαραμετρική προσέγγιση για την εκτίμηση του $R(s)$, είναι να μοντελοποιήσουμε την πιθανότητα εμφάνισης ασθένειας για ένα συγκεκριμένο αποτέλεσμα και να μην μοντελοποιήσουμε απευθείας τις σ.κ. F και G

για τα αποτελέσματα του test, δεδομένης της κατάστασης του ασθενή. Σε αυτήν την προσέγγιση, ένα μοντέλο λογιστικής παλινδρόμησης εξηγεί καλά τα αποτελέσματα.

Έστω ότι $D=1$ υποδεικνύει την κατάσταση ενός ασθενούς και $D=0$ την κατάσταση ενός υγιούς. Για ένα συγκεκριμένο αποτέλεσμα του test με $X=x$, το μοντέλο λογιστικής παλινδρόμησης είναι:

$$P\left(D=1\middle|X=x\right)=\frac{\exp\left\{\alpha+\beta^T r(x)\right\}}{1+\left\{\exp\alpha+\beta^T r(x)\right\}} \qquad (1)$$

όπου α μια βαθμιδωτή παράμετρος, β ένα $p\times 1$ διάνυσμα παράμετρος και $r(x)$ ένα $p\times 1$ διάνυσμα συνάρτηση του x. Τότε, $F(x)=P\left(X\le x\middle|D=1\right)$ και $G(x)=P\left(X\le x\middle|D=0\right)$. Εάν $f(x),\ g(x)$ οι αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας, ισχύει ότι:

$$\frac{f(x)}{g(x)}=\exp\left\{\alpha^*+\beta^T r(x)\right\}$$

όπου $\alpha^*=\alpha+\log\left[\left\{1-P(D=1)\right\}/P(D=1)\right]$.

Αν $F\sim N(\mu_1,\sigma_1^2)$ και $G\sim N(\mu_2,\sigma_2^2)$, το μοντέλο (1) ικανοποιείται για $r(x)=x$.

Η αθροιστική κατανομή του test δίνεται από τις παρακάτω σχέσεις:

$$P(X<x\middle|D=0)=\frac{e^x}{1+e^x} \qquad \text{για τον υγιή πληθυσμό}$$

και (2)

$$P(X<x\middle|D=1)=\frac{e^{-\alpha+\beta x}}{1+e^{-\alpha+\beta x}} \qquad \text{για τον ασθενή πληθυσμό}$$

Αν το διαχωριστικό όριο του test είναι λ, τότε από το μοντέλο (2), η πιθανότητα ενός ψευδώς θετικού (FP) αποτελέσματος είναι $\frac{1-e^{\lambda}}{1+e^{\lambda}}$. Δηλαδή, $\text{logit}(FPR)=-\lambda$.
Αντίστοιχα, $\text{logit}(TPR)=\alpha-\beta\lambda$.

1.4.11 Η καμπύλη ROC ως τυχαίος περίπατος

Για δείγματα μεγέθους n και m που προέρχονται από υγιείς και ασθενείς αντίστοιχα, η εμπειρική καμπύλη ROC που σχεδιάζεται από αυτά, αντιστοιχεί στον τυχαίο περίπατο από το σημείο (1,1) προς το (0,0) με βήμα $\frac{1}{m}$ προς τα κάτω όταν η

επόμενη, από τις διατεταγμένες, μέτρηση προέρχεται από το δείγμα των ασθενών και βήμα $\frac{1}{n}$ προς τα αριστερά, όταν η επόμενη μέτρηση προέρχεται από το δείγμα των υγιών.

Παράδειγμα:

Έστω οι μετρήσεις 2, 3, 5 να είναι το δείγμα των υγιών με n=3 και έστω οι μετρήσεις 1, 4, 7, 8 το δείγμα των ασθενών με m=4. Κατασκευάζουμε την εμπειρική ROC με των παρακάτω τρόπο:

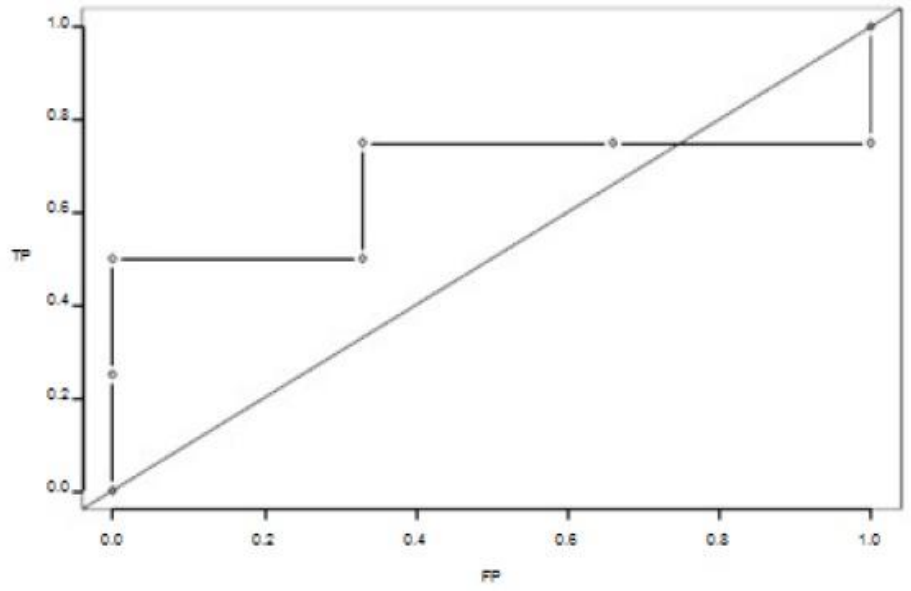
- Διατάσουμε όλες τις μετρήσεις: 1, **2, 3**, 4, **5**, 6, 7, 9 (οι υγιείς είναι οι έντονες μετρήσεις).
- Παίρνουμε από ένα σημείο απόφασης μεταξύ κάθε ζεύγους μετρήσεων και δύο επιπλέον σημεία, ένα πριν και ένα μετά από όλες τις μετρήσεις:

$$c = (0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 7.5, 9.5).$$

- Με αυτόν τον τρόπο, ορίζονται τα ζεύγη των εσφαλμένα θετικών και ορθώς θετικών ποσοστών (FP,TP) για το διάνυσμα c των σημείων απόφασης:

$$\{(1,1),(1,0.75),(0.66,0.75),(0.33,0.75),(0.33,0.5),(0,0.25),(0,0)\}$$

- Η πολυγωνική γραμμή που ορίζουν τα παραπάνω ζεύγη (FP,TP) είναι η εμπειρική καμπύλη ROC που προκύπτει:



***Εικόνα 1.17**: Παράδειγμα εμπειρικής καμπύλης ROC*

Η καμπύλη ROC της παραπάνω εικόνας, είναι ένας τυχαίος περίπατος από το σημείο (1,1) στο (0,0) με βήμα 1/4 προς τα κάτω, αν η επόμενη μέτρηση είναι από ασθενή, και βήμα 1/3 όταν η επόμενη μέτρηση είναι από υγιή.

1.4.12 Σύνδεση της καμπύλης ROC με το PP-plot

Ένα PP-Plot είναι το γράφημα των percentiles μιας κατανομής F_0 έναντι των percentiles μιας άλλης κατανομής F_1 με το ίδιο πεδίο ορισμού. Η συναρτησιακή μορφή του PP-Plot είναι:

$$\mathrm{PP}(p) = F_1\left(F_0^{-1}(p)\right) \quad , \quad 0 < p < 1$$

Όταν έχουμε δείγματα μεγέθους n και m από τις F_0 , F_1 αντίστοιχα, τότε το PP-Plot γράφεται στη μορφή:

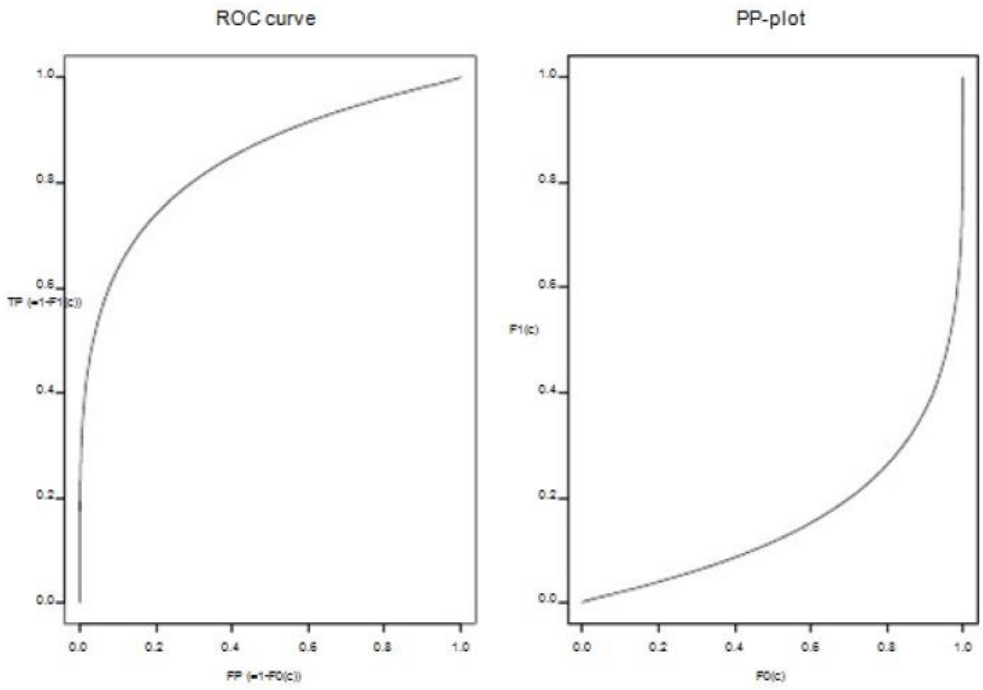
$$\mathrm{PP}_{mn}(p) = F_{1m}\left(F_{0n}^{-1}(p)\right) \quad , \quad 0 < p < 1$$

με F_{0n} , F_{1m} να είναι οι εμπειρικές κατανομές που προκύπτουν από τα δύο δείγματα. Οι δυο κατανομές F_0 και F_1 συμπίπτουν όταν το PP-Plot συμπίπτει με την κύρια διαγώνιο στο τετράγωνο $[0,1]\times[0,1]$, ενώ διαφέρουν όταν αυτό απέχει από την κύρια διαγώνιο.

Η συναρτησιακή μορφή της καμπύλης ROC που αντιστοιχεί στις κατανομές F_0 , F_1 είναι:

$$\mathrm{ROC}(p) = 1 - F_1\left(F_0^{-1}(1-p)\right) \quad , \quad 0 < p < 1.$$

Από τις δυο συναρτήσεις $\mathrm{PP}(p)$ και $\mathrm{ROC}(p)$ συμπεραίνουμε ότι η καμπύλη ROC είναι ένα PP-Plot με αντίθετες διαβαθμίσεις στους άξονες. Δηλαδή, το σημείο 0 στον άξονα των $x'x$ στο PP-Plot είναι το σημείο 1 του άξονα των $x'x$ στην καμπύλη ROC. Αντίστοιχα στον άξονα των $y'y$.



Εικόνα 1.18: *Καμπύλη ROC και PP-Plot που αντιστοιχούν σε δείγματα 1000 μετρήσεων υγιών και ασθενών από κατανομές* $N(0,1)$ *και* $N(1.8,1.5^2)$.

Για ένα δείγμα μεγέθους n το QQ-Plot είναι το γράφημα των ταξινομημένων τιμών του δείγματος έναντι των quantiles της υπό έλεγχο κατανομής που αντιστοιχούν στις τιμές $\frac{i-c}{n-2c+1}$, όπου c μια σταθερά με $0 \le c \le 1$ (συνήθως επιλέγουμε $c=0$ και $c=0.5$).

Λόγω της σχέσης των καμπύλων ROC με τα PP-Plot, οι ROC μπορούν να χρησιμοποιηθούν για την εκτίμηση της κατανομής προέλευσης ενός δείγματος και έχουν καλύτερα αποτελέσματα από τα QQ-Plots τα οποία και συνήθως χρησιμοποιούνται για αυτό το σκοπό. Οι απλοί στατιστικοί έλεγχοι με βάση το εμβαδό κάτω από την καμπύλη ROC, μπορούν να χρησιμοποιηθούν για τον τυπικό υπολογισμό της σημαντικότητας της απόκλισης του δείγματος από την υπό έλεγχο κατανομή και για αυτό το λόγο τα PP-Plots είναι αρκετά χρήσιμα.

1.5 Η μέθοδος Cross – Validation

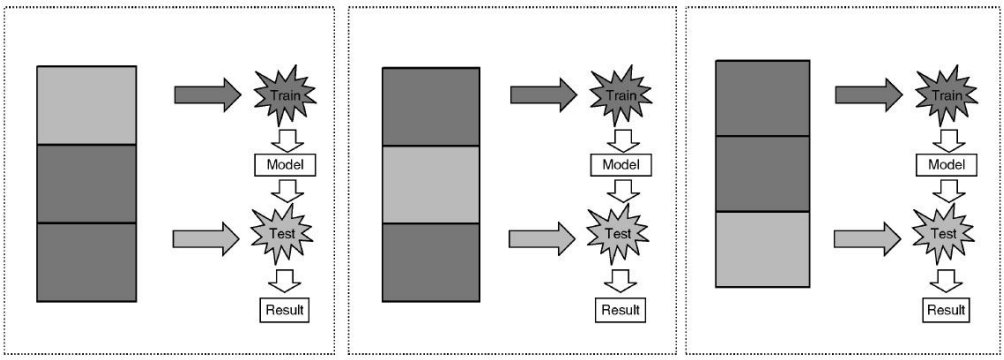
Η μέθοδος Cross-validation είναι μια στατιστική μέθοδος για την αξιολόγηση και σύγκριση αλγορίθμων μάθησης, η οποία χωρίζει τα δεδομένα σε δύο σύνολα: ένα για τη μάθηση ή την εκπαίδευση ενός μοντέλου και το άλλο για την επικύρωση του μοντέλου. Στην τυπική cross-validation μέθοδο, τα σύνολα εκπαίδευσης και

επικύρωσης εναλλάσσονται διαδοχικά ώστε κάθε στοιχείο από τα δεδομένα να έχει πιθανότητα να επικυρωθεί.

1.5.1 Η μέθοδος της k-fold Cross-Validation

Η βασική μορφή της μεθόδου cross-validation είναι η cross-validation σε k-πτυχές (k-fold cross validation). Άλλες μορφές της μεθόδου cross-validation είναι ειδικές περιπτώσεις της k-fold cross validation ή η k-fold cross validation σε επαναλαμβανόμενους γύρους.

Στην cross-validation σε k-πτυχές, τα δεδομένα κατ' αρχήν χωρίζονται σε k κομμάτια, ή αλλιώς σε k πτυχές, ίσου (ή σχεδόν ίσου) μεγέθους. Οι πτυχές κατασκευάζονται ύστερα από «στρωμάτωση» (stratification) των δεδομένων, δηλαδή τοποθέτηση των δεδομένων με τρόπο που κάθε πτυχή θα είναι αντιπροσωπευτική του αρχικού συνόλου των δεδομένων. Στη συνέχεια, k υπολογισμοί για εκπαίδευση και επικύρωση πραγματοποιούνται έτσι ώστε σε κάθε υπολογισμό μια πτυχή να χρησιμοποιείται για επικύρωση και οι υπόλοιπες k-1 πτυχές να χρησιμοποιούνται για την εκπαίδευση. Στο Σχήμα 1.19 φαίνεται ένα παράδειγμα cross validation για k= 3. Η πιο συνηθισμένη cross-validation μέθοδος η οποία χρησιμοποιείται στην διαδικασία εξόρυξης δεδομένων (data mining) και στη μηχανική μάθηση (machine learning) είναι η 10-πτυχη cross-validation (k=10).



Σχήμα 1.19: *Η διαδικασία της 3-fold cross-validation*

Η μέθοδος cross-validation αξιολογεί και συγκρίνει αλγόριθμους μάθησης με τον παρακάτω τρόπο: Σε κάθε υπολογισμό, ένας ή περισσότεροι αλγόριθμοι μάθησης χρησιμοποιούν k-1 πτυχές για να μάθουν ένα ή περισσότερα μοντέλα και στη συνέχεια αυτά τα μοντέλα κάνουν προβλέψεις για τα δεδομένα στην πτυχή που έχει μείνει. Η απόδοση κάθε αλγόριθμου μάθησης σε κάθε πτυχή παρακολουθείται από κάποιο προκαθορισμένο μέτρο απόδοσης, όπως για παράδειγμα η ευστοχία. Με την ολοκλήρωση της μεθόδου, k δείγματα απόδοσης θα είναι διαθέσιμα για κάθε αλγόριθμο. Διαφορετική μεθοδολογία, όπως η χρήση του μέσου όρου, μπορεί να χρησιμοποιηθεί για να υπολογιστεί ένα συνολικό μέτρο για κάθε δείγμα ή τα

δείγματα μπορούν να υποστούν έναν στατιστικό έλεγχο υποθέσεως για να διαπιστωθεί ποιος αλγόριθμος είναι ανώτερος.

1.5.2 Η επαναλαμβανόμενη k-fold Cross-Validation

Για να υπάρξει ένα αξιόπιστο συμπέρασμα σχετικά με την απόδοση ενός αλγόριθμου, είναι επιθυμητή η ύπαρξη ενός μεγάλου αριθμού εκτιμήσεων. Η απλή cross validation με k-πτυχές, αποδίδει μόνο k το πλήθος εκτιμήσεις. Μια αρκετά διαδεδομένη μέθοδος για να εξασφαλιστεί μεγαλύτερο πλήθος εκτιμήσεων είναι να επαναληφθεί η k-fold cross-validation πολλαπλές φορές, με τα δεδομένα να ανακατεύονται και να επαναστρωματώνονται πριν από κάθε επανάληψη.

1.5.3 Η 10-fold Cross – Validation

Η επιλογή της τιμής για τον αριθμό k των πτυχών είναι πολύ σημαντική. Ένας μεγάλος αριθμός για το k μπορεί να φαίνεται καλός καθώς πρώτον, αποδίδει πολλές εκτιμήσεις και δεύτερον, το σύνολο εκπαίδευσης είναι σε μέγεθος πιο κοντά στο συνολικό δείγμα και άρα τα συμπεράσματα που αφορούν τον αλγόριθμο εκπαίδευσης πιθανώς να μπορούν να γενικευτούν πιο εύκολα στην υπόθεση ότι για την εκπαίδευση του μοντέλου μάθησης χρησιμοποιήθηκαν όλα τα δεδομένα. Ωστόσο, όταν το k αυξάνεται, η επικάλυψη ανάμεσα στα σύνολα εκπαίδευσης αυξάνεται επίσης. Για παράδειγμα, στην 5-fold cross-validation, κάθε σύνολο εκπαίδευσης μόνο τα 3/4 των περιπτώσεων με τα υπόλοιπα σύνολα εκπαίδευσης, ενώ με την 10-fold cross-validation, το κάθε σύνολο εκπαίδευσης μοιράζεται τα 8/9 των περιπτώσεων με τα υπόλοιπα 8 σύνολα εκπαίδευσης. Επίσης, η αύξηση του k μειώνει το μέγεθος του συνόλου ελέγχου και έτσι μειώνεται η ευστοχία του μέτρου απόδοσης, όπως αυτό έχει προκαθοριστεί. Για παράδειγμα, σε ένα σύνολο ελέγχου με 10 περιπτώσεις, η ευστοχία μπορεί να μετρηθεί στο 10%, ενώ σε ένα σύνολο με 20 περιπτώσεις, η ευστοχία μπορεί να μετρηθεί στο 5%.

Όλα τα παραπάνω έχουν τεθεί υπ’ όψη της επιστημονικής κοινότητας που ασχολείται με το data mining, η οποία έχει αποφανθεί πως το k=10 είναι μια καλή τιμή για την k-fold cross-validation μέθοδο. Για την συγκεκριμένη τιμή, οι προβλέψεις γίνονται με χρήση του 90% των δεδομένων και άρα είναι πιο πιθανό να μπορούν να γενικευτούν στο σύνολο των δεδομένων.

1.6 Η μέθοδος Bootstrapping

Το bootstrapping είναι μια γενική προσέγγιση στην στατιστική συμπερασματολογία η οποία βασίζεται στην κατασκευή μιας κατανομής δειγματοληψίας για ένα στατιστικό, μέσω της επανα-δειγματοληψίας των υπαρχόντων δεδομένων.

1.6.1 Η βασική ιδέα της μεθόδου Bootstrapping

Έστω ότι εξάγουμε ένα δείγμα $S = \{X_1, X_2, ..., X_n\}$ από ένα πληθυσμό $p = \{x_1, x_2, ..., x_N\}$. Υποθέτουμε ότι το N είναι πολύ μεγαλύτερο του n και ότι το S είναι είτε ένα απλό τυχαίο δείγμα από το P. Ας υποθέσουμε επίσης ότι ενδιαφερόμαστε για ένα στατιστικό $T = t(S)$ σαν μια εκτίμηση της αντίστοιχης παραμέτρου $\theta = t(P)$ για τον πληθυσμό. Το θ μπορεί να είναι και ένα διάνυσμα παραμέτρων και το T να είναι το αντίστοιχο διάνυσμα-εκτίμηση. Η παραδοσιακή προσέγγιση στην στατιστική συμπερασματολογία είναι να κάνουμε υποθέσεις για τη δομή του πληθυσμού και συναρτήσει αποτελεσμάτων τυχαίας δειγματοληψίας να χρησιμοποιήσουμε αυτές τις υποθέσεις για να εξάγουμε την κατανομή δειγματοληψίας για το T. Υπάρχουν περιπτώσεις όπου η ακριβής κατανομή δεν μπορεί να προσδιοριστεί και προσδιορίζουμε την ασυμπτωτική της κατανομή. Αυτή η προσέγγιση έχει δυο ενδεχομένως σοβαρά ελαττώματα:

- Εάν οι υποθέσεις για τον πληθυσμό είναι λαθεμένες, τότε και η αντίστοιχη κατανομή δειγματοληψίας για το στατιστικό μπορεί να είναι σοβαρά λαθεμένη. Από την άλλη, εάν βασιζόμαστε σε ασυμπτωτικά αποτελέσματα, αυτά μπορεί να μην ανταποκρίνονται στον απαιτούμενο βαθμό ακρίβειας σε ένα σχετικά μικρό δείγμα.

- Αυτή η μέθοδος είναι μαθηματικά αρκετά απαιτητική για να εξαχθεί η κατανομή δειγματοληψίας του στατιστικού που μας ενδιαφέρει.

Από την άλλη, η η-παραμετρική μέθοδος bootstrap μας επιτρέπει να εκτιμήσουμε τν κατανομή δειγματοληψίας ενός στατιστικού εμπειρικά, χωρίς να χρειαστεί να κάνουμε υποθέσεις σχετικά με τη μορφή του πληθυσμού και χωρίς να εξάγουμε την κατανομή της δειγματοληψίας του με σαφήνεια. Η βασική ιδέα πίσω από την μη-παραμετρική μέθοδο bootstrap είναι η ακόλουθη:

Εξάγουμε ένα δείγμα μεγέθους n από τα στοιχεία του S με αντικατάσταση. Θέτουμε αυτό το bootstrap δείγμα που προέκυψε σαν $S_1^* = \{X_{11}^*, X_{12}^*, ..., X_{1n}^*\}$. Είναι απαραίτητο να κάνουμε τη δειγματοληψία με αντικατάσταση διότι, διαφορετικά, θα αναπαράγαμε το αρχικό δείγμα S. Στην πραγματικότητα, αντιμετωπίζουμε το δείγμα S σαν μια εκτίμηση του πληθυσμού P. Δηλαδή, κάθε στοιχείο X_i του S

επιλέγεται από το δείγμα bootstrap με πιθανότητα 1/n , ακριβώς σαν την αρχική επιλογή του δείγματος S από τον πληθυσμό P. Επαναλαμβάνουμε αυτή τη διαδικασία, για ένα μεγάλο αριθμό επαναλήψεων R, επιλέγοντας πολλά bootstrap δείγματα. Το b -οστό τέτοιο bootstrap δείγμα συμβολίζεται $S_b^* = \left\{ X_{b1}^*, X_{b2}^*, ..., X_{bn}^* \right\}$.

Η ιδανική bootstrap αναλογία συνοψίζεται ως εξής:

Ο πληθυσμός είναι για το δείγμα, ότι είναι το δείγμα για τα bootstrap δείγματα.

Στη συνέχεια, υπολογίζουμε το στατιστικό T για κάθε ένα από τα bootstrap δείγματα: $T_b^* = t\left(S_b^*\right)$. Η κατανομή του T_b^* γύρω από την αρχική εκτίμηση T είναι ανάλογη της κατανομής δειγματοληψίας του εκτιμητή T γύρω από την παράμετρο πληθυσμού θ. Για παράδειγμα, το μέσο των bootstrapped στατιστικών

$$\bar{T}^* = \hat{E}^*\left(T^*\right) = \frac{\sum_{b=1}^{R} T_b^*}{R}$$

εκτιμά την αναμενόμενη τιμή των bootstrapped στατιστικών και το $\hat{B}^* = \bar{T}^* - T$ είναι μια εκτίμηση της μεροληψίας (bias) του T.

Αντίστοιχα, η εκτιμούμενη διακύμανση του T^* είναι

$$\hat{V}^*(T^*) = \frac{\sum_{b=1}^{R} (T_b^* - \bar{T}^*)^2}{R-1}$$

εκτιμά την διακύμανση δειγματοληψίας του T.

Η τυχαία επιλογή bootstrap δειγμάτων δεν είναι υποχρεωτική στην μη-παραμετρική μέθοδο bootstrap. Κατ' αρχήν, μπορούμε να απαριθμήσουμε όλα τα bootstrap δείγματα μεγέθους n. Έπειτα, μπορούμε να υπολογίσουμε τα $E^*(T^*)$ και $V^*(T^*)$ επακριβώς, χωρίς να χρειαστεί να τα εκτιμήσουμε. Ωστόσο, ο αριθμός των bootstrap δειγμάτων είναι αστρονομικά μεγάλος, εκτός εάν n είναι πολύ μικρό. Επομένως, υπάρχουν δύο πηγές λαθών στην συμπερασματολογία με τη μέθοδο bootstrap: Πρώτον, το σφάλμα που προκαλείται από την χρήση συγκεκριμένου δείγματος S για την αντιπροσώπευση του πληθυσμού και δεύτερον, το σφάλμα δειγματοληψίας που προκαλείται από την αποτυχία απαρίθμησης όλων των bootstrap δειγμάτων. Το δεύτερο σφάλμα μπορεί να ελεγχθεί με το να ορίσουμε τον αριθμό των bootstrap επαναλήψεων R επαρκώς μεγάλο.

1.6.2 Διαστήματα εμπιστοσύνης για τη μέθοδο Bootstrapping

Υπάρχουν διάφορες προσεγγίσεις για την κατασκευή διαστημάτων εμπιστοσύνης για τη μέθοδο bootstrap. Η θεωρία κανονικών διαστημάτων (normal-theory interval)

υποθέτει ότι το στατιστικό T κατανέμεται κανονικά (υπόθεση που προσεγγίζει συχνά τις περιπτώσεις στατιστικών σε επαρκώς μεγάλα δείγματα), και χρησιμοποιεί την bootstrap εκτίμηση της διακύμανσης δειγματοληψίας, ίσως και της μεροληψίας, για την κατασκευή 100(1-α) % διαστήματα εμπιστοσύνης της μορφής:

$$\theta = (T - \hat{B}^*) \pm z_{1-a/2} SE^*(T^*)$$

Το $SE^*(T^*) = \sqrt{\hat{V}^*(T^*)}$ είναι η εκτίμηση bootstrap του τυπικού σφάλματος του T και $z_{1-a/2}$ είναι το 1-α/2 ποσοστιαίο σημείο της τυπικής κανονικής κατανομής.

Μια εναλλακτική προσέγγιση που ονομάζεται bootstrap διαστήματα ποσοστιαίων σημείων (bootstrap percentile interval) χρησιμοποιεί τα εμπειρικά ποσοστιαία σημεία του T_b^* για να σχηματιστεί ένα διάστημα εμπιστοσύνης του θ:

$$T^*_{(\text{lower})} < \theta < T^*_{(\text{upper})}$$

,όπου $T^*_{(1)}, T^*_{(2)}, ..., T^*_{(R)}$ είναι οι bootstrap επαναλήψεις του στατιστικού, με $\text{lower} = [(R+1)a/2]$ και $\text{upper} = [(R+1)(1-a/2)]$, και οι αγκύλες συμβολίζουν τη στρογγυλοποίηση στον πλησιέστερο ακέραιο. Για παράδειγμα, εάν $a = .05$ και αντιστοιχεί σε ένα 95% διάστημα εμπιστοσύνης και $R = 999$, τότε lower=25 και upper=975.

Παρ' όλο που τεχνητά δεν υποθέτουν κανονικότητα, τα ποσοστιαία διαστήματα εμπιστοσύνης συχνά δεν αποδίδουν ικανοποιητικά. Προτιμώνται τα λεγόμενα BC_a (bias-corrected, accelerated percentile intervals).

Για να βρούμε το διάστημα BC_a για το θ ακολουθούμε τα παρακάτω βήματα:

- Υπολογίζεται το

$$z = \Phi^{-1}\frac{\#_{b=1}^{R}(T_b^* \le T)}{R+1}$$

, όπου $\Phi^{-1}(\cdot)$ είναι η τυπική κανονική ποσοστιαία συνάρτηση και $\#(T_b^* \le T)/(R+1)$ είναι η προσαρμοσμένη αναλογία των bootstrap επαναλήψεων πάνω στην ή κάτω από τν αρχική εκτίμηση T του θ. Εάν η bootstrap κατανομή δειγματοληψίας είναι συμμετρική και εάν το T είναι αμερόληπτο, τότε η αναλογία θα είναι κοντά στο .5 και ο «συντελεστής διόρθωσης» z θα είναι κοντά στο 0.

- Έστω ότι το $T_{(-i)}$ αναπαριστά την τιμή του T που παράγεται όταν η i-οστή παρατήρηση σβήνεται από το δείγμα (υπάρχουν n τέτοια ποσοστιαία σημεία). Έστω ότι το $\bar{T}$ αναπαριστά το μέσο $T_{(-i)}$, δηλαδή $\bar{T} = \sum_{i=1}^{n} T_{(-i)}/n$. Τότε υπολογίζουμε το

$$a = \frac{\sum_{i=1}^{n} (T_{(-i)} - \bar{T})^3}{6\left[\sum_{i=1}^{n} (T_{(-i)} - \bar{T}^2\right]^{\frac{3}{2}}}$$

- Χρησιμοποιώντας τους συντελεστές διόρθωσης z και a, υπολογίζουμε τις ποσότητες:

$$a_1 = \Phi\left[z + \frac{z - z_{1-a/2}}{1 - a(z - z_{1-a/2})} \right]$$

και

$$a_2 = \Phi\left[z + \frac{z + z_{1-a/2}}{1 - a(z + z_{1-a/2})} \right]$$

, όπου $\Phi(\cdot)$ είναι η τυπική κανονική αθροιστική συνάρτηση κατανομής. Οι τιμές a_1 και a_2 χρησιμοποιούνται για να προσδιοριστούν τα τελικά σημεία (endpoints) του διορθωμένου εκατοστιαίου διαστήματος εμπιστοσύνης:

$$T^*_{(\text{lower}^*)} < \theta < T^*_{(\text{upper}^*)}$$

, όπου

$\text{lower}^* = [R_{a_1}]$ και $\text{upper}^* = [R_{a_2}]$. Όταν οι συντελεστές διόρθωσης a και z είναι και οι δύο μηδέν, τότε $a_1 = \Phi(-z_{1-a/2}) = \Phi(z_{a/2}) = a/2$ και $a_2 = \Phi(z_{1-a/2}) = 1 - a/2$ που αντιστοιχούν εκατοστιαίο διάστημα πριν την διόρθωση.

Για να αποκομίσουμε επαρκώς ακριβή 95% bootstrap εκατοστιαία σημεία ή BC_a διαστήματα εμπιστοσύνης, ο αριθμός R των bootstrap δειγμάτων πρέπει να είναι της τάξης των χιλιάδων ή περισσότερο. Για τα normal-theory bootstrap, μπορούμε να χρησιμοποιήσουμε και μικρότερο αριθμό R, της τάξης των εκατοντάδων ή περισσότερο, αφού το μόνο που χρειάζεται να κάνουμε είναι να εκτιμήσουμε το τυπικό σφάλμα του στατιστικού.

Κεφάλαιο 2:

2.1 Εισαγωγή

Υπάρχει μεγάλο ερευνητικό ενδιαφέρον για την σωστή πρόβλεψη της ασθένειας σε ανθρώπους που εμφανίζουν κάποια συμπτώματα ή για την πρόβλεψη της πιθανότητας να συμβούν αρνητικά περιστατικά σε ασθενείς που έχουν εισαχθεί για φαρμακευτική ή χειρουργική θεραπεία. Η πρόβλεψη τέτοιων περιστατικών μπορεί να βοηθήσει τους ασθενείς με το να υποβάλλονται σε έγκαιρη, καλύτερη, πιο αποτελεσματική θεραπεία.

Η λογιστική παλινδρόμηση είναι η πιο συνηθισμένη μέθοδος για την πρόβλεψη τέτοιων φαινομένων. Αρκετά διαδεδομένη είναι και η χρήση των δέντρων ταξινόμησης/παλινδρόμησης (CART) για την πρόβλεψη περιστατικών ή και για την πρόβλεψη συγκεκριμένων διαγνώσεων. Οι αλγόριθμοι CART θεωρείται ότι μπορούν να δώσουν απλές, κατανοητές οδηγίες απόφασης, οι οποίες μπορούν εύκολα να εφαρμοστούν στην πράξη. Επιπλέον, οι αλγόριθμοι CART, είναι ιδιαίτερα χρήσιμοι στο να μπορούν να αναγνωρίσουν της αλληλεπιδράσεις μεταξύ δεδομένων και να κατηγοριοποιούν το δείγμα σε ομάδες υψηλού και χαμηλού κινδύνου.

Συγκρίσεις της προβλεπτικής ικανότητας της λογιστικής παλινδρόμησης και των δέντρων ταξινόμησης/παλινδρόμησης έχουν διεξαχθεί σε διάφορες μελέτες. Αυτές οι μελέτες μπορούν να ταξινομηθούν στις παρακάτω τρεις κατηγορίες:

- Στις μελέτες που σύγκριναν τις μεταβλητές που έδωσε η λογιστική παλινδρόμηση σαν σημαντικούς predictors για την εμφάνιση περιστατικού και στις μεταβλητές που έδωσαν τα δέντρα παλινδρόμησης σαν σημαντικούς predictors.
- Στις μελέτες που σύγκριναν την ευαισθησία (sensitivity) και την ειδικότητα (specificity) ανάμεσα στη λογιστική παλινδρόμηση και στα δέντρα παλινδρόμησης.
- Στις μελέτες που σύγκριναν την προβλεπτική ικανότητα των δυο μεθόδων, με βάση το εμβαδό κάτω από την καμπύλη ROC που παράγεται από κάθε μια από αυτές.

Στην πρώτη περίπτωση, η μέθοδος συγκρίνει εάν υπάρχει συμφωνία ανάμεσα στις δυο μεθόδους για το ποιες μεταβλητές είναι προγνωστικά σημαντικές. Ωστόσο, επειδή κάθε μέθοδος αξιοποιεί τις μεταβλητές με διαφορετικό τρόπο, η συμφωνία ανάμεσα στην σημαντικότητα των μεταβλητών δεν σημαίνει απαραίτητα ότι οι μέθοδοι δεν μπορεί να έχουν διαφορετική προβλεπτική ικανότητα.

Στη δεύτερη περίπτωση, για να υπολογιστούν η ευαισθησία και η ειδικότητα ενός μοντέλου λογιστικής παλινδρόμησης, πρέπει πρώτα να οριστεί ένα όριο πιθανότητας και να υποθέσουμε ότι μια απόκριση θεωρείται θετική εάν η προβλεπόμενη

πιθανότητά της ξεπεράσει αυτό το όριο. Αυτή η μέθοδος είναι εξαρτημένη από το όριο το οποίο θα επιλεγεί και επιπλέον είναι μη-ευαίσθητη και όχι ικανοποιητική στην μέτρηση της ακρίβειας της πρόβλεψης.

Ελάχιστες είναι οι περιπτώσεις όπου η σύγκριση ανάμεσα στις δυο μεθόδους έχει γίνει με βάση το εμβαδό κάτω από την καμπύλη ROC. Επιπλέον, τα αποτελέσματα που παρήγαγαν δεν ήταν συνεπή μεταξύ τους. Κάποιες μελέτες κατέληξαν στο συμπέρασμα ότι η λογιστική παλινδρόμηση και τα δέντρα παλινδρόμησης έχουν ισάξια αποτελέσματα, άλλες ότι η λογιστική παλινδρόμηση είναι πιο έγκυρη και μια έρευνα το αντίθετο.

Μια πρακτική που συνιστάται είναι τα αποτελέσματα να επικυρώνονται χρησιμοποιώντας επαναλαμβανόμενα τη μέθοδο σε διαφορετικούς διαμερισμούς του δείγματος.

Παρακάτω θα παρουσιαστούν δυο εφαρμογές στις οποίες γίνεται σύγκριση ανάμεσα στην λογιστική παλινδρόμηση και στα δέντρα παλινδρόμησης χρησιμοποιώντας το εμβαδό κάτω από μια καμπύλη ROC.

2.2 Σύγκριση δέντρων παλινδρόμησης, λογιστικής παλινδρόμησης, γενικευμένων αθροιστικών μοντέλων και προσαρμοσμένων πολυμεταβλητών splines παλινδρόμησης για την πρόβλεψη θνησιμότητας από οξύ έμφραγμα του μυοκαρδίου (AMI)

Ο P. C. Austin (2007) σε συγκεκριμένη μελέτη που διεξήχθη είχε δύο στόχους: Πρώτον, να συγκρίνει την προβλεπτική ικανότητα της συμβατικής λογιστικής παλινδρόμησης με αυτήν των δέντρων παλινδρόμησης για την πρόβλεψη της θνησιμότητας 30 ημερών από οξύ έμφραγμα του μυοκαρδίου (Acute Myocardial Infraction –AMI) και δεύτερον, να εξετάσει την σχετική απόδοση δυο ακόμα μεθόδων, των γενικευμένων αθροιστικών μοντέλων (GAM) και των προσαρμοσμένων πολυμεταβλητών splines παλινδρόμησης (MARS).

Η παρούσα εργασία πραγματεύεται κυρίως το πρώτο σκέλος, το σχετικό με τη λογιστική παλινδρόμηση και τα δέντρα παλινδρόμησης και με τα συμπεράσματα που προέκυψαν σε σχέση με αυτές τις δύο μεθόδους. Το σκέλος που αφορά τις μεθόδους GAM και MARS θα περιγραφεί συνοπτικά.

Prevalence of dichotomous variables	
Female	36.0 per cent
Presenting signs and symptoms	
Acute congestive heart failure (acute CHF)/pulmonary oedema	5.7 per cent
Cardiogenic shock	1.6 per cent
Classic cardiac risk factors	
Diabetes	26.3 per cent
History of hypertension	46.0 per cent
Smoking history	32.3 per cent
History of cerebrovascular accident or transient ischaemic attack	10.3 per cent
History of hyperlipidaemia	30.6 per cent
Family history of coronary artery disease	30.2 per cent
Comorbid conditions	
Angina	33.0 per cent
Cancer	3.1 per cent
Dementia	3.9 per cent
Peptic ulcer disease	5.5 per cent
Previous acute myocardial infarction	23.1 per cent
Asthma	5.5 per cent
Depression	7.2 per cent
Peripheral arterial disease	7.7 per cent
Previous PCI	3.2 per cent
Congestive heart failure (chronic)	5.0 per cent
Hyperthyroidism	1.3 per cent
Previous CABG surgery	6.7 per cent
Aortic stenosis	1.7 per cent
Medians of continuous variables (25th percentile–75th percentile)	
Age	69 (57–78)
Vital signs on admission	
Systolic blood pressure on admission	146 (126–168)
Diastolic blood pressure on admission	82 (70–95)
Heart rate on admission	81 (68–98)
Respiratory rate on admission	20 (18–23)
Laboratory test results	
Haemoglobin	139 (127–151)
White blood count	9.6 (7.7–12.2)
Sodium levels	139 (137–141)
Potassium levels	4.1 (3.7–4.4)
Glucose levels	7.85 (6.4–10.9)
Urea level	6.5 (5.1–8.7)
Creatinine levels	93 (78–115)

Πίνακας 2.1:Μεταβλητές και χαρακτηριστικά του δείγματος που χρησιμοποιήθηκαν στη μελέτη

2.2.1 Το αρχικό μοντέλο πρόβλεψης της AMI θνησιμότητας

Το δείγμα που χρησιμοποιήθηκε αποτελείται από 9484 ασθενείς από 102 νοσοκομεία του Ontario σε έρευνα που πραγματοποιήθηκε από τη 1 Απρίλη του 1999 μέχρι και

31 Μάρτη του 2001. Από το δείγμα συγκεντρώθηκαν δεδομένα σχετικά με το ιατρικό ιστορικό, τους παράγοντες καρδιακού ρίσκου, συννοσηρές συνθήκες, αγγειακό ιστορικό, ζωτικές ενδείξεις και εργαστηριακές εξετάσεις. Τα επικρατή σημεία διχοτόμων μεταβλητών και τα 25^{α} και 75^{α} εκατοστιαία σημεία των συνεχών μεταβλητών που λήφθηκαν υπόψη για τη μελέτη παρουσιάζονται στον Πίνακα 2.1. Συνολικά 1065 (11,2%) ασθενείς απεβίωσαν μέσα σε διάστημα 30 ημερών από την εισαγωγή τους.

Το αρχικό μοντέλο για την πρόβλεψη 30 ημερών AMI θνησιμότητας αναπτύχθηκε σε πιο παλιά έρευνα χρησιμοποιώντας επαναλαμβανόμενη επιλογή bootstrap δειγμάτων. Το μοντέλο που προέκυψε αποτελούνταν από τους παρακάτω παράγοντες: ηλικία, παρουσία καρδιογεννητικού σοκ κατά την εισαγωγή, συστολική πίεση αίματος κατά την εισαγωγή, οικογενειακό ιστορικό για νόσο της στεφανιαίας αρτηρίας (CAD), τιμή του αναπνευστικού κατά την εισαγωγή, επίπεδο γλυκόζης, αριθμός λευκών αιμοσφαιρίων και επίπεδο κρεατίνης. Τα μοντέλα για την πρόβλεψη της AMI θνησιμότητας προέκυψαν με την εφαρμογή της μεθόδου της προς τα πίσω εξάλειψης (backward elimination) των μεταβλητών σε κάθε bootstrap δείγμα. Οι μεταβλητές που κρίθηκαν σημαντικές για την πρόβλεψη της AMI θνησιμότητας σε ποσοστό τουλάχιστον 60% των bootstrap δειγμάτων επιλέχτηκαν για το τελικό μοντέλο. Η παρουσία καρδιογεννητικού σοκ και το οικογενειακό ιστορικό CAD είναι διχοτομικές μεταβλητές και οι υπόλοιπες μεταβλητές είναι συνεχείς και εισήχθησαν γραμμικά στο μοντέλο παλινδρόμησης. Το αρχικό αυτό μοντέλο, χρησιμοποιήθηκε σαν βάση για τη κατασκευή κάποιων μοντέλων παλινδρόμησης που θα θεωρηθούν σε αυτή τη μελέτη.

2.2.2 Τα μοντέλα πρόβλεψης της AMI θνησιμότητας

Πρέπει να σημειωθεί ότι για την προσαρμογή και την επικύρωση των μοντέλων που χρησιμοποιήθηκαν στην παρούσα έρευνα χρησιμοποιήθηκε η στατιστική γλώσσα προγραμματισμού της R.

2.2.2.1 Η λογιστική παλινδρόμηση για τη πρόβλεψη της AMI θνησιμότητας

Για τη μέθοδο της λογιστικής παλινδρόμησης χρησιμοποιήθηκαν τρία διαφορετικά μοντέλα.

Το $1^{ο}$ μοντέλο περιείχε τις εξής οκτώ μεταβλητές: ηλικία, παρουσία καρδιογεννητικού σοκ κατά την εισαγωγή, συστολική πίεση αίματος κατά την εισαγωγή, οικογενειακό ιστορικό για νόσο της στεφανιαίας αρτηρίας (CAD), τιμή του αναπνευστικού κατά την εισαγωγή, επίπεδο γλυκόζης, αριθμός λευκών αιμοσφαιρίων και επίπεδο κρεατίνης.

Το $2^{ο}$ μοντέλο κατασκευάστηκε με τη μέθοδο της προς τα πίσω εξάλειψης. Στο αρχικό μοντέλο των οκτώ μεταβλητών προστέθηκαν όλες οι αλληλεπιδράσεις $2^{ης}$ τάξης ανάμεσα στις οκτώ βασικές. Από τις οκτώ βασικές που υποχρεώθηκαν να μείνουν στο μοντέλο και μαζί με όσες αλληλεπιδράσεις παρέμειναν από τη διαδικασία της προς τα πίσω εξάλειψης, προέκυψε το μοντέλο που χρησιμοποιήθηκε.

Το $3^{ο}$ μοντέλο κατασκευάστηκε χρησιμοποιώντας όλες τις 34 μεταβλητές του Πίνακα 2.1 και εφαρμόστηκε στη συνέχεια η μέθοδος της προς τα πίσω εξάλειψης.

Η προσαρμογή των μοντέλων πραγματοποιήθηκε με την συνάρτηση glm της R και η μέθοδος της προς τα πίσω εξάλειψης με την συνάρτηση step της R.

2.2.2.2. Τα δέντρα παλινδρόμησης για τη πρόβλεψη της AMI θνησιμότητας

Για την κατασκευή των δέντρων παλινδρόμησης χρησιμοποιήθηκαν μέθοδοι δυαδικού αναδρομικού διαμερισμού. Σε κάθε κόμβο επιλέχθηκε ο διαμερισμός που μεγιστοποιεί τη μείωση της απόκλισης. Το δέντρο αρχικά κατασκευάστηκε χρησιμοποιώντας και τις 34 μεταβλητές του Πίνακα 2.1 και στη συνέχεια κλαδεύτηκε. Το σύνολο δεδομένων που χρησιμοποιήθηκε για τον καθορισμό του αριθμού των φύλλων του δέντρου, κατασκευάστηκε με χρήση 10-fold cross-validation. Το μέγεθος του δέντρου που επιλέχθηκε ήταν αυτό που ελαχιστοποιούσε την απόκλιση όταν εφαρμόστηκε η 10-fold cross-validation. Οι προβλέψεις πραγματοποιήθηκαν από το παραγόμενο σύνολο του κλαδεμένου δέντρου. Το μοντέλο προσαρμόστηκε χρησιμοποιώντας την συνάρτηση TREE της R και το κλάδεμα πραγματοποιήθηκε με τη συνάρτηση prune.tree. Στην περίπτωση που δυο διαφορετικά μεγέθη δέντρου κατέληγαν στην ίδια ελάχιστη απόκλιση, τότε επιλεγόταν το δέντρο με το μικρότερο μέγεθος. Το αρχικό δέντρο παλινδρόμησης κλαδεύτηκε και το τελικό δέντρο προσαρμόστηκε στο παραγόμενο δείγμα και χρησιμοποιήθηκε για τις προβλέψεις των στοιχείων του συνόλου επικύρωσης.

2.2.2.3 Generalized Additive Models (GAM)

Ένα μοντέλο GAM είναι ένα αθροιστικό μοντέλο παλινδρόμησης της μορφής:

$$\eta(x) = \alpha + f_1(x_1) + f_2(x_2) + ... + f_p(x_p)$$

με τα f_i να είναι οι συναρτήσεις των μεταβλητών πρόβλεψης (predictors). Αυτές οι συναρτήσεις μπορούν να είναι μη-παραμετρικοί scatterplot εξομαλυντές (smoothers) ή splines παλινδρόμησης. Ένα πλεονέκτημα των μοντέλων GAM είναι ότι μπορεί κάποιος να χαλαρώσει την υπόθεση της γραμμικότητας ανάμεσα στην τιμή της μεταβλητής πρόβλεψης και στο αποτέλεσμα. Επιπλέον, μπορεί να μην καθορίσει καθόλου το είδος της σχέσης και να αφήσει τα δεδομένα να το καθορίσουν.

Θεωρήθηκαν τρία διαφορετικά μοντέλα GAM για την πρόβλεψη της θνησιμότητας σε 30 μέρες λόγω AMI. Πρώτον, το λιτό μοντέλο παλινδρόμησης που αναφέρθηκε και προηγουμένως. Το οικογενειακό ιστορικό CAD και η παρουσία καρδιογενούς σοκ κατά την εισαγωγή θεωρήθηκαν σαν δυαδικές μεταβλητές. Οι μεταβλητές ηλικία, η συστολική αρτηριακή πίεση κατά την εισαγωγή, ο ρυθμός του αναπνευστικού κατά την εισαγωγή, η γλυκόζη, ο αριθμός λευκών αιμοσφαιρίων και η κρεατινίνη μοντελοποιήθηκαν χρησιμοποιώντας λείες splines με πέντε βαθμούς ελευθερίας η κάθε μια. Το δεύτερο μοντέλο που χρησιμοποιήθηκε αποτελείται από το αρχικό μοντέλο GAM με χρήση επιπλέον, όλων των αλληλεπιδράσεων $2^{ης}$ τάξης. Το τρίτο μοντέλο GAM περιλαμβάνεις και τις 34 μεταβλητές του Πίνακα 2.1. Οι 22 διχοτομικές μεταβλητές έχουν εισαχθεί σαν δυαδικές μεταβλητές πρόβλεψης και οι 12 συνεχείς μοντελοποιήθηκαν με λείες splines με πέντε βαθμούς ελευθερίας η κάθε μια. Οι οροί εξομάλυνσης (smoothing terms) μοντελοποιήθηκαν με splines παλινδρόμησης με πέντε βαθμούς ελευθερίας και για τα τρία μοντέλα GAM που χρησιμοποιήθηκαν.

2.2.2.4 Multivariate Adaptive Regression Spline models (MARS)

Η MARS είναι μια προσαρμοστική διαδικασία παλινδρόμησης που ενδείκνυται για προβλήματα με μεγάλο αριθμό μεταβλητών πρόβλεψης. Κάνει χρήση μιας επέκτασης που βασίζεται σε γραμμικές συναρτήσεις spline. Για μια δοσμένη μεταβλητή πρόβλεψης X_j και για μια δοσμένη τιμή t της μεταβλητής, μπορούν να οριστούν δύο γραμμικές συναρτήσεις spline:

$$\left(X_j - t\right)_+ \text{ και } \left(t - X_j\right)_-$$

όπου το «+» αναφέρεται στο θετικό μέρος.

Ένα μοντέλο MARS κατασκευάζεται χρησιμοποιώντας ένα υποσύνολο όλων των πιθανών γραμμικών spline συναρτήσεων. Επιπλέον, μπορεί να χρησιμοποιηθούν και οι παράγωγοι αυτών των γραμμικών spline συναρτήσεων.

Τα MARS χρησιμοποιούνται για συνεχή αποτελέσματα. Άρα, η θνησιμότητα σε 30 ημέρες αρχικά μοντελοποιήθηκε σαν συνεχές αποτέλεσμα. Ο πίνακας σχεδιασμού που προέκυψε από την ανάλυση MARS χρησιμοποιήθηκε στη συνέχεια σαν πίνακας σχεδιασμού για ένα μοντέλο λογιστικής παλινδρόμησης που αντιμετωπίζει την θνησιμότητα σε 30 μέρες λόγω AMI σαν δυαδικό αποτέλεσμα.

Ερευνήθηκαν τρία διαφορετικά μοντέλα MARS που όλα χρησιμοποίησαν τις 34 μεταβλητές του Πίνακα 2.1. Το πρώτο ήταν ένα αθροιστικό μοντέλο που δεν περιελάμβανε αλληλεπιδράσεις μεταξύ των μεταβλητών πρόβλεψης. Το δεύτερο μοντέλο περιελάμβανε τις αλληλεπιδράσεις $2^{ης}$ τάξης και το τρίτο μοντέλο περιελάμβανε κάθε είδους αλληλεπιδράσεις, ακόμα και αυτές της $34^{ης}$ τάξης. Τα μοντέλα κατασκευάστηκαν χρησιμοποιώντας γενικευμένη cross-validation για να

καθοριστεί ο βέλτιστος αριθμός όρων για το μοντέλο. Η γενικευμένη cross-validation βοηθάει να αποφευχθεί η υπέρ-προσαρμογή του μοντέλο στο derivation δείγμα.

2.2.3 Σύγκριση των μοντέλων πρόβλεψης

Για τη σύγκριση της προβλεπτικής ικανότητας των στατιστικών μοντέλων χρησιμοποιήθηκε η μέθοδος της επικύρωσης των αποτελεσμάτων μέσω επαναλαμβανόμενων διαχωρισμών του δείγματος.

Τα στοιχεία του δείγματος χωρίζονταν τυχαία σε δύο σύνολα. Τα 2/3 του δείγματος (6323 στοιχεία) αποτελούσαν το σύνολο που χρησιμοποιούνταν για την παραγωγή αποτελεσμάτων (derivation sample) και το υπόλοιπο 1/3 του δείγματος (3161 στοιχεία) αποτελούσε το σύνολο για την επικύρωση (validation sample). Η διαδικασία επαναλήφθηκε 1000 φορές. Κάθε ένα μοντέλο προσαρμόστηκε στο derivation σύνολο και με αυτό τον τρόπο προέκυψε το παραγόμενο μοντέλο. Στη συνέχεια, μέσω του παραγόμενου μοντέλου, παρήχθησαν προβλέψεις για κάθε στοιχείο του συνόλου επικύρωσης.

Η προβλεπτική ικανότητα κάθε μοντέλου συνοψίζεται από το εμβαδό της επιφάνειας κάτω από την καμπύλη ROC. Το εμβαδό της επιφάνειας κάτω από την ROC χρησιμοποιήθηκε και για τα derivation σύνολα και για τα validation σύνολα. Επιπλέον, στη μελέτη η προβλεπτική ικανότητα συνοψίστηκε χρησιμοποιώντας και τον δείκτη R_N^2 των Nagelkerke, Cragg και Uhler, και η τιμή Brier (Brier's score).

Για την διαμέτρηση (calibration) των μοντέλων εκτός αυτού των δέντρων παλινδρόμησης, στα derivation και calibration σύνολα χρησιμοποιήθηκε το test καλής προσαρμογής των Hosmer-Lemeshow. Αυτό το test χωρίζει τα στοιχεία σε δέκατα ρίσκου, όπως αυτό προκύπτει από τα δέκατα της εκτιμώμενης πιθανότητας για θνησιμότητα στις 30 μέρες. Όλα όμως τα δέντρα παλινδρόμησης που προέκυψαν είχαν λιγότερους από δέκα τερματικούς κόμβους και έτσι δεν είναι εφικτό να χωριστούν οι προβλεπόμενες πιθανότητες για θνησιμότητα 30 ημερών σε δέκατα.

2.2.4 Αποτελέσματα

2.2.4.1 Απόδοση της προβλεπτικής ικανότητας

Το μέσο εμβαδό κάτω από την καμπύλη ROC για κάθε μοντέλο και για τα 1000 derivation και validation δείγματα, παρουσιάζονται στον Πίνακα 2.2.

Στο validation δείγμα, το μέσο εμβαδό κάτω από την καμπύλη ROC για το μοντέλο του δέντρου παλινδρόμησης είναι 0.762, ενώ για το μοντέλο της απλής λογιστικής παλινδρόμησης είναι 0.845. Η διαφορά στο εμβαδό κάτω από την

καμπύλη ROC ανάμεσα στα δύο αυτά μοντέλα κυμαίνεται από 0.041 έως 0.165 στα 1000 derivation δείγματα, με μέση διαφορά 0.083.

Για τα μοντέλα MARS και GAM, το μέσο εμβαδό κάτω από την καμπύλη ROC στο validation δείγμα κυμαινόταν από 0.851 το ελάχιστο (μοντέλο MARS που περιλαμβάνει κάθε είδους αλληλεπιδράσεις) έως 0.851 το μέγιστο (μοντέλο GAM όλων των κύριων επιδράσεων). Τα δύο μοντέλα MARS που περιελάμβαναν αλληλεπιδράσεις είχαν χαμηλότερη προβλεπτική ικανότητα από όλα τα μοντέλα λογιστικής παλινδρόμησης, τα μοντέλα GAM και το αθροιστικό MARS. Αυτό ενδεχομένως να οφείλεται στο ότι τα μοντέλα MARS με αλληλεπιδράσεις, υπέρ-προσαρμόστηκαν στα derivation δείγματα.

Model	ROC area: derivation sample	ROC area: validation sample	Hosmer–Lemeshow GOF: derivation sample	Hosmer–Lemeshow GOF: validation sample	R_N^2: validation sample	Brier's score: validation sample
Regression tree	0.779	0.762			0.198	0.087
Logistic regression (eight main effects)	0.846	0.845	0.2271	0.2363	0.319	0.078
Logistic regression (two-way interactions)	0.849	0.844	0.2255	0.2109	0.313	0.078
Logistic regression (backwards elimination from full model)	0.853	0.846	0.2243	0.2137	0.321	0.078
GAM (eight main effects)	0.857	0.850	0.3642	0.2493	0.333	0.076
GAM (two-way interactions)	0.861	0.849	0.5526	0.1984	0.328	0.077
GAM (full model)	0.869	0.851	0.2263	0.1316	0.332	0.077
MARS (additive)	0.858	0.848	0.0820	0.1139	0.326	0.077
MARS (two-way interactions)	0.867	0.837	0.0947	0.0167	0.275	0.080
MARS (all interactions)	0.868	0.831	0.0748	0.0051	0.244	0.082

Πίνακας 2.2: Αποτελέσματα (κατά μέσο όρο) για κάθε μοντέλο στα 1000 επαναλαμβανόμενα split-samples

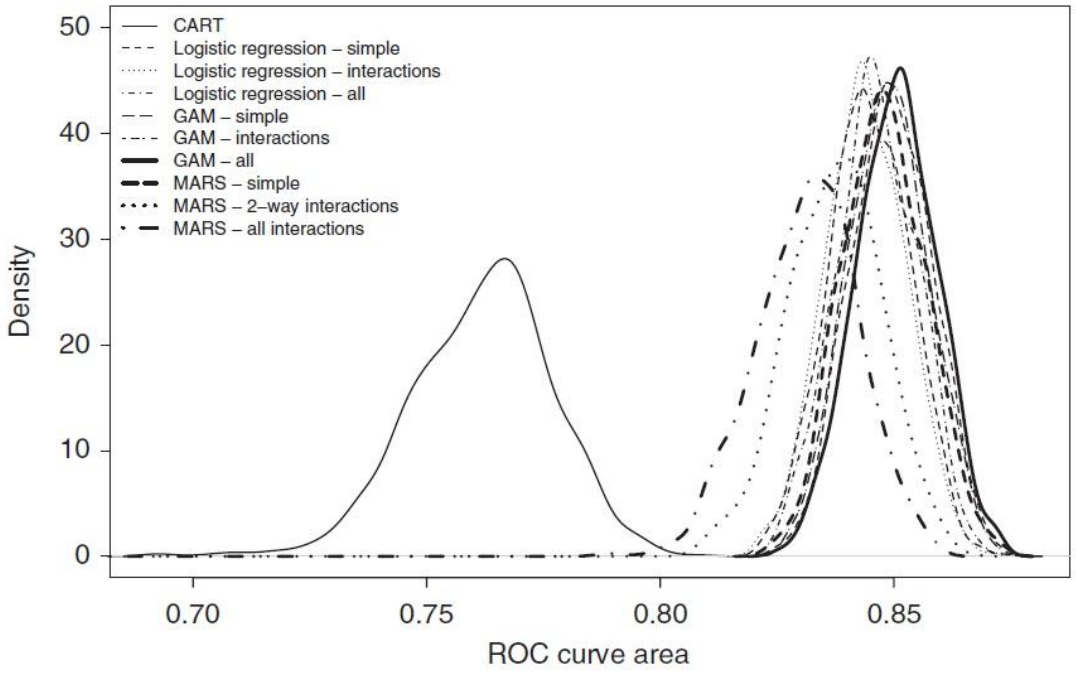
Το μέσο εμβαδό κάτω από την καμπύλη ROC για το μοντέλο των δέντρων παλινδρόμησης μειώθηκε κατά 0.017 (από 0.779 σε 0.762) από το derivation δείγμα στο validation δείγμα. Η μείωση του μέσου εμβαδού ανάμεσα στα derivation και validation δείγματα για την απλή λογιστική παλινδρόμηση , είναι αμελητέα (από 0.846 σε 0.845). Αντίστοιχα και για την απλή λογιστική παλινδρόμηση που περιλάμβανε και τις αλληλεπιδράσεις $2^{ης}$ τάξης (από 0.849 σε 0.844), για το μοντέλο της λογιστικής παλινδρόμησης που προέκυψε από backward elimination (από 0.853 σε 0.846) και για το απλό GAM (από 0.857 σε 0.850). Για τα υπόλοιπα μοντέλα, η μείωση κυμάνθηκε από 0.010 (αθροιστικό MARS) έως 0.037 (MARS που περιλαμβάνει κάθε είδους αλληλεπίδραση).

Η κατανομή του εμβαδού κάτω από την ROC, στα 1000 validation σύνολα, για κάθε μοντέλο περιγράφεται στο Γράφημα 1. Για την καλύτερη ανάλυση στο

γράφημα, αυτό περιορίστηκε στο να παρουσιάζει μόνο εκείνα τα density plots τα οποία είναι πάνω από 0.69. Με αυτό τον τρόπο προκύπτουν κάποιες παρατηρήσεις. Κατ' αρχήν, η κατανομή των εμβαδών κάτω από τη ROC για τα δέντρα παλινδρόμησης μετατοπίστηκε προς τα κάτω σε σχέση με τα υπόλοιπα μοντέλα.

Χρησιμοποιώντας τα 1000 derivation και validation δείγματα, προέκυψαν τα παρακάτω :

1. Τα μοντέλα των δέντρων παλινδρόμησης είχαν σταθερά φτωχότερη απόδοση σε σχέση με τα άλλα μοντέλα.
2. Οι κατανομές των εμβαδών για όλα τα μοντέλα της λογιστικής παλινδρόμησης (απλής, με αλληλεπιδράσεις και ύστερα από χρήση backward elimination), ήταν σχεδόν παρόμοιες.
3. Το δεύτερο συμπέρασμα ισχύει και για το απλό GAM μοντέλο και το GAM που περιελάμβανε αλληλεπιδράσεις.
4. Η μεγαλύτερη διαφοροποίηση ήταν ανάμεσα στα δέντρα παλινδρόμησης και τις υπόλοιπες μεθόδους, παρ' όλο που υπήρχαν διαφοροποιήσεις και μεταξύ αυτών.
5. Τα δύο μοντέλα MARS που περιελάμβαναν αλληλεπιδράσεις παρουσίασαν μεγαλύτερη διαφοροποίηση στο εμβαδόν κάτω από την καμπύλη ROC στο validation δείγμα σε σχέση με τα μοντέλα λογιστικής παλινδρόμηση, τα GAM και το αθροιστικό MARS.
6. Η επαναλαμβανόμενη χρήση της split-sample επικύρωσης επέτρεψε να απεικονιστεί η κατανομή των εμβαδών κάτω από τις ROC για τα validation δείγματα. Ενώ το μέσο εμβαδό για τη μέθοδο των δέντρων παλινδρόμησης ήταν 0.762, τα παρατηρούμενα εμβαδά στα derivation δείγματα κυμάνθηκαν από 0.692 έως 0.808. Παλαιότερες μελέτες που δεν χρησιμοποιούσαν επαναλαμβανόμενα τη μέθοδο της split-sample επικύρωσης δεν μπορούσαν να εξετάσουν το βαθμό στον οποίο τα αποτελέσματα ήταν εξαρτημένα, σε κάποιο βαθμό, από τον τρόπο με τον οποίο το δείγμα χωριζόταν σε derivation και validation δείγματα.



Σχήμα 2.1:*Κατανομή των εμβαδών κάτω από την καμπύλη ROC για τα 1000 validation δείγματα*

2.2.4.2 Έλεγχος για την προσαρμογή των μοντέλων

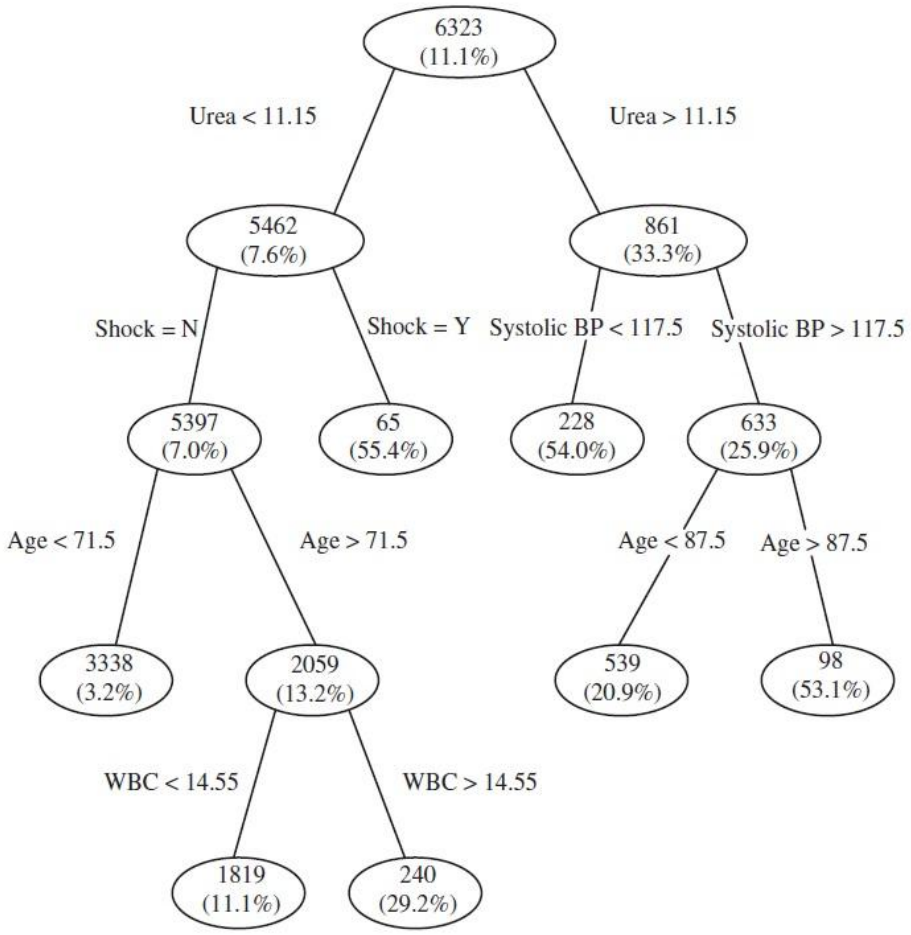
Τα αποτελέσματα για τον έλεγχο προσαρμογής των Hosmer-Lemeshow παρουσιάζονται στον Πίνακα 2.2. Γενικά, τα μοντέλα της λογιστικής παλινδρόμησης και τα GAMS εμφάνισαν καλή προσαρμογή, με $p > 0.22$ για το derivation δείγμα και $p > 0.13$ για το validation δείγμα. Το αθροιστικό MARS έδειξε δείγματα έλλειψης προσαρμογής ($p = 0.0820$ και 0.1139 στα derivation και validation δείγματα αντίστοιχα). Τα υπόλοιπα μοντέλα MARS με αλληλεπιδράσεις παρουσίασαν χαμηλή προσαρμογή. Ο αντίστοιχος έλεγχος δεν μπορούσε να πραγματοποιηθεί για το μοντέλο των δέντρων παλινδρόμησης καθώς σε πολλές περιπτώσεις δεν μπορούσαν τα στοιχεία να χωριστούν σε δεκατημόρια, αφού τα δέντρα είχαν λιγότερους από δέκα τερματικούς κόμβους. Έτσι δεν υπήρχε αρκετή ποικιλία στις προβλεπόμενες πιθανότητες ώστε να υπολογιστούν δεκατημόρια για το ρίσκο.

2.2.4.3 Λοιπά αποτελέσματα

Ο μέσος αριθμός τερματικών κόμβων για τα δέντρα παλινδρόμησης στα 1000 derivation δείγματα ήταν 6,4. Ο αριθμός των τερματικών κόμβων κυμαινόταν από 4 έως 9, με 1^{α} και 3^{α} τεταρτημόρια να είναι 6 και 7 αντίστοιχα. Ο αριθμός των μεταβλητών που χρησιμοποιήθηκαν στην κατασκευή των δέντρων κυμάνθηκε από 3

έως 6, με μέσο 4,4 με 1^{α} και 3^{α} τεταρτημόρια να είναι 4 και 5 αντίστοιχα. Ένα δέντρο παλινδρόμησης όπως αυτό προέκυψε από ένα από τα derivation δείγματα παρουσιάζεται στο Σχήμα 2. Σε αυτό παρουσιάστηκαν 7 τερματικοί κόμβοι και χρησιμοποιήθηκαν 5 μεταβλητές για τη δημιουργία του.

Ο μέσος αριθμός αλληλεπιδράσεων που αναγνωρίστηκαν στα 1000 derivation δείγματα με χρήση του μοντέλου λογιστικής παλινδρόμησης που περιλάμβανε και αλληλεπιδράσεις ήταν 7,8. Ο αριθμός τους κυμαινόταν από 2 έως 15 με 1^{α} και 3^{α} τεταρτημόρια 6 και 9 αντίστοιχα. Ο μέσος αριθμός κύριων επιδράσεων που αναγνωρίστηκαν στα 1000 derivation δείγματα με χρήση του μοντέλου λογιστικής παλινδρόμησης ύστερα από backward elimination των μεταβλητών τους πλήρους μοντέλου ήταν 17,3. Ο αριθμός των κύριων επιδράσεων κυμαινόταν από 12 έως 23 με 1^{α} και 3^{α} τεταρτημόρια 16 και 18 αντίστοιχα.



Σχήμα 2.2:*Δέντρο παλινδρόμησης για την πρόβλεψη της θνησιμότητας σε 30 ημέρες. Σε κάθε κόμβο αναγράφεται το πλήθος των στοιχείων που περιέχει και το ποσοστό θνησιμότητας για αυτά.*

2.2.5 Συμπεράσματα

Η μελέτη κατέληξε στο συμπέρασμα ότι η μέθοδος των δέντρων παλινδρόμησης δεν μπορεί να προβλέψει την θνησιμότητα σε 30 ημέρες τόσο καλά όσο η συμβατική λογιστική παλινδρόμηση. Τα δύο μοντέλα λογιστικής παλινδρόμηση, αυτό με χρήση μόνο κύριων επιδράσεων και αυτό με χρήση κύριων επιδράσεων και $2^{ης}$ τάξης αλληλεπιδράσεων, είχαν παρόμοια προβλεπτική ικανότητα στα validation δείγματα. Επιπλέον, διαπιστώθηκε ότι η συμβατική λογιστική παλινδρόμηση παρουσιάζει συγκρίσιμα αποτελέσματα με μοντέρνες, ευέλικτες μεθόδους παλινδρόμησης όπως τα μοντέλα GAM και MARS. Το εμβαδό της περιοχής κάτω από την καμπύλη ROC ήταν, έστω και ελάχιστα, υψηλότερο για τις πιο ευέλικτες μεθόδους σε σχέση με τη λογιστική παλινδρόμηση. Τη μέγιστη τιμή (0.851) την παρουσίασε το μοντέλο GAM με όλες τις 34 κύριες επιδράσεις.

Ο P. C. Austin (2007) προτείνει σε ερευνητές που θέλουν να αξιολογήσουν την προβλεπτική ικανότητα συγκεκριμένων μοντέλων λογιστικής παλινδρόμησης ή να συγκρίνουν την προβλεπτική ικανότητα ανάμεσα σε διαφορετικά μοντέλα λογιστικής παλινδρόμησης, να χρησιμοποιούν ένα αθροιστικό μέτρο όπως το εμβαδό κάτω από την καμπύλη ROC σε ένα ανεξάρτητο validation δείγμα. Επιπλέον, προτείνει να χρησιμοποιείται επαναλαμβανόμενα η μέθοδος της split-sample επικύρωσης. Για κάθε μοντέλο παλινδρόμησης, παρουσιάστηκε σημαντική μεταβλητότητα στο εμβαδό κάτω από τη ROC για τα 1000 validation δείγματα. Η επαναλαμβανόμενη χρήση split-sample επικύρωσης βοηθάει στο να εξεταστεί η μεταβλητότητα στην προβλεπτική ικανότητα ενός συγκεκριμένου μοντέλου παλινδρόμησης. Επίσης, με αυτόν τον τρόπο, μπορεί κάποιος να αναπτύξει έναν πιο σωστό χαρακτηρισμό για τη φύση της σχέσης ανάμεσα σε συγκεκριμένες μεταβλητές πρόβλεψης και στο τελικό αποτέλεσμα. Η μελέτη κατάφερε έτσι να καθορίσει τη μέση σχέση παλινδρόμησης μέσα από 1000 derivation δείγματα. Έτσι μπορεί κάποιος να καθορίσει τη μέση σχέση που είναι λιγότερο πιθανή να επηρεαστεί από παρατηρήσεις που ίσως περιληφθούν σε κάποια λίγα derivation δείγματα.

2.3 Σύγκριση δέντρων παλινδρόμησης και λογιστικής παλινδρόμησης για την πρόβλεψη της θνησιμότητας υπό νοσηλεία ασθενών που εισήχθησαν λόγω καρδιακής ανεπάρκειας

Οι Austin et al. (2010) σε συγκεκριμένη μελέτη που διεξήχθη είχε τρεις στόχους: Πρώτον, να συγκρίνουν την ικανότητα των δέντρων παλινδρόμησης σε σχέση με τη λογιστική παλινδρόμηση για την πρόβλεψη της θνησιμότητας υπό νοσηλεία ενός δείγματος ασθενών που έχουν εισαχθεί λόγω καρδιακής ανεπάρκειας. Δεύτερον, να εξετάσουν την σταθερότητα ή αναπαραγωγή των δέντρων παλινδρόμησης που προκύπτουν από την πρόβλεψη της θνησιμότητας στο

συγκεκριμένο δείγμα. Τρίτον, να ερευνήσουν τη φύση και τη σχέση ανάμεσα σε διάφορες σημαντικές κλινικές μεταβλητές και στην πιθανότητα για θνησιμότητα μετά την εισαγωγή για περίθαλψη λόγω οξείας καρδιακής ανεπάρκειας.

2.3.1 Μέθοδοι που χρησιμοποιήθηκαν

Τα δεδομένα για την έρευνα πάρθηκαν από το μητρώο της έρευνας EFFECT (Effective Cardiac Treatment). Η έρευνα EFFECT διεξήχθη σε δύο φάσεις: Στην πρώτη φάση συλλέχθηκαν αναλυτικά ιατρικά δεδομένα από ασθενείς που νοσηλεύτηκαν λόγω καρδιακής ανεπάρκειας σε 103 νοσοκομεία του Ontario, κατά τη χρονική περίοδο 1 Απρίλη του 1999 έως και 31 Μάρτη του 2001. Κατά τη δεύτερη φάση εξήχθησαν δεδομένα από ασθενείς 96 νοσοκομείων του Ontario που νοσηλεύονταν λόγω καρδιακής ανεπάρκειας κατά τη χρονική περίοδο 1 Απρίλη του 2004 με 31 Μάρτη του 2005. Τα δεδομένα που εξήχθησαν αφορούσαν δημογραφικά στοιχεία των ασθενών, ζωτικές ενδείξεις, φυσικές εξετάσεις, ιατρικό ιστορικό και κλινικές εξετάσεις.

Οι ασθενείς για τους οποίους υπήρχαν κενά-ελλειπούσες τιμές όσον αφορά κάποια από τα παραπάνω στοιχεία, αποκλείστηκαν από το δείγμα για την έρευνα. Συνολικά από τους 9.945 ασθενείς που νοσηλεύτηκαν κατά την περίοδο της πρώτης φάσης και τους 8.339 ασθενείς που νοσηλεύτηκαν κατά την περίοδο της δεύτερης, συμπεριλήφθησαν οι 8.240 στο δείγμα της πρώτης φάσης και οι 7.609 για το δείγμα της δεύτερης. Το πρώτο δείγμα από την EFFECT αποτέλεσε το derivation δείγμα και το δεύτερο αποτέλεσε το validation δείγμα της έρευνας στο οποίο και αξιολογήθηκε η προβλεπτική ικανότητα των διάφορων στατιστικών μεθόδων.

Για τη σύγκριση ανάμεσα στην λογιστική παλινδρόμηση και στα δέντρα παλινδρόμησης χρησιμοποιήθηκαν δυο παλιότερα μοντέλα λογιστικής παλινδρόμησης και ένα παλιότερο μοντέλο δέντρων παλινδρόμησης. Επιπλέον, αναπτύχθηκαν και εξετάστηκαν τρία νέα μοντέλα λογιστικής παλινδρόμησης και ένα νέο μοντέλο δέντρων παλινδρόμησης.

2.3.2 Μοντέλα που χρησιμοποιήθηκαν

2.3.2.1 Μοντέλα λογιστικής παλινδρόμησης

Για την μελέτη της προβλεπτικής ικανότητας της λογιστικής παλινδρόμησης, χρησιμοποιήθηκαν τα παρακάτω μοντέλα:

- Το μοντέλο EFFECT-HF (Enhanced feedback for Effective Cardiac Treatment in Heart Failure), είναι ένα μοντέλο λογιστικής παλινδρόμησης που χρησιμοποιεί τις εξής μεταβλητές: Ηλικία (συνεχής μεταβλητή), συστολική αρτηριακή πίεση (συνεχής), αναπνευστικός ρυθμός (συνεχής), συγκέντρωση του νατρίου στον ορό

(<136 ή ≥ 136 mEq/L), συγκέντρωση BUN στον ορό (συνεχής), ιστορικό αγγειακής εγκεφαλικής νόσου, ιστορικό άνοιας, ιστορικό χρόνιας αποφρακτικής πνευμονοπάθειας, ιστορικό ηπατικής κίρρωσης και ιστορικό καρκίνου. Αρχικά το μοντέλο αυτό αναπτύχθηκε για την πρόβλεψη θνησιμότητας σε 30 μέρες από την εισαγωγή. Στη συγκεκριμένη μελέτη, οι συντελεστές παλινδρόμησης του EFFECT-HF υπολογίστηκαν για το δείγμα EFFECT. Προβλέψεις για την θνησιμότητα υπό νοσηλεία υπολογίστηκαν για κάθε στοιχείο του επόμενου EFFECT δείγματος, χρησιμοποιώντας τους συντελεστές που εκτιμήθηκαν στο δείγμα EFFECT.

Χρησιμοποιήθηκαν δύο παραλλαγές του αρχικού EFFECT-HF μοντέλου. Η πρώτη παραλλαγή περιλαμβάνει επιπλέον όλες τις αλληλεπιδράσεις $2^{ης}$ τάξης. Η δεύτερη παραλλαγή χρησιμοποιεί το γενικευμένο αθροιστικό μοντέλο (GAM).

- Το μοντέλο λογιστικής παλινδρόμησης με χρήση backward elimination. Το αρχικό μοντέλο περιλάμβανε 28 μεταβλητές οι οποίες σταδιακά αποκλείονταν μέχρι να παραμείνουν οι σημαντικές, αυτές με *p-value* <0.05.

- Το μοντέλο λογιστικής παλινδρόμησης ADHERE που αναπτύχθηκε χρησιμοποιώντας το μητρώο ADHERE. Αυτό το μοντέλο περιελάμβανε τις εξής μεταβλητές: BUN, συστολική αρτηριακή πίεση, καρδιακός ρυθμός και ηλικία. Κάθε μια από αυτές τις μεταβλητές θεωρήθηκε ότι σχετίζεται γραμμικά με τα log odds της θνησιμότητας υπό νοσηλεία. Αρχικά χρησιμοποιήθηκε το ADHERE μοντέλο με τις μεταβλητές από το μητρώο ADHERE. Στη συνέχεια, το μοντέλο ADHERE ρυθμίστηκε για το δείγμα EFFECT. Προβλέψεις υπολογίστηκαν για κάθε στοιχείο του δεύτερου EFFECT validation δείγματος.

2.3.2.2 Μοντέλα δέντρων παλινδρόμησης

Χρησιμοποιήθηκαν τρία διαφορετικά μοντέλα δέντρων παλινδρόμησης. Πιθανότητες για θνησιμότητα υπό νοσηλεία υπολογίστηκαν για κάθε στοιχείο του δεύτερου EFFECT validation δείγματος.

- Το ADHERE δέντρο παλινδρόμησης περιλαμβάνει τις μεταβλητές: BUN, συστολική αρτηριακή πίεση και κρεατινίνη ορού. Έχει πέντε τερματικούς κόμβους ή φύλλα με προβλεπόμενες πιθανότητες για θνησιμότητα υπό νοσηλεία που κυμαίνονται από 2,14% έως 21,94%. Αυτό το δέντρο χρησιμοποίησε τις προβλεπόμενες πιθανότητες που υπολόγισε ο Fonarrow για το δείγμα του που περιλαμβάνει 32.046 περιστατικά. Για αυτό το προηγούμενο δέντρο, οι προβλέψεις υπολογίστηκαν απ' ευθείας από κάθε στοιχείο του δεύτερου EFFECT validation δείγματος.
- Το επαναβαθμονομημένο (recalibrated) δέντρο παλινδρόμησης Fonarrow. Πρόκειται για το ADHERE δέντρο παλινδρόμησης, τροποποιημένο για το δείγμα EFFECT. Με αυτήν τη προσέγγιση, οι προβλεπόμενες πιθανότητες για κάθε τερματικό κόμβο του ADHERE δέντρου αντικαταστάθηκαν από την εκτιμώμενη

πιθανότητα για τα στοιχεία του EFFECT δείγματος που βρίσκονται στον τερματικό κόμβο. Στη συνέχεια, οι προβλεπόμενες πιθανότητες υπολογίστηκαν για κάθε στοιχείο του δεύτερου EFFECT validation δείγματος.

- Δέντρο παλινδρόμησης που προέκυψε από το δείγμα EFFECT. Για τη δημιουργία του, το αρχικό δείγμα EFFECT χωρίστηκε τυχαία σε δύο μέρη: Το βασικό EFFECT (A) και το βασικό EFFECT (B). Το μέρος (A) περιέχει τα 2/3 των στοιχείων του αρχικού EFFECT δείγματος και το μέρος (B) περιέχει το υπόλοιπο 1/3. Ένα πρώτο δέντρο αναπτύχθηκε για το μέρος (A) χρησιμοποιώντας και τις 28 μεταβλητές. Για την ανάπτυξη του δέντρου χρησιμοποιήθηκε επαναλαμβανόμενος δυαδικός διαμερισμός. Τα κριτήρια για την ανάπτυξη ήταν τα ακόλουθα: Σε κάθε κόμβο, επιλεγόταν ο διαμερισμός που μεγιστοποιούσε την μείωση της απόκλισης. Το μικρότερο αποδεκτό μέγεθος κόμβου ήταν 10. Ένας κόμβος δεν διαμεριζόταν εάν η απόκλιση μέσα σε αυτόν ήταν μικρότερη από το 0.01 του κόμβου-ρίζα. Μόλις κατασκευάστηκε το πρώτο δέντρο, ακολούθησε κλάδεμα. Ο βέλτιστος αριθμός φύλλων καθορίστηκε αναγνωρίζοντας το μέγεθος δέντρου που ελαχιστοποιούσε την απόκλιση στο δέντρο όταν το μέρος (B) χρησιμοποιούνταν σαν validation δείγμα. Το πρώτο δέντρο κλαδευόταν έως ότου φτάσει στο επιθυμητό μέγεθος. Το δέντρο που προέκυψε χρησιμοποιήθηκε για να προβλεφθεί η πιθανότητα θνησιμότητας υπό νοσηλεία για κάθε στοιχείο στο δεύτερο EFFECT validation δείγμα.

Demographic and clinical characteristics of the 8,236 heart failure patients in the EFFECT baseline derivation study sample

Variable	EFFECT baseline sample ($N = 8,236$)	In-hospital death: No ($N = 7,613$)	In-hospital death: Yes ($N = 623$)	P-value
Demographic characteristics				
Age, years	77 (70–84)	77 (69–83)	82 (76–88)	<0.001
Female	4,154 (50.4%)	3,820 (50.2%)	334 (53.6%)	0.099
Vital signs on admission				
Systolic blood pressure, mm Hg	146 (126–170)	148 (128–171)	130 (110–150)	<0.001
Heart rate, beats per minute	92 (76–110)	92 (76–110)	92 (78–110)	0.664
Respiratory rate, breaths per minute	24 (20–30)	24 (20–30)	26 (20–32)	<0.001
Presenting signs and physical examination				
Neck vein distension	4,516 (54.8%)	4,202 (55.2%)	314 (50.4%)	0.021
S3	785 (9.5%)	750 (9.9%)	35 (5.6%)	<0.001
S4	302 (3.7%)	293 (3.8%)	9 (1.4%)	0.002
Rales >50% of lung field	902 (11.0%)	791 (10.4%)	111 (17.8%)	<0.001
Findings on chest X-ray				
Pulmonary edema	4,215 (51.2%)	3,909 (51.3%)	306 (49.1%)	0.285
Cardiomegaly	2,944 (35.7%)	2,737 (36.0%)	207 (33.2%)	0.172
Past medical history				
Diabetes	2,871 (34.9%)	2,675 (35.1%)	196 (31.5%)	0.064
Cerebrovascular accident (CVA)/Transient ischemic attack (TIA)	1,372 (16.7%)	1,220 (16.0%)	152 (24.4%)	<0.001
Previous MI	3,021 (36.7%)	2,804 (36.8%)	217 (34.8%)	0.319
Atrial fibrillation	2,402 (29.2%)	2,205 (29.0%)	197 (31.6%)	0.161
Peripheral vascular disease	1,082 (13.1%)	986 (13.0%)	96 (15.4%)	0.081
Chronic obstructive pulmonary disease	1,404 (17.0%)	1,265 (16.6%)	139 (22.3%)	<0.001
Dementia	642 (7.8%)	513 (6.7%)	129 (20.7%)	<0.001
Cirrhosis	63 (0.8%)	54 (0.7%)	9 (1.4%)	0.043
Cancer	948 (11.5%)	854 (11.2%)	94 (15.1%)	0.004
Electrocardiogram—first available within 48 hr				
Left bundle branch block	1,232 (15.0%)	1,127 (14.8%)	105 (16.9%)	0.168
Laboratory tests				
Hemoglobin, g/L	124 (110–138)	125 (110–138)	121 (105–136)	<0.001
White blood count, 10E9/L	9 (7–12)	9 (7–12)	10 (8–13)	<0.001
Sodium, mmol/L	139 (136–141)	139 (136–141)	138 (134–141)	<0.001
Potassium, mmol/L	4 (4–5)	4 (4–5)	4 (4–5)	<0.001
Glucose, mmol/L	8 (6–11)	8 (6–11)	8 (6–11)	0.007
BUN, mmol/L	8 (6–12)	8 (6–12)	12 (9–18)	<0.001
Creatinine, μmol/L	106 (83–145)	105 (83–142)	129 (95–185)	<0.001

Continuous variables are reported as medians (25th–75th percentiles) and dichotomous variables are reported as N (%).
Abbreviations: EFFECT, Enhanced Feedback for Effective Cardiac Treatment; BUN, blood urea nitrogen.

***Πίνακας 2.3**: Δημογραφικά και κλινικά στοιχεία των ασθενών που χρησιμοποιήθηκαν στη μελέτη*

2.3.3 Μέτρηση της προβλεπτικής ικανότητας

Για τη μέτρηση της προβλεπτικής ικανότητας κάθε μεθόδου, υπολογίστηκαν οι προβλεπόμενες πιθανότητες θνησιμότητας υπό νοσηλεία για τα στοιχεία του δεύτερου EFFECT validation δείγματος χρησιμοποιώντας όλες τις μεθόδους. Η προβλεπτική ικανότητα του συνοψίστηκε από το εμβαδό κάτω από την καμπύλη ROC. Η μελέτη εκφράζει επιπλέον την προβλεπτική ικανότητα χρησιμοποιώντας τον γενικευμένο δείκτη R_N^2 των Nagelkerke, Cragg και Uhler και τον δείκτη Brier.

2.3.4 Εξέταση της σταθερότητας των μεθόδων ανάλυσης που βασίζονται στα δεδομένα

Χρησιμοποιήθηκαν δύο μέθοδοι που βασίζονται στα δεδομένα για να αναπτυχθούν προβλεπτικά μοντέλα: Η μέθοδος backward elimination και τα δέντρα παλινδρόμησης. Εξετάστηκε η σταθερότητα των μοντέλων που προέκυψαν από αυτές τις μεθόδους.

2.3.4.1 Σταθερότητα των δέντρων παλινδρόμησης για την πρόβλεψη της θνησιμότητας υπό νοσηλεία ασθενών που εισήχθησαν λόγω καρδιακής ανεπάρκειας

Το αρχικό δείγμα EFFECT χωρίστηκε τυχαία σε δύο μέρη: Το πρώτο μέρος περιέχει τα 2/3 των στοιχείων του αρχικού EFFECT δείγματος και το δεύτερο μέρος περιέχει το υπόλοιπο 1/3. Χρησιμοποιώντας παρόμοια μεθοδολογία με αυτήν που παρουσιάστηκε στην ενότητα 2.3.2.2 για την $3^{η}$ περίπτωση, ένα αρχικό δέντρο αναπτύχθηκε για το πρώτο μέρος και κλαδεύτηκε ώστε να ελαχιστοποιηθεί η απόκλιση στο δεύτερο μέρος. Η προβλεπτική ικανότητα του δέντρου παλινδρόμησης που προέκυψε καθορίστηκε χρησιμοποιώντας το δεύτερο EFFECT validation δείγμα. Για το δέντρο παλινδρόμησης που αναπτύχθηκε στο τυχαίο derivation δείγμα που προέκυψε από το αρχικό EFFECT δείγμα, σημειώθηκαν τα ακόλουθα χαρακτηριστικά: Αριθμός φύλλων και τερματικών κόμβων, αριθμός μεταβλητών που χρησιμοποιήθηκαν για την κατασκευή του δέντρου, την πρώτη μεταβλητή στην οποία πραγματοποιήθηκε διαχωρισμός στο δέντρο που προέκυψε, η τιμή για την οποία πραγματοποιήθηκε ο διαχωρισμός για αυτήν τη μεταβλητή. Η διαδικασία επαναλήφθηκε 1000 φορές και συνοψίστηκε για τα 1000 τυχαία επιλεγμένα derivation συστατικά.

2.3.4.2 Σταθερότητα των μοντέλων λογιστικής παλινδρόμησης που αναπτύχτηκαν με χρήση backward elimination

Για την μελέτη της σταθερότητας αυτών των μοντέλων, χρησιμοποιήθηκαν bootstrapping μέθοδοι. Εξήχθησαν 1000 bootstrap δείγματα από το αρχικό EFFECT

δείγμα. Σε κάθε bootstrap δείγμα, χρησιμοποιήθηκε backward elimination για την ανάπτυξη ενός ελάχιστου μοντέλου για την πρόβλεψη της θνησιμότητας υπό νοσηλεία. Σημειώθηκαν οι μεταβλητές που επιλέχτηκαν για να περιληφθούν στο τελικό μοντέλο. Επιπλέον, εξετάστηκε η προβλεπτική ικανότητα του τελικού μοντέλου στο δεύτερο EFFECT validation δείγμα. Τα αποτελέσματα σταθμίστηκαν στα 1000 bootstrap δείγματα.

2.3.4.3 Χαρακτηρισμός της σχέσης ανάμεσα στις σημαντικές συνεχείς μεταβλητές και στην θνησιμότητα ασθενών υπό νοσηλεία λόγω καρδιακής ανεπαρκείας

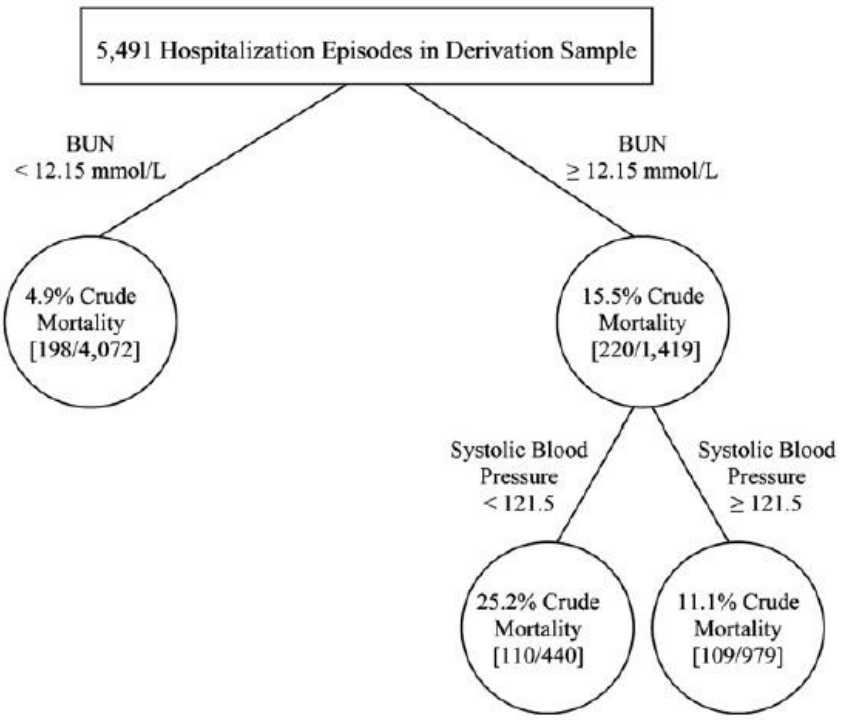
Η σχέση ανάμεσα στις συνεχείς μεταβλητές πρόβλεψης και στα log odds της θνησιμότητας των ασθενών υπό νοσηλεία περιγράφηκε με τη χρήση της GAM. Από το αρχικό EFFECT derivation δείγμα, εξάχθηκαν 1000 αυτοδύναμα δείγματα και στη συνέχεια το μοντέλο GAM προσαρμόστηκε σε κάθε ένα από αυτά. Για κάθε τιμή των τεσσάρων συνεχών μεταβλητών πρόβλεψης καθορίστηκαν οι προβλεπόμενες log odds για τη θνησιμότητα υπό νοσηλεία, κρατώντας τις υπόλοιπες συνεχείς μεταβλητές σταθερές στην διάμεσο του αρχικού EFFECT δείγματος και τις διχοτομικές μεταβλητές πρόβλεψης στο mode του αρχικού EFFECT δείγματος. Αφήνοντας κάθε μια από τις τέσσερις μεταβλητές πρόβλεψης να αυξηθούν σε όλο το παρατηρούμενο εύρος τους στο αρχικό EFFECT δείγμα, υπολογίστηκαν τα μέσα log odds της υπό νοσηλεία θνησιμότητας και τα 2.5^{α} και 97.5^{α} εκατοστιαία σημεία των log odds της υπό νοσηλεία θνησιμότητας σε όλα τα 1000 μοντέλα που προσαρμόστηκαν στα αυτοδύναμα δείγματα.

2.3.5 Αποτελέσματα

Τα δημογραφικά και κλινικά χαρακτηριστικά των ασθενών στο αρχικό EFFECT derivation δείγμα περιγράφονται στον Πίνακα 2.3. Η επικρατούσα τιμή των διχοτομικών μεταβλητών και οι διάμεσοι, τα 25^{α} και 75^{α} εκατοστιαία σημεία των συνεχών μεταβλητών που λήφθηκαν υπ' όψη για τη συγκεκριμένη, μελέτη αναφέρονται για όλο το αρχικό EFFECT δείγμα και ξεχωριστά για τους ασθενείς που απεβίωσαν πριν βγουν από το νοσοκομείο και για αυτούς που επιβίωσαν μέχρι να βγουν. Η μέση ηλικία των ασθενών ήταν τα 77 χρόνια (διάστημα 70 – 84) και το 50,4% τω ασθενών ήταν γυναίκες. Συνολικά, 623 ασθενείς (το 7,6%) απεβίωσαν πριν βγουν από το νοσοκομείο. Οι έλεγχοι Kruskal – Wallis και chi-squared χρησιμοποιήθηκαν για να συγκριθούν τα συνεχή και κατηγορικά χαρακτηριστικά αντίστοιχα, ανάμεσα στους ασθενείς που απεβίωσαν πριν βγουν από το νοσοκομείο και σε αυτούς που επιβίωσαν μέχρι να βγουν. Υπήρχαν στατιστικά σημαντικές διαφορές σε μια σειρά βασικά χαρακτηριστικά ανάμεσα σε αυτούς που απεβίωσαν και σε αυτούς που επιβίωσαν μέχρι να βγουν από το νοσοκομείο.

2.3.5.1 Σύγκριση της προβλεπτικής ικανότητας ανάμεσα στα μοντέλα λογιστικής παλινδρόμησης και στα δέντρα παλινδρόμησης για την πρόβλεψη της θνησιμότητας υπό νοσηλεία ασθενών που εισήχθησαν λόγω καρδιακής ανεπάρκειας

Στον Πίνακα 2.4 παρουσιάζονται τα στοιχεία για το εμβαδό κάτω από την καμπύλη ROC για κάθε μέθοδο στο δεύτερο EFFECT validation δείγμα. Σε αυτό το δείγμα, το εμβαδό της ROC για το ADHERE δέντρο παλινδρόμησης ήταν 0,650, ενώ για το επαναβαθμονομημένο ADHERE δέντρο παλινδρόμησης το εμβαδό ήταν 0,620. Το δέντρο παλινδρόμησης για το αρχικό EFFECT derivation δείγμα φαίνεται στο Σχήμα 2.3 και η εφαρμογή του στο δεύτερο EFFECT validation δείγμα, παρουσίασε εμβαδό με τιμή 0,633. Το εμβαδό για το EFFECT-HF μοντέλο πρόβλεψης ήταν 0,772. Η μέθοδος GAM αύξησε ελάχιστα το εμβαδό στο 0,773. Η παραλλαγή του EFFECT-HF που περιλαμβάνει όλες τις αλληλεπιδράσεις $2^{ης}$ τάξης ανάμεσα στις μεταβλητές, εμφάνισε στο εμβαδό κάτω από την ROC την τιμή 0,765. Το επαναβαθμονομημένο ADHERE μοντέλο λογιστικής παλινδρόμησης εμφάνισε εμβαδό για τη ROC 0,751 στο δεύτερο EFFECT validation δείγμα, ενώ το μοντέλο με τους αρχικούς συντελεστές παλινδρόμησης είχε εμβαδό ROC 0,747. Ο Πίνακας 2.4 περιλαμβάνει επιπλέον τις τιμές του δείκτη R_N^2 και το Brier's score για κάθε μοντέλο.



Σχήμα2.3: *Δέντρο παλινδρόμησης για το αρχικό EFFECT derivation δείγμα*

2.3.5.2 Αναπαραγωγιμότητα των μεθόδων ανάλυσης που βασίζονται στα δεδομένα

Παρουσιάστηκε σημαντική ανομοιογένεια στα δέντρα παλινδρόμησης που αναπτύχθηκαν στα 1000 derivation δείγματα που εξήχθησαν τυχαία από το αρχικό EFFECT derivation δείγμα. Ο αριθμός των τερματικών κόμβων ή φύλλων κυμαινόταν από 1 το ελάχιστο έως 6 το μέγιστο (ένα δέντρο με 1 μόνο τερματικό κόμβο σημαίνει ότι δεν υπήρχαν δυαδικοί διαμερισμοί και ότι όλα τα στοιχεία βρίσκονται στον τερματικό κόμβο που είναι η ρίζα του δέντρου). Τα ποσοστά των δέντρων με 1, 2, 3, 4, 5 και 6 κόμβους ήταν 0.2%, 8.0%, 49.2%, 32.2%, 8.9% και 1.6% αντίστοιχα. Ο αριθμός των μεταβλητών που χρειάστηκαν για την κατασκευή των δέντρων παλινδρόμησης κυμαινόταν από 0 το ελάχιστο έως 5 το μέγιστο. Τα ποσοστά των δέντρων που χρησιμοποίησαν 0, 1, 2, 3, 4 και 5 μεταβλητές είναι 0.1%, 8.0%, 50.4%, 31.7%, 8.4% και 1.4% αντίστοιχα.

Στο 96.9% των δέντρων παλινδρόμησης, η πρώτη μεταβλητή που χρησιμοποιήθηκε για να καθορίσει μια διχοτόμηση ήταν το BUN. Στο 0.5% ήταν η άνοια, στο 2.5% ήταν η συστολική αρτηριακή πίεση και στο 0.1% των δέντρων δεν έγινε καμία διχοτόμηση. Στα 969 δέντρα παλινδρόμησης όπου πρώτη μεταβλητή για τη διχοτόμηση ήταν το BUN, η τιμή της μεταβλητής για την οποία έγινε η διχοτόμηση κυμαινόταν από 8.75 το ελάχιστο, έως 17.05 το μέγιστο (το $25^{ο}$ ποσοστιαίο σημείο και η διάμεσος ήταν 12.15 και το $75^{ο}$ ποσοστιαίο σημείο 12.15). Στα 25 δέντρα παλινδρόμησης στα οποία πρώτη μεταβλητή για τη διχοτόμηση ήταν η συστολική αρτηριακή πίεση, η τιμή για αυτή τη μεταβλητή ώστε να γίνει η διχοτόμηση κυμάνθηκε από 120.5 το ελάχιστο έως 121.5 το μέγιστο ($25^{ο}$ ποσοστιαίο σημείο 121.5, διάμεσος και $75^{ο}$ ποσοστιαίο σημείο 121.5).

Η μεταβλητή «ηλικία» χρησιμοποιήθηκε στο 17.8% των δέντρων παλινδρόμησης που αναπτύχθηκαν στα 1000 δείγματα που εξήχθησαν τυχαία από το αρχικό EFFECT derivation δείγμα. Άλλες μεταβλητές που χρησιμοποιήθηκαν σε τουλάχιστον ένα από τα δέντρα παλινδρόμησης ήταν: η συστολική αρτηριακή πίεση στο 80.9% των δέντρων, οι ρόγχοι στο 2.9%, η άνοια στο 18.7%, η αιμοσφαιρίνη στο 0.4%, ο αριθμός λευκών αιμοσφαιρίων στο 1.3%, το κάλιο σε ποσοστό 22.5%, το BUN στο 99.6% και το νάτριο σε ποσοστό 0.4%. Παρ' όλο που το BUN και η συστολική αρτηριακή πίεση χρησιμοποιήθηκαν στην πλειοψηφία των δέντρων παλινδρόμησης, καμία μεταβλητή δεν χρησιμοποιήθηκε σε όλα.

Στο Σχήμα 2.4 παρουσιάζονται τέσσερα από τα δέντρα παλινδρόμησης που προέκυψαν. Το δέντρο του Σχήματος 2.4(a) είχε 2 τερματικούς κόμβους και κατηγοριοποίησε τους ασθενείς σε χαμηλού κινδύνου (BUN < 12.15mmol/L με

προβλεπόμενη θνησιμότητα υπό περίθαλψη σε ποσοστό 4.6%) και σε ασθενείς υψηλού κινδύνου (BUN ≥12.15mmol/L και προβλεπόμενη θνησιμότητα σε ποσοστό 16.0%).Το δέντρο του Σχήματος 2.4(b) έχει τρεις τερματικούς κόμβους και χρησιμοποιεί δυο μεταβλητές. Σε αυτό το δέντρο, η πρώτη διχοτόμηση γίνεται για την μεταβλητή BUN (BUN < 12.15mmol/L έναντι BUN ≥ 12.15mmol/L). Οι προβλεπόμενες πιθανότητες για θνησιμότητα χωρίζονται σε τρεις κατηγορίες με ποσοστά 5.0%, 10.2% και 24.7% αντίστοιχα. Τα δέντρα παλινδρόμησης του Σχήματος 2.4 (a) και (d), έχουν 5 και 6 τερματικούς κόμβους αντίστοιχα. Στο τελευταίο δέντρο, οι προβλεπόμενες πιθανότητες θνησιμότητας κυμαίνονται σε ποσοστά από 4.2% το ελάχιστο έως 79.0% το μέγιστο.

Τα 1000 δέντρα παλινδρόμησης που προέκυψαν από τυχαία δείγματα που εξήχθησαν από το αρχικό EFFECT derivation δείγμα χρησιμοποιήθηκαν για να υπολογιστούν προβλέψεις για τη θνησιμότητα υπό νοσηλεία στο δεύτερο EFFECT validation δείγμα. Το μέσο εμβαδό των καμπύλων ROC για το δεύτερο EFFECT validation δείγμα ήταν 0.637. Οι τιμές του εμβαδού κυμαίνονταν από 0.5 έως 0.676. Τα 25^{α} και 75^{α} εκατοστιαία σημεία ήταν 0.633 και 0.637 αντίστοιχα. Ακόμα και το δέντρο παλινδρόμησης με την καλύτερη προβλεπτική ικανότητα στο δεύτερο EFFECT validation δείγμα έχει μικρότερη προβλεπτική ικανότητα από το EFFECT-HF μοντέλο για την πρόβλεψη θνησιμότητας.

Τα 1000 μοντέλα λογιστικής παλινδρόμησης που προέκυψαν με χρήση backward elimination στα αυτοδύναμα δείγματα που εξήχθησαν από το αρχικό EFFECT derivation δείγμα χρησιμοποιήθηκαν για να υπολογιστούν προβλέψεις της θνησιμότητας υπό νοσηλεία στο δεύτερο EFFECT validation δείγμα. Το μέσο εμβαδό της ROC στο δεύτερο EFFECT validation δείγμα ήταν 0.772. Οι τιμές του εμβαδού κάτω από την ROC κυμαίνονταν από 0.756 έως 0.783. Το 25^{o} και 75^{o} ποσοστιαίο σημείο ήταν 0.769 και 0.774 αντίστοιχα. Επομένως ακόμα και το μοντέλο λογιστικής παλινδρόμησης με το μικρότερο εμβαδό κάτω από την ROC στο δεύτερο EFFECT validation δείγμα, είχε εμβαδό ROC μεγαλύτερο από τις τρεις διαφορετικές μεθόδους δέντρων παλινδρόμησης.

Παρ' όλο που το εμβαδό κάτω από την ROC είχε μια σχετική σταθερότητα για τα 1000 μοντέλα λογιστικής παλινδρόμησης που κατασκευάστηκαν με χρήση της backward elimination μεθόδου, τα μοντέλα διέφεραν όσον αφορά τις μεταβλητές τις οποίες διατηρούσαν. Ο αριθμός των μεταβλητών που παρέμεναν κυμαινόταν από 9 έως 21 με διάμεσο 14 (25^{o} και 75^{o} ποσοστιαίο σημείο 13 και 16 αντίστοιχα). Κάθε μία από τις 28 υποψήφιες μεταβλητές περιλήφθηκαν τουλάχιστον στο 2.9% των τελικών μοντέλων παλινδρόμησης. Οι μεταβλητές ηλικία, συστολική αρτηριακή πίεση, άνοια και BUN περιλήφθηκαν και στα 1000 μοντέλα λογιστικής παλινδρόμησης που προέκυψαν από backward elimination. Οι μεταβλητές κολπική μαρμαρυγή, θηλυκό γένος, διαβήτης, αποκλεισμός του αριστερού σκέλους και πνευμονικό οίδημα διατηρήθηκαν στο 9.7%, 8.9%, 6.9%, 3.5% και 2.9% των τελικών μοντέλων παλινδρόμησης αντίστοιχα.

Predictive accuracy of the different models in the EFFECT follow-up validation sample

Model	ROC curve area	Generalized R^2_N index	Brier score
Logistic regression models			
EFFECT-HF mortality model	0.772	0.154	0.062
EFFECT-HF model with two-way interactions	0.765	0.141	0.063
Generalized additive model	0.773	0.154	0.062
Logistic regression—backward variable elimination	0.775	0.161	0.061
Logistic regression (ADHERE model—original coefficients)	0.747	0.136	0.063
Logistic regression (ADHERE model recalibrated)	0.751	0.142	0.061
Regression trees			
Regression tree (ADHERE tree)	0.651	0.065	0.066
Regression tree (recalibrated ADHERE tree)	0.620	0.043	0.059
Regression tree (grown to sample)	0.633	0.055	0.064

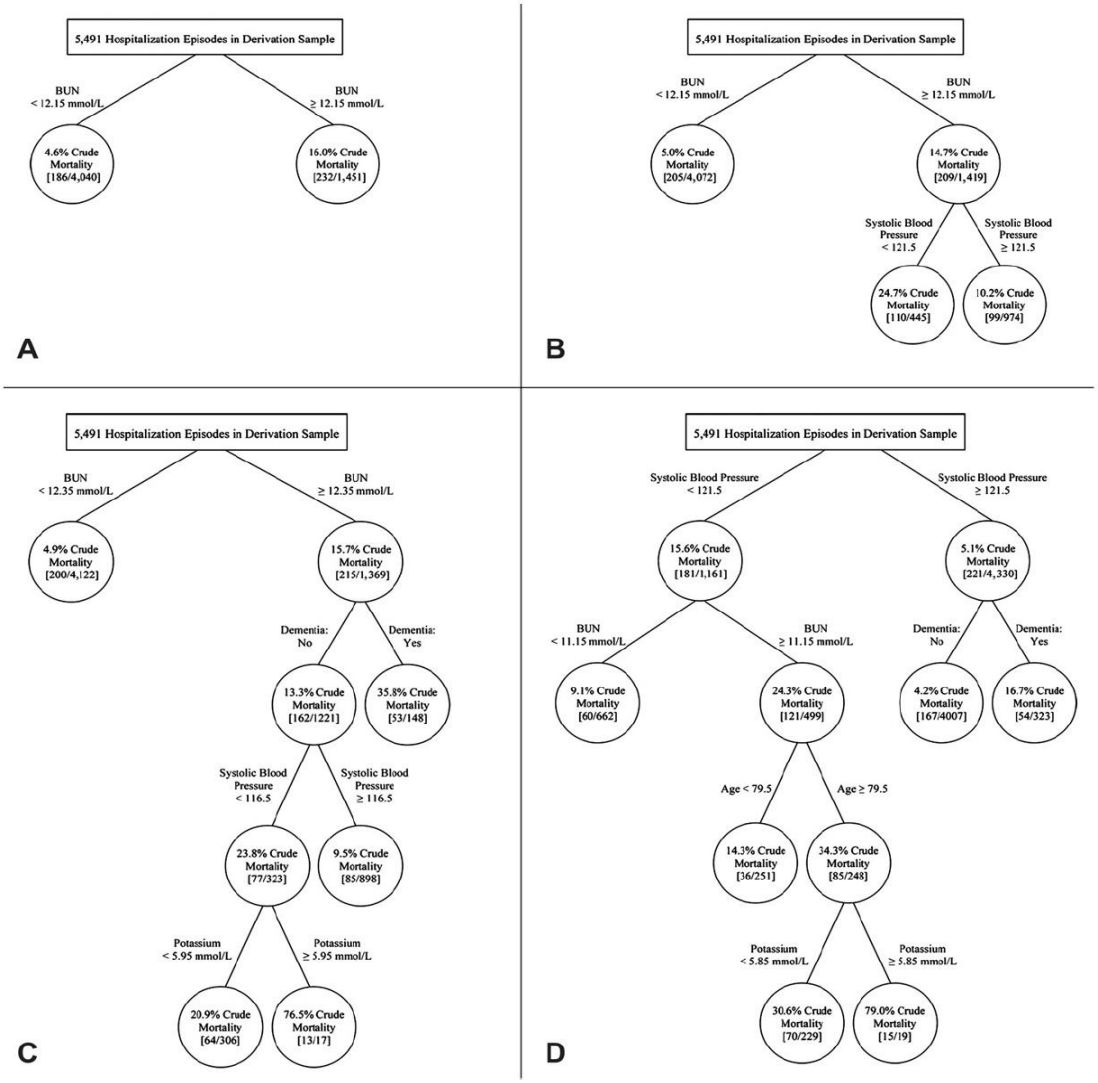
Abbreviations: ROC, receiver operating characteristic; EFFECT-HF, Enhanced Feedback for Effective Cardiac Treatment in Heart Failure.

Πίνακας 2.4: *στοιχεία για το εμβαδό κάτω από την καμπύλη ROC για κάθε μέθοδο στο δεύτερο EFFECT validation δείγμα*

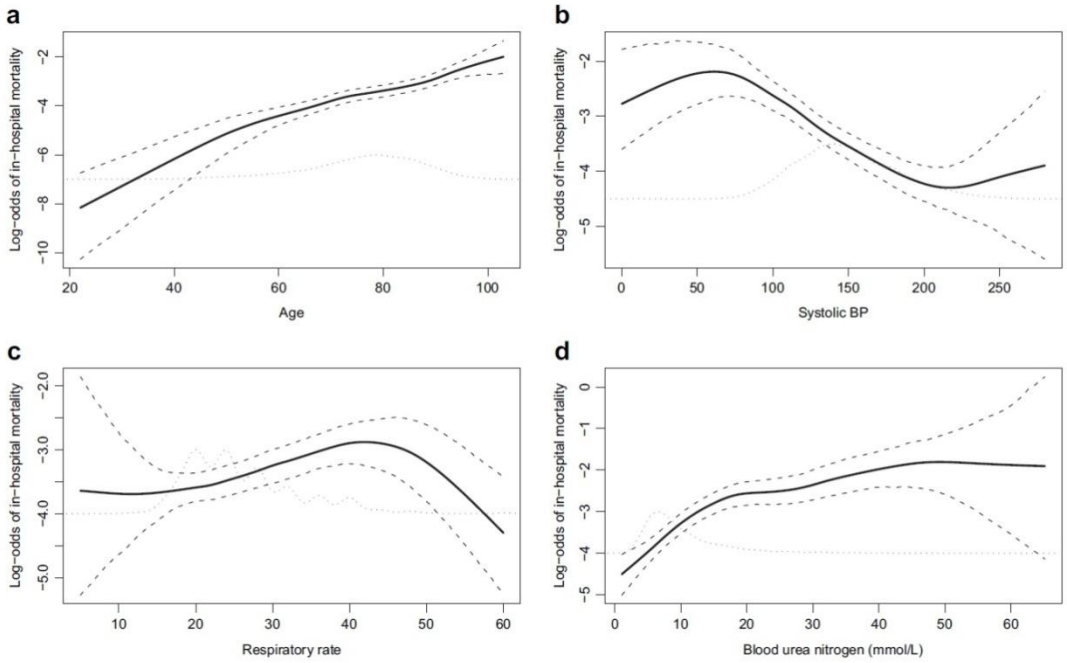
2.3.5.3 Σχέση ανάμεσα σε κρίσιμες συνεχείς μεταβλητές και στην θνησιμότητα ασθενών υπό νοσηλεία λόγω καρδιακής ανεπαρκείας

Στο Σχήμα 2.5 περιγράφεται η σχέση ανάμεσα στις μεταβλητές ηλικία, συστολική αρτηριακή πίεση, ρυθμός αναπνευστικού, BUN και στα μέσα log odds της υπό νοσηλεία θνησιμότητας. Επιπλέον παρουσιάζονται τα εμπειρικά 2.5^{α} και 97.5^{α} εκατοστιαία σημεία των προβλεπόμενων log odds για τη θνησιμότητα σε όλα τα 1000 αυτοδύναμα δείγματα που εξήχθησαν από το αρχικό EFFECT derivation δείγμα. Σε κάθε ένα από τα τέσσερα μέρη του σχήματος, υπερτίθεται η συνάρτηση πυκνότητας που περιγράφει την κατανομή της δεδομένης κατανομής στο αρχικό EFFECT δείγμα. Παρατηρείται ότι η σχέση ανάμεσα στην ηλικία και τα log odds της υπό νοσηλεία θνησιμότητας προσεγγίζει τη γραμμικότητα σε όλο το φάσμα των ηλικιών. Για της υπόλοιπες τρεις συνεχείς μεταβλητές, είναι φανερό ότι οι σχέση είναι μη-γραμμική. Παρ' όλα αυτά, για τη συστολική αρτηριακή πίεση, το ρυθμό του αναπνευστικού και

το BUN, οι σχέσεις ήταν κατά κύριο λόγο γραμμικές για ένα φάσμα της κατανομής των στοιχείων.



Σχήμα 2.4 *:Τέσσερα διαφορετικά δέντρα παλινδρόμησης για την πρόβλεψη θνησιμότητας υπό νοσηλεία ασθενών με καρδιακή ανεπάρκεια*



Σχήμα 2.5: *Σχέση μεταξύ των συνεχών μεταβλητών και των log odds της θνησιμότητας υπό νοσηλεία σε 1.000 αυτοδύναμα δείγματα EFFECT. (a) Ηλικία και log odds της θνησιμότητας (b) Συστολική αρτηριακή πίεση και log odds της θνησιμότητας (c) Ρυθμός αναπνευστικού και log odds της θνησιμότητας (d) Άζωτο στην ουρίας του αίματος και log odds της θνησιμότητας*

2.3.6 Συμπεράσματα

Η μελέτη είχε τρία βασικά συμπεράσματα. Κατ' αρχήν, τα δέντρα παλινδρόμησης δεν προέβλεψαν την θνησιμότητα υπό νοσηλεία λόγω καρδιακής ανεπάρκειας με τόση ακρίβεια όση τα μοντέλα λογιστικής παλινδρόμησης. Δεύτερον, ότι πολλαπλά δέντρα παλινδρόμησης μπορούν να αναπτυχθούν σε δείγματα τα οποία δεν διαφέρουν συστηματικά μεταξύ τους. Τρίτον, ότι ορισμένες βασικές αρχικές μεταβλητές είχαν σχέση με τη θνησιμότητα υπό νοσηλεία, η οποία ήταν προσεγγιστικά γραμμική σε κάποιο φάσμα τιμών αυτών των μεταβλητών στα περισσότερα στοιχεία.

Παρά το γεγονός ότι η μέθοδος των δέντρων παλινδρόμησης είναι αρκετά απλή, τα μοντέλα που χρησιμοποιήθηκαν είχαν μικρότερη προβλεπτική ακρίβεια σε σχέση με όλα τα μοντέλα λογιστικής παλινδρόμησης. Τα αποτελέσματα έδειξαν ότι η πρόβλεψη της θνησιμότητας με δέντρα παλινδρόμησης είναι δυνητικά λαθεμένη λόγω της χαμηλής διακριτικής ικανότητας. Το εμβαδό κάτω από την ROC για το EFFECT-HF μοντέλο πρόβλεψης ήταν 0.772 στο δεύτερο EFFECT validation δείγμα, ενώ για το ADHERE δέντρο παλινδρόμησης, το εμβαδό ήταν 0.651. Τα αποτελέσματα δείχνουν ότι η επιλογή για τη μοντελοποίηση της θνησιμότητας από καρδιακή ανεπάρκεια (και πιθανών και από άλλες ασθένειες) μέσω λογιστικής παλινδρόμησης ή δέντρων παλινδρόμησης , είναι μια επιλογή ανάμεσα σε ένα

μοντέλο που είναι απλό ή είναι ακριβές στην πρόβλεψη. Ωστόσο, αυτές οι δύο παράμετροι δεν είναι συγκρίσιμες, καθώς η προβλεπτική ακρίβεια είναι σημαντική για μοντέλα θνησιμότητας, ενώ η απλότητα είναι μια ιδιότητα που είναι ευνοϊκή για γενικότερη κλινική χρήση.

Τα δέντρα παλινδρόμησης που αναπτύχθηκαν σε διαφορετικά, τυχαία σχεδιασμένα δείγματα που προήλθαν από το ίδιο αρχικό derivation δείγμα, εμφάνισαν διαφορές μεταξύ τους. Ο αριθμός των τερματικών κόμβων κυμαινόταν από 0 έως 6 στα 1000 διαφορετικά δέντρα παλινδρόμησης που αναπτύχθηκαν. Επίσης, τρεις διαφορετικές μεταβλητές χρησιμοποιήθηκαν ως πρώτες μεταβλητές για διαμερισμό στα 1000 διαφορετικά δέντρα παλινδρόμησης. Τέλος, ακόμα και ανάμεσα στα δέντρα που χρησιμοποίησαν το BUN σαν πρώτη μεταβλητή για διχοτόμηση, υπήρχαν κλινικά σημαντικές διαφορές στις τιμές για τις οποίες αυτή η μεταβλητή χρησιμοποιήθηκε για τη διχοτόμηση. Η τιμή του BUN κυμαινόταν από 8.75 έως 17.05mmol/L. Η μη-αναπαραγωγή των μοντέλων που προέρχονται από μεθόδους ανάλυσης που βασίζονται στα δεδομένα έχει παρατηρηθεί και στο παρελθόν. Τα αποτελέσματα αυτής της έρευνας συμφωνούν με την παρατήρηση ότι τυχαίες μεταβολές ανάμεσα σε δείγματα μπορεί να έχουν σαν αποτέλεσμα οι μέθοδοι ανάλυσης που προέρχονται από δεδομένα να αποφέρουν διαφορετικά μοντέλα. Όταν χρησιμοποιήθηκε η μέθοδος του επαναλαμβανόμενου αποκλεισμού μεταβλητών στα 1000 αυτοδύναμα δείγματα που εξήχθησαν από το derivation δείγμα, τα μοντέλα που προέκυψαν διέφεραν από δείγμα σε δείγμα. Ωστόσο, όλα τα μοντέλα εμφάνισαν μεγαλύτερο εμβαδό κάτω από την ROC στο validation δείγμα σε σχέση με τα δέντρα παλινδρόμησης. Οι καμπύλες ROC ήταν παρόμοιες με αυτές του μοντέλου EFFECT-HF.

Η σχέση ανάμεσα στα log odds της υπό νοσηλεία θνησιμότητας και των μεταβλητών ηλικία, συστολική αρτηριακή πίεση, ρυθμός αναπνευστικού και BUN ήταν προσεγγιστικά γραμμική για το μεγαλύτερο μέρος της κατανομής της κάθε μίας από τις μεταβλητές. Αυτό το συμπέρασμα εξηγεί το λόγο που το μοντέλο EFFECT-HF έχει καλύτερη απόδοση από τα δέντρα παλινδρόμησης. Το EFFECT-HF κατάφερε να αποκαλύψει την ισχυρή γραμμική σχέση των δεδομένων. Τα δέντρα παλινδρόμησης βασίζονται σε δυαδικούς κανόνες και έτσι χάνουν την ικανότητα να επεξεργάζονται αυτή τη σχέση. Ένα πλεονέκτημα των δέντρων παλινδρόμησης είναι η ικανότητά τους να αναγνωρίζουν της αλληλεπιδράσεις ανάμεσα στις βασικές μεταβλητές. Χρησιμοποιήθηκαν δύο διαφορετικά μοντέλα λογιστικής παλινδρόμησης. Ένα που περιείχε μόνο τις κύριες επιδράσεις και ένα που περιλάμβανε και τις αλληλεπιδράσεις $2^{ης}$ τάξης. Οι επιδόσεις των δύο μοντέλων ήταν σχεδόν παρόμοιες. Παρ' όλο που υπάρχουν ισχυρές αλληλεπιδράσεις, η χρήση τους δεν βελτιώνει την πρόβλεψη του μοντέλου για τη θνησιμότητα. Τα δέντρα παλινδρόμησης έχουν δυσκολία να αντιληφθούν τις αθροιστικές σχέσεις. Αυτό μπορεί να συνέβαλε στην χαμηλή απόδοση αυτής της μεθόδου στο δείγμα που χρησιμοποιήθηκε.

Τα αποτελέσματα είναι παρόμοια με αυτά παλαιότερης μελέτης που σύγκρινε την απόδοση ανάμεσα στη λογιστική παλινδρόμηση και σε σύγχρονες μεθόδους παλινδρόμησης που προέρχονται από δεδομένα για την πρόβλεψη της θνησιμότητας μετά από νοσηλεία λόγω AMI. Η απόδοση της λογιστικής παλινδρόμησης ήταν ανώτερη από αυτή των δέντρων παλινδρόμησης. Ωστόσο, η απόδοση της λογιστικής παλινδρόμησης ήταν ελάχιστα κατώτερη από αυτή των GAM μοντέλων παλινδρόμησης.

Η έρευνα είχε ορισμένους περιορισμούς. Κατ' αρχήν, ο σκοπός δεν ήταν η σύγκριση ανάμεσα στη λογιστική παλινδρόμηση και τα δέντρα παλινδρόμησης γενικά. Ήταν η σύγκρισή τους στην πρόβλεψη της θνησιμότητας υπό νοσηλεία ατόμων με καρδιακή ανεπάρκεια. Τα αποτελέσματα μπορεί να μην αρμόζουν σε ασθενείς που πάσχουν από άλλες ασθένειες. Δεύτερον, η έρευνα επικεντρώθηκε στην πρόβλεψη για θνησιμότητα των ασθενών και όχι στο να κατηγοριοποιήσει την ζωτική κατάσταση κάθε ασθενούς. Παρ' όλο που τα δέντρα παλινδρόμησης μπορούν να χρησιμοποιηθούν για πρόβλεψη και κατηγοριοποίηση, η λογιστική παλινδρόμηση έχει σαν αποτέλεσμα την πρόβλεψη της θνησιμότητας για κάθε ασθενή. Το να εξαχθεί μια κατηγοριοποίηση μέσω ενός μοντέλου δέντρων παλινδρόμησης θα απαιτούσε ένα κανόνα για τη διχοτόμηση της προβλεπόμενης πιθανότητας. Λόγω της έλλειψης συμφωνίας για το ποια είναι η καλύτερη μέθοδος για τη διχοτόμηση μιας προβλεπόμενης πιθανότητας, ο σκοπός της έρευνας ήταν η σύγκριση της ευστοχίας κάθε μεθόδου στην πρόβλεψη της θνησιμότητας για κάθε ασθενή. Ένας επιπλέον περιορισμός για δυαδική κατηγοριοποίηση είναι ότι σε ορισμένες κλινικές περιπτώσεις, οι ερευνητές θέλουν να κατατάξουν τους ασθενείς σε πάνω από δύο επίπεδα ρίσκου με βάση την προβλεπόμενη πιθανότητα θνησιμότητας. Η δυαδική κατηγοριοποίηση κατατάσσει τους ασθενείς σε δύο ομάδες: αυτούς που προβλέπεται να πεθάνουν και σε αυτούς που προβλέπεται να επιβιώσουν. Ένας τρίτος περιορισμός της μελέτης ήταν ο αποκλεισμούς των ατόμων με κενά κάποια από τα στοιχεία μελέτης που χρησιμοποιήθηκαν στα μοντέλα παλινδρόμησης. Αυτά τα άτομα μπορεί να είχαν βρεθεί συστηματικά διαφορετικά από αυτά με πλήρη στοιχεία. Η έρευνα όμως είχε πρωταρχικά μεθοδολογική φύση με σκοπό τη σύγκριση της σχετικής απόδοσης δυο διαφορετικών μεθόδων εκτίμησης πιθανοτήτων θνησιμότητας. Στη συνέχεια διεξήχθησαν επιπλέον αναλύσεις για να ερευνηθεί ο λόγος που στο συγκεκριμένο δείγμα η λογιστική παλινδρόμηση ήταν ανώτερη από τα δέντρα παλινδρόμησης.

Η έρευνα περιέλαβε μοντέλα πρόβλεψης από προηγούμενες μελέτες. Το μοντέλο ADHERE και το μοντέλο EFFECT-HF. Το μοντέλο ADHERE χρησιμοποιήθηκε για το λόγο ότι το ADHERE δέντρο παλινδρόμησης ήταν το μόνο μοντέλο που ήξεραν η συγγραφείς ότι χρησιμοποιείται για την πρόβλεψη αποτελεσμάτων της καρδιακής ανεπάρκειας σε ασθενείς υπό νοσηλεία. Η εστίαση στα ADHERE μοντέλα περιόρισε τους συγγραφείς στο ποια άλλα παλαιότερα μοντέλα μπορούσε να περιλάβει η έρευνα. Το EFFECT-HF μοντέλο μπορεί να χρησιμοποιηθεί για την βραχυπρόθεσμη πρόβλεψη για τη θνησιμότητα ασθενών υπό

νοσηλεία με καρδιακή ανεπάρκεια. Τα μοντέλα ADHERE και EFFECT-HF χρησιμοποιούν ένα μικρό σύνολο στοιχείων για τους ασθενείς που είναι διαθέσιμο σε σύντομο διάστημα μετά την νοσηλεία. Υπάρχουν μοντέλα ρίσκου για θνησιμότητα ατόμων με καρδιακή ανεπάρκεια που δεν περιλήφθηκαν. Το μοντέλο MUSIC (MUerte Subita en Insuficiencia Cardiace) είναι για την πρόβλεψη θνησιμότητας κατά τη μετακομιδή, ενώ τα μοντέλα EFFECT-HF και ADHERE είναι για τους ασθενείς υπό νοσηλεία. Το OPTIMIZE-HF (Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure) μοντέλο για θνησιμότητα υπό νοσηλεία απαιτεί τη γνώση παρουσίας ή όχι συστολικής λειτουργίας της αριστερής κοιλίας, ενώ αυτή η πληροφορία είναι διαθέσιμη μόνο για ένα υποσύνολο των ασθενών του μητρώου EFFECT. Τα μοντέλα DIG (Digitalis Investigation Group) χρησιμοποιούνται για ασθενείς με υποπεριπτώσεις συστολικής λειτουργίας και προβλέπει τη θνησιμότητα σε ένα βάθος χρόνου και όχι υπό περίθαλψη. Τέλος το μοντέλο Seattle HF προβλέπει τη θνησιμότητα σε ένα, δύο ή τρία χρόνια και όχι όταν ο ασθενής είναι υπό περίθαλψη. Για τους παραπάνω λόγους, τα μοντέλα που αναφέρθηκαν προηγουμένως, δεν περιλήφθηκαν στην μελέτη.

2.4 Λογιστική παλινδρόμηση, δέντρα ταξινόμησης και παλινδρόμησης για τη μοντελοποίηση δεδομένων από οξύ έμφραγμα του μυοκαρδίου

Οι Faltus και Monhart (2008) σε μελέτη που διενέργησαν σύγκριναν τα δέντρα ταξινόμησης και παλινδρόμησης (CART) με τη λογιστική παλινδρόμηση για τη μοντελοποίηση της θνησιμότητας ασθενών υπό νοσηλεία λόγω οξέως στεφανιαίου συνδρόμου. Οι μεταβλητές πρόβλεψης που χρησιμοποιήθηκαν είναι οι πέντε παραδοσιακοί παράγοντες ρίσκου (σακχαρώδης διαβήτης, υπέρταση, υπερλιπιδαιμία, κάπνισμα και προηγούμενη κατάσταση IM) και έξι ομάδες φαρμάκων (ηπαρίνη, ασπιρίνη, betablocker, στατίνη, ACEI/ARB και θειενοπυριδίνη). Επίσης, συγκρίνεται η ικανότητα πρόβλεψης της λογιστικής παλινδρόμησης με αυτήν των δέντρων παλινδρόμησης.

2.4.1 Μέθοδοι που χρησιμοποιήθηκαν

Τα δεδομένα προήλθαν από ένα σύνολο ασθενών με οξύ έμφραγμα του μυοκαρδίου που εισήχθησαν διαδοχικά σε έξι κρατικά νοσοκομεία στην Τσεχία κατά τη χρονική περίοδο 2003-2006. Το δείγμα συλλέχτηκε από ετήσια διαγράμματα ανασκόπησης. Τα νοσοκομεία είναι του Caslav, Kutna Hora και Znojmo για τις χρονιές 2003-2006, τα Jindrichuv Hradec και Pisek το 2004 και το Chrudim τις χρονιές 2005-2006. Συνολικά περιέχονται 2415 ασθενείς (244 παραλείφτηκαν). Οι 1057 (43,77%) είναι γυναίκες και είναι κατά μέσο όρο μεγαλύτερες σε ηλικία και λιγότερο συχνά καπνιστές από τους 1358 (56,23%) άντρες. Οι μεταβλητές «κάπνισμα» και «φύλλο»

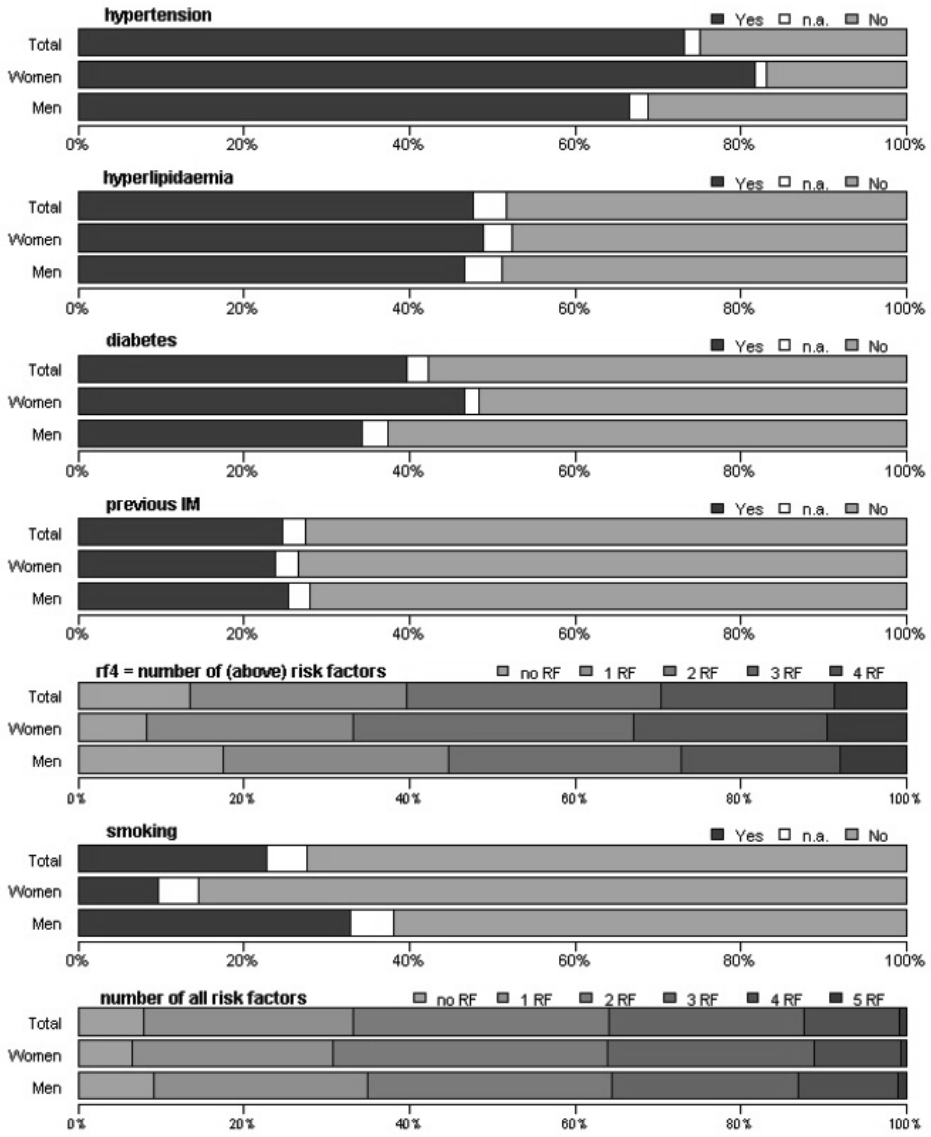
είναι σημαντικά συσχετισμένες με την μεταβλητή «ηλικία» και επίσης είναι συχνά μη-σημαντικές στην μοντελοποίηση της θνησιμότητας υπό νοσηλεία με την λογιστική παλινδρόμηση και έτσι δεν θεωρούμε την μεταβλητή «φύλλο» σαν μεταβλητή πρόβλεψης και χρησιμοποιήθηκε ο παράγοντας ρίσκου «κάπνισμα» μόνο πέρα από τον αριθμό των παρόντων παραγόντων ρίσκου. Ορίστηκε σαν rf4 ο αριθμός των παρόντων παραγόντων ρίσκου όπως φαίνονται στο Σχήμα 2.6. Λόγω του μεγάλου αριθμού κενών στα δεδομένα για τη μεταβλητή thienopyridin δεν χρησιμοποιείται όταν λαμβάνονται υπ' όψη ξεχωριστοί predictors. Ορίζονται επίσης σαν am5 και am6 ο αριθμός των χορηγούμενων φαρμάκων όπως φαίνονται στο Σχήμα 2.7.

2.4.2 Σύγκριση των μοντέλων πρόβλεψης

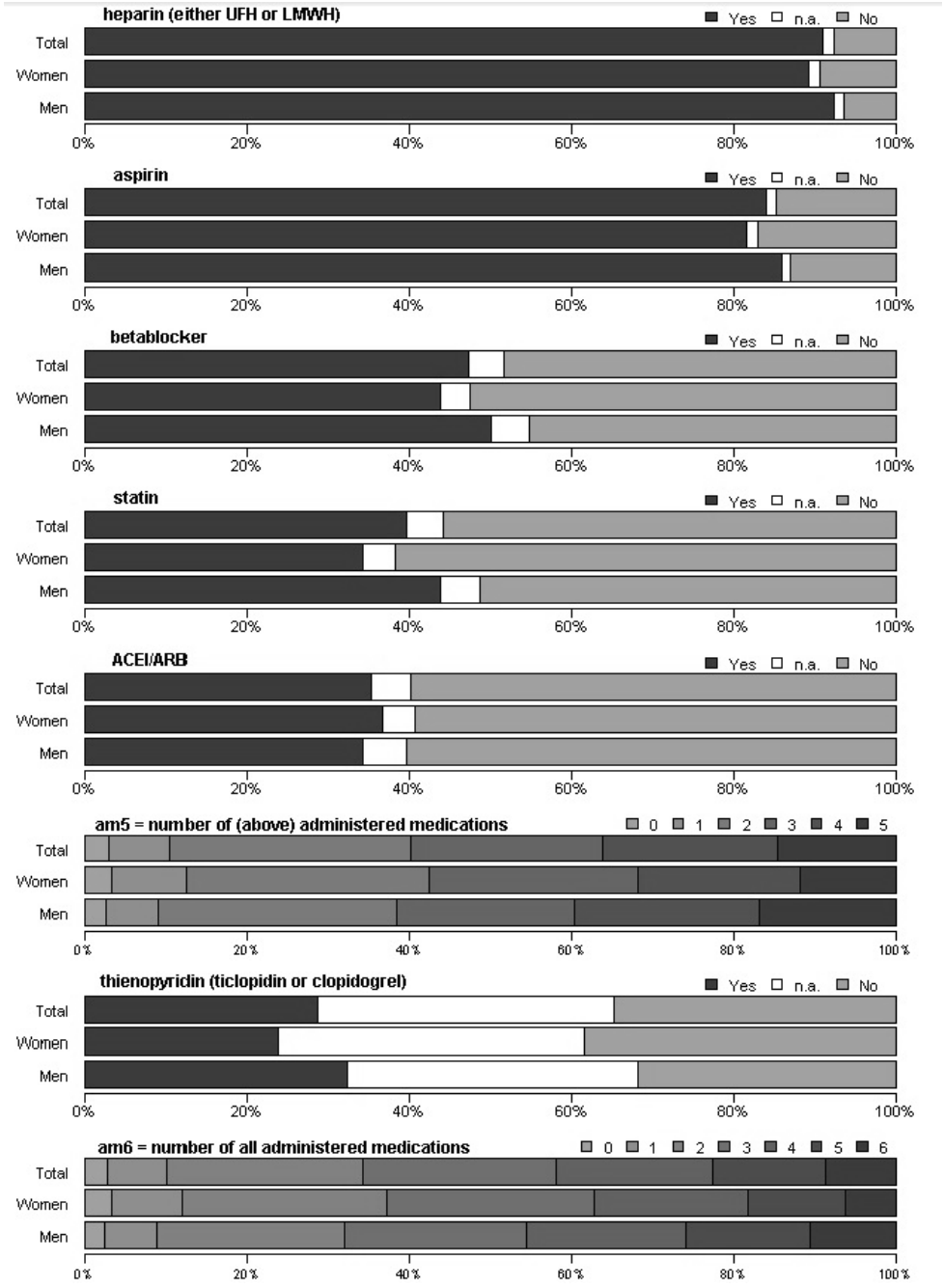
Το μοντέλα που χρησιμοποιήθηκαν είναι, για τα δέντρα παλινδρόμησης τα εξής:
$CART(counts): exitus \sim age + rf4 + am6 + smoking$,
$CART(separate): exitus \sim age + all\ RF(excepts\ moking) + all\ AM(except\ thenopyridin)$
και για τη λογιστική παλινδρόμηση:
$Logistic.reg(counts): exitus \sim age + rf4 + am6 + smoking$ και
$Logistic.reg(separate): exitus \sim age + all\ RF(except\ smoking) + all\ AM(except\ thienopyridin)$

Χρησιμοποιήθηκε η μέθοδος της επικύρωσης με επαναλαμβανόμενο διαχωρισμό για τη σύγκριση της ικανότητας πρόβλεψης των δέντρων CART και της λογιστικής παλινδρόμησης. Τα δεδομένα χωρίστηκαν τυχαία σε δείγματα validation και derivation. Τα derivation και validation δείγματα αποτελούνται από το 70% και το 30% των δεδομένων αντίστοιχα. Στη συνέχεια, κάθε μοντέλο προσαρμόστηκε στο derivation δείγμα και υπολογίστηκαν προβλέψεις για κάθε στοιχείο του validation δείγματος χρησιμοποιώντας το μοντέλο που προέκυψε από το derivation δείγμα. Η προβλεπτική ικανότητα των μοντέλων συνοψίστηκε από το εμβαδό κάτω από την καμπύλη ROC.

Το εμβαδό κάτω από την καμπύλη ROC υπολογίστηκε και για τα derivation και για τα validation δείγματα. Παρουσιάζονται και κάποια επιπλέον χαρακτηριστικά της προβλεπτικής ικανότητας των μοντέλων. Χρησιμοποιήθηκε και ο γενικευμένος δείκτης R_N^2 του Nagelkerke και το Brier's score. Όλοι οι υπολογισμοί και οι προσαρμογές των μοντέλων έγιναν με τη στατιστική γλώσσα προγραμματισμού R.



Σχήμα 2.6: *Παραδοσιακοί καρδιακοί παράγοντες ρίσκου. Το 5^{ο} διάγραμμα αναπαριστά την predictor μεταβλητή rf4 που κατασκευάστηκε.*



Σχήμα 2.7*: Οξεία φαρμακευτική θεραπεία (χορηγούμενη μέσα σε 24 ώρες από την εισαγωγή). Το 6^{ο} και 8^{ο} διάγραμμα αναπαριστούν τις μεταβλητές am5 και am6 αντίστοιχα που κατασκευάστηκαν.*

2.4.3 Αποτελέσματα

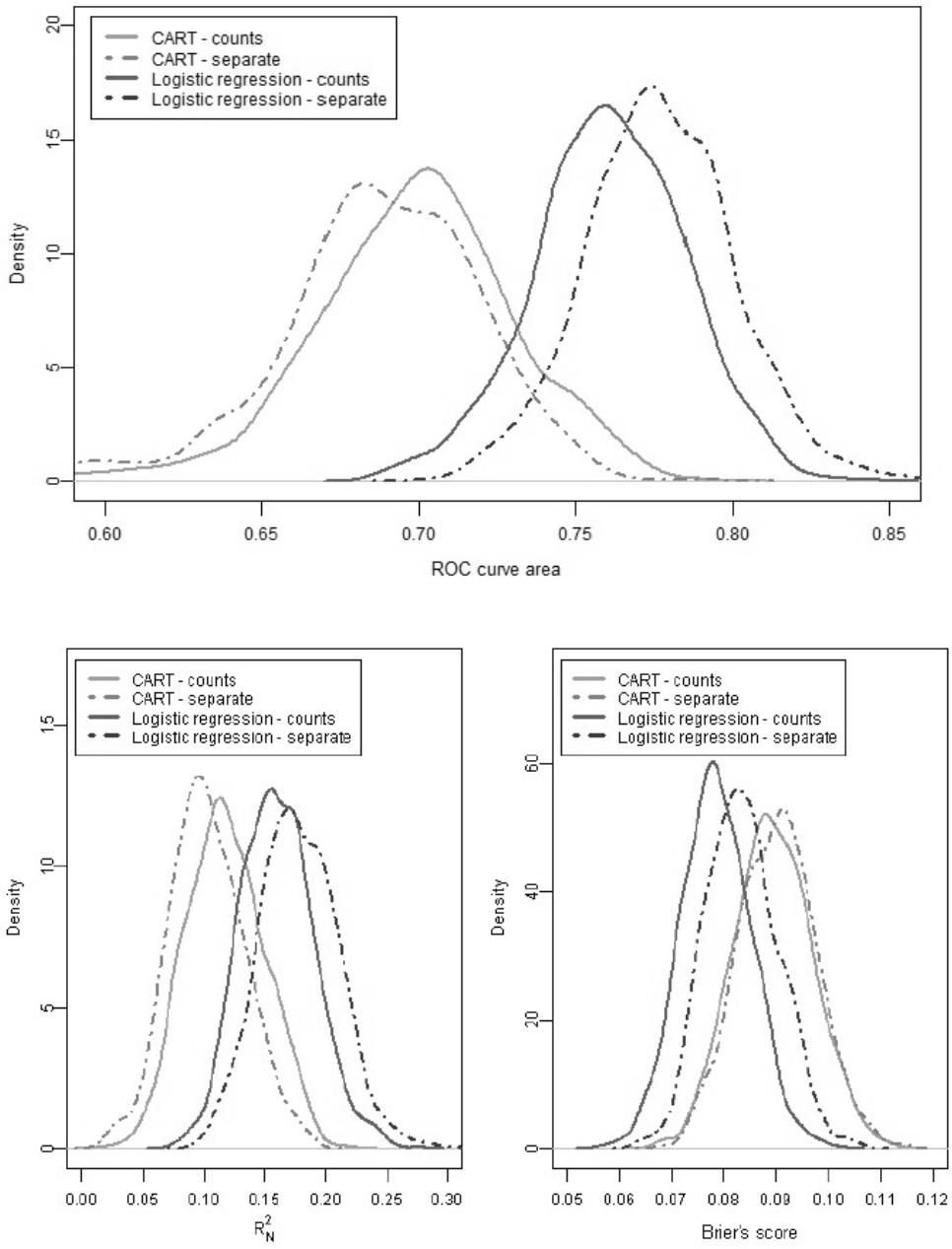
Ο Πίνακας 2.5 παρουσιάζει τα αποτελέσματα για το εμβαδό κάτω από την καμπύλη ROC, τον δείκτη R_N^2 και το Brier's score, όπως υπολογίστηκαν για τα 1000 validation δείγματα. Το μέσο εμβαδό κάτω από την ROC όσον αφορά το μοντέλο δέντρων παλινδρόμησης που περιλαμβάνει το πλήθος των χορηγούμενων φαρμάκων και το

πλήθος των παραγόντων ρίσκου είναι 0.698, ενώ για το μοντέλο που χρησιμοποιεί τις μεταβλητές πρόβλεψης ξεχωριστά είναι 0.687. Το μέσο εμβαδό κάτω από την καμπύλη ROC για το μοντέλο της λογιστικής παλινδρόμησης με χρήση του πλήθους των παραγόντων για τα φάρμακα που χορηγούνται και τους παράγοντες ρίσκου είναι 0.782, ενώ για το μοντέλο λογιστικής παλινδρόμησης που χρησιμοποιεί τις μεταβλητές πρόβλεψης είναι 0.795. Και τα δυο μοντέλα λογιστικής παλινδρόμησης ξεπερνάνε τα δέντρα παλινδρόμησης στην προβλεπτική ικανότητα. Το μοντέλο λογιστικής παλινδρόμησης με τις μεταβλητές πρόβλεψης ξεχωριστά είναι ελαφρά πιο ακριβές από το άλλο με τα πλήθη. Από τα μοντέλα δέντρων παλινδρόμησης, αυτό που χρησιμοποιεί τα πλήθη είναι πιο ακριβές από αυτά με τις μεταβλητές πρόβλεψης ξεχωριστά.

	ROC: derivation sample	ROC: validation sample	R_N^2: validation sample	Brier's score: validation sample
CART: (counts)	0.716	0.698	0.116	0.090
CART: (separate)	0.705	0.687	0.103	0.090
Logistic reg.: (counts)	0.782	0.761	0.159	0.079
Logistic reg.: (separate)	0.795	0.777	0.179	0.084

Πίνακας 2.5: *Μέσες τιμές του εμβαδού κάτω από την ROC καμπύλη, του δείκτη* R_N^2 *και του Brier's score για κάθε μοντέλο.*

Το Σχήμα 2.8 παρουσιάζει τις εκτιμήσεις για την πυκνότητα του εμβαδού κάτω από την καμπύλη ROC, τον δείκτη R_N^2 και το Brier's score στα 1000 validation δείγματα. Οι κατανομές του εμβαδού της ROC για τα δέντρα παλινδρόμησης μετατοπίζεται σε μικρότερες τιμές κάτι που αντιπροσωπεύει χαμηλότερη απόδοση σε σχέση με τα μοντέλα λογιστικής παλινδρόμησης.



Σχήμα 2.8: *Εκτιμήσεις πυκνότητας για το εμβαδό κάτω από την καμπύλη ROC, τον δείκτη* R^2_N *, και το Brier's score. Τα διαγράμματα έχουν περικοπεί σε διαστήματα με μη-μηδενικές εκτιμήσεις πυκνότητας.*

2.4.4 Συμπεράσματα

Η έρευνα κατέληξε ότι η μέθοδος των δέντρων παλινδρόμησης δεν αποδίδει όσο καλά αποδίδει η μέθοδος της λογιστικής παλινδρόμησης στην πρόβλεψη της θνησιμότητας υπό νοσηλεία. Αυτό το συμπέρασμα παρέμεινε ανεξάρτητα εάν οι μεταβλητές πρόβλεψης χρησιμοποιήθηκαν ξεχωριστά ή σαν πλήθος. Η σχετικά καλύτερη απόδοση της λογιστικής παλινδρόμησης σημαίνει ότι υπάρχει μια γραμμική

σχέση ανάμεσα στα log-odds της θνησιμότητας υπό νοσηλεία και των μεταβλητών πρόβλεψης που χρησιμοποιήθηκαν. Διαφορετικές αναλύσεις στα ίδια δεδομένα έδειξαν ότι υπάρχουν ισχυρές αλληλεπιδράσεις αλλά η χρήση τους δεν βελτίωσε το μοντέλο λογιστικής παλινδρόμησης και έτσι δεν συμπεριλήφθησαν.

Το μοντέλο των δέντρων παλινδρόμησης που χρησιμοποίησε τις μεταβλητές σαν πλήθη, έδειξε ελαφρώς καλύτερη απόδοση σε σχέση με εκείνο που τις χρησιμοποίησε ξεχωριστά. Είναι γνωστό ότι τα δέντρα παλινδρόμησης είναι προβληματικά στο να αντιλαμβάνονται τις προσθετικές σχέσεις και θεωρείται ότι το γεγονός αυτό συνέβαλε σε αυτή τη διαφορά στην απόδοση ανάμεσα στα δύο μοντέλα CART.

2.5 Χρήση των μεθόδων των Δέντρων Ταξινόμησης και της Λογιστικής Παλινδρόμησης για την διάγνωση του Εμφράγματος του Μυοκαρδίου

Οι Tsien et al. (1998) σε έρευνα τους εξετάζουν την χρήση των μεθόδων των δέντρων ταξινόμησης και της λογιστικής παλινδρόμησης για την κατασκευή ενός διαγράμματος ροής και ενός μοντέλου εξίσωσης αντίστοιχα, που θα μπορούν να βοηθήσουν στη διάγνωση εμφράγματος του μυοκαρδίου σε ασθενείς που επισκέπτονται τη μονάδα επειγόντων περιστατικών με πόνο στο στήθος. Οι πληροφορίες που χρησιμοποιούνται και στις δύο μεθόδους περιλαμβάνουν κλινικά και ηλεκτροκαρδιογραφικά δεδομένα (ECG) διαθέσιμα από τη στιγμή που οι ασθενείς θα παρουσιαστούν στα επείγοντα περιστατικά.

2.5.1 Μέθοδοι που χρησιμοποιήθηκαν

Τα δεδομένα για τον πόνο στο στήθος που χρησιμοποιήθηκαν χωρίζονται σε 1752 περιπτώσεις και προέρχονται από 1252 ασθενείς που παρουσιάστηκαν στο Edinburg Royal Infirmary της Σκωτίας και 500 ασθενείς που παρουσιάστηκαν στο Northern General Hospital στο Sheffield της Αγγλίας. Για κάθε περίπτωση, υπάρχουν 45 μεταβλητές διαθέσιμες που εμφανίζονται στον Πίνακα 2.6. Η μέθοδοι των δέντρων και της λογιστικής παλινδρόμησης χρησιμοποίησαν από τα δεδομένα του Εδιμβούργου τις 630 υποθέσεις (στο 23% αυτών παρουσιάστηκε έμφραγμα) και οι υπόλοιπες 622 χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου. Από τα δεδομένα του Sheffield χρησιμοποιήθηκαν και οι 500 περιπτώσεις (το 31% αυτών παρουσίασε έμφραγμα) για την αξιολόγηση των μοντέλων σαν υποθέσεις από άλλο νοσοκομείο, άλλης περιοχής.

age	smoker	ex-smoker
family history of MI	diabetes	high blood pressure
lipids	retrosternal pain	chest pain major symptom
left chest pain	right chest pain	back pain
left arm pain	right arm pain	pain affected by breathing
postural pain	chest wall tenderness	sharp pain
tight pain	sweating	shortness of breath
nausea	vomiting	syncope
episodic pain	worsening of pain	duration of pain
previous angina	previous MI	pain worse than prev. Angina
crackles	added heart sounds	hypoperfusion
heart rhythm	left vent. hypertrophy	left bundle branch block
ST elevation	new Q waves	right bundle branch block
ST depression	T wave changes	ST or T waves abnormal
old ischemia	old MI	sex

Πίνακας 2.6: *Οι 45 μεταβλητές που σχετίζονται με τον πόνο στο στήθος*

2.5.2 Μοντέλα που χρησιμοποιήθηκαν

2.5.2.1 Για τα δέντρα ταξινόμησης

Το δέντρο ταξινόμησης κατασκευάστηκε με χρήση του πρόγραμμα Quinlan's C4.5 για συστήματα UNIX. Οι ερευνητές πειραματιστήκανε με τον ελάχιστο αριθμό των περιπτώσεων που χρειάζονται σε τουλάχιστον δύο αποτελέσματα ενός κόμβου προκειμένου να συμπεριληφθεί αυτός ο κόμβος στο δέντρο, με το επίπεδο εμπιστοσύνης για κάθε προβλεπόμενο ρυθμό λάθους σε κάθε φύλλο και υπό-δέντρο, το οποίο χρησιμοποιείται για να βρεθεί το άνω φράγμα της πιθανότητας λάθους σε ένα φύλλο ή υπό-δέντρο κατά τη διαδικασία κλαδέματος και με τον αριθμό των δέντρων που θα παραχθούν από το δοσμένο σύνολο εκπαίδευσης , το καλύτερο από τα οποία θα επιλεγεί.

Το τελικό δέντρο FT επιλέχτηκε ύστερα από κλινική επιλογή ανάμεσα σε αυτά με τα καλύτερα αριθμητικά αποτελέσματα. Η επιλογή περιελάμβανε την εξέταση εάν τα προτεινόμενα γνωρίσματα από το δέντρο είναι συνεπή με το σκοπό. Επίσης, εάν τα γνωρίσματα κάθε κλάδου έχουν νόημα μεταξύ τους και επιπλέον, στις περιπτώσεις όπου ορισμένα γνωρίσματα επαναλαμβάνονταν πάνω από μία φορά σε κάποιο κλάδο, εξετάστηκε εάν κάτι τέτοιο είναι κλινικά εφικτό. Τα δέντρα FT και τα δέντρα Goldman και Long συγκρίθηκαν μεταξύ τους. Τα δέντρα Goldman (Goldman et al. 1982) και Long (Long et al. 1993) είναι αποτελέσματα προηγούμενων μελετών.

2.5.2.2. Για τη λογιστική παλινδρόμηση

Για να επιλέξουν ποιες μεταβλητές θα περιλαμβάνει το μοντέλο λογιστικής παλινδρόμησης από τις 45 συνολικά που σχετίζονται με τον πόνο στο στήθος, οι Tsien et al. (1998) προτίμησαν να αφήσουν το πρόγραμμα C4.5 να κάνει την επιλογή για τις βέλτιστες από αυτές. Οι μεταβλητές «ηλικία» και «διάρκεια» χρησιμοποιήθηκαν σαν συνεχείς μεταβλητές παρ' όλο που θα μπορούσαν μέσω του προγράμματος να τις μετατρέψουν σε διχοτομικές. Τα τελικά μοντέλα που

χρησιμοποιήθηκαν για την λογιστική παλινδρόμηση ήταν το μοντέλο FT LR που προέκυψε από την παραπάνω διαδικασία και το μοντέλο του Kennedy που είχε προκύψει από προηγούμενη μελέτη (Kennedy et al. 1996).

2.5.3 Μέτρα απόδοσης για τη σύγκριση των μοντέλων

Οι υπολογισμοί που έγιναν για την σύγκριση των μοντέλων περιλαμβάνουν τα εξής μέτρα απόδοσης: την ευαισθησία (sensitivity), την ειδικότητα (specificity), την θετική προβλεπτική τιμή (Positive Predictive Value) όπου είναι ο αριθμός των σωστά διαγνωσμένων ασθενών με MI δια τον αριθμό όλων των ασθενών πιο διαγνώστηκαν με MI είτε διαγνώστηκαν σωστά από το μοντέλο είτε λάθος , την ευστοχία (accuracy) και το εμβαδό κάτω από την καμπύλη ROC. Το εμβαδό και το τυπικό λάθος κάτω από την ROC υπολογίστηκαν με τη μέθοδο Hanley-McNeil.

2.5.4 Αποτελέσματα

2.5.4.1 Για τα δέντρα ταξινόμησης

Το Σχήμα 2.9 αναπαριστά το τελικό FT δέντρο. Σε κάθε φύλλο εμφανίζεται ο αριθμός των σωστά κατηγοριοποιημένων υποθέσεων προς τον συνολικό αριθμό υποθέσεων στο φύλλο (τα κλάσματα οφείλονται στο κλάδεμα). Οι τιμές για τις μεταβλητές που χρησιμοποιήθηκαν στην κατασκευή των δέντρων ήταν: Ελάχιστος αριθμός περιπτώσεων σε τουλάχιστον δύο κλάδους ενός κόμβου $m = 5$, επίπεδο εμπιστοσύνης $c = 15\%$, δέντρα που κατασκευάζονται με χρήση διαχωρισμού του δοσμένου συνόλου εκπαίδευσης $t = 10$. Ο Πίνακας 2.7 παρουσιάζει τα αποτελέσματα για τα δέντρα FT, Goldman και Long, το καθένα για το δικό του σύνολο ελέγχου. Ο Πίνακας 2.8 συγκρίνει τα χαρακτηριστικά για το κάθε δέντρο. Περίπου τα μισά χαρακτηριστικά του δέντρου Goldman και το 1/3 των χαρακτηριστικών του Long δέντρου είναι κοινά με το FT.

```
ST elevation = 1: 1 (40.7/49.0 = 83.1%)
ST elevation = 0:
|  New Q waves = 1: 1 (4.1/7.0 = 58.6%)
|  New Q waves = 0:
|  |  ST depression = 0: 0 (329.4/345.0 = 95.5%)
|  |  ST depression = 1:
|  |  |  Old ischemia = 1: 0 (3.2/6.0 = 53.3%)
|  |  |  Old ischemia = 0:
|  |  |  |  Family history of MI = 1: 1 (6.8/11.0 = 61.8%)
|  |  |  |  Family history of MI = 0:
|  |  |  |  |  age <= 61 : 1 (4.0/8.0 = 50.0%)
|  |  |  |  |  age > 61 :
|  |  |  |  |  |  Duration of pain (hours) <= 2 : 0 (14.1/22.0 = 64.1%)
|  |  |  |  |  |  Duration of pain (hours) > 2 :
|  |  |  |  |  |  |  T wave changes = 1: 1 (7.0/10.0 = 70.0%)
|  |  |  |  |  |  |  T wave changes = 0:
|  |  |  |  |  |  |  |  Right arm pain = 1: 0 (3.4/5.0 = 68.0%)
|  |  |  |  |  |  |  |  Right arm pain = 0:
|  |  |  |  |  |  |  |  |  Crackles = 0: 0 (3.0/8.0 = 37.5%)
|  |  |  |  |  |  |  |  |  Crackles = 1: 1 (4.9/9.0 = 54.4%)
```

Σχήμα 2.9: *Το τελικό FT δέντρο που χρησιμοποιήθηκε*

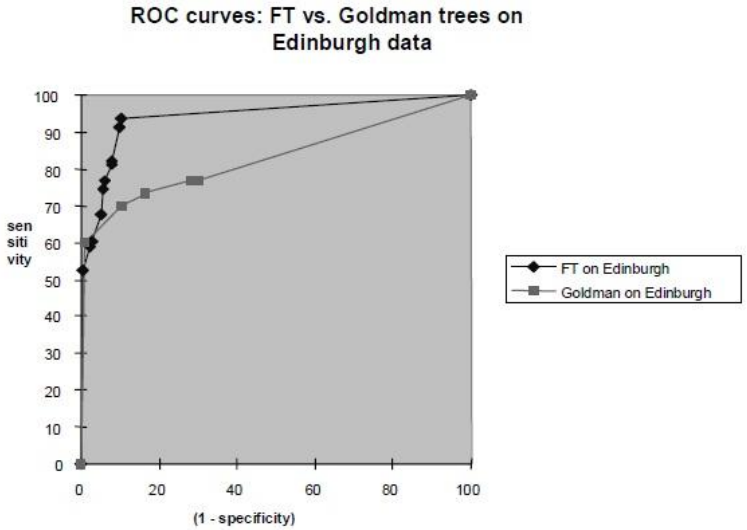
	Goldman	**FT Tree**	**Long**
Sensitivity =	90.9%	81.4%	66.1%
Specificity =	69.7%	92.1%	85.8%
PPV =	35.4%	72.9%	68.3%
Accuracy =	73.1%	89.9%	80.1%

Πίνακας 2.7: *Αποτελέσματα για τα τρία δέντρα που χρησιμοποιήθηκαν, όπως προέκυψαν για το σύνολο ελέγχου του καθενός.*

Goldman	FT	Long
ST elevation or Q waves	ST elevation New Q waves	ST change Q waves
Duration	Duration	
ST or T wave	T wave	T wave
Shoulder, neck, arm	Right arm	Arm,neck,shoulder
Age	**Age**	**Age**
Local Pressure	ST depression	Stomach pain
Previous angina	Old ischemia	Sex
Left shoulder	Family history	Systolic BP
Pain worse	Crackles	Heart rate
Diaphoresis		Rapid/skipping beats
		Chest pain
		History of MI
		Nitroglycerin use
		Shortness of breath
		Fainted, dizzy, lightheaded

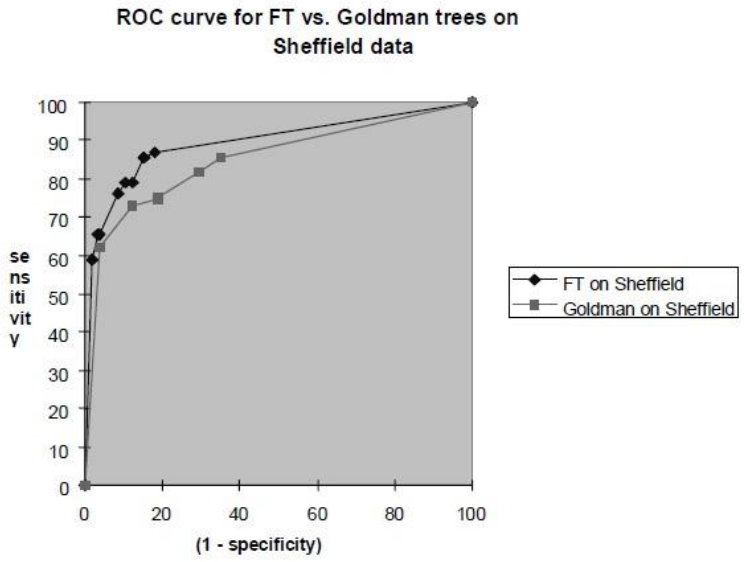
Πίνακας 2.8*: Σύγκριση των χαρακτηριστικών κάθε δέντρου. Πάνω από την γραμμή είναι τα κοινά χαρακτηριστικά.*

Το εμβαδό κάτω από την καμπύλη ROC για το FT δέντρο στο δικό του σύνολο ελέγχου ήταν 94.04% (με τυπικό σφάλμα 0.72%), ενώ του δέντρου Long ήταν 86% από την προηγούμενη μελέτη. Η μελέτη για το δέντρο Goldman δεν ανέφερε αποτέλεσμα για τη ROC, οπότε υπολογίστηκε για τα δεδομένα από το Εδιμβούργο όπου το εμβαδό ήταν 84.03% (με τυπικό σφάλμα 2.28%). Το Σχήμα 2.10 δείχνει τις καμπύλες ROC για τα δέντρα FT και Goldman στα δεδομένα του Εδιμβούργου, όπου τα εμβαδά έχουν στατιστικά σημαντική διαφορά ($p < 0.0001$).



Σχήμα 2.10*: Δέντρο Goldman VS Δέντρο FT στα δεδομένα του Εδιμβούργου*

Από τη στιγμή που το δέντρο FT αναπτύχθηκε με βάση τα δεδομένα από το Εδιμβούργο, μια πιο αυστηρή σύγκριση είναι να χρησιμοποιήσουμε τα δεδομένα για να εξετάσουμε ένα διαφορετικό σύνολο δεδομένων, όπως αυτά από το Sheffield. Το Σχήμα 2.11 δείχνει τα αποτελέσματα για τις ROC καμπύλες των δέντρων FT και Goldman. Οι διαφορά ανάμεσα στα εμβαδά των δύο δέντρων είναι και πάλι στατιστικά σημαντική ($p = 0.006$).



Σχήμα 2.11: *Δέντρο Goldman VS Δέντρο FT στα δεδομένα του Sheffield.*

2.5.4.2 Για τη λογιστική παλινδρόμηση

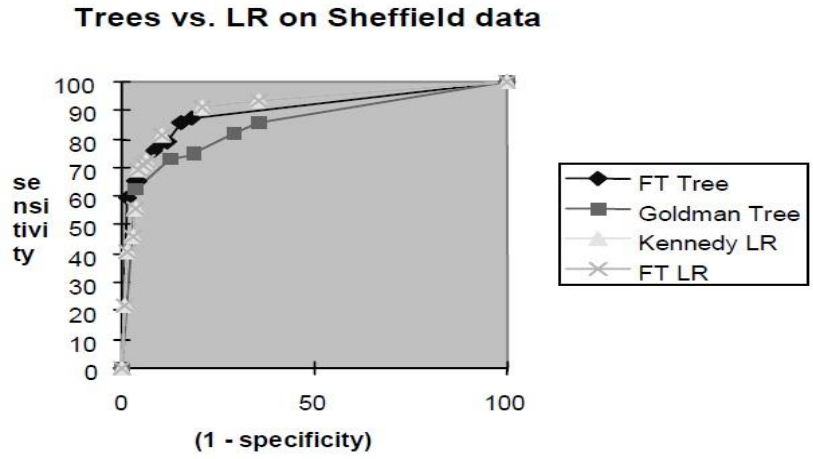
Ο Πίνακας 2.8 συγκρίνει τα χαρακτηριστικά των μοντέλων παλινδρόμησης FT, Kennedy και Long. Τα μοντέλα Kennedy και FT χρησιμοποιούν αρκετά κοινά χαρακτηριστικά για τον πόνο στο στήθος, ενώ τα μοντέλα Long και FT όχι. Η μελέτη για το μοντέλο Long εμφάνιζε εμβαδό κάτω από την ROC της τάξης του 89%. Τα μοντέλα Kennedy και FT εξετάστηκαν στα δεδομένα από το Εδιμβούργο και το Sheffield. Στα δεδομένα του Εδιμβούργου, το μοντέλο Kennedy απέδωσε ένα εμβαδό ROC 94.30% (με τυπικό σφάλμα 0.92%), ενώ το FT μοντέλο απέδωσε εμβαδό ROC 94.28% (με τυπικό σφάλμα 1.16%). Στα δεδομένα του Sheffield, το μοντέλο Kennedy απέδωσε εμβαδό ROC 91.25% (με τυπικό σφάλμα 1.32%) ενώ το FT μοντέλο απέδωσε εμβαδό ROC 89.28% (με τυπικό σφάλμα 1.59%). Δεν υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στα μοντέλα ($p = 0.50$ και $p = 0.17$ αντίστοιχα).

	Kennedy	FT LR	Long LR
Constant	-3.07	-2.14	
ST elevation	3.16	2.96	
New Q waves	1.37	2.00	
ST depression	1.95	1.76	
LV Failure (Crackles)	1.54	0.807	
Old ischemia		-0.86	
Family hx		0.43	
Age		-0.016	
Duration		-.0046	
T wave		0.805	
Right arm		-0.22	
Vomiting		0.68	
Hypoperfusion		0.47	
Chest pain #1 Sx			0.71
Chest pain/24h			1.00
T wave nl/flat			1.13
Nitro use			0.51
Previous MI		0.42	
STchange nl/flat			0.77
STchange normal			0.83

***Πίνακας 2.8**: Σύγκριση των χαρακτηριστικών των τριών μοντέλων λογιστικής παλινδρόμησης*

2.5.4.3 Δέντρα ταξινόμησης έναντι λογιστικής παλινδρόμησης

Η απόδοση του δέντρου FT ήταν παρόμοια με αυτή του καλύτερου μοντέλου λογιστική παλινδρόμησης, τόσο στα δεδομένα του Εδιμβούργου ($p = 0.41$ για τη διαφορά τους), όσο και στα δεδομένα του Sheffield ($p = 0.17$ για τη διαφορά τους). Το δέντρο Goldman είχε αρκετά χαμηλή απόδοση από όλα τα υπόλοιπα μοντέλα και στα δύο σετ δεδομένων. Το Σχήμα 2.12 δείχνει τις καμπύλες ROC για όλα τα μοντέλα στα δεδομένα του Sheffield.



Σχήμα 2.12: Καμπύλες ROC όλων των μοντέλων για τα δεδομένα από το Sheffield

2.5.5 Συμπεράσματα

Οι συγγραφείς συμπεραίνουν από τα αποτελέσματα ότι το FT δέντρο αποδίδει το ίδιο καλά με τα μοντέλα λογιστικής παλινδρόμησης. Το δέντρο FT είχε καλύτερα αποτελέσματα από τα δέντρα Long και Goldman και επιπλέον, είναι σχετικά μικρό και πιο λογικό κλινικά. Το δέντρο Goldman είναι επίσης μικρό, αλλά δεν είχε την ίδια απόδοση. Επιπλέον, ορισμένα μονοπάτια του δέντρου Goldman, ειδικά αυτά που ρωτάνε για την ηλικία του ασθενή ή τη διάρκεια του πόνου πολλαπλές φορές πριν καταλήξουν σε ένα φύλλο, δεν είναι κλινικά και τόσο ενδεδειγμένα. Το δέντρο Long είναι αρκετά μεγάλο, δεν απέδωσε τόσο καλά και επίσης περιέχει μονοπάτια που δεν είναι κλινικά ενδεδειγμένα (π.χ. συγκρίνει το σφυγμό με όριο 77 παλμούς το λεπτό και σε ακόλουθους κόμβους συγκρίνει με όριο 89 παλμούς το λεπτό).

Η μελέτη θα μπορούσε να περιλαμβάνει χρήση του μοντέλου λογιστικής παλινδρόμησης του Long και του μοντέλου του Selker (το οποίο μοντέλο μπορεί να προβλέψει και στηθάγχη εκτός από MI). Μια προσέγγιση που θα μπορούσε να βοηθήσει στο να γίνει αυτό, θα ήταν να αντιμετωπιστεί το πρόβλημα με τις ελλείπουσες τιμές.

Η δυσκολία στο να κατασκευαστεί ένα καλό μοντέλο λογιστικής παλινδρόμησης έγκειται στο να αποφασιστεί ποιά χαρακτηριστικά θα συμπεριληφθούν στο μοντέλο και ποιά όχι. Δεδομένου ότι το FT μοντέλο απέδωσε παρόμοια καλά με το Kennedy μοντέλο, μια χρήση των δέντρων κατηγοριοποίησης θα μπορούσε να είναι το να επιλέξουν τα χαρακτηριστικά για το μοντέλο λογιστικής παλινδρόμησης. Μια άλλη χρήση θα μπορούσε να είναι το να καθορίσουν τα οριακά σημεία για τις συνεχείς μεταβλητές, σε περίπτωση που απαιτούνται διχοτομικές μεταβλητές. Ένας πρωταρχικός πειραματισμός για κατασκευή ενός μοντέλου λογιστικής παλινδρόμησης, όπου οι μεταβλητές «ηλικία» και «διάρκεια πόνου» θα περιληφθούν σαν διχοτομικές (με οριακά σημεία τα 61 χρόνια και τις 2 ώρες αντίστοιχα), δεν είχε κάποια σημαντική αλλαγή στις καμπύλες ROC που προέκυψαν.

Οι υπολογισμοί για όλα τα εμβαδά κάτω από την καμπύλη ROC έγιναν με χρήση της μεθόδου Hanley-McNeil. Αυτή η μέθοδος μπορεί να μην είναι τόσο ακριβής όσο το πρόγραμμα υπολογισμού πιθανοφάνειας των Dorfman και Alf ή όσο η κλίση (slope) και η τομή (intercept) των αρχικών δεδομένων όταν σχεδιάζονται σε ένα binormal διάγραμμα. Στον υπολογισμό του εμβαδού κάτω από τη ROC των μοντέλων λογιστικής παλινδρόμησης, αυτό ίσως να μην είναι πρόβλημα, καθώς πολλαπλές threshold τιμές μπορούν να χρησιμοποιηθούν για να υπολογιστούν πολλά ζευγάρια ευαισθησίας-ειδικότητας. Ωστόσο, μπορεί να αποτελεί πρόβλημα στα δέντρα ταξινόμησης όπου ο αριθμός των ζευγαριών ευαισθησίας-ειδικότητας περιορίζεται από τον αριθμό των φύλλων στο δέντρο. Με λιγότερα σημεία για τη σχεδίαση της καμπύλης ROC μπορεί πιο εύκολα να υποτιμηθεί το εμβαδό κάτω από την καμπύλη σε σχέση με την πραγματική του τιμή και έτσι να ελαττωθεί η απόδοση των δέντρων ταξινόμησης. Ακόμα και έτσι όμως, το δέντρο FT αποδίδει συγκρίσιμα με τις μεθόδους λογιστικής παλινδρόμησης.

Η διαδοχική μέθοδος Bayes και τα νευρωνικά δίκτυα έχουν επίσης εξεταστεί για τη διάγνωση MI. Η μέθοδος του Bayes απέδωσε υποσχόμενα αποτελέσματα, αλλά απαιτεί μια heuristic μέθοδο για να χειριστεί την αλληλεξάρτηση ανάμεσα στα δεδομένα. Τα νευρωνικά δίκτυα επίσης απέδωσαν υποσχόμενα αποτελέσματα αλλά σαν μέθοδος παραμένει αρκετά δυσνόητη για τους κλινικούς ερευνητές ώστε να γενικευτεί η χρήση τους. Τα μοντέλα λογιστικής παλινδρόμησης είναι επίσης λιγότερο κατανοητά σε σχέση με τα διαγράμματα ροής όπως είναι τα δέντρα παλινδρόμησης. Όλες οι μέθοδοι περιέχουν τη δυσκολία του να παράγουν μοντέλα που να μπορούν να γενικευτούν και σε άλλα νοσοκομεία. Για το λόγο αυτό, συστήνεται η εξέταση των υπαρχόντων μοντέλων σε περισσότερα σετ δεδομένων.

Βιβλιογραφία

1. P. C. Austin, (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariable adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*, **26**, 2937-2957.
2. P. C. Austin, J. V. Tu, D. S. Lee, (2010). Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *Journal of Clinical Epidemiology*, **63**, 1145-1155.
3. V. Faltus, Z. Monhart, (2008). L*ogistic regression and classification and regression trees (CART) in acute myocardial infarction data modeling*, *ISCB 2008, 29th Annual Conference of the International Society for Clinical Biostatistics*, Copenhagen, Denmark.
4. T. Fawcett, (2003) "*ROC Graphs*: *Notes* and *Practical Considerations* for *Data Mining Researchers*", Technical Report HPL-2003-4, Intelligent EnterpriseTechnologies Laboratory.
5. J. Fox, (2002). *Bootstrapping Regression Models: Appendix to An R and S-PLUS Companion to Applied Regression*.
6. L. Goldman, M. Weinberg , M. Weisberg, R. Olshen , E. Cook, R. Sargent, G. Lamas, C. Dennis, C. Wilson, L. Deckelbaum, H. Fineberg, R. Stiratelli, (1982). A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *N Engl J Med*, (307), 588-596.
7. R.L. Kennedy, A.M. Burton, H.S. Fraser, L.N. McStay, R.F. Harrison, (1996). Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *EHJ*, (17), 1181-1191.
8. W.J. Long, J.L. Griffith , H.P. Selker, R.B. DAgostino. A comparison of logistic regression to decision-tree induction in a medical domain. *Comput Biomed Res*, 26, 74-97.
9. R.H. Myers, D.C. Montgomery, G.G. Vining, (2002). *Generalized Linear Models: With Applications Engineering and the Sciences.* John Wiley and Sons, New York.
10. M.S. Pepe, (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford Statistical Science Series, Oxford University Press.
11. M.S. Pepe, T. Cai, G. Longton, (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, **62**, 221-229.
12. P. Refaeilzadeh, L. Tang, H. Liu, (2008). *Notes for Cross-validation*, Arizona State University.
13. C. L. Tsien, H. S. F. Fraser, W. J. Long, R.L. Kennedy, (1998). Using Classification Tree and Logistic Regression Methods to Diagnose Myocardial Infarction. *MEDINFO 98*, B. Cesnik et al. (Eds), IOS Press, Amsterdam.

Πρόσθετες Βιβλιογραφικές Πηγές

1. Ιωάννα Μπλίντζιου, (2010). "*Ανάλυση ROC καμπύλων και εφαρμογή σε πραγματικά ιατρικά δεδομένα*", Εθνικό Μετσόβιο Πολυτεχνείο.
2. Χρήστος Θ. Νάκας (2002). "*Προσαρμογή καμπύλης, στατιστική συμπερασματολογία, επεκτάσεις και εφαρμογές στην ανάλυση των καμπύλων λειτουργικού χαρακτηριστικού δέκτη (ROC)*", Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
3. "*Clementine 12.0 Algorithms Guide*".
4. Wikipedia