ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Τεχνικές Εξόρυξης Δεδομένων, Λογιστική Παλινδρόμηση και χρήση ROC Καμπυλών για την Ανάλυση Πραγματικών Ιατρικών Δεδομένων

ΗΡΑΚΛΗ ΔΗΜΟΥΛΑ

Επιβλέπων: Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2012

ΠΕΡΙΛΗΨΗ

Η τεχνολογία του data mining είναι μια σχετικά καινούρια περιοχή η οποία περιλαμβάνει τεχνικές επεξεργασίας και ανάλυσης μεγάλων βάσεων δεδομένων. Ο στόχος αυτών των τεχνικών, είναι η ανακάλυψη νέων προτύπων μεταξύ των δεδομένων και η εξαγωγή χρήσιμων πληροφοριών. Στην παρούσα διπλωματική εργασία μελετήθηκαν οι εξής : τα Δέντρα Αποφάσεων, τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ), η Λογιστική Παλινδρόμηση και οι Καμπύλες ROC.

Το πρώτο κεφάλαιο περιέχει μια εισαγωγή στις τεχνικές του data mining η οποία περιλαμβάνει τους σκοπούς και τη διαδικασία του, καθώς επίσης και τους τομείς στους οποίους εφαρμόζεται.

Στο δεύτερο κεφάλαιο αναλύονται οι τεχνικές εξόρυξης γνώσης. Στα Δέντρα Αποφάσεων μελετήθηκε ο Αλγόριθμος C&RT και στα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) αναλύθηκαν τα κύρια χαρακτηριστικά τους καθώς και δύο κύρια μοντέλα των τεχνητών νευρωνικών δικτύων, τα δίκτυα MLP και RBFN. Ακολούθως, παρουσιάζεται η Λογιστική Παλινδρόμηση και τρόποι με τους οποίους εκτιμούνται οι παράμετροι. Τέλος, παρουσιάζεται η βασική ορολογία της ROC Ανάλυσης και περιγράφεται η έννοια και η χρήση του Εμβαδού Κάτω από την Καμπύλη ROC.

Το τρίτο κεφάλαιο περιλαμβάνει την εφαρμογή που πραγματοποιήθηκε πάνω σε πραγματικά ιατρικά δεδομένα (8862 ασθενείς με τραύματα) με τη βοήθεια του προγράμματος Clementine SPSS, ενός λογισμικού εξόρυξης δεδομένων. Τα δεδομένα αυτά εφαρμόστηκαν στον Αλγόριθμο C&RT, στις νευρωνικές μεθόδους MLP και RBFN και στο μοντέλο της Λογιστικής Παλινδρόμησης με σκοπό να προβλέψουμε και να συγκρίνουμε τα αποτελέσματα από την άποψη της ακρίβειας (classification accuracy), της ευαισθησίας (sensitivity), της ειδικότητας (specificity), της θετικής προγνωστικής αξίας (positive predictive value), της αρνητικής προγνωστικής αξίας (negative predictive value) και των γραφικών παραστάσεων που δείχνουν την περιοχή κάτω από την ROC καμπύλη (AUC).

ABSTRACT

Data mining is a relatively new field which includes techniques for processing and analysing large databases. The goal of these techniques is to discover new patterns among the data and extract useful knowledge. In my diploma dissertation, I have investigated the following: the technology of Decision Trees, Artificial Neural Networks (ANNs), Logistic Regression and ROC curves.

The first chapter constitutes an introduction to the data mining techniques and examines their purposes, the general data mining procedure as well as their applications.

The second chapter analyses the data mining techniques. Firstly, as far as Decision trees are concerned, I focused on C&RT Algorithm. On the other hand, in the case of Artificial Neural Networks (ANNs), I analysed the main characteristics as well as the two most common classes of ANNs, namely; the Multilayer Perceptrons (MLPs) and the Radial Basis Function Networks (RBFNs). This chapter also deals with Logistic Regression and ways in which the parameters are estimated. Finally, the basic terminals for ROC Analysis are presented and the meaning and utility of the Area Under the ROC Curve are also described.

The objective of the third chapter is to examine an application on actual medical data (8862 trauma patients) using Clementine SPSS, a data mining software. This data was applied to the Classification and Regression Trees (C&RT), Multilayer Perceptron Neural Networks (MLPs), Radial Basis Function Neural Networks (RBFNs) and Logistic Regression (LR) to predict and compare the results in terms of overall classification accuracy, sensitivity , specificity, positive predictive value, the negative predictive value and area under the ROC curve (AUC).

ΕΥΧΑΡΙΣΤΙΕΣ

Η εκπόνηση της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε υπό την επίβλεψη του Καθηγητή του Ε.Μ.Π., κ. Χρήστου Κουκουβίνου, τον οποίο θα ήθελα να ευχαριστήσω θερμά για τη δυνατότητα που μου έδωσε να ασχοληθώ με ένα θέμα το οποίο ανήκει στα ερευνητικά μου ενδιαφέροντα.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτορα Χριστίνα Παρπούλα, για την πολύτιμη βοήθεια της και το συνεχές ενδιαφέρον κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Θα ήθελα επίσης να ευχαριστήσω την οικογένεια μου για την υπομονή και την υποστήριξή τους. Τέλος, αισθάνομαι την ανάγκη να ευχαριστήσω τους φίλους και συμφοιτητές μου, για την αμέριστη βοήθεια και τη συμπαράσταση τους.

Ηράκλη Δημούλα

Αθήνα, 2012

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1: DATA MINING

1.1 Εισαγωγή

Η συνεχής ανάπτυξη στον τομέα της πληροφορικής σε συνδυασμό με τα σύγχρονα εργαλεία αυτοματοποιημένης συλλογής δεδομένων έχουν οδηγήσει στη δημιουργία μιας κοινωνίας ικανής να παρέχει διαρκώς νέες πληροφορίες. Το υλικό που συγκεντρώνεται καταγράφεται διαρκώς, με αποτέλεσμα τη δημιουργία τεράστιων βάσεων δεδομένων. Το ζήτημα λοιπόν που προκύπτει είναι κατά πόσο μπορούμε να διαχειριστούμε αυτές τις βάσεις δεδομένων.

Η ανάγκη εξόρυξης πληροφοριών από αυτό τον τεράστιο όγκο δεδομένων , όπου οι μέχρι πρότινος κλασσικές μέθοδοι της στατιστικής, δεν αποδεικνύονταν επαρκείς και ικανοποιητικές, οδήγησε στη διαδικασία της Εξόρυξης Δεδομένων (Data Mining). Πρόκειται για μια σειρά από τεχνικές που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς κλάδους όπως η οικονομία, η βιοστατιστική , η δημογραφία και η μετεωρολογία.

Η λέξη «Data» είναι μια λατινική λέξη που σημαίνει «τα πράγματα που έχουν δοθεί». Στην πληροφορική αναφέρεται ως μια συλλογή από αριθμούς ή σύμβολα σε τέτοια μορφή που είναι εύκολα επεξεργάσιμη από ηλεκτρονικούς υπολογιστές. Οι συλλογές αυτές από δεδομένα δεν έχουν καμία αξία, αν δεν μπορούν να μετασχηματιστούν σε γνώση.

Γενικά υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων (ΕΔ). Ωστόσο, αποδεχόμαστε σαν ορισμό του data mining τον εξής :

«Εξόρυξη Δεδομένων είναι η ανάλυση , συνήθως τεράστιων , παρατηρούμενων συνόλων δεδομένων , έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο δεδομένων».

Ο κύριος στόχος του data mining είναι η εξαγωγή νέων πληροφοριών από τα δεδομένα. Η ανακάλυψη της γνώσης γίνεται με τεχνικές οι οποίες διακρίνονται σε δυο βασικές κατηγορίες:

- **μέθοδοι με επίβλεψη (supervised methods)** : Αλγόριθμοι εκμάθησης με επίβλεψη είναι εκείνοι που χρησιμοποιούνται στην ταξινόμηση και στην πρόβλεψη. Ουσιαστικά μοντελοποιούν μια μεταβλητή απόκρισης βασιζόμενοι σε μια ή περισσότερες επεξηγηματικές μεταβλητές (input variable). Μερικές από αυτές τις supervised τεχνικές είναι και τα νευρωνικά δίκτυα (neural networks), δέντρα αποφάσεων (decision trees), λογιστική παλινδρόμηση (logistic regression) με τις οποίες θα ασχοληθούμε εκτεταμένα στη συνέχεια της παρούσας εργασίας.

- **μέθοδοι χωρίς επίβλεψη (unsupervised methods)** : Αλγόριθμοι εκμάθησης χωρίς επίβλεψη είναι εκείνοι που χρησιμοποιούνται όταν δεν υπάρχει μια μεταβλητή απόκρισης να προβλεφθεί ή να ταξινομηθεί. Ουσιαστικά, οι unsupervised τεχνικές χρησιμοποιούνται όταν δεν υπάρχει κάποιο πεδίο να προβλεφθεί αλλά οι σχέσεις των δεδομένων εξερευνούνται ώστε να ανακαλυφθεί η γενική δομή τους. Μερικές από τις τεχνικές αυτές είναι οι kohonen networks, two step , k-means.

Τα έξι βασικά αποτελέσματα που αναμένεται να λάβουμε ανάλογα με τους στόχους που έχουμε θέσει (tasks) είναι :

- Ταξινόμηση (classification): εξέταση των χαρακτηριστικών ενός νέου αντικειμένου και η ταξινόμησή του σε ήδη προκαθορισμένες κλάσεις.
- Εκτίμηση (estimation): εύρεση τιμών για μια άγνωστη μεταβλητή, με δεδομένα κάποια δεδομένα εισόδου.
- Πρόβλεψη (prediction): παρόμοια με την ταξινόμηση και την εκτίμηση, αλλά οι εγγραφές ταξινομούνται με βάση κάποιες προβλεπόμενες μελλοντικές τάσεις ή εκτιμώμενες μελλοντικές τιμές.

- Ομαδοποίηση (grouping): καθορισμός των αντικειμένων που ανήκουν σε συγκεκριμένη ομάδα.
- Συσταδοποίηση (clustering): κατάτμηση ενός πληθυσμού σε ένα αριθμό υποομάδων ή συστάδων.
- Περιγραφή και οπτικοποίηση (description and visualization): διερευνητικό ή οπτικό data mining.

1.2 Διαδικασία KDD

Επεξεργαζόμενοι μια τεράστια βάση δεδομένων είναι πιθανό να ανακαλύψουμε την ύπαρξη «κρυμμένης γνώσης». Δηλαδή, μπορεί να εντοπίσουμε συσχετίσεις, αλληλεξαρτήσεις ή ομαδοποιήσεις μεταξύ των δεδομένων, πράγματα τα οποία να μην είναι άμεσα εμφανή. Το είδος αυτής της γνώσης θεωρείται ότι δεν είναι εκ των προτέρων διαθέσιμο αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Την ανάγκη αυτή ανάκτησης γνώσης έρχεται να καλύψει η ΕΔ, η οποία αποτελεί τον πυρήνα της γενικότερης μεθοδολογίας της ανακάλυψης της γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases - KDD).

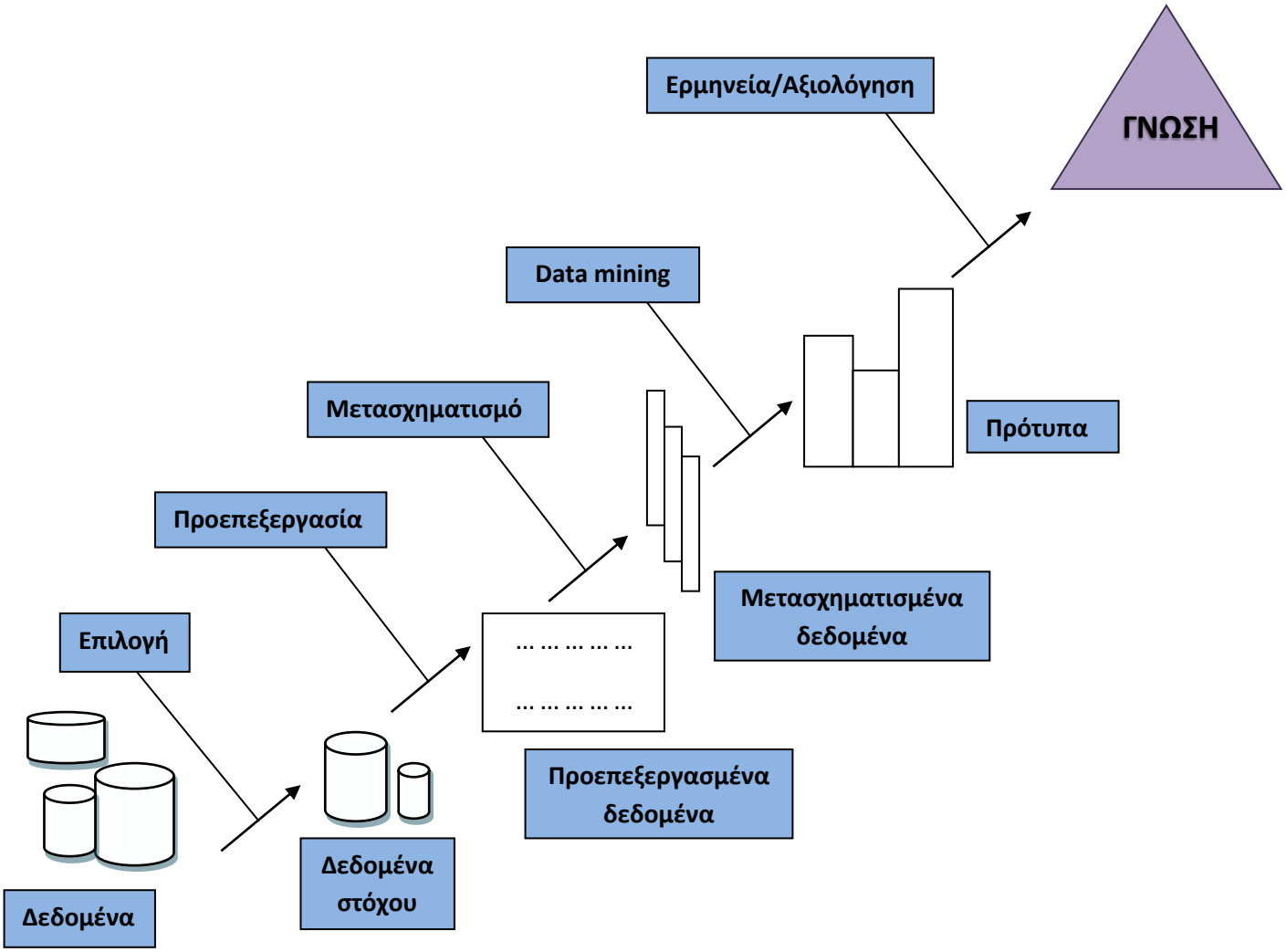
Η KDD είναι μια αυτοματοποιημένη διαδικασία ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Πρόκειται για μια συγκροτημένη μεθοδολογία αναγνώρισης έγκυρων και πρωτότυπων προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων , με στόχο τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά.

Τα βασικά βήματα της KDD διαδικασίας είναι τα ακόλουθα (Σχήμα 1):

- **Ανάπτυξη και κατανόηση του πεδίου της εφαρμογής** συμπεριλαμβανόμενης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα καθώς επίσης και των στόχων / προσδοκιών των τελικών χρηστών.
- **Δημιουργία του στοχευόμενου συνόλου δεδομένων (target data)**, το οποίο θα περιλαμβάνει τα δεδομένα από τα οποία πρόκειται να εξαχθεί η γνώση.

Το βήμα αυτό είναι εξαιρετικά κρίσιμο καθώς η ποιότητα των δεδομένων επηρεάζει την απόδοση του συστήματος αποκάλυψης γνώσης.

- **Καθαρισμός και επεξεργασία δεδομένων (data cleaning).** Το βήμα αυτό περιλαμβάνει βασικές λειτουργίες όπως η απομάκρυνση του θορύβου, η αντιμετώπιση του προβλήματος των δεδομένων με ελλιπείς τιμές κ.α
- **Μείωση της ποσότητας των δεδομένων (data reduction).** Το βήμα αυτό περιλαμβάνει την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων του προβλήματος ανάλογα με τους στόχους της ανακάλυψης γνώσης, τη μείωση του πλήθους αυτών των χαρακτηριστικών κ.α.
- **Επιλογή των εργασιών εξόρυξης γνώσης (data mining)** που θα χρησιμοποιηθούν για τις ανάγκες του προβλήματος π.χ ταξινόμηση, πρόβλεψη , ομαδοποίηση κ.α
- **Επιλογή των αλγορίθμων εξόρυξης γνώσης (data mining)** που θα χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Το βήμα αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου, την επιλογή των κατάλληλων παραμέτρων του μοντέλου κ.α
- **Data Mining**: αναζήτηση στα δεδομένα των προτύπων που μας ενδιαφέρουν.
- **Ερμηνεία των προτύπων** που ανακαλύφθηκαν από την KDD διαδικασία – πιθανόν να χρειαστεί να επιστρέψουμε και πάλι σε κάποια από τα παραπάνω βήματα.
- **Ενοποίηση της γνώσης που έχει εξαχθεί**: Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα και χρησιμοποιούνται κάποιες τεχνικές αντιπροσώπευσης αυτής προκειμένου να παρουσιαστεί ευκρινώς στο χρήστη.



Σχήμα 1: Διαδικασία του Data mining

1.3 Εφαρμογές του Data mining

Το data mining χρησιμοποιείται σε μια πληθώρα πεδίων και εφαρμογών καθώς βοηθάει στη λήψη ολοκληρωμένων αποφάσεων. Για να επιτευχθεί αυτό όμως, πρέπει αρχικά τα δεδομένα να συγκεντρωθούν και να οργανωθούν με ένα συνεπή και χρήσιμο τρόπο (data warehousing). Οι data warehouses πρέπει να έχουν ακριβή ιστορικά δεδομένα, μια και η διαδικασία των data mining, γεννά μοντέλα από ιστορικά δεδομένα που χρησιμοποιούνται για προβλέψεις, ανίχνευση τάσεων κ.α.

Το data mining χρησιμοποιείται σήμερα σε πολλούς επιστημονικούς κλάδους όπως: η **ιατρική**, η **οικονομία**, οι **τηλεπικοινωνίες**, το **marketing**. Συνοπτικά παραθέτουμε μερικές εφαρμογές του data mining σε συνδυασμό με παραδείγματα :

1) <u>Ανάλυση εταιριών και διαχείριση ρίσκου:</u>
    i. Προβλέψεις
    ii. Διατήρηση πελατολογίου
    iii. Βελτιωμένη χρηματοδότηση

<u>Π.χ (1)</u> Κατασκευή δένδρων αποφάσεων από ιστορικά στοιχεία τραπεζικών δανείων για την παραγωγή αλγόριθμων, ώστε να αποφασίζεται αν πρέπει ή όχι να δοθεί ένα δάνειο σε έναν υποψήφιο πελάτη.

<u>Π.χ (2)</u> Ιατρικά εργαστήρια θέλουν να συσχετίσουν ασθένειες με χαρακτηριστικά των ασθενών, όπως τόπος διαμονής, διατροφικές συνήθειες, παλαιότερες ασθένειες, κ.α., ώστε να καταφέρουν να βγάλουν κάποια ιατρικά συμπεράσματα και καινούρια γνώση, με τη βοήθεια των συγκεκριμένων χαρακτηριστικών.

2) <u>Ανάλυση αγοράς και διαχείριση:</u>
    i. Target marketing
    ii. Customer relation Management
    iii. Market basket analysis (supermarket)
    iv. Cross selling

<u>Π.χ (1)</u> τράπεζες

Έλεγχος ποιότητας και Ανάλυση ανταγωνισιμότητας

<u>Π.χ (2)</u> Η περίπτωση «Diapers and beer». Η παρατήρηση ότι πελάτες που αγοράζουν πάνες αγοράζουν και μπύρα επιτρέπουν στα καταστήματα να τοποθετούν αυτά τα είδη σχετικά κοντά , γνωρίζοντας ότι οι πελάτες θα κάνουν τη διαδρομή μεταξύ των ραφιών με τις πάνες και αυτών με τις μπύρες. Τοποθετώντας ανάμεσά τους και πατατάκια αυξάνουν τις πωλήσεις και στα τρία είδη.

<u>Π.χ (3)</u> Εταιρία πωλήσεως ηλεκτρονικών συσκευών θέλει να μελετήσει τις αγοραστικές συνήθειες των πελατών της, ώστε να προγραμματίσει ανάλογα την επόμενη διαφημιστική καμπάνια.

3) <u>Εντοπισμός απάτης και διαχείριση ρίσκου:</u>

Άλλες εφαρμογές που χρησιμοποιούν Εξόρυξη Δεδομένων:

i. Εξόρυξη κειμένου (newsgroup, Email, documents) και Web analysis
ii. Ευφυείς απαντήσεις σε ερωτήματα

Π.χ (1) Άτομα που σκηνοθετούν ατυχήματα για να εισπράξουν από τις ασφαλιστικές εταιρίες ή κάποιοι που κάνουν ξέπλυμα «βρώμικου χρήματος» εντοπίζοντας ύποπτες μεταφορές χρημάτων ή κάποιοι που κλέβουν τους πάροχους τηλεπικοινωνιών και κάνουν τηλεφωνήματα που έχουν κάποια επαναλαμβανόμενα σχέδια είτε προς μια κλειστή ομάδα ατόμων (κινητά) είτε κάποια συγκεκριμένη ώρα της ημέρας κλπ.
Π.χ (2) Εντοπισμός ακατάλληλων ιατρικών μεθόδων και θεραπειών.

1.4 Ταξινόμηση

1.4.1 Εισαγωγή ταξινόμησης

Η ταξινόμηση αποτελεί μία από τις βασικές τεχνικές εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένου) το οποίο με βάση τα χαρακτηριστικά αυτά, αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η διαδικασία της κατηγοριοποίησης χαρακτηρίζεται από ένα σαφή καθορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προκαθορισμένα παραδείγματα. Η ταξινόμηση δεδομένων είναι μια διαδικασία η οποία βρίσκει τις κοινές ιδιότητες μεταξύ ενός συνόλου αντικειμένων σε μια βάση δεδομένων και ταξινομεί τα αντικείμενα αυτά σε διαφορετικές κλάσεις (τάξεις) σύμφωνα με ένα μοντέλο ταξινόμησης.

1.4.2 Μαθηματική περιγραφή του προβλήματος ταξινόμησης

Έστω μια βάση δεδομένων $D = \{t_1, t_2, \dots, t_n\}$, όπου t_i είναι πλειάδες της μορφής $< t_{i1}, t_{i2}, \dots, t_{ip} >$ (που καλούνται στοιχεία ή εγγραφές ή παραδείγματα) και ένα

σύνολο κλάσεων $\mathrm{C} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{m}\}$. Το πρόβλημα της κατηγοριοποίησης συνίσταται στον προσδιορισμό της απεικόνισης

$$f: D \to C$$

όπου κάθε t_i αντιστοιχεί σε μια κλάση C_j. Η απεικόνιση αυτή ονομάζεται και *μοντέλο*.

Έτσι μια κλάση C_j ορίζεται ως το σύνολο των παραδειγμάτων που κατατάσσονται σ' αυτήν:

$$C_j = \{ \frac{t_i}{f(t_i)} = C_j, 1 \leq i \leq n, t_i \in D \}$$

Όπου κάθε παράδειγμα t_i θεωρείται ως ένα διάνυσμα . Τα t_{ik} , $k = 1$, p είναι τιμές (διακριτές ή αριθμητικές), που αναφέρονται σε αντίστοιχα φυσικά χαρακτηριστικά (features) $X_1, X_2, \ldots, X_p$. Γι ' αυτό και ένα τέτοιο διάνυσμα ονομάζεται διάνυσμα χαρακτηριστικών (feature vector). Κάθε χαρακτηριστικό X_k μπορεί να πάρει κάποιες τιμές $D_{xk} = \{ x_{ki}, i = 1, r \}$. Επομένως σ' ένα παράδειγμα κάθε t_{ik} είναι μια από τις x_{ki} , δηλ. $t_{ik} \in D_{xk}$.

Οι κλάσεις αναφέρονται και αυτές σ' ένα χαρακτηριστικό X_f , που ονομάζεται χαρακτηριστικό στόχου (target feature). Πιο συγκεκριμένα , οι κλάσεις αντιστοιχούν στις διαφορετικές τιμές που μπορεί να πάρει το χαρακτηριστικό στόχου.

ΚΕΦΑΛΑΙΟ 2: ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Για την επιτυχή διεκπεραίωση των διαφόρων εργασιών data mining έχουν αναπτυχθεί πολλές τεχνικές. Κάποιες από τις πιο σημαντικές τεχνικές είναι οι ακόλουθες:

- Τα δέντρα απόφασης (decision trees)
- Τα νευρωνικά δίκτυα (neural networks)
- Λογιστική παλινδρόμηση (logistic regression)

Οι παραπάνω τεχνικές διαφέρουν ως προς την ακρίβεια και τη δυνατότητα κατανόησής τους. Στη συνέχεια αναλύουμε κάθε επιμέρους τεχνική.

2.1 Δέντρα Αποφάσεων

Τα δέντρα απόφασης είναι πολύ ισχυρά εργαλεία που χρησιμοποιούνται ευρέως για τις περιπτώσεις της ταξινόμησης και της πρόβλεψης. Ένα δέντρο απόφασης αντιπροσωπεύει μια σειρά από IF THEN κανόνες ξεκινώντας από τη ρίζα του δέντρου και καταλήγοντας στα φύλλα του.

Οι εσωτερικοί κόμβοι ενός δέντρου απόφασης περιέχουν τα γνωρίσματα του προβλήματος, οι ακμές περιέχουν τις δυνατές τιμές των γνωρισμάτων και τα φύλλα περιέχουν τις πιθανές κλάσεις του προβλήματος. Απαραίτητο για την κατασκευή ενός δέντρου απόφασης είναι ένα σύνολο από στιγμιότυπα εκπαίδευσης, κάθε στιγμιότυπο του οποίου περιγράφεται από κάποια γνωρίσματα και την κλάση του προβλήματος στην οποία ανήκει.

Η διαδικασία που ακολουθούν οι αλγόριθμοι κατασκευής ενός δέντρου απόφασης συνοψίζεται στα ακόλουθα βήματα :

- Ξεκινώντας από τη ρίζα του δέντρου ο αλγόριθμος διασπά το σύνολο των στιγμιότυπων εκπαίδευσης σε υποσύνολα με βάση τη βέλτιστη ιδιότητα (best attribute) του κόμβου – η βέλτιστη ιδιότητα ενός κόμβου καθορίζεται

από κάποιο κριτήριο όπως το information gain, το gain ratio, δείκτη Gini (Index Gini). Επομένως, μπορούμε να πούμε συμπερασματικά ότι ως ρίζα επιλέγουμε εκείνο το χαρακτηριστικό που δίνει το μέγιστο κέρδος πληροφορίας και για να το ποσοτικοποιήσουμε θα χρησιμοποιήσουμε την έννοια της εντροπίας. Έτσι προκύπτει ένα πλήθος υποσυνόλων που το καθένα περιέχει λιγότερα παραδείγματα από το αρχικό σύνολο. Για καθένα από αυτά τα επιμέρους υποσύνολα εφαρμόζεται επαναληπτικά η παραπάνω διαδικασία χρησιμοποιώντας τα εναπομείναντα γνωρίσματα, οπότε η διάσπαση των στιγμιότυπων προχωρά και σταματά όταν όλα τα στιγμιότυπα του υποσυνόλου ανήκουν στην ίδια κλάση ή έχουν εξαντληθεί όλα τα γνωρίσματα. Στην ουσία πρόκειται για εφαρμογή της μεθόδου «Διαίρει και βασίλευε».

Εκτός από το σύνολο των στιγμιότυπων εκπαίδευσης, υπάρχει και το σύνολο ελέγχου με βάση τα οποία ελέγχεται η απόδοση του δέντρου, δηλαδή η ακρίβεια με την οποία το κατασκευασμένο δέντρο απαντά στο πρόβλημα της ταξινόμησης. Στην περίπτωση αυτή δίνουμε ως είσοδο στο δέντρο τις τιμές των γνωρισμάτων του στιγμιότυπου ελέγχου και περιμένουμε ως απάντηση την τάξη του στιγμιότυπου. Το πλήθος των λανθασμένων απαντήσεων (δηλαδή τα στιγμιότυπα στα οποία το δέντρο απάντησε διαφορετική κλάση από την πραγματική) καθορίζει την ακρίβεια του δέντρου.

Τα δέντρα απόφασης χρησιμοποιούνται ευρέως τόσο από την επιστημονική κοινότητα όσο και από τη βιομηχανία και αρκετοί αλγόριθμοι έχουν αναπτυχθεί για το σκοπό αυτό. Οι γνωστότεροι αλγόριθμοι εκπαίδευσης (ID3, C4.5, C&RT) χρησιμοποιούν μια top-down, εξαντλητική αναζήτηση στο χώρο των πιθανών δέντρων απόφασης. Αρχίζουν με ένα κενό δέντρο και προοδευτικά θέτουν πιο περίπλοκες προθέσεις με στόχο την εύρεση ενός δέντρου που ταξινομεί σωστά τα δεδομένα εκπαίδευσης.

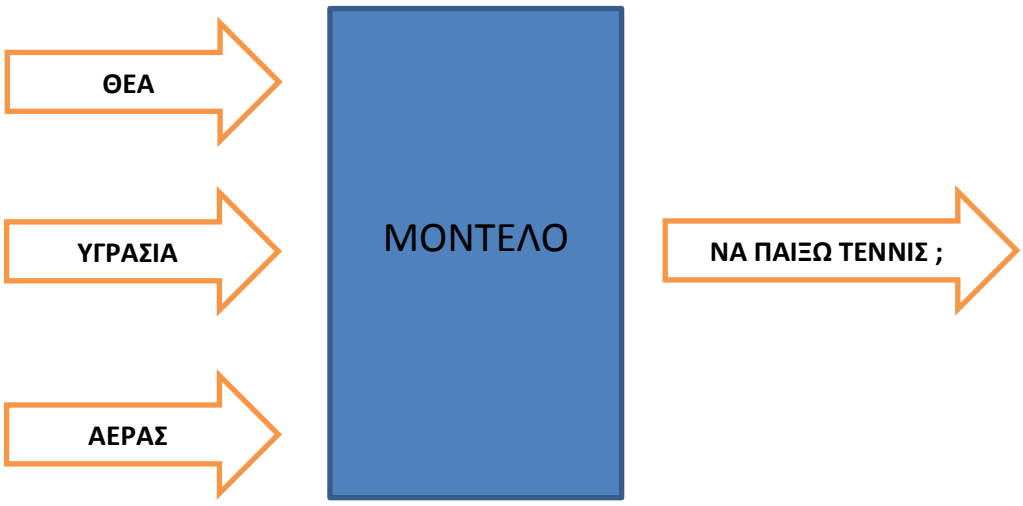
Έτσι, η διαδικασία κατασκευής δέντρου απόφασης είναι η εξής:

- Επιλογή χαρακτηριστικού για τη θέση του αρχικού κόμβου (ρίζας) και δημιουργία κλάδων για κάθε πιθανή τιμή του χαρακτηριστικού.
- Διάσπαση υποδειγμάτων σε υποσύνολα, ένα για κάθε κλάδο που εκτείνεται από τη ρίζα.
- Επανάληψη των παραπάνω για κάθε κλάδο με χρήση μόνο του υποσυνόλου των υποδειγμάτων κάθε κλάδου.
- Ολοκλήρωση της διαδικασίας όταν όλα τα υποδείγματα σε ένα κόμβο ανήκουν στην ίδια τάξη.

Όταν έχει ολοκληρωθεί η διαδικασία ανακάλυψης γνώσης με χρήση του αλγορίθμου, τότε το δέντρο μπορεί να αναπαρασταθεί ως σύνολο κανόνων της μορφής:

« If <ΣΥΝΟΛΟ ΣΥΝΘΗΚΩΝ> then <ΣΥΜΠΕΡΑΣΜΑ>» .

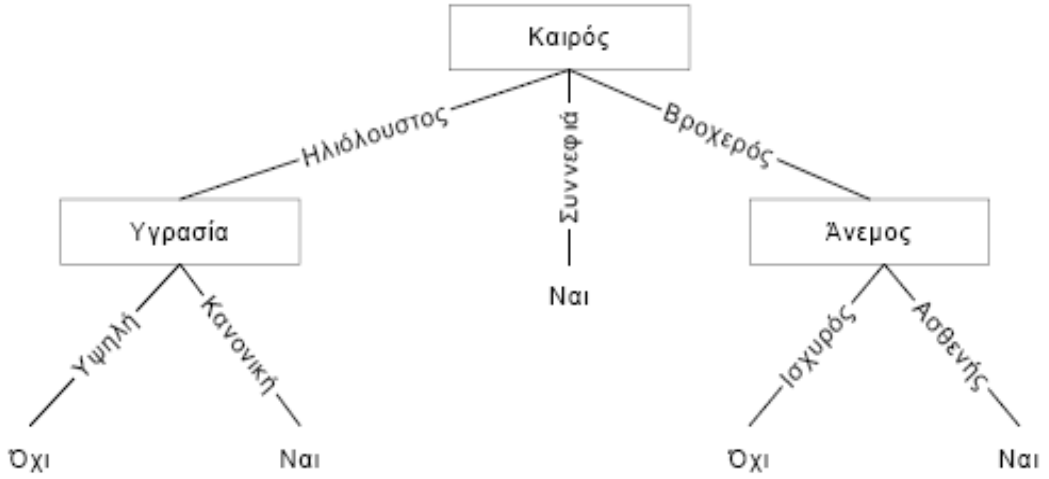
Η ανακάλυψη γνώσης με χρήση αλγορίθμων δέντρων απόφασης αποτελεί μια από τις πλέον δημοφιλείς τεχνικές επαγωγικής εκμάθησης και έχει μεγάλη εφαρμογή στη διάγνωση ιατρικών περιπτώσεων, στην εκτίμηση πιθανού ρίσκου από πιστοληπτικές τραπεζικές εργασίες κ.α.



Σχήμα 2: «Παίζεις τέννις»

Έστω για παράδειγμα το κλασικό πρόβλημα που προσπαθεί να απαντήσει στο ερώτημα «Παίζεις τέννις» και το οποίο έχει δυο κλάσεις: «Ναι» και «Όχι». Η απάντηση στο πρόβλημα εξαρτάται από τους εξής παράγοντες: τον Καιρό (με πιθανές τιμές: ήλιος, βροχή, συννεφιά), την Υγρασία (με πιθανές τιμές: υψηλή, κανονική) και τον Αέρα (με πιθανές τιμές: δυνατός, αδύνατος).

Στο Σχήμα 3 φαίνεται το δέντρο απόφασης του προβλήματος. Περιέχει τρεις εσωτερικούς κόμβους, σε κάθε κόμβο γίνεται έλεγχος ως προς κάποιο από τα γνωρίσματα του προβλήματος, ενώ στα φύλλα του περιέχονται οι κλάσεις του προβλήματος.



Σχήμα 3: Δέντρο απόφασης για το πρόβλημα «παίξε τέννις»

2.1.1 Αλγόριθμος C&RT

Οι C&RT αλγόριθμοι βασίζονται στη θεωρία των δέντρων ταξινόμησης και παλινδρόμησης που διατυπώθηκε από τον Breiman et al (1984). Σ' ένα αλγόριθμο C&RT τα δεδομένα χωρίζονται σε δύο υποσύνολα, ξεκινώντας βέβαια με τον αρχικό κόμβο-ρίζα ο οποίος περιέχει ολόκληρο το δείγμα εκπαίδευσης, με τρόπο ώστε κάθε υποσύνολο να εξασφαλίζει περισσότερη ομοιογένεια απ' ότι το προηγούμενο. Η διαδικασία αυτή επαναλαμβάνεται έως ότου να επιτευχθεί το κριτήριο ομοιογένειας ή κάποιο άλλο κριτήριο διακοπής. Το πεδίο πρόβλεψης μπορεί να χρησιμοποιηθεί αρκετές φορές σε διαφορετικά επίπεδα στο δέντρο.

Πλεονεκτήματα αλγορίθμου C&RT:

- Είναι αρκετά ευέλικτος
- Καλύτερη διαχείρηση δεδομένων με ελλειπούσες τιμές χρησιμοποιώντας υποκατάστατο διαχωρισμό
- Δίνει στο χρήστη τη δυνατότητα να καθορίσει την προηγούμενη (prior) κατανομή πιθανότητας σε ένα πρόβλημα ταξινόμησης
- Ο χρήστης μπορεί να εφαρμόσει ένα αυτόματο κλάδεμα, που θα παρουσιάζει πολυπλοκότητα ως προς τα κόστη με στόχο να αποκομίσει ένα πιο γενικευμένο δέντρο

2.1.1.1 Πεδία συχνότητας και πεδία βάρους

Για την κατασκευή του μοντέλου είναι απαραίτητο να γίνουν κάποιοι υπολογισμοί. Για παράδειγμα, για τη μείωση του μεγέθους του συνόλου δεδομένων χρειάζεται υπολογισμός των:

- Πεδίων συχνότητας
- Πεδίων βάρους

Είναι πολύ σημαντικό να γίνει ο σωστός διαχωρισμός μεταξύ των πεδίων βάρους και συχνότητας γιατί διαφορετικά θα προκύψουν λανθασμένα αποτελέσματα. Στην περίπτωση όπου τα πεδία συχνότητας ή βάρους δεν ορίζονται τότε η συχνότητα και τα βάρη για όλες τις καταχωρήσεις παίρνουν τις τιμές 1,0.

2.1.1.1.1 Πεδία συχνότητας

Ένα πεδίο συχνότητας αναπαριστά το συνολικό αριθμό των παρατηρήσεων που αντιπροσωπεύονται από κάθε καταχώρηση. Στην ανάλυση των συνολικών δεδομένων είναι σημαντικό να γνωρίζουμε σε ποιο πεδίο μια καταχώρηση (συνδυασμός περιπτώσεων) αναπαριστά περισσότερες από μια παρατηρήσεις. Ο συνολικός αριθμός των παρατηρήσεων μέσα στο δείγμα πρέπει πάντοτε να είναι

ίσος με το άθροισμα των τιμών στο πεδίο συχνότητας. Το αποτέλεσμα που προκύπτει αν χρησιμοποιήσουμε πεδίο συχνότητας είναι ίδιο με αυτό που λαμβάνουμε χρησιμοποιώντας δεδομένα case by case.

Στον παρακάτω πίνακα παρατηρούμε ένα υποθετικό παράδειγμα, με τα πεδία πρόβλεψης «φύλλο», «απασχόληση» και το πεδίο στόχος -κάπνισμα- «απόκριση» . Το πεδίο συχνότητας μας λέει, για παράδειγμα ότι 11 εργαζόμενοι άντρες ανταποκρίθηκαν με «ναι» στην ερώτηση αν καπνίζουν και 18 άνεργες γυναίκες ανταποκρίθηκαν με «όχι».

ΦΥΛΛΟ	ΕΡΓΑΣΙΑ	ΑΠΟΚΡΙΣΗ	ΣΥΧΝΟΤΗΤΑ
Άντρας	**ΝΑΙ**	**ΝΑΙ**	**11**
Άντρας	**ΝΑΙ**	**ΟΧΙ**	**17**
Άντρας	**ΟΧΙ**	**ΝΑΙ**	**12**
Άντρας	**ΟΧΙ**	**ΟΧΙ**	**21**
Γυναίκα	**ΝΑΙ**	**ΝΑΙ**	**11**
Γυναίκα	**ΝΑΙ**	**ΟΧΙ**	**15**
Γυναίκα	**ΟΧΙ**	**ΝΑΙ**	**15**
Γυναίκα	**ΟΧΙ**	**ΟΧΙ**	**18**

Πίνακας 1: Πίνακας με Πεδίο Συχνότητας

Στο συγκεκριμένο παράδειγμα, χρησιμοποιώντας το πεδίο συχνότητας, επεξεργαζόμαστε ένα πίνακα 8 καταχωρήσεων ενώ, έαν χρησιμοποιούσαμε δεδομένα case by case θα ήταν απαραίτητες 120 καταχωρήσεις.

2.1.1.1.2 Πεδία βάρους

Κάνοντας χρήση ενός πεδίου βάρους οδηγούμαστε σε μια άνιση μεταχείριση στις καταχωρήσεις σε ολόκληρο το σύνολο δεδομένων. Έτσι, η συνεισφορά μιας καταχώρησης στην ανάλυση είναι σταθμισμένη (weighted) σε αναλογία με τον πληθυσμό των μονάδων που η καταχώρηση αναπαριστά μέσα στο δείγμα. Για παράδειγμα, στην ερώτηση μιας έρευνας σε δείγμα 100.000 καταναλωτών, για λογαριασμό μιας επώνυμης βιομηχανίας παραγωγής καπνού, κατά πόσο καπνίζουν ή όχι, 20.000 ερωτηθέντες απάντησαν θετικά και 80.000 αρνητικά. Σε μια προσπάθεια να μείωσουμε το μέγεθος των δεδομένων, πιθανότατα θα

συμπεριλάβουμε όλους όσους είναι καπνιστές και μόνο 25% του δείγματος (20.000) που δεν είναι καπνιστές. Κάτι τέτοιο μπορούμε να το κάνουμε αν ορίσουμε μια περίπτωση βάρους ίση με 1 γι' αυτούς που καπνίζουν και 4 γι' αυτούς που δεν καπνίζουν.

2.1.1.2 Κατασκευή ενός C&RT δέντρου

Η βασική ιδέα κατασκευής ενός δέντρου είναι να επιλέξουμε ένα διαχωρισμό (split) μεταξύ όλων των πιθανών διαχωρισμών σε κάθε κόμβο έτσι ώστε οι θυγατρικοί κόμβοι που θα προκύψουν ως αποτέλεσμα να είναι οι καθαρότεροι. Με τον όρο καθαρότητα αναφερόμαστε στην ομοιότητα των τιμών του πεδίου στόχος. Σ' ένα εντελώς καθαρό κόμβο, όλες οι καταχωρήσεις έχουν την ίδια τιμή στο πεδίο στόχος. Ο C&RT αλγόριθμος μετρά την καθαρότητα ενός διαχωρισμού σε ένα κόμβο ορίζοντας ένα μέτρο καθαρότητας.

Τα βήματα που χρησιμοποιούνται για την κατασκευή ενός C&RT δέντρου είναι τα ακόλουθα (ξεκινώντας βέβαια με τον αρχικό κόμβο-ρίζα ο οποίος περιέχει όλες τις καταχωρήσεις):

1) Βρίσκουμε τον καλύτερο διαχωρισμό για κάθε πεδίο πρόβλεψης (predictor field). Για κάθε πεδίο πρόβλεψης, βρίσκουμε τον καλύτερο δυνατό διαχωρισμό για αυτό το πεδίο ως ακολούθως:

    - *Αριθμητικά πεδία (range):* Ταξινομούμε τις τιμές των πεδίων στον κόμβο από την μικρότερη στη μεγαλύτερη. Επιλέγουμε κάθε σημείο με τη σειρά σαν σημείο διαχωρισμού και υπολογίζουμε το στατιστικό μη καθαρότητας για τους θυγατρικούς κόμβους που προκύπτουν σαν αποτέλεσμα του διαχωρισμού. Έπειτα διαλέγουμε σαν σημείο διαχωρισμού για το πεδίο, αυτό το οποίο αποδίδει τη μεγαλύτερη μείωση στη μη καθαρότητα σε σύγκριση με την μη καθαρότητα του κόμβου ο οποίος διαχωρίζεται.

- *Κατηγορικά πεδία (συμβολικά):* Εξετάζουμε τον κάθε πιθανό συνδυασμό των τιμών σαν δυο υποσύνολα. Για κάθε συνδυασμό, υπολογίζουμε τη μη καθαρότητα των θυγατρικών κόμβων για το διαχωρισμό που βασίζεται σε αυτό το συνδυασμό. Επιλέγουμε σαν καλύτερο σημείο διαχωρισμού για το πεδίο, αυτό το οποίο αποδίδει τη μεγαλύτερη μείωση στη μη καθαρότητα σε σύγκριση με τη μη καθαρότητα του κόμβου ο οποίος διαχωρίζεται.

2) Βρίσκουμε τον καλύτερο διαχωρισμό για τον κόμβο και προσδιορίζουμε το πεδίο του οποίου ο καλύτερος διαχωρισμός δίνει τη μεγαλύτερη μείωση στη μη καθαρότητα για τον κόμβο. Στη συνέχεια επιλέγουμε αυτό τον καλύτερο διαχωρισμό του πεδίου ως το βέλτιστο συνολικό διαχωρισμό για τον κόμβο.
3) Ελέγχουμε εάν ικανοποιούνται οι κανόνες διακοπής και επαναλαμβάνουμε. Εάν οι κανόνες διακοπής δεν ικανοποιούνται από το διαχωρισμό ή από τον γεννήτορα κόμβο, εφαρμόζουμε το διαχωρισμό για να δημιουργήσουμε δύο θυγατρικούς κόμβους. Επαναλαμβάνουμε όλη τη διαδικασία σε κάθε θυγατρικό κόμβο.

2.1.1.3 Μέτρα μη καθαρότητας

Για την εύρεση διαχωρισμών στα C&RT μοντέλα υπάρχουν τρία γνωστά διαφορετικά μέτρα μη καθαρότητας τα οποία όμως εξαρτώνται από τον τύπο του πεδίου στόχου. Οι τύποι *Gini* και *Towing* χρησιμοποιούνται για συμβολικά πεδία στόχου ενώ η LSD (Least Squared deviation) ή μέθοδος απόκλισης ελαχίστων τετραγώνων χρησιμοποιείται για συνεχείς στόχους. Στην παρούσα εργασία θα ασχοληθούμε με τα *Gini* και *Towing* μέτρα μη καθαρότητας.

Gini:

Ο δείκτης ακαθαρσίας Gini $g(t)$ σε έναν κόμβο t ενός CART δέντρου, ορίζεται ως:

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

όπου i και j είναι κατηγορίες στο πεδίο στόχου

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

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

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

όπου :

$\pi(j)$ είναι η τιμή της prior πιθανότητας για την κατηγορία j

$N_j(t)$ είναι το πλήθος των καταχωρήσεων στην κατηγορία j του κόμβου t

N_j είναι το πλήθος των των καταχωρήσεων στην κατηγορία j στον αρχικό κόμβο-ρίζα.

Σημειώνουμε ότι όταν χρησιμοποιείται ο δείκτης Gini για την εύρεση της βελτίωσης για έναν διαχωρισμό κατά την διάρκεια της ανάπτυξης του δέντρου, για να υπολογιστούν το N_j και το $N_j(t)$ χρησιμοποιούνται οι καταχωρήσεις στον αρχικό κόμβο-ρίζα και στον κόμβο t αντίστοιχα που έχουν έγκυρες τιμές για το πεδίο διαχωρισμού (split-predictor).

Μια άλλη μορφή του δείκτη μη καθαρότητας Gini είναι:

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

Έτσι, όταν οι καταχωρήσεις σε ένα κόμβο διανέμονται ομαλά δια μέσου των κατηγοριών, ο δείκτης Gini λαμβάνει τη μεγαλύτερη τιμή του $1 - \frac{1}{k}$, όπου k είναι το πλήθος των κατηγοριών για το πεδίο στόχος. Ο δείκτης Gini ισούται με 0 όταν όλες οι καταχωρήσεις σε ένα κόμβο ανήκουν στην ίδια κατηγορία.

Για το διαχωρισμό s στον κόμβο t η συνάρτηση του κριτηρίου Gini $\Phi(s,t)$ ορίζεται:

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

όπου :

p_L είναι η μερίδα των καταχωρήσεων στον κόμβο t οι οποίες στέλνονται στον αριστερό θυγατρικό κόμβο

p_R είναι η μερίδα των καταχωρήσεων στον κόμβο t οι οποίες στέλνονται στο δεξιό θυγατρικό κόμβο.

Οι λόγοι p_Lκαι p_R ορίζονται ως εξής:

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

και

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

Επιλέγεται ο κατάλληλος διαχωρισμός s ούτως ώστε να μεγιστοποιηθεί η τιμή της $\Phi(s,t)$ συνάρτησης.

Twoing:

Ο δείκτης Twoing είναι βασισμένος στο διαχωρισμό των κατηγοριών στόχου σε δύο υπερκλάσεις και ακολούθως στην εύρεση του βέλτιστου διαχωρισμού στο πεδίο πρόβλεψης και στηρίζεται στις δύο υπερκλάσεις. Οι υπερκλάσεις C_1και C_2 ορίζονται ως εξής:

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

και

$$C_2 = C - C_1$$

όπου:

C είναι το σύνολο των κατηγοριών του πεδίου στόχος

$p(j/t_R)$, $p(j/t_L)$ είναι τα $p(j/t)$ όπως ορίζονται στο κριτήριο Gini για τους δεξιούς και αριστερούς θυγατρικούς κόμβους αντίστοιχα.

Η συνάρτηση του κριτηρίου του Twoing για το διαχωρισμό s στον κόμβο t ορίζεται ως:

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

όπου:

t_Lκαι t_R είναι οι κόμβοι που δημιουργούνται από το διαχωρισμό s.

Ο διαχωρισμός που επιλέγεται είναι αυτός ο οποίος μεγιστοποιεί το Twoing κριτήριο.

2.1.1.4 Κανόνες διακοπής – ολοκλήρωσης της διαδικασίας

Οι κανόνες διακοπής-ολοκλήρωσης της διαδικασίας ελέγχουν αν η διαδικασία κατασκευής δέντρου πρέπει να σταματήσει ή όχι. Χρησιμοποιούνται οι εξής κανόνες διακοπής:

- Αν ο κόμβος γίνει καθαρός: δηλαδή αν όλες οι περιπτώσεις μέσα σε ένα κόμβο έχουν πανομοιότυπες τιμές της εξαρτημένης μεταβλητής τότε ο κόμβος δε θα διαχωριστεί.
- Αν όλες οι περιπτώσεις μέσα σε ένα κόμβο έχουν πανομοιότυπες τιμές για κάθε μεταβλητή πρόβλεψης τότε ο κόμβος δε θα διαχωριστεί.
- Αν το βάθος του πρόσφατου δέντρου πλησιάζει την τιμή του μέγιστου ορίου βάθους το οποίο καθορίζεται από το χρήστη, η διαδικασία κατασκευής δέντρου θα σταματήσει.
- Αν το μέγεθος ενός κόμβου είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από τον χρήστη τότε ο κόμβος δεν θα διαχωριστεί.
- Αν ο διαχωρισμός ενός κόμβου έχει σαν αποτέλεσμα ένα θυγατρικό κόμβο του οποίου το μέγεθος είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από το χρήστη τότε ο κόμβος δε θα διαχωριστεί.
- Ο καλύτερος διαχωρισμός για ένα κόμβο αποδίδει μια μείωση στη μη καθαρότητα η οποία είναι μικρότερη από την ελάχιστη αλλαγή στη μη καθαρότητα που ορίζεται από το χρήστη.

2.1.1.5 *Διαδικασία κλαδέματος*

Το κλάδεμα (pruning) αναφέρεται στη διαδικασία του ελέγχου ενός πλήρους αναπτυσσόμενου δέντρου και της αφαίρεσης των διαχωρισμών των κάτω επιπέδων που δεν έχουν σημαντική συνεισφορά στην ακρίβεια του δέντρου. Το λογισμικό στο κλάδεμα του δέντρου προσπαθεί να δημιουργήσει το μικρότερο δέντρο του οποίου το ρίσκο λανθασμένης ταξινόμησης δεν είναι πολύ μεγαλύτερο από το ρίσκο λανθασμένης ταξινόμησης του μεγαλύτερου πιθανού δέντρου. Η διαδικασία αφαιρεί ένα κλαδί δέντρου, αν το κόστος το οποίο σχετίζεται με τη μεγαλύτερη πολυπλοκότητα του δέντρου είναι μεγαλύτερο από το κέρδος το οποίο σχετίζεται με το εάν έχουμε ένα άλλο επίπεδο κόμβων (κλαδί). Χρησιμοποιεί ένα δείκτη ο οποίος μετρά το ρίσκο λανθασμένης ταξινόμησης και την πολυπλοκότητα του δέντρου αφού στόχος μας είναι να ελαχιστοποιήσουμε και τα δύο.

Το μέτρο κόστους πολυπλοκότητας (cost complexity) ορίζεται ως εξής:

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

όπου:

$R(T)$ είναι το ρίσκο λανθασμένης ταξινόμησης του δέντρου T ,

$\left|\tilde{T}\right|$είναι το πλήθος των τερματικών κόμβων για το δέντρο T ,

a είναι το κόστος πολυπλοκότητας ανά τερματικό κόμβο για το δέντρο

Η τιμή a υπολογίζεται από τον αλγόριθμο κατά την διάρκεια του κλαδέματος.

Κάθε δέντρο που μπορούμε να παράγουμε έχει ένα μέγιστο μέγεθος (T_{max}), όπου σε κάθε τερματικό κόμβο περιέχεται μόνο μια καταχώρηση. Στην περίπτωση που το κόστος πολυπλοκότητας είναι μηδενικό $(\alpha = 0)$, το μέγιστο δέντρο έχει το χαμηλότερο ρίσκο, αφού κάθε εγγραφή προβλέπεται τέλεια. Επομένως, όσο μεγαλύτερη είναι η τιμή του α, τόσο μικρότερος είναι ο αριθμός των τερματικών κόμβων στο $T(a)$, δηλαδή το δέντρο με το μικρότερο κόστος πολυπλοκότητας για το δοσμένο α. Όταν το α αυξάνεται από το 0 τότε παράγει μια πεπερασμένη ακολουθία από υποδέντρα (T_1, T_2, T_3), το καθένα με λιγότερους τερματικούς κόμβους από το προηγούμενο. Το κλάδεμα κόστους πολυπλοκότητας δουλεύει

αφαιρώντας τον πιο αδύναμο διαχωρισμό. Οι εξισώσεις που ακολουθούν εκφράζουν το κόστος πολυπλοκότητας για τον κόμβο $\{t\}$, που είναι ένας οποιοσδήποτε ξεχωριστός-μόνος κόμβος και για T_t, τον υπο-κλάδο του $\{t\}$:

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

και

$$R_a(T_t) = R(T_t) + a\left|\widetilde{T}_t\right|$$

Στην περίπτωση που το $R_\alpha(T_t)$ είναι μικρότερο από το $R_\alpha(\{t\})$, το κλαδί T_t έχει μικρότερο κόστος πολυπλοκότητας από αυτό του ξεχωριστού κόμβου $\{t\}$.

Η διαδικασία ανάπτυξης του δέντρου εξασφαλίζει ότι για $(\alpha = 0)$ ισχύει

$$R_a(\{t\}) \geq R_a(T_t) \qquad (\mathbf{1})$$

Καθώς το α αυξάνεται από το 0, τα $R_\alpha(\{t\})$ και $R_\alpha(T_t)$ αυξάνονται γραμμικά με το $R_\alpha(T_t)$ να αυξάνεται με ταχύτερο ρυθμό. Τελικά, βρίσκουμε ένα κάτω όριο α' τέτοιο ώστε $R_a(\{t\}) < R_a(T_t)$ για όλα τα $\alpha > \alpha'$. Συμπεραίνουμε ότι όταν το α γίνεται μεγαλύτερο από το α', το κόστος πολυπλοκότητας του δέντρου μειώνεται αν κόψουμε το υποκλάδι (sub branch) T_t κάτω από το $\{t\}$.

Μπορούμε εύκολα να υπολογίσουμε το κάτω όριο λύνοντας την $(\mathbf{1})$ ούτως ώστε να βρούμε τη μεγαλύτερη τιμή του α για την οποία ισχύει η ανισότητα, η οποία συμβολίζεται και ως $g(t)$. Συνεπώς, προκύπτει:

$$\alpha \leq g(t) = \frac{R(t) - R(T_t)}{\left|\widetilde{T}_t\right| - 1}$$

Μπορούμε να ορίσουμε σαν το πιο αδύναμο (link) σύνδεσμο (t) στο δέντρο T τον κόμβο ο οποίος παίρνει τη μικρότερη τιμή του $g(t)$:

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

Κατά συνέπεια, καθώς το α αυξάνεται, ο $\bar{t}$ είναι ο πρώτος κόμβος για τον οποίο ισχύει

$R_a(\{t\}) = R_a(T_t)$. Στο σημείο αυτό, το $\{t\}$ προτιμάται από το $T_{\bar{t}}$ και το υποκλάδι κλαδεύεται.

Συνοπτικά ο αλγόριθμος κλαδέματος βασίζεται στα εξής βήματα:

- Ορίζουμε το $\alpha_1 = 0$ και ξεκινούμε με το δέντρο για το οποίο $T_1 = T(0)$ δηλαδή το πλήρως αναπτυσσόμενο δέντρο.
- Αυξάνουμε το α μέχρι το κλάδεμα ενός κλαδιού. Έπειτα, κλαδεύουμε το κλαδί από το δέντρο και υπολογίζουμε την εκτίμηση του ρίσκου του δέντρου το οποίο έχουμε κλαδέψει.
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να απομείνει μόνο ο αρχικός κόμβος ρίζα, αποδίδοντας μια σειρά από υποδέντρα $T_1, T_2, \dots, T_k$.
- Στην περίπτωση που επιλέξουμε τον κανόνα του τυπικού σφάλματος, τότε διαλέγουμε το μικρότερο δέντρο T_{opt} για το οποίο
$$R\big(T_{opt}\big) \leq \min_k R(T_k) + m \times SE\big(R(T)\big)$$
- Στην περίπτωση που δεν επιλέγουμε τον κανόνα τυπικού σφάλματος τότε διαλέγουμε το δέντρο με τη μικρότερη ρίσκου $R(T)$.

2.2 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι ένα εργαλείο που έχει πολλές και ποικίλες εφαρμογές στον κλάδο του data mining λόγω της δυναμικής τους, της ευελιξίας τους και της ευκολίας στη χρήση τους. Ο όρος νευρωνικά δίκτυα χρησιμοποιείται για μια αόριστη οικογένεια μοντέλων, που χαρακτηρίζονται από ένα μεγάλο χώρο παραμέτρων και ευέλικτη δομή. Η ιδέα για την μελέτη και την ανάπτυξη των νευρωνικών δικτύων, προήλθε από τη λειτουργία και τη δομή του ανθρώπινου εγκεφάλου και των διαδικασιών του σχετικά με τη μάθηση, τη μνήμη, τη γενίκευση, την ομαδοποίηση προτύπων κ.λ.π.

Οι ορισμοί για τα νευρωνικά δίκτυα, ποικίλουν όσο και οι τομείς στους οποίους χρησιμοποιούνται. Από την στιγμή που δεν υπάρχει ένας συγκεκριμένος ορισμός

που να καλύπτει απόλυτα ολόκληρη την οικογένεια των μοντέλων, αποδεχόμαστε την ακόλουθη περιγραφή (Haykin , 1998):

Νευρωνικό δίκτυο είναι ένας μαζικός παράλληλος διανεμημένος επεξεργαστής ο οποίος εκ φύσεως αποθηκεύει εμπειρική γνώση και την καθιστά διαθέσιμη για χρήση. Προσομοιάζει τον ανθρώπινο εγκέφαλο σε δυο τομείς:

- Η γνώση αποκτάται από το δίκτυο μέσω μιας διαδικασίας μάθησης
- Οι ενδονευρωνικές συνδέσεις, γνωστές και ως συναπτικά βάρη, χρησιμοποιούνται για τη φύλαξη γνώσης.

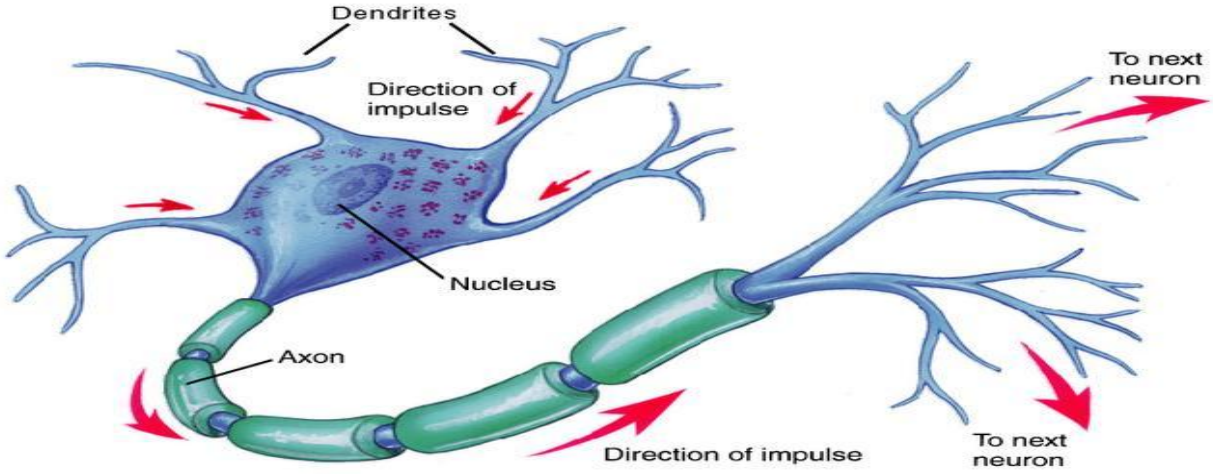
2.2.1 Νευρώνας-Λειτουργία του βιολογικού νευρώνα

Ο ανθρώπινος εγκέφαλος είναι ένας ιδιαίτερα πολύπλοκος, μη γραμμικός και παράλληλος ηλεκτρονικός υπολογιστής (σύστημα επεξεργασίας πληροφοριών). Έχει την ικανότητα να οργανώνει τους νευρώνες με τέτοιο τρόπο ώστε να κάνει συγκεκριμένους υπολογισμούς, όπως για παράδειγμα την αναγνώριση προτύπων (pattern recognition), την αντίληψη (perception) και την κίνηση, πολύ πιο γρήγορα από τον γρηγορότερο ηλεκτρονικό υπολογιστή που υπάρχει.

Ο ανθρώπινος εγκέφαλος αποτελείται από περίπου 100 δισεκατομμύρια νευρικά κύτταρα ή νευρώνες. Ο νευρώνας είναι ένα μεγάλο κύτταρο του οποίου η δομή περιλαμβάνει τέσσερα κύρια τμήματα που λειτουργικά παίζουν διαφορετικούς ρόλους:

- Το **σώμα**, που περιέχει τον πυρήνα και αποτελεί την καρδιά του κυττάρου.
- Οι **δενδρίτες**, που είναι οι πύλες εισόδου του νευρώνα και δέχονται ηλεκτρικά σήματα από άλλους νευρώνες.
- Ο **άξονας**, που είναι η πύλη εξόδου του νευρώνα. Μοιάζει με μακρόστενη κλωστή και ο σκοπός του είναι να μεταδώσει τα ηλεκτρικά σήματα, που δημιουργούνται στο νευρώνα, στους άλλους νευρώνες.

- Οι **συνάψεις** οι οποίες αποτελούν την περιοχή της σύνδεσης μεταξύ δύο νευρώνων. Είναι τα σημεία ένωσης των διακλαδώσεων του άξονα ενός νευρώνα-αποστολέα και των δενδριτών των νευρώνων-παραληπτών.



Σχήμα 4: Ο Βιολογικός νευρώνας

Στους βιολογικούς νευρώνες, φορείς πληροφορίας είναι ηλεκτρικοί παλμοί που ταξιδεύουν στον άξονα κάθε νευρώνα και μέσω των συνάψεων διαδίδονται στους δενδρίτες των παραληπτών νευρώνων. Κάθε νευρώνας A συλλέγει όλο το ηλεκτρικό φορτίο που δέχεται από κάθε σύναψη στους δενδρίτες του, ζυγίζοντας το εισερχόμενο φορτίο με το αντίστοιχο συνοπτικό βάρος. Έτσι, όσο πιο ισχυρή είναι η συναπτική ζεύξη τόσο πιο πολύ έντονα συμμετέχει το συγκεκριμένο φορτίο εισόδου στο συνολικό άθροισμα. Αν το άθροισμα του φορτίου ξεπερνάει κάποιο κατώφλι τότε ο άξονας του A αρχίζει να παράγει ηλεκτρικούς παλμούς με μεγάλη συχνότητα οπότε λέμε ότι ο νευρώνας *πυροβολεί (fires).* Αν όμως το φορτίο δεν περνάει το συγκεκριμένο αυτό όριο, τότε ο νευρώνας παράγει πολύ αραιά παλμούς, σε τυχαίες στιγμές οπότε λέμε ότι ο νευρώνας είναι *αδρανής.* Κάθε παλμός έχει συγκεκριμένο χρονικό πλάτος t_p και μετά από κάθε παλμό ο νευρώνας χρειάζεται ένα ελάχιστο χρόνο ανάπαυσης t_r. Έτσι, ο μέγιστος αριθμός των παλμών δεν ξεπερνάει το όριο :

$$Firing\ frequency\ < 1/(t_p + t_r)$$

Τελικά, οι παλμοί που παράγονται ταξιδεύουν κατά μήκος του άξονα και τροφοδοτούν τους άλλους νευρώνες με τους οποίος συνδέεται ο A.

2.2.2 Το μαθηματικό μοντέλο

Είναι πρακτικά αδύνατο για τα τεχνητά νευρωνικά δίκτυα να προσομοιάσουν πλήρως την πολυπλοκότητα του ανθρώπινου εγκεφάλου, αφού αποτελούνται, το πολύ, από μερικές εκατοντάδες (ή χιλιάδες) νευρώνες και περιορισμένο αριθμό συνδέσεων μεταξύ τους. Παρόλα αυτά κάποια δίκτυα έχουν χρησιμοποιηθεί για την επίλυση αρκετά περίπλοκων υπολογιστικών προβλημάτων.

Για τη μοντελοποίηση ενός βιολογικού νευρώνα σε ένα μαθηματικό μοντέλο, λαμβάνουμε υπόψη τρεις βασικές συνιστώσες. Αρχικά, οι συνάψεις των βιολογικών νευρώνων μοντελοποιούνται σαν **συναπτικά βάρη (synaptic weights)**. Ας θυμηθούμε πως οι συνάψεις των βιολογικών νευρώνων είναι υπεύθυνες για τη διασύνδεση του νευρωνικού δικτύου και δίνουν τη δύναμη των συνδέσεων. Για ένα τεχνητό νευρώνα, τα βάρη είναι πραγματικοί αριθμοί και αντιπροσωπεύουν τις συνάψεις. Ένα αρνητικό βάρος εκφράζει μια ανασταλτική σύνδεση, ενώ ένα θετικό μια διεγερτική σύνδεση. Πολλά μοντέλα νευρώνων περιλαμβάνουν επίσης και ένα εξωτερικό βάρος, που ονομάζεται **μεροληψία (bias)**. Σκοπός της μεροληψίας είναι η αύξηση ή η μείωση της τιμής που δίνει το δίκτυο σαν είσοδο στη συνάρτηση ενεργοποίησης ανάλογα με το αν είναι αρνητικό ή θετικό.

Οι υπόλοιπες συνιστώσες του μοντέλου αντιπροσωπεύουν τη δραστηριότητα του νευρώνα. Οι **είσοδοι (inputs)** του νευρώνα αθροίζονται και τροποποιούνται από τα συναπτικά βάρη. Τέλος, μια **συνάρτηση ενεργοποίησης (activation function)** ελέγχει το εύρος του εξερχόμενου φορτίου.

Θα δούμε τώρα πιο αναλυτικά τη διαδικασία μοντελοποίησης ενός βιολογικού νευρώνα σε ένα μαθηματικό μοντέλο. Έστω,

$w_{k1}, w_{k2}, \dots, w_{kp}$ είναι τα συναπτικά βάρη,

$x_1, \dots, x_p$ οι είσοδοι του νευρώνα k και

b_k η μεροληψία.

Τότε, το άθροισμα v_k του φορτίου που δέχεται ο νευρώνας εκφράζεται ως (Ρίζος, 1996):

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

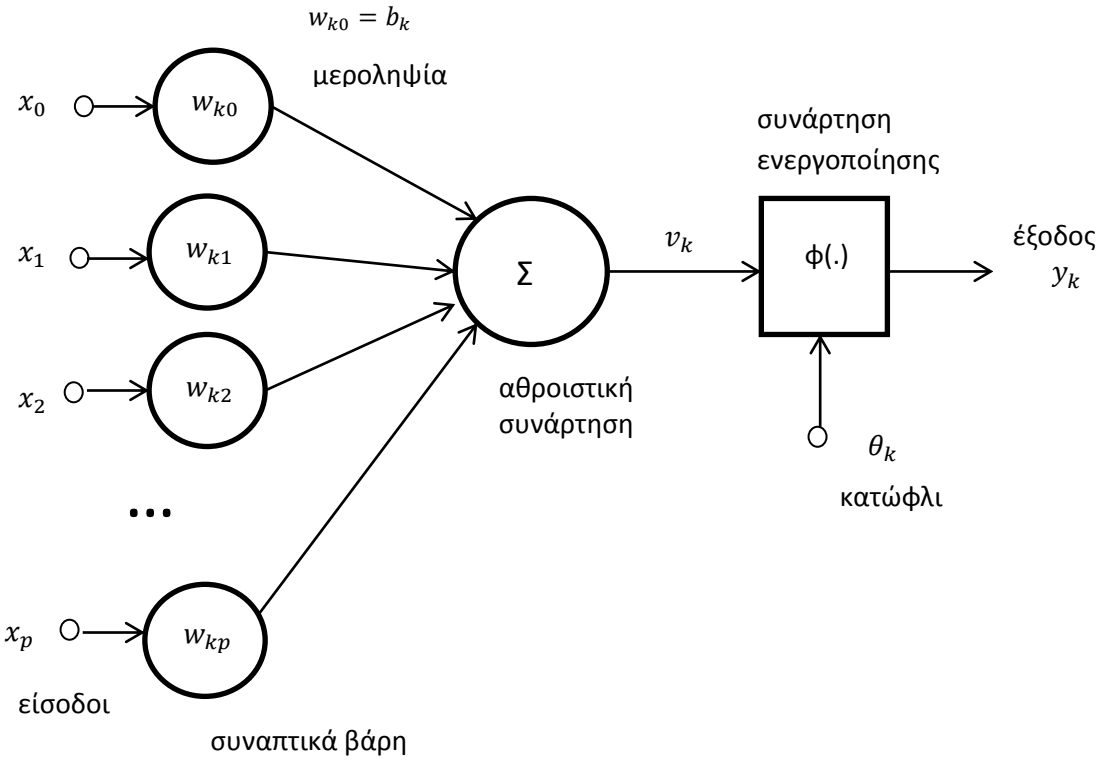
Η έξοδος του νευρώνα y_k, θα είναι το αποτέλεσμα της εφαρμογής μιας συνάρτησης ενεργοποίησης $\varphi(.)$ στη τιμή του v_k:

$$y_k = \varphi(v_k - b_k).$$

Η μεροληψία b_k είναι μια εξωτερική παράμετρος του νευρώνα που δεν εξαρτάται από καμιά τιμή εισόδου και μπορούμε να την εντάξουμε στο μοντέλο του νευρώνα ως μια νέα σύναψη που έχει σαν είσοδο $x_0 = \pm 1$ (ανάλογα αν αυξάνει ή μειώνει την τιμή εισόδου στο δίκτυο) και βάρος $w_{k0} = b_k$. Έτσι, θέτοντας: $u_k = v_k - b_k$ οι εξισώσεις που περιγράφουν το νευρώνα γίνονται τελικά:

$$u_k = \sum_{j=0}^{p} w_{kj} x_j$$

$$y_k = \varphi(u_k)$$



Σχήμα 5: Ο τεχνητός νευρώνας

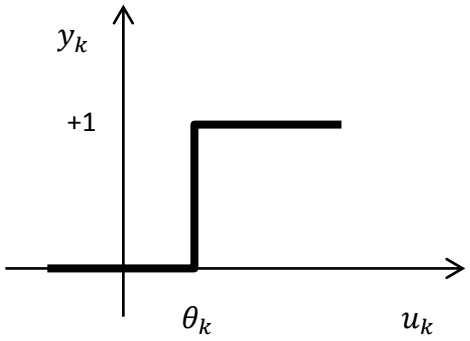
2.2.3 Συνάρτηση Ενεργοποίησης

Η συνάρτηση ενεργοποίησης λειτουργεί σαν *συμπιεστική* συνάρτηση, ώστε η έξοδος του νευρώνα σε ένα νευρωνικό δίκτυο να είναι μεταξύ συγκεκριμένων τιμών (συνήθως μεταξύ 0 και 1, ή -1 και 1). Ακολουθούν μερικοί τύποι συναρτήσεων ενεργοποίησης $\boldsymbol{\varphi}(.)$:

- Η **συνάρτηση κατώφλι (threshold function)** η οποία έχει σαν έξοδο 0 αν το εισερχόμενο άθροισμα είναι μικρότερο από μια καθορισμένη τιμή-κατώφλι θ_k, και 1 αν είναι μεγαλύτερο ή ίσο με αυτό.

$$\upsilon = u_k - \theta_k$$

$$y_k = \varphi(\upsilon) = \begin{cases} 0, & \alpha\nu\ \upsilon < 0 \\ 1, & \alpha\nu\ \upsilon \geq 0 \end{cases}$$

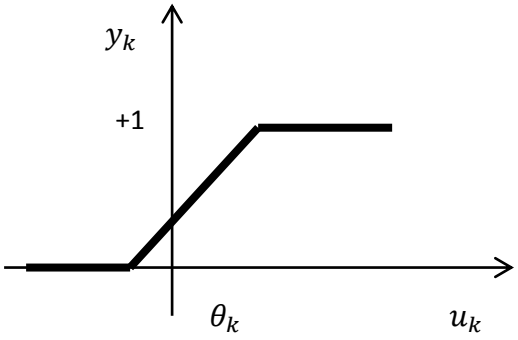


Γράφημα 1: Συνάρτηση κατώφλι

- Η **τμηματική γραμμική συνάρτηση (pricewise-linear function).** Όπως και η συνάρτηση κατώφλι έχει σαν έξοδο 0 ή 1, καθώς επίσης και τιμές μεταξύ αυτών που εξαρτώνται από τον παράγοντα ενίσχυσης μέσα στη γραμμική περιοχή της συνάρτησης.

$$\upsilon = u_k - \theta_k$$

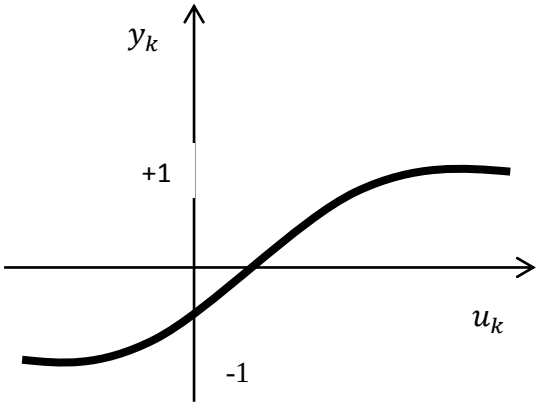
$$\varphi(\upsilon) = \begin{cases} 1, & \alpha\nu \quad \upsilon \geq \frac{1}{2} \\ \upsilon, & \alpha\nu \quad -\frac{1}{2} < \upsilon < \frac{1}{2} \\ 0, & \alpha\nu \quad \upsilon \leq -\frac{1}{2} \end{cases}$$



Γράφημα 2: Τμηματική γραμμική συνάρτηση

- Η **σιγμοειδής συνάρτηση (sigmoid function)**. Είναι μια γνησίως αύξουσα συνάρτηση που είναι ομαλή και ασυμπτωτική. Αντίθετα με τη συνάρτηση κατώφλι είναι διαφορίσιμη και είναι η πλέον χρησιμοποιούμενη συνάρτηση ενεργοποίησης για τα νευρωνικά δίκτυα. Ένα παράδειγμα της σιγμοειδής συνάρτησης είναι η υπερβολική εφαπτόμενη:

$$\varphi(\upsilon) = \tanh(\upsilon) = \frac{1 - e^{-\upsilon}}{1 + e^{-\upsilon}}$$

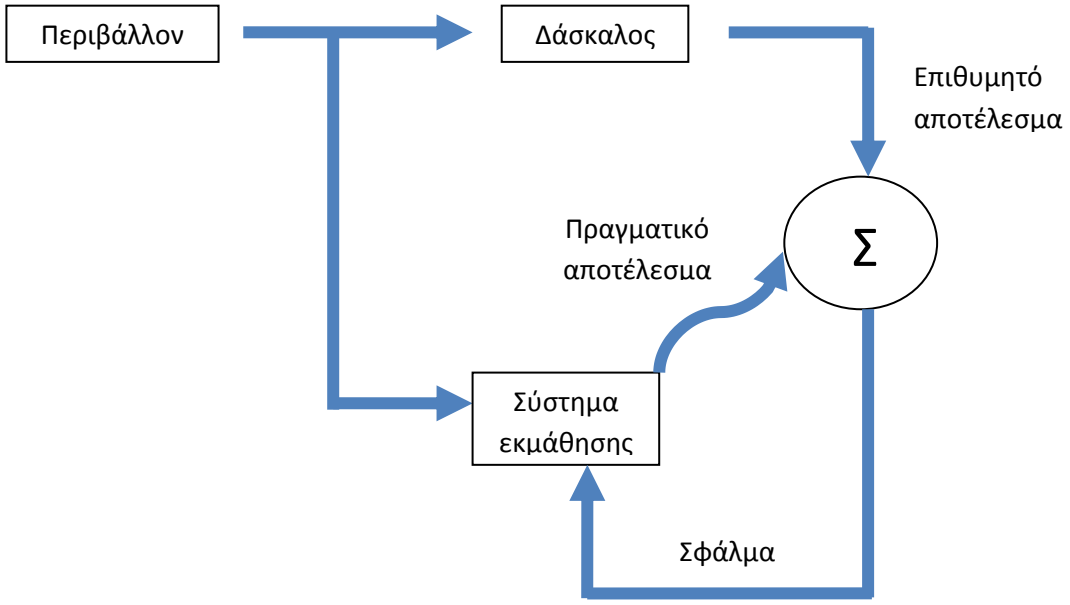


Γράφημα 3: Σιγμοειδής συνάρτηση

2.2.4 Ταξινόμηση Νευρωνικών αλγορίθμων

Ένα νευρωνικό δίκτυο είναι απαραίτητο να ρυθμιστεί έτσι ώστε όταν ένα set δεδομένων δοθεί σαν είσοδος, να παράγει το επιθυμητό set δεδομένων εξόδου. Υπάρχουν διάφορες μέθοδοι για την προσαρμογή των δυνάμεων των συνδέσεων σε ένα δίκτυο. Ένας τρόπος για να επιτευχθεί η μάθηση του δικτύου, είναι η ρύθμιση των συναπτικών βαρών, χρησιμοποιώντας μια *a priori* γνώση. Μια άλλη μέθοδος είναι η εκπαίδευση του νευρωνικού δικτύου, τροφοδοτώντας το, με πρότυπα διδασκαλίας και επιτρέποντάς του να αλλάξει τα συναπτικά βάρη ανάλογα με κάποιο διδακτικό κανόνα. Μπορούμε να κατηγοριοποιήσουμε τις διαδικασίες μάθησης σε τρεις κλάσεις:

- **Μάθηση με επίβλεψη (supervised learning).** Στη μάθηση με επίβλεψη είναι απαραίτητη η παρουσία ενός εξωτερικού, ως προς το δίκτυο, παράγοντα που μπορούμε να ονομάσουμε «δάσκαλο». Στο σχήμα που ακολουθεί παρουσιάζεται το πώς επιδρά ο δάσκαλος στο δίκτυο και το περιβάλλον κατά τη διαδικασία της μάθησης. Ο δάσκαλος έχει την απαραίτητη γνώση για το περιβάλλον, που πρακτικά είναι ένα σύνολο από παραδείγματα εισόδου και την αντίστοιχη επιθυμητή έξοδο. Το Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ) δεν έχει καμιά γνώση για το περιβάλλον. Αν υποθέσουμε ότι παρουσιάζουμε στο δάσκαλο και το δίκτυο ένα πρότυπο από το περιβάλλον, τότε λόγο της προηγούμενης γνώσης του δασκάλου για το περιβάλλον, θα είναι σε θέση να παρέχει στο δίκτυο την επιθυμητή απάντηση – έξοδο. Στη συνέχεια οι παράμετροι του δικτύου προσαρμόζονται ανάλογα με το πρότυπο που χρησιμοποιείται για την εκπαίδευση και το σφάλμα του δικτύου (δηλαδή τη διαφορά μεταξύ της επιθυμητής εξόδου και της εξόδου που στην πράξη δίνει το δίκτυο). Η προσαρμογή αυτών των παραμέτρων, γίνεται επαναληπτικά, βήμα προς βήμα, με στόχο το δίκτυο να μπορεί να προσομοιάσει το δάσκαλο. Αν αυτό γίνει εφικτό, τότε μπορούμε να επιτρέψουμε στο δίκτυο να αλληλεπιδράσει με το περιβάλλον χωρίς την παρουσία του δασκάλου.

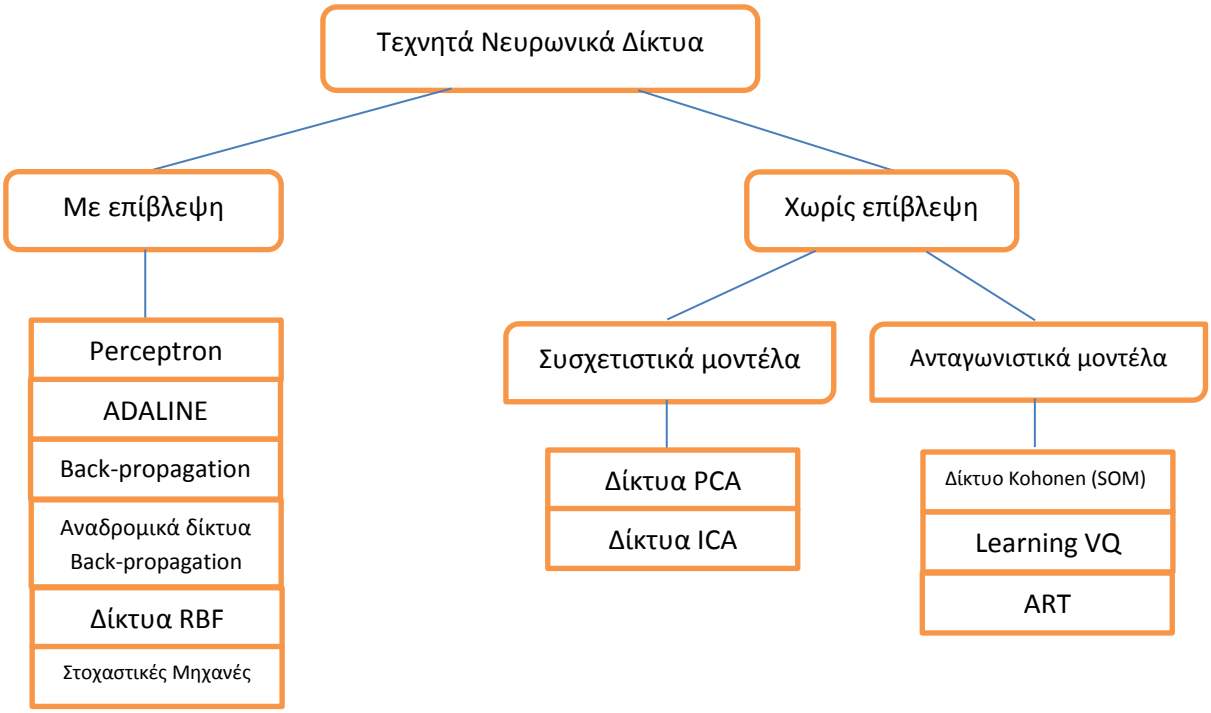


Σχήμα 6: Εκπαίδευση με επίβλεψη

- **Μάθηση χωρίς επίβλεψη (unsupervised learning) ή Μάθηση με αυτο-οργάνωση (self-organization).** Στην περίπτωση της μάθησης χωρίς επίβλεψη, δεν υπάρχει κάποιος εξωτερικός παράγοντας που επιβλέπει τη διαδικασία μάθησης. Αυτό σημαίνει ότι δεν υπάρχουν παραδείγματα της συνάρτησης που πρέπει να μάθει το δίκτυο. Υπάρχει όμως ένα μέτρο, ανεξάρτητο από το εκάστοτε έργο που πρέπει να φέρει εις πέρας το ΤΝΔ, που μετράει την ποιότητα της αναπαράστασης που πρέπει να μάθει το δίκτυο. Οι ελεύθερες παράμετροι του δικτύου βελτιστοποιούνται ως προς αυτό το μέτρο. Όταν το δίκτυο «μάθει» τις στατιστικές ιδιότητες των προτύπων που του δίνονται σαν είσοδος, αναπτύσσει την ικανότητα να δημιουργεί εσωτερικές αναπαραστάσεις για την κωδικοποίηση των χαρακτηριστικών των προτύπων. Αποκτά δηλαδή την ικανότητα να δημιουργεί νέες κλάσεις αυτόματα.

- **Ενισχυτική μάθηση (reinforcement learning)**. Αυτός ο τύπος μάθησης θεωρείται μια ενδιάμεση μορφή των δύο προηγούμενων τύπων. Εδώ, το σύστημα μάθησης αξιολογεί τις ενέργειές του ως καλές (επιβράβευση) ή κακές (αξιόποινη) βασισμένο σε κάποιες αντιδράσεις από το περιβάλλον και

ανάλογα προσαρμόζει τις παραμέτρους του δικτύου. Γενικά, η προσαρμογή των παραμέτρων συνεχίζεται μέχρι να επιτευχθεί μια κατάσταση ισορροπίας, στην οποία, αν εφαρμοστεί ο μηχανισμός μάθησης δεν θα υπάρξουν περαιτέρω αλλαγές στις παραμέτρους.



Σχήμα 7: Διάγραμμα Τεχνητών Νευρωνικών Δικτύων

2.2.5 Το δίκτυο Perceptron

Ο όρος «Perceptrons» επινοήθηκε από τον Frank RosenBlatt το 1962 και χρησιμοποιείται για να περιγράψει την σύνδεση των απλών νευρώνων σε ένα δίκτυο. Αυτά τα δίκτυα είναι απλοποιημένες μορφές του βιολογικού νευρικού συστήματος, όπου όμως κάποιες ιδιότητες του μεγαλοποιούνται και κάποιες αγνοούνται. Προς το παρόν θα επικεντρωθούμε στο Perceptron ενός στρώματος

(Single Layer Perceptrons), δηλαδή το δίκτυο Perceptron όπου δεν υπάρχουν κρυφά επίπεδα νευρώνων. Η λέξη δίκτυο εδώ χρησιμοποιείται καταχρηστικά, αφού δεν υπάρχουν περισσότεροι από ένα νευρώνες για να συνδεθούν μεταξύ τους, παρά μόνο υπάρχουν οι συνδέσεις μεταξύ των εισόδων του νευρώνα.

Γενικά, τα δεδομένα εισάγονται στο στρώμα εισόδου και το δίκτυο, στην συνέχεια τα επεξεργάζεται, πολλαπλασιάζοντάς τα με τα συναπτικά βάρη. Το αποτέλεσμα αυτού του πολλαπλασιασμού, μεταβάλλεται από το τελικό στρώμα, το στρώμα εξόδου, χρησιμοποιώντας μια συνάρτηση που καθορίζει καταπόσον ο κόμβος εξόδου «πυροβολά» ή όχι.

Η διαδικασία κατά την οποία το δίκτυο «εκπαιδεύεται», ο κανόνας εκπαίδευσης δηλαδή, περιλαμβάνει την εύρεση των σωστών τιμών των συναπτικών βαρών. Πρώτα όμως, ο πίνακας των βαρών αρχικοποιείται με τυχαίους αριθμούς μεταξύ -1 και +1. Έπειτα, όσο το δίκτυο «μαθαίνει», αυτές οι τιμές μεταβάλλονται μέχρι να αποφασιστεί ότι το δίκτυο έχει επιλύσει το πρόβλημα. Για να εκπαιδευτεί το δίκτυο χρησιμοποιούνται set δεδομένων ως πρότυπα εισόδου για τα οποία οι σωστές έξοδοι είναι γνωστές. Ξεκινώντας από τυχαία συναπτικά βάρη, ένα πρότυπο εισόδου παρουσιάζεται στο δίκτυο, το οποίο κάνει μια αρχική υπόθεση για το ποια πρέπει να είναι η σωστή έξοδος.

Κατά τη διάρκεια της φάσης εκπαίδευσης, η διαφορά μεταξύ της υπόθεσης που κάνει το δίκτυο και της σωστής τιμής της εξόδου αξιολογείται και τα συναπτικά βάρη αλλάζουν έτσι ώστε το λάθος να ελαχιστοποιηθεί.

Το απλό perceptron υλοποιείται όπως το βασικό μοντέλο που περιγράφεται πιο πάνω και έχει σαν συνάρτηση ενεργοποίησης μια απλή συνάρτηση κατώφλι:

$$f(x) = \begin{cases} 1, & x > t \\ 0, & διαφορετικά \end{cases}$$

όπου:

x είναι η έξοδος του νευρώνα και

t μια σταθερά-κατώφλι (threshold).

Αν συμβολίσουμε τα συναπτικά βάρη με τον πίνακα W_{ij} , όπου i είναι ο αριθμός των εισόδων και j ο αριθμός των εξόδων, και το διάνυσμα εισόδου με $\boldsymbol{I}$ τότε η έξοδος $\boldsymbol{O}$ του νευρώνα υπολογίζεται ως εξής:

$$O = f(IW_{ij})$$

Ο κανόνας εκπαίδευσης του απλού perceptron είναι σχετικά απλός. Ξεκινώντας από έναν πίνακα από τυχαία συναπτικά βάρη, παρουσιάζουμε στο δίκτυο ένα πρότυπο εκπαίδευσης και υπολογίζουμε την έξοδο του δικτύου όπως πιο πάνω. Καθορίζουμε έτσι μια **συνάρτηση λάθους E**:

$$E(O) = (T - O)$$

Όπου σ' αυτή την περίπτωση, το **T** είναι το επιθυμητό διάνυσμα εξόδου για την είσοδο του προτύπου εκπαίδευσης. Για να καθορίσουμε λοιπόν το πώς πρέπει να προσαρμοστούν τα συναπτικά βάρη, ώστε το δίκτυο να παράγει την επιθυμητή έξοδο για τη συγκεκριμένη είσοδο, θα πρέπει η συνάρτηση λάθους να ελαχιστοποιηθεί.

Στα νευρωνικά δίκτυα, ο στόχος είναι να εκτιμηθεί η επίδραση των συναπτικών βαρών στην ολική συνάρτηση λάθους. Συνδυάζοντας τα πιο πάνω έχουμε ότι η συνάρτηση λάθους εκφράζεται:

$$E(O) = (T - O) = T - f(IW_{ij})$$

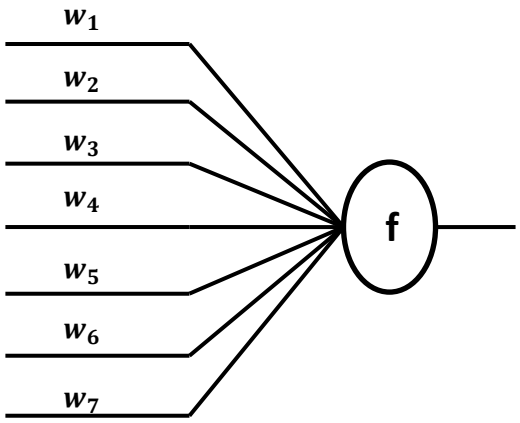
Στη συνέχεια παραγωγίζουμε τη συνάρτηση αυτή και εκτιμούμε έτσι τον πίνακα συναπτικών βαρών που την ελαχιστοποιεί. Η διαδικασία αυτή θα περιγραφεί εκτενέστερα στα επόμενα κεφάλαια και συγκεκριμένα στα δίκτυα Perceptron πολλαπλών στρωμάτων (MLP). Η συνάρτηση που ελαχιστοποιεί το λάθος στον αλγόριθμο του Perceptron ενός στρώματος είναι ιδιαίτερα απλή. Σε κάθε κόμβο εξόδου, υπολογίζεται το λάθος και προστίθεται στο συναπτικό βάρος που τροφοδοτεί αυτό τον κόμβο.

2.2.5.1 *Αλγόριθμος μάθησης*

Για να περιγράψουμε τον αλγόριθμο μάθησης του απλού Perceptron, ορίζουμε τις ακόλουθες μεταβλητές:

- $y = f(z)$ είναι η έξοδος του δικτύου για ένα διάνυσμα εισόδου z
- b είναι η μεροληψία, η οποία στο παράδειγμα παρακάτω θεωρείται 0
- $D = \{(x_1, d_1), \dots, (x_s, d_s)\}$ είναι το set δεδομένων εκπαίδευσης που αποτελείται από s δυάδες προτύπων:
  - x_j είναι ένα n-διάστατο διάνυσμα εισόδου
  - d_j είναι η επιθυμητή έξοδος του δικτύου για αυτή την είσοδο
- $x_{j,i}$ είναι η τιμή του i-οστού κόμβου του j-οστού διανύσματος εισόδου εκπαίδευσης ($x_{j,0} = 1$)
- w_i είναι η i-οστή τιμή του διανύσματος συναπτικών βαρών, που θα πολλαπλασιαστεί με την i-οστή τιμή του διανύσματος εισόδου
- Για τη χρονική εξάρτηση του w θεωρούμε $w_i(t)$ το i-συναπτικό βάρος τη χρονική στιγμή t
- α είναι μια σταθερά που ονομάζεται ρυθμός μάθησης

Μπορούμε επίσης να προσθέσουμε μια επιπλέον διάσταση, με δείκτη $n + 1$, τέτοια ώστε $x_{j,n+1} = 1$ και $w_{n+1} = b$ και με το τρόπο αυτό εισαγάγουμε τη μεροληψία στο δίκτυο.



Σχήμα 8: Απλό Perceptron

Ο αλγόριθμος εκτελεί τα πιο κάτω βήματα:

1) *Αρχικοποίησε τα συναπτικά βάρη και το κατώφλι στη συνάρτηση ενεργοποίησης. Τα βάρη μπορούν να αρχικοποιηθούν θέτοντας τα* $w_i(0) = 0$ *ή κάποια μικρή τυχαία τιμή.*
2) *Για κάθε πρότυπο j στο σετ εκπαίδευσης D, εκτέλεσε τα πιο κάτω βήματα για τη είσοδο* x_j *και την επιθυμητή έξοδο* d_j*:*
   - *Υπολόγισε την έξοδο:*

     $$y_j(t) = f\left[w(t) \cdot x_j\right] = f\left[w_0(t) + w_1(t)x_{j,1} + w_2(t)x_{j,2} + \cdots + w_n(t)x_{j,n}\right]$$

   - *Προσάρμοσε τα συναπτικά βάρη ως εξής:*

     $w_i(t+1) = w_i(t) + a\left(d_j - y_i(t)\right)x_{j,i}$, *για όλους τους κόμβους* $0 \le i \le n$

- *Το βήμα 2 επαναλαμβάνεται μέχρι το λάθος* $d_j - y_j(t)$ *να είναι μικρότερο από μια προκαθορισμένη τιμή* γ *ή όταν συμπληρωθεί ένας προκαθορισμένος αριθμός επαναλήψεων.*

Αυτή η διαδικασία αποτελεί το γενικό αλγόριθμο του perceptron. Μπορεί να δειχθεί ότι αυτή η τεχνική ελαχιστοποιεί τη συνάρτηση λάθους. Ο χρόνος που χρειάζεται για να καταλήξουμε σε λύση (ο χρόνος δηλαδή για να βρεθεί η ελάχιστη τιμή του λάθους) μπορεί να είναι απρόβλεπτος. Αυτό συμβαίνει γιατί, αν η συνάρτηση λάθους προσεγγίζεται με μεγάλα βήματα, η ελάχιστη τιμή ενδέχεται να βρεθεί πιο αργά. Αν γίνονται μικρότερα βήματα είναι πιο πιθανόν να χτυπήσουμε την ελάχιστη τιμή της συνάρτησης. Έτσι λοιπόν, για να ελέγξουμε το ρυθμό σύγκλισης και να μειώσουμε τον αριθμό των βημάτων που γίνονται, χρησιμοποιούμε την παράμετρο που ονομάζεται **ρυθμός μάθησης (learning rate)** $\boldsymbol{\alpha}$. Αυτή η παράμετρος ορίζεται στο διάστημα [0,1] και έτσι τα συναπτικά βάρη αλλάζουν με μικρότερα βήματα.

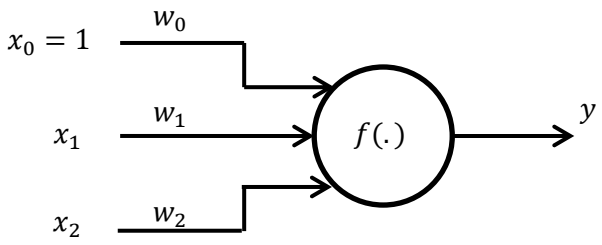
2.2.5.2 *Παράδειγμα του αλγόριθμου μάθησης στον απλό Perceptron*

Στο πιο κάτω παράδειγμα θα διακρίνουμε πώς ένας απλός perceptron με μια διαδικασία μάθησης μπορεί να χρησιμοποιηθεί ώστε να μιμείται τη λογική πύλη AND. Στο πίνακα που ακολουθεί φαίνονται όλες οι πιθανές είσοδοι και η αντίστοιχη επιθυμητή έξοδος που θέλουμε να παράγει το δίκτυο.

x_1	x_2	z
0	0	0
0	1	0
1	0	0
1	1	1

Πίνακας 3: Η λογική πύλη AND

Τα δεδομένα που φαίνονται στον πίνακα αποτελούν το set εκπαίδευσης και δείχνουν την τοπολογία που πρέπει να έχει το δίκτυο. Αυτό θα περιέχει τρεις κόμβους εισόδου (δύο για τα x_1, x_2 και ένα για το $x_0 = 1$) και έναν κόμβο εξόδου.



Σχήμα 9: Διάταξη νευρωνικού δικτύου που υλοποιεί τη λογική πύλη AND

Είσοδοι: x_0, x_1, x_2 όπου η είσοδος x_0 παραμένει σταθερή και ίση με 1

Κατώφλι συνάρτησης ενεργοποίησης: $t = 0.5$

Μεροληψία: $b = 0$

Ρυθμός μάθησης: $\alpha = 0.1$

Αθροιστική συνάρτηση: $s = x_0 w_0 + x_1 w_1 + x_2 w_2$

Έξοδος του δικτύου: $n = \begin{cases} 1, & s > t \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$

Συνάρτηση λάθους: $e = z - n$

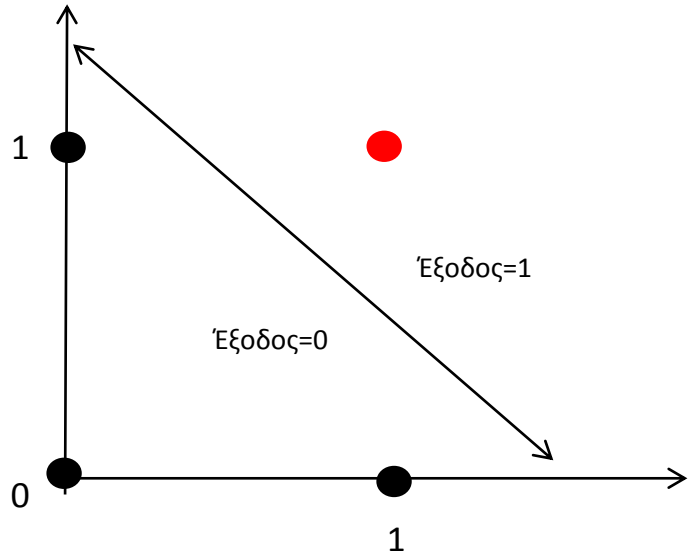
Τελική διόρθωση στα συναπτικά βάρη: $d = \alpha * e$

είσοδος				**αρχικά βάρη**			**έξοδος**		**λάθος**	**διόρθωση**	**τελικά βάρη**		
x_0	x_1	x_2	z	w_0	w_1	w_2	s	n	e	d	w_0	w_1	w_2
1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	1	+0.1	0.1	0.1	0.1
1	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0.1	0.1	0.1
1	0	1	0	0.1	0.1	0.1	0.2	0	0	0	0.1	0.1	0.1
1	1	0	0	0.1	0.1	0.1	0.2	0	0	0	0.1	0.1	0.1
1	1	1	1	0.1	0.1	0.1	0.3	0	1	+0.1	0.2	0.2	0.2
1	0	0	0	0.2	0.2	0.2	0.2	0	0	0	0.2	0.2	0.2
1	0	1	0	0.2	0.2	0.2	0.4	0	0	0	0.2	0.2	0.2
1	1	0	0	0.2	0.2	0.2	0.4	0	0	0	0.2	0.2	0.2
1	1	1	1	0.2	0.2	0.2	0.6	1	0	0	**0.2**	**0.2**	**0.2**

Πίνακας 4: Βήματα αλγορίθμου μάθησης του δικτύου για τη λογική πύλη AND

Ο αλγόριθμος τερματίζεται, αφού μετά την τρίτη είσοδο των δεδομένων εκπαίδευσης, το λάθος μηδενίζεται. Τελικά, τα συναπτικά βάρη με τα οποία το δίκτυο μας προσομοιάζει τη λογική πύλη AND είναι: $w_0, w_1, w_2 = 0.2$.

Το παραπάνω πρόβλημα μπορούμε να το εξετάσουμε και γραφικά. Στο Γράφημα 1.2.8.1 φαίνεται η χωρική διάταξη των δεδομένων εισόδου. Παρατηρούμε ότι είναι δυνατόν να σχεδιάσουμε μια ευθεία γραμμή μεταξύ των συντεταγμένων των τιμών εισόδου που έχουν σαν έξοδο τη τιμή 1 και αυτών των οποίων απαιτείται έξοδος ίση με 0. Τα προβλήματα που έχουν την ιδιότητα αυτή ονομάζονται **γραμμικά διαχωρίσιμα**.



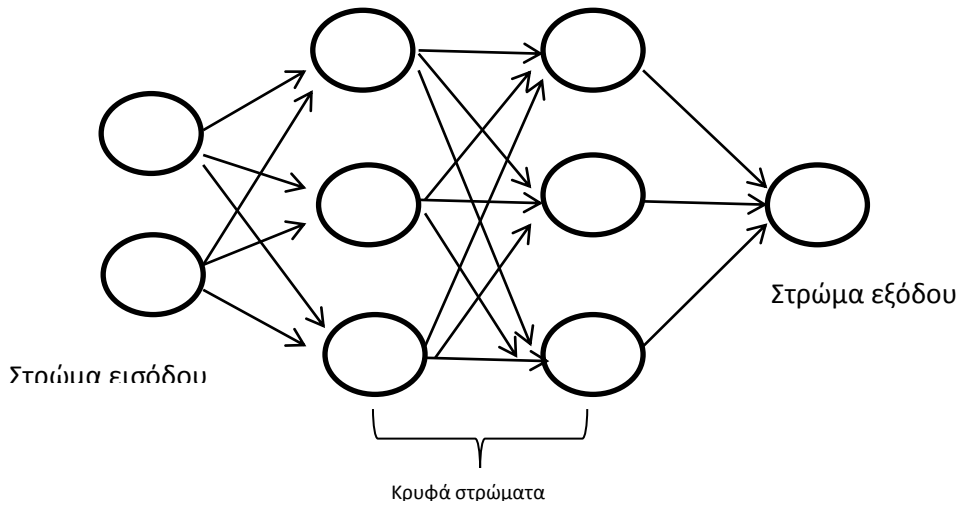
Σχήμα 10: Χωρική διάταξη δεδομένων εισόδου για τη λογική πύλη AND

Το απλό δίκτυο perceptron με ένα νευρώνα και μια συνάρτηση ενεργοποίησης/κατώφλι, μπορεί μόνο να λύσει προβλήματα που είναι γραμμικά διαχωρίσιμα. Τα πιο πολλά προβλήματα τα οποία μπορούν να αντιμετωπιστούν με τη βοήθεια των νευρωνικών δικτύων δεν είναι γραμμικά διαχωρίσιμα και γι' αυτό το απλό δίκτυο perceptron δεν μπορεί φυσικά να αντεπεξέλθει στις απαιτήσεις για τις οποίες η τεχνολογία των νευρωνικών δικτύων αναπτύχθηκε.

2.2.6 Perceptron πολλών στρωμάτων(Multilayer Perceptrons-MLP)

Γύρω στο 1986 αναπτύχθηκαν τα **perceptron πολλών στρωμάτων** ή αλλιώς **multilayer perceptrons (MLP)** για την επίλυση μη γραμμικών διαχωρίσιμων προβλημάτων. Ένα MLP είναι ένα δίκτυο αποτελούμενο από πολλούς νευρώνες, κατανεμημένους σε στρώματα (layers). Τα στρώματα αυτά χωρίζονται ως εξής:

- Το **στρώμα εισόδου (input layer)**, από όπου εισέρχονται τα δεδομένα στο δίκτυο. Ο αριθμός των νευρώνων από τους οποίους αποτελείται το στρώμα αυτό, εξαρτάται από τον αριθμό των εισόδων που θέλουμε να πάρει το δίκτυο.
- Ένα ή περισσότερα **κρυφά στρώματα (hidden layers)**. Αυτά τα στρώματα, παρεμβάλλονται μεταξύ του στρώματος εισόδου και εξόδου και ο αριθμός τους ποικίλει. Η λειτουργία των κρυφών στρωμάτων είναι να κωδικοποιούν τις εισόδους και να καθορίζουν τις εξόδους του δικτύου. Έχει αποδειχθεί ότι ένα MLP δίκτυο μπορεί να προσεγγίσει οποιαδήποτε συνάρτηση που συνδέει τις εισόδους της με τις εξόδους της, δεδομένου ότι μια τέτοια συνάρτηση υπάρχει.
- Το **στρώμα εξόδου (output layer)** στο οποίο παρουσιάζεται η έξοδος του δικτύου. Ο αριθμός των νευρώνων στο στρώμα αυτό, εξαρτάται από το πρόβλημα που θέλουμε να μάθει το κάθε δίκτυο.



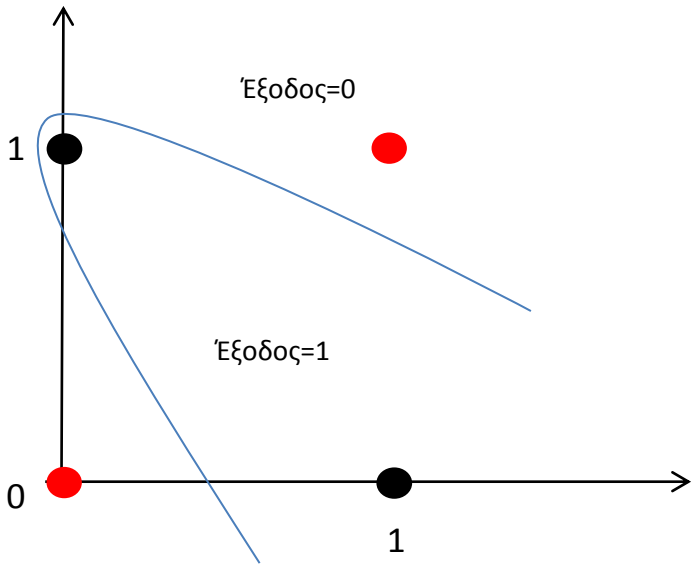
Σχήμα 11: MLP δίκτυο

2.2.6.1 *Παράδειγμα MLP δικτύου*

Η λογική πύλη XOR είναι ένα μη γραμμικά διαχωρίσιμο πρόβλημα, όπως βλέπουμε και στο γρφημα που ακολουθεί. Θα δούμε πώς μπορούμε να υλοποιήσουμε τη λογική πύλη XOR με τη βοήθεια ενός MLP δικτύου.

x_1	x_2	z
0	0	0
0	1	1
1	0	1
1	1	0

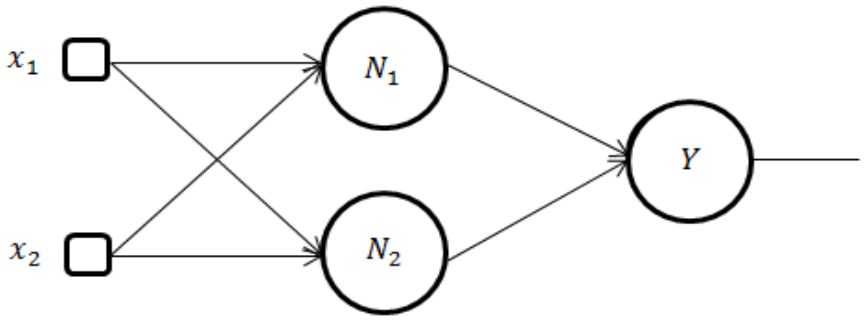
Πίνακας 5: Λογική πύλη XOR



Γράφημα 4: Χωρική διάταξη δεδομένων εισόδου για τη λογική πύλη XOR

Το γράφημα, που δείχνει τη διάταξη των προτύπων, υποδεικνύει πως για το διαχωρισμό αυτών είναι απαραίτητη μια καμπύλη και όχι ευθεία γραμμή. Γι' αυτό και ο αλγόριθμος του απλού perceptron δεν μπορεί να εφαρμοστεί αποτελεσματικά

στο συγκεκριμένο πρόβλημα. Θεωρούμε λοιπόν τη διάταξη ενός MLP δικτύου όπως φαίνεται στο σχήμα.



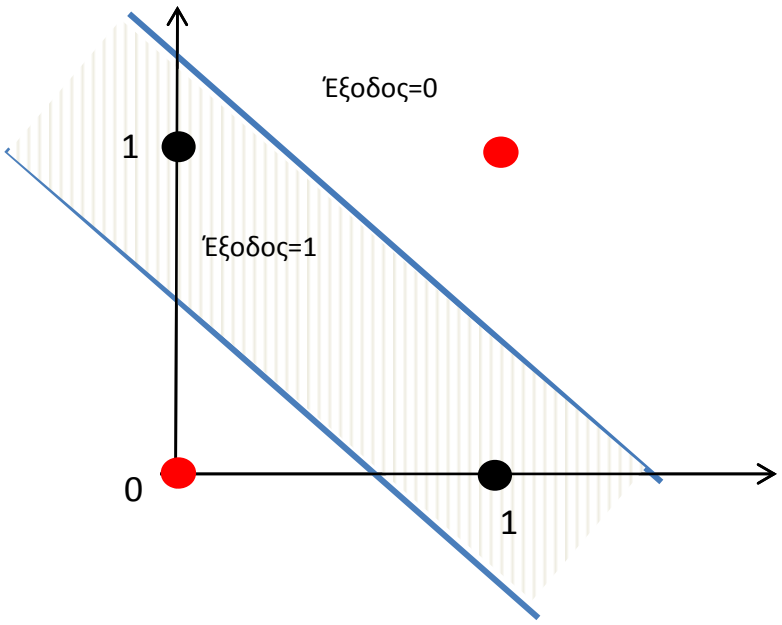
Σχήμα 12: Διάταξη MLP δικτύου που υλοποιεί τη λογική πύλη XOR

Οπότε, εφαρμόζοντας τον αλγόριθμο του perceptron χρησιμοποιώντας μια κλασική βηματική συνάρτηση μπορούμε να επιτύχουμε τις εξόδους που φαίνονται στον Πίνακα 6 για τους τρεις νευρώνες του δικτύου.

Είσοδοι		N_1	N_2	Y
x_1	x_2			
0	0	0	0	0
0	1	1	0	1
1	0	1	0	1
1	1	1	1	0

Πίνακας 6: Έξοδοι των νευρώνων

Από τον πίνακα μπορούμε να δούμε, ότι οι κρυφοί νευρώνες N_1, N_2 υλοποιούν τις λογικές πύλες OR και AND αντίστοιχα και ο νευρώνας εξόδου Y συνδυάζει τις εξόδους τους δίνοντας το επιθυμητό αποτέλεσμα του δικτύου. Σχηματικά, το πιο πάνω φαίνεται με το Γράφημα 5.



Γράφημα 5: Χωρική διάταξη των εισόδων για τη λογική πύλη XOR και οι έξοδοι του MLP δικτύου

Στα προβλήματα που απαιτούνται περισσότερες διαχωριστικές γραμμές, χρησιμοποιούνται και περισσότεροι κρυφοί νευρώνες. Συνδυάζοντας τις ευθείες, μπορούμε να πάρουμε μια μεγάλη ποικιλία περιοχών τις οποίες τα δίκτυα MLP μπορούν να διαχωρίσουν.

2.2.6.2 *Εφαρμογές MLP*

Τα δίκτυα MLP είναι ο πιο συνηθισμένος αλγόριθμος των νευρωνικών δικτύων για προβλήματα ταξινόμησης προτύπων και εκτίμησης συναρτήσεων. Κάποιοι τομείς που εφαρμόζονται τα MLP δίκτυα είναι :

- Ιατρική (διάγνωση ασθενειών, βιοπληροφορικής κ.λ.π)
- Ηλεκτρονικούς υπολογιστές (ασφάλεια, ηλεκτρονικά παιχνίδια, αναγνώριση προτύπων κ.τ.λ)
- Βιομηχανία (βιομηχανικός έλεγχος, ρομποτική κ.τ.λ)
- Εμπόριο και Οικονομία (πρόβλεψη οικονομικών μεγεθών, εκτίμηση αξίας ακινήτων κ.τ.λ)

2.2.7 Δίκτυα Συναρτήσεων Βάσης Ακτινικού Τύπου (RBFN)

Τα δίκτυα συναρτήσεων βάσης ακτινικού τύπου (RBFN) ανήκουν στη κατηγορία των τεχνητών νευρωνικών δικτύων από τα τέλη της δεκαετίας του 80. Τα RBFN δίκτυα αποτελούνται από τρία μόνο στρώματα και έχουν τη δυνατότητα να προσεγγίζουν οποιαδήποτε συνεχή συνάρτηση.

2.2.7.1 *Συναρτήσεις Βάσης Ακτινικού Τύπου*

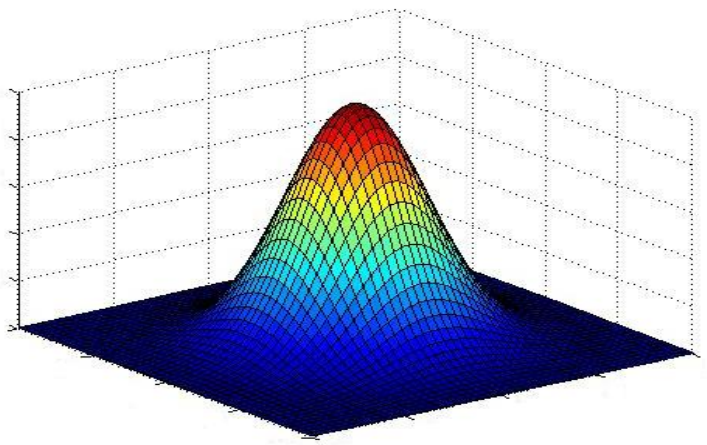
Συνάρτηση βάσης ακτινικού τύπου(radial function) είναι μια συνάρτηση $\varphi(x)$ για την οποία υπάρχει κάποιο διάνυσμα c το οποίο ονομάζουμε ***κέντρο*** (center or centroid) και η τιμή της συνάρτησης εξαρτάται μόνο από την απόσταση του x από το κέντρο αυτό (Ρίζος, 1996). Δηλαδή :

$$\varphi(x) = \varphi(\|x - c\|)$$

όπου νόρμα, η ευκλείδια απόσταση.

Κάθε κρυφός νευρώνας, υλοποιεί διαφορετική συνάρτηση με δικό του κέντρο c_i και εύρος συνάρτησης σ_i . Η πιο συχνά χρησιμοποιούμενη συνάρτηση βάσης ακτινικού τύπου στα δίκτυα RBF είναι η **συνάρτηση Gauss**:

$$\varphi(x) = e^{-\frac{\|x-c\|^2}{\sigma^2}}$$



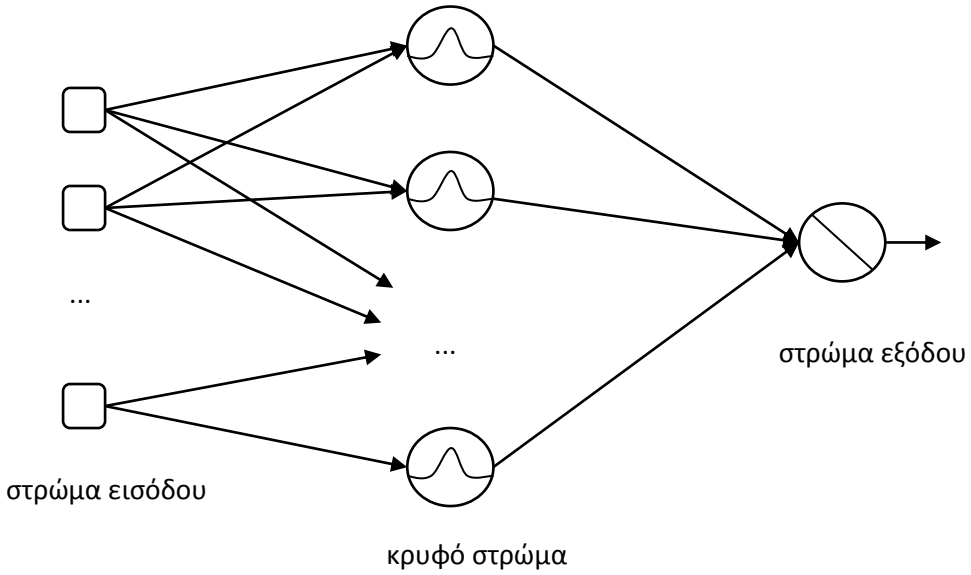
Γράφημα 6: Η συνάρτηση Gauss

Η συνάρτηση Gauss θεωρείται «τοπική» διότι για τιμές εισόδου σε ένα νευρώνα, που είναι μακριά από το κέντρο του, η αλλαγή που γίνεται στις παραμέτρους του νευρώνα είναι πολύ μικρές.

2.2.7.2 *Αρχιτεκτονική Δικτύων RBF*

Στα δίκτυα RBF οι συναρτήσεις ακτινικού τύπου, ενσωματώνονται σε ένα εμπρόσθιας τροφοδότησης νευρωνικό δίκτυο δύο στρωμάτων, το στρώμα *εισόδου* και το στρώμα *εξόδου*. Ανάμεσα στα στρώματα αυτά, περιέχεται **ένα μόνο κρυφό στρώμα**, το οποίο αποτελείται από νευρώνες που υλοποιούν τις ακτινικού τύπου συναρτήσεις.

Διακρίνουμε την τοπολογία των δικτύων συναρτήσεων ακτινικής βάσης, που χρησιμοποιούνται για την αναγνώριση προτύπων στο παρακάτω σχήμα:



Σχήμα 13: Τοπολογία RBFN

Οι είσοδοι του δικτύου, προωθούνται από το στρώμα εισόδου στους νευρώνες του κρυφού στρώματος, οι οποίοι υλοποιούν συναρτήσεις ακτινικής βάσης. Δεν συναντούμε συναπτικά βάρη στις συνδέσεις του στρώματος εισόδου με το κρυφό στρώμα. Έπειτα, τα σήματα στις εξόδους των νευρώνων του κρυφού στρώματος

συνδυάζονται με συναπτικά βάρη, μέσω μιας γραμμικής συνάρτησης ενεργοποίησης στο στρώμα εξόδου.

Τα δίκτυα RBF, υλοποιούν μια μη γραμμική χαρτογράφηση εισόδων-εξόδων, συνδυάζοντας τις μη γραμμικές εξόδους των νευρώνων στο κρυφό στρώμα, σύμφωνα με την εξίσωση (Tao, 1993):

$$y = \sum_{i=1}^{m} w_i \varphi_i(x)$$

όπου :

x είναι το διάνυσμα εισόδου,

y το διάνυσμα εξόδου,

w_i είναι τα συναπτικά βάρη με τα οποία συνδυάζονται γραμμικά οι έξοδοι.

Οι παράμετροι που εκπαιδεύονται στο κρυφό στρώμα, είναι τα κέντρα c_i και τα εύρη σ_i των συναρτήσεων φ_i, ενώ στο στρώμα εξόδου τα συναπτικά βάρη w_i. Τα δύο προβλήματα εκπαίδευσης (των παραμέτρων c_i, σ_i και των συναπτικών βαρών) αντιμετωπίζονται ξεχωριστά.

Η βασική φιλοσοφία της ανάπτυξης των δικτύων RBF και ταυτόχρονα σημαντικό πλεονέκτημα των RBF έναντι των MLP, είναι το γεγονός ότι δεν γενικεύουν σε μεγάλο βαθμό την κατηγοριοποίηση των προτύπων σε περίπτωση που δέχονται σαν είσοδο πρότυπα εκτός κάποιας κατηγορίας. Σ' αντίθεση ένα MLP δίκτυο, μπορεί να εκπαιδευτεί για να έχει μεγάλη ακρίβεια στην κατηγοριοποίηση από ένα set γνωστών κατηγοριών, αλλά την ίδια στιγμή κατηγοριοποιεί κάθε πρότυπο που δεν ανήκει σε κάποια κατηγορία εκπαίδευσης. Το γεγονός αυτό δημιουργεί ποικίλα προβλήματα στις πραγματικές εφαρμογές.

Αντίθετα, ένα RBF δίκτυο με kernel τη συνάρτηση Gauss, εκπαιδεύεται βάση της συνάρτησης πυκνότητας πιθανότητας των προτύπων, αντί να διαιρεί το χώρο όπως τα MLP δίκτυα. Συνεπώς, όταν αντιμετωπίζουν ένα πρότυπο εκτός κάποιας κατηγορίας, πιθανότατα να το εντάξουν σε κάποια κατηγορία αγνώστων. Έτσι,

μπορούμε να καταλήξουμε στο συμπέρασμα ότι τα RBF δίκτυα έχουν λιγότερη ανοχή στις λάθος τιμές, από τα δίκτυα MLP.

2.2.7.3 *Εκπαίδευση του Δικτύου*

Ένας γενικός αλγόριθμος κατασκευής ενός δικτύου RBF είναι ο εξής:

1. Χρησιμοποίησε τα εισερχόμενα δεδομένα εκπαίδευσης ως κέντρα των RBF kernels.
2. Έστω χρησιμοποιούμε σαν kernel την συνάρτηση Gauss. Προσδιόρισε τα εύρη των συναρτήσεων σ_i , είτε με δοκιμές στο set δεδομένων για καλύτερα αποτελέσματα, είτε χρησιμοποιώντας κάποιο είδος εσωτερικής απόστασης πυρήνα.
3. Βρες τα βέλτιστα συναπτικά βάρη w_i , επιλύοντας ένα πρόβλημα ελαχιστοποίησης του μέσου τετραγωνικού σφάλματος.

Όταν όλα τα δεδομένα εκπαίδευσης που εισάγονται, «απομνημονευθούν» στα kernels του δικτύου και το πρόβλημα ελαχιστοποίησης του μέσου τετραγωνικού σφάλματος καθοριστεί, το RBF δίκτυο μπορεί να υλοποιήσει τη χαρτογράφηση όπως και στα πρότυπα που του παρουσιάστηκαν.

2.2.7.4 *Εφαρμογές RBFN*

Οι ιδιότητες των RBFN όπως η προσέγγιση συναρτήσεων, η παρεμβολή, η ομαδοποίηση και ο εντοπισμός τα κάνουν πολύ δημοφιλή σε πολλές περιοχές όπως:

- Τηλεπικοινωνίες
- Αναγνώριση σήματος και εικόνας
- Μοντελοποίηση χαοτικών χρονοσειρών
- Αναγνώριση ομιλίας

- Μοντελοποίηση τρισδιάστατων αντικειμένων
- Μηχανικός έλεγχος
- Αποκατάσταση εικόνας
- Συγχώνευση δεδομένων

2.3 Λογιστική Παλινδρόμηση

Το μοντέλο της Λογιστικής παλινδρόμησης (logistic regression) αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων. Άρχισε να χρησιμοποιείται ευρέως κατά την δεκαετία του $50'$, κυρίως με εφαρμογές στη βιοστατιστική. Είναι μια μέθοδος στατιστικής ανάλυσης που χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής.

Η Λογιστική παλινδρόμηση είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης.

Σε πολλές εφαρμογές η εξαρτημένη μεταβλητή παίρνει δυο μόνο τιμές, οι οποίες αντιστοιχούν σε δύο ενδεχόμενα. Για παράδειγμα , το αν ο ασθενής ζει ή απεβίωσε, το αν ο άνεργος βρίσκει εργασία ή όχι, το αν ραγίζει ή αντέχει το δοκάρι. Οι τιμές της μεταβλητής αποτελούν μια αυθαίρετη κωδικοποίηση των δύο ενδεχομένων, συνήθως 0 και 1.

Εάν ορίσουμε την τιμή $y = 1$ σαν «επιτυχία» και την τιμή $y = 0$ σαν «αποτυχία» , τότε η y είναι τ.μ της κατανομής Bernoulli, δηλαδή $y \sim B(p)$, με μέση τιμή $E(y) = p$ και διασπορά $V(y) = p(1-p)$.

Γενικεύοντας σε μια σειρά από n επαναλήψεις (δηλαδή πραγματοποιήσεων των ενδεχομένων), ορίζουμε την τ.μ:

$$y = αριθμός\ επιτυχιών\ σε\ n\ δοκιμές$$

Υπό την υπόθεση ότι η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή και οι δοκιμές είναι ανεξάρτητες μεταξύ τους, τότε ισχύει η Διωνυμική (binomial) κατανομή :

$$y \sim b(n, p)$$

Με συνάρτηση πυκνότητας

$$f(y) = \binom{n}{p} p^y (1-p)^{n-y}\ , y = 0,1,2, \dots, n$$

όπου:

p η πιθανότητα επιτυχίας η οποία είναι παράμετρος της κατανομής.

Η Διωνυμική κατανομή αποτελεί τη βασική κατανομή για την περιγραφή και ανάλυση μιας μεταβλητής αυτής της φύσης. Η μέση τιμή της y είναι ίση με $E(y) = np$ και η διασπορά με $V(y) = np(1-p)$. Στην ειδική περίπτωση που $n = 1$ μιλάμε για *δυαδικά δεδομένα ,* αλλιώς για *διωνυμικά δεδομένα.*

Σε πολλές περιπτώσεις η τ.μ y ενδέχεται να εξαρτάται από κάποιες επεξηγηματικές μεταβλητές. Η εξάρτηση της y από τις επεξηγηματικές μεταβλητές x (ανεξάρτητες μεταβλητές ή συμμεταβλητές) εισάγεται μέσω της εξάρτησης της πιθανότητας επιτυχίας p από τις x (π.χ η πιθανότητα να μείνει κάποιος άνεργος εξαρτάται από το φύλο, την ηλικία, το μορφωτικό επίπεδο κ.α). Πιο συγκεκριμένα, κατασκευάζεται το αποκαλούμενο *μοντέλο λογιστικής παλινδρόμησης*, το οποίο είναι ένα γενικευμένο γραμμικό μοντέλο και εκφράζεται μέσω της σχέσης:

$$n_x = g\big(E(y_x)\big) = g(\mu_x) = \boldsymbol{x'\beta}$$

με την ακόλουθη δομή:

1. $y_x \sim b(n_x, \mu_x)\ (n_x > 1\ ,\ διωνυμικά\ δεδομένα)$

ή $y_x \sim B(n_x, \mu_x)\ (n_x = 1\ , δυαδικά\ δεδομένα)$

2. $n_x = g(\mu_x) = \ln \frac{\mu_x}{n_{x-\mu_x}} = \ln \frac{p_x}{1-p_x} = logit(p_x) = \boldsymbol{x'\beta}$ $(συνάρτηση\ Logit)$
3. Ανεξαρτησία μεταξύ των παρατηρήσεων y_x **,**

όπου:

n_x είναι ο αριθμός των επαναλήψεων της τιμής του διανύσματος $\boldsymbol{x}$ των επεξηγηματικών μεταβλητών.

Αντιστρέφοντας τη συνάρτηση σύνδεσης προκύπτει:

$$p_x = e^{n_x}/(1 + e^{n_x})$$

για την οποία ισχύει ο περιορισμός $0 < p_x < 1$.

Για κάθε παρατήρηση i το μοντέλο γράφεται ως:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}\ , i = 1, \dots, n$$

όπου:

$$p_i = p_{x_i} = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} = \frac{1}{1 + e^{-\boldsymbol{x_i'\beta}}} \qquad (1)$$

η πιθανότητα «επιτυχίας»

$$x_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

είναι ο linear predictor.

$$E(y_i) = n_i p_i = n_i \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}}$$

2.3.1 Εκτίμηση παραμέτρων με τη μέθοδο μέγιστης πιθανοφάνειας (maximum likelihood)

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 δώσει μια συγκεκριμένη δοσολογία φαρμάκου). Το μοντέλο μας βάση της εξίσωσης $(\mathbf{1})$, γράφεται στη μορφή:

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

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

$$var(y_i) = n_i P(x_i)[1 - P(x_i)]$$

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 $\boldsymbol{\beta}$, οπότε δεν μπορεί να χρησιμοποιηθεί. Επομένως, η log πιθανοφάνεια για το λογιστικό μοντέλο παλινδρόμησης δίνεται από τον τύπο:

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

Είναι εφικτό 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 γράφεται ως:

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

Σαν αποτέλεσμα η log πιθανοφάνεια της εξίσωσης (2) γράφεται ως:

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

Στη συνέχεια η εξίσωση (3) πρέπει να μεγιστοποιηθεί ως προς τον όρο β_j. Σε μορφή πινάκων , η εξίσωση (3) γράφεται ως:

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

όπου:

X είναι ο κλασσικός πίνακας του μοντέλου που συναντάμε και στη γραμμική παλινδρόμηση και

y το διάνυσμα της απόκρισης.

Παραγωγίζουμε τώρα την εξίσωση (4) ως προς β:

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

Γνωρίζοντας ότι ισχύει:

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

Προκύπτει ότι:

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

Εφόσον, ο όρος $n_i P(\mathbf{x_i})$ αποτελεί τον μέσο της διωνυμικής τυχαίας μεταβλητής το δεξί μέλος, της παραπάνω σχέσης γράφεται σε μορφή πινάκων ως $X'(y - \mu)$ όπου:

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

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

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0} \qquad (5)$$

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

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

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

Όπου $\mu_i = n_i P(\boldsymbol{x_i})$ και ${\sigma_i}^2$ είναι η διωνυμική διακύμανση στο i-οστό σημείο δεδομένων με:

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

Ελαχιστοποιούμε το S:

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

Η διακύμανση ${\sigma_i}^2$είναι σταθερή, επομένως παραγωγίζουμε μόνο τον αριθμητή του S και παίρνουμε:

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

Ισχύει ότι:

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

Επομένως , η λύση που παίρνουμε από την ελαχιστοποίηση του σταθμισμένου αθροίσματος τετραγώνων των υπολοίπων με σταθερό $\sigma_i{}^2$είναι:

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

η οποία είναι παρόμοια με την εξίσωση $\mathrm{X}'(\mathrm{y} - \mu) = \mathbf{0}$, εξίσωση (5). Άρα, μια επαναληπτική μέθοδος όπως η παραπάνω μπορεί να χρησιμοποιηθεί για να προσδιοριστούν οι αριθμητικές τιμές των $\mathrm{b}_0, \mathrm{b}_1, \ldots, \mathrm{b}_\mathrm{k}$ δηλαδή των εκτιμητών μέγιστης πιθανοφάνειας.

2.3.2 Άλλες μορφές στατιστικής συμπερασματολογίας για τις οποίες γίνεται χρήση λογιστικής παλινδρόμησης

Όπως έχουμε δει 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 εξετάζεται η σημασία της κάθε μεταβλητής. Σε πολλές περιπτώσεις τα δεδομένα που έχουμε δεν είναι ομαδοποιημένα δηλαδή $n_i = 1$. Όταν όμως 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 ασθενή ανάλογα με την δοσολογία που του χορηγείται. Σε μια τέτοια περίπτωση, ισχύει ότι $k = 1$ και $p = 2$ και το μοντέλο παίρνει τη μορφή:

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

Σε αρκετές περιπτώσεις τα όρια εμπιστοσύνης για τα β_0 και β_1 καθώς επίσης και τα όρια εμπιστοσύνης για τους συντελεστές της $P(x_i)$ είναι σημαντικά για τους ερευνητές. Το ενδιαφέρον για την μελέτη του κάθε συντελεστή παλινδρόμησης ξεχωριστά πηγάζει από την ανάγκη να προσδιορίσουμε τους λόγους πιθανοτήτων (odd ratios).

Για παράδειγμα, αρκετά συχνά, η ανεξάρτητη μεταβλητή x είναι κατηγορική και ότι μια ομάδα από τα πειραματικά μας υποκείμενα χωρίζονται σε αυτά που τους χορηγήθηκε μεγάλη ποσότητα βιταμίνης C $(x = 0)$ και σε αυτά που δεν τους χορηγήθηκε τίποτα $x = 1$. Έτσι, η απόκριση μπορεί να είναι η μόλυνση του αναπνευστικού συστήματος ή όχι , παίρνει τις τιμές $y = 1$ και $y = 0$ αντίστοιχα.

Η ιδέα προσδιορισμού του λόγου πιθανοτήτων 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 $logit[P(\boldsymbol{x_i})]$ δίνεται από τη σχέση:

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

και μέσω του μετασχηματισμού του P, γραμμικοποιείται η λογιστική συνάρτηση. Αν χρησιμοποιήσουμε την εξίσωση (7) για το μοντέλο της εξίσωσης (6), τότε στο παραπάνω παράδειγμα προκύπτει η σχέση:

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

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:

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

Χρησιμοποιώντας την παραπάνω ερμηνεία του β_0 μπορούμε να βρούμε την αντίστοιχη *odd ratio* ερμηνεία του β_1. Για την ομάδα υποκειμένων που δε 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 τον 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.

2.3.3 Ιδιότητες της διασποράς των εκτιμητών μέγιστης πιθανοφάνειας στη λογιστική παλινδρόμηση

Είναι ευρέως γνωστό ότι οι εκτιμητές μέγιστης πιθανοφάνειας παρουσιάζουν ασυμπτωτικές ιδιότητες στη διακύμανση και τη συνδιακύμανση, στον πίνακα πληροφορίας. Στην περίπτωση ενός γραμμικού μοντέλου με κανονικά, ανεξάρτητα και ομοιόμορφα κατανεμημένα (iid) σφάλματα, ο πίνακας πληροφορίας για τους εκτιμώμενους συντελεστές παλινδρόμησης εκφράζεται από τον τύπο:

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

Όπου, σ^2 είναι η διακύμανση του σφάλματος. Σαν αποτέλεσμα, σε αυτήν την περίπτωση, ο πίνακας διακύμανσης – συνδιακύμανσης (variance-covariance matrix) των εκτιμώμενων συντελεστών είναι:

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

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. Μπορούμε να υπολογίσουμε τον πίνακα πληροφορίας με διάφορες μεθόδους, όπως με τη βοήθεια της εξίσωσης (5):

$$\boldsymbol{I}(\boldsymbol{b}) = var(score)$$

$$= var[\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu})]$$

Όπου, var συμβολίζουμε τον πίνακα διακύμανσης-συνδιακύμανσης (variance-covariance matrix).

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

$$var[\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{\mu})] = \boldsymbol{X}'var[(\boldsymbol{y}-\boldsymbol{\mu})]\boldsymbol{X}$$

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

$$\boldsymbol{V} = diag\{{\sigma_1}^2, {\sigma_2}^2, \dots, {\sigma_m}^2\}$$

Και

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

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

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

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

2.3.4 Συμπερασματολογία με χρήση της μεθόδου Wald στη λογιστική παλινδρόμηση

Η πρώτη εφαρμογή της μεθόδου Wald πραγματεύεται 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 $\boldsymbol{\chi_i}'\boldsymbol{\beta}$ 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 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 $\hat{\sigma} b_j$. Ο έλεγχος που διαξάγουμε, είναι ο συνηθισμένος μονομερής ή διμερής έλεγχος (one or two sided test). Για τον υπολογισμό των τιμών της ${\chi_1}^2$ και της $p - τιμής$ για κάθε συντελεστή του απαιτούμενου μοντέλου γίνεται χρήση διαφόρων στατιστικών πακέτων.

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 ύπαρξης του linear predictor $\boldsymbol{\chi_i}'\boldsymbol{\beta}$ στο λογιστικό μοντέλο ακολουθείται μια εναλακτική διαδικασία υπολογισμού των διαστημάτων εμπιστοσύνης.

Στη λογιστική παλινδρόμηση πρέπει να έχουμε υπόψην ότι η μέση απόκριση στο $\boldsymbol{x} = \boldsymbol{x_i}$ δίνεται από τον τύπο $\frac{1}{1+e^{-\boldsymbol{x_i}'\boldsymbol{\beta}}}$ και άρα είναι πιθανότητα. Για παράδειγμα ένας μηχανικός πιθανόν να απαιτεί ένα 95% διάστημα εμπιστοσύνης για την πιθανότητα «ελαττωματικού» προϊόντος σε μια βιομηχανία όπου οι συνθήκες παραγωγής ορίζονται ως $\boldsymbol{x} = \boldsymbol{x_i}$.

Η σημειακή εκτίμηση της πιθανότητας δίνεται από το $\widehat{y_i} = \hat{P}(\boldsymbol{x_i})$

Στο λογιστικό μοντέλο $P = \frac{1}{1+e^{-x_i\beta}}$ το P είναι μια μονότονη εξίσωση του $\boldsymbol{x}'$. Μπορούμε να ορίσουμε ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης στο P, χρησιμοποιώντας ένα διάστημα εμπιστοσύνης στο $\boldsymbol{x}'\boldsymbol{\beta}$**.** Προφανώς, ο linear predictor περιέχει όρους που είναι γραμμικοί στο $\boldsymbol{\beta}$ και μπορούμε να εκμεταλλευτούμε το γεγονός ότι ο $\boldsymbol{b}$ (εκτιμητής μέγιστης πιθανοφάνειας του **β**) είναι ασυμπτωτικά κανονικός. Συνεπώς, ένα άνω διάστημα εμπιστοσύνης για το $\boldsymbol{x}'\boldsymbol{\beta}$, παράγει ένα άνω διάστημα για το P.

Ασυμπτωτικά ισχύει ότι:

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

Έτσι, το διάστημα εμπιστοσύνης για το $\boldsymbol{x}'\boldsymbol{\beta}$ δίνεται από τον τύπο:

$$\boldsymbol{x}'\boldsymbol{b} \pm z_{\frac{\alpha}{2}}\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}}$$

Σε διάφορα παραδείγματα από τον τομέα της βιολογίας και της χημείας, όπου τα δεδομένα είναι ομαδοποιημένα και η $i-$ οστή παρατηρούμενη απόκριση y_i είναι διωνυμική με παραμέτρους $P(\boldsymbol{x_i})$ και n_i είναι πολύ ενδιαφέρον να υπολογίσουμε το διάστημα πρόβλεψης για το y_i. Για το σκοπό αυτό είναι απαραίτητη μια έκφραση για τη διακύμανση του:

$$\hat{P}(\boldsymbol{x_i}) = \frac{1}{1+e^{-\boldsymbol{x_i}'\boldsymbol{b}}} \qquad i = 1,2,\dots,m$$

Μπορούμε επίσης να χρησιμοποιήσουμε την μέθοδο Δέλτα, όπου προκύπτει η σχέση:

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

Μια πολύ σημαντική ιδιότητα της λογιστικής παλινδρόμησης θεωρείται η παρακάτω:

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

ή γενικότερα:

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

Από την παραπάνω σχέση προκύπτει:

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

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

Κατ' αρχήν:

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

ασυμπτωτικά.

Επομένως, ένα κατάλληλο $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το y_i, μπορεί να βρεθεί από τη σχέση:

$$\hat{P}(\boldsymbol{x_i}) \pm z_{\frac{\alpha}{2}}\{n_i[P(\boldsymbol{x_i})][1 - P(\boldsymbol{x_i})]\}\sqrt{1 + \boldsymbol{x_i}'(\boldsymbol{X'VX})^{-1}\boldsymbol{x_i}} \quad , \qquad i = 1,2,\dots,m$$

Φυσικά, στην πράξη πρέπει να αντικαταστήσουμε το $\hat{P}(\boldsymbol{x_i})$ στον πίνακα **V**.

2.3.5 Συμπερασματολογία με χρήση πιθανοφάνειας στη λογιστική παλινδρόμηση

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

$$SS_E(reduced) - SS_E(full)$$

Η διαφορά στο άθροισμα τετραγώνων του σφάλματος αντικαθίσταται, στη λογιστική παλινδρόμηση, από τη διαφορά της log πιθανοφάνειας.

Ασυμπτωτικά ισχύει:

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

όπου:

το $\mathcal{L}(.)$ είναι η πιθανοφάνεια και στην περίπτωσή μας , ζητούμε 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.

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

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

$$2[\ln\mathcal{L}[b_0,b_1,b_2,b_3]-\ln\mathcal{L}[b_0^*,b_3^*]]$$

όπου:

$\mathcal{L}(b_0^*,b_3^*)$ είναι η πιθανοφάνεια για το λογιστικό μοντέλο στο οποίο έχουμε επικαλεστεί τη μηδενική υπόθεση (δηλαδή $\beta_1=\beta_2=0$)

Σαν αποτέλεσμα, η υπόθεση απορρίπτεται στην περίπτωση που 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, χρησιμοποιείται για ελέγχους ενός set υποθέσεων.

2.3.6 Παράδειγμα εφαρμογής της λογιστικής παλινδρόμησης

Ο σκοπός του παραδείγματος είναι να μελετήσουμε τη χρήση της λογιστικής παλινδρόμησης για την ανάλυση της επίδρασης μιας ουσίας σε ένα πείραμα τοξικότητας. Ο παρακάτω πίνακας δείχνει την επίδραση διαφορετικών δόσεων νικοτίνης στην κοινή μύγα των φρούτων.

Συγκέντρωση x (g/100cc)	**Αριθμός εντόμων n**	**Αριθμός εντόμων που απεβίωσαν 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 7: Επίδραση διαφορετικών δόσεων νικοτίνης

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 μιας πειραματικής διαδικασίας. Θα εκτιμήσουμε την ED_{50} , όπου ED_p είναι η τιμή του x για την οποία η πιθανότητα του θανάτου μιας μύγας των φρούτων παίρνει την τιμή P.

Το στατιστικό πακέτο (PROC LOGIST from SAS) δίνει τα ακόλουθα 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*
INTERCPT	*1*	*3.1236*	*0.3349*	*86.9818*	*0.0001*	
LOGX	*1*	*2.1279*	*0.2214*	*92.3628*	*0.0001*	*0.898802*

Πίνακας 8: Ανάλυση Μέγιστης Πιθανοφάνειας

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

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

Σε τέτοιου είδους πειράματα συχνά αντικαθιστούμε το x με το $\ln x$. Αυτό είναι ιδιαίτερα χρήσιμο όταν το x έχει μεγάλο εύρος τιμών. Οι $p - τιμές$ των παραμέτρων που παράχθηκαν από τα στατιστικά χ^2 του Wald ελέγχου είναι αρκετά σημαντικές και για τα δυο μοντέλα, έτσι έχουμε δύο υποψήφια μοντέλα. Μια μέθοδος για να συγκρίνουμε τα δύο μοντέλα είναι να συγκρίνουμε τα εύρη των διαστημάτων εμπιστοσύνης γύρω από το $\hat{y}$ (Lewis , Montgomery and Myers 2001). Μια άλλη σχετική μέθοδος είναι να παρατηρήσουμε το τυπικό σφάλμα του εκτιμώμενου predictor $\boldsymbol{x'b}$ για τα δύο μοντέλα. Στον παρακάτω πίνακα 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 9: Τυπικά σφάλματα των linear predictors

Στην περίπτωσή μας είναι δύσκολο να επιλέξουμε ανάμεσα στα δύο μοντέλα χρησιμοποιώντας τις πιο πάνω πληροφορίες, παρόλο 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.

Υπολογίζουμε το ED_{50} χρησιμοποιώντας και τα δυο μοντέλα για το Linear predictor:

- Για το μοντελο $b_0 + b_1 x$, ο $\widehat{ED}_{50}$ δίνεται από την εξίσωση:

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

Στο παράδειγμα ισούται με 0.277g/100cc.

- Για το μοντέλο $b_0' + b_1' \ln x$, το $\widehat{ED}_{50}$ δίνεται από την εξίσωση

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

2.4 Καμπύλες ROC

Η πραγματοποίηση προβλέψεων αποτελεί ένα από τα σημαντικότερα μελήματα κάθε επιχείρησης και επιστημονικού πεδίου σχετικά με την αναζήτηση πληροφορίας. Το γεγονός αυτό καθιστά την εξασφάλιση προγνωστικής ακρίβειας απαραίτητη στον σχεδιασμό και την σύγκριση μοντέλων, αλγορίθμων και τεχνολογιών που παράγουν προβλέψεις. Οι καμπύλες ROC (Receiver Operating Characteristic-Λειτουργικό Χαρακτηριστικό Δέκτη) συμβάλλουν στην εξασφάλιση της επιθυμητής ακρίβειας στις προβλέψεις. Έτσι, αποτελούν χρήσιμη τεχνική για την οργάνωση, επιλογή και απεικόνιση ταξινομητών με βάση τη γραφική τους παράσταση.

Ιστορικά οι καμπύλες ROC χρονολογούνται στις αρχές τις δεκαετίας του 50, όταν οι φοιτητές του τμήματος ηλεκτρολόγων μηχανικών του πανεπιστημίου του Michigan -WW Peterson και TG Birdsall - εφάρμοσαν τη στατιστική θεωρία αποφάσεων σε προβλήματα της θεωρίας λήψης σημάτων (signal detection theory). Αρχικά, προτάθηκαν ως γραφική μέθοδος μέτρησης της ποιότητας λήψης σήματος από ένα δείκτη σε ατελή διαγνωστικά συστήματα.

Η ROC καμπύλη ορίζεται ως το μοναδιαίο τετράγωνο $[0,1] \times [0,1]$, το οποίο ξεκινά από το σημείο $(0,0)$- όταν το σημείο απόφασης είναι μεγαλύτερο από όλες τις μετρήσεις θορύβου και σήματος – για να καταλήξει στο σημείο $(1,1)$ – στην περίπτωση που το σημείο απόφασης είναι μικρότερο από όλες τις μετρήσεις. Το εμβαδόν που ορίζεται κάτω από την καμπύλη αποτελεί ένα μέτρο της ποιότητας

διαχωρισμού θορύβου και σήματος και χρησιμοποιείται συχνα στην στατιστική συμπερασματολογία των ROC καμπυλών.

2.4.1 Βασικές Έννοιες

Συχνά στην ιατρική έρευνα αναπτύσσονται διαγνωστικοί έλεγχοι σε συνεχή ή διακριτή κλίμακα για το διαχωρισμό των πληθυσμών υγιών και ασθενών. Η κλινική χρησιμότητα μιας διαγνωστικής δοκιμασίας (εργαστηριακό αποτέλεσμα ή κλινικό εύρημα) προσδιορίζεται κατά κύριο λόγο από τη διακριτική της ικανότητα, δηλαδή από την ακρίβεια με την οποία διακρίνει αρρώστους με ή χωρίς το υπό διερεύνηση νόσημα, για το οποίο αυτή επιτελείται.

Στην θεωρία ανίχνευσης σημάτων, μια ROC καμπύλη είναι μια γραφική παράσταση της ευαισθησίας(sensitivity) ή των αληθώς θετικών, έναντι του 1- ειδικότητα (specificity) ή ψευδώς θετικών, για ένα σύστημα δυαδικής ταξινόμησης, καθώς το όριο ταξινόμησης ποικίλει. Ο δείκτης λειτουργικού χαρακτηριστικού μπορεί ισοδύναμα να εκπροσωπηθεί με τη γραφική παράσταση του ποσοστού των αληθώς θετικών (TPR = True Positive Rate) έναντι του ποσοστού των ψευδώς θετικών (FPR =False Positive Rate). Είναι ακόμη γνωστή ως ***καμπύλη σχετικού λειτουργικού χαρακτηριστικού***, αφού αποτελεί τη σύγκριση δύο λειτουργικών χαρακτηριστικών (TPR, FPR) καθώς το κριτήριο αλλάζει.

Τώρα, ας εξέτασουμε ενα πρόβλημα πρόβλεψης διπλής κλάσης (δυαδική ταξινόμηση), στο οποίο το αποτέλεσμα χαρακτηρίζεται ως θετική (**p**) ή αρνητική (**n**) κλάση. Υπάρχουν τέσσερις πιθανές εκβάσεις για ένα δυαδικό ταξινομητή. Αν το αποτέλεσμα της πρόβλεψης είναι p και η πραγματική τιμή είναι επίση p, αυτό ονόμαζεται αληθώς θετικό. Ωστόσο, εάν η πραγματική τιμή είναι n, λέγεται ψευδώς θετικό. Αντίθετα, ένα αληθώς αρνητικό έχει προκύψει όταν τόσο το αποτέλεσμα της πρόβλεψης όσο και η πραγματική τιμή είναι n και ένα ψευδώς αρνητικό όταν το αποτέλεσμα πρόβλεψης είναι n, ενώ η πραγματική τιμή είναι p.

Ορίζουμε ένα πείραμα με P θετικές και N αρνητικές περιπτώσεις. Τα τέσσερα αποτελέσματα μπορούν να παρουσιαστούν με ένα 2×2 πίνακα συνάφειας ως εξής:

		Πραγματική τιμή		
		p	**N**	
Αποτέλεσμα Πρόβλεψης	**p'**	Αληθώς Θετικό	Ψευδώς Θετικό	**P'**
	n'	Ψευδώς Αρνητικό	Αληθώς Αρνητικό	**N'**
	Σύνολο	**P**	**N**	

Πίνακας 10: Πίνακας Συνάφειας

Βασική Ορολογία

- Ευαισθησία ή ποσοστό θετικών (TPR):

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

- Ποσοστό ψευδώς θετικών (FPR):

$$FPR = \frac{FP}{P} = \frac{FP}{FP + TN}$$

- Ειδικότητα ή ποσοστό αληθώς αρνητικών (SPC):

$$SPR = \frac{TN}{N} = \frac{TP}{FP + TN} = 1 - FPR$$

- Ακρίβεια (Accuracy):

$$ACC = \frac{TP + TN}{P + N}$$

- Θετική προβλεπόμενη τιμή (PPV):

$$PPV = \frac{TP}{TP + FP}$$

- Αρνητική προβλεπόμενη τιμή (NPV):

$$NPV = \frac{TN}{TN + FN}$$

- Θετικός λόγος πιθανοφανειών (LR+ ή L):

$$L = \frac{TPR}{FPR}$$

- Αρνητικός λόγος πιθανοφανειών (LR- ή λ):

$$\lambda = \frac{FNR}{TNR}$$

- Επιπολασμός (Prevalence):

$$PRV = \frac{TP + FN}{P + N}$$

- Συντελεστή συσχέτισης του Matthews (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{PNP'N'}}$$

- F1 Score:

$$F1 = \frac{2TP}{P + P'}$$

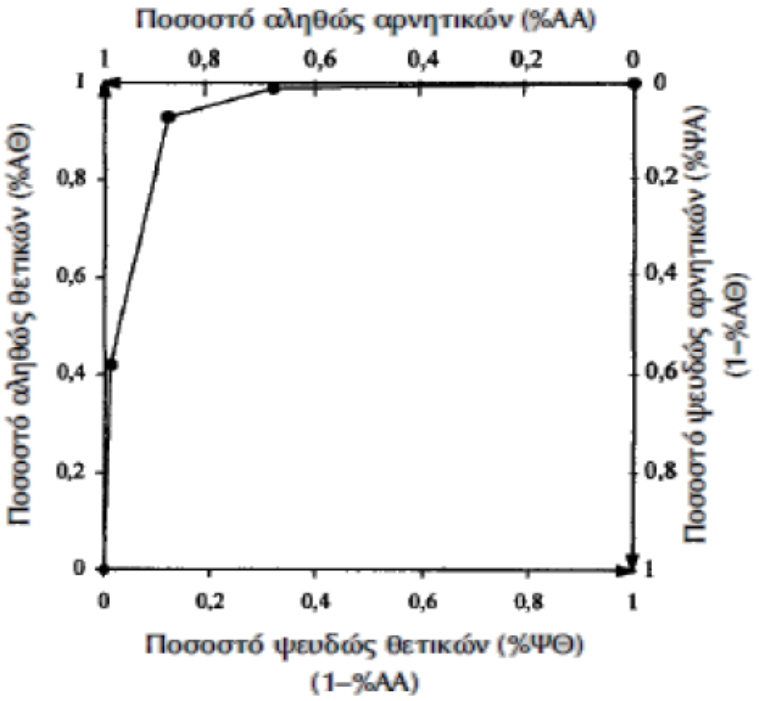
Από ένα πίνακα συνάφειας μπορούμε να βρούμε πολλές μετρικές :

- Η *ευαισθησία (TPR)* καθορίζει ένα διαγνωστικό test που ταξινομεί σωστά τα θετικά περιστατικά μεταξύ όλων των διαθέσιμων θετικών δειγμάτων κατά τη διάρκεια δοκιμής. Ο όρος ευαισθησία χρησιμοποιείται για να ορίσει τη μικρότερη συγκέντρωση μιας ουσίας που μπορεί η μέθοδος να ανιχνεύσει και να μετρήσει. Από την άλλη πλευρά, ο όρος 1- ειδικότητα (FPR) ορίζει πόσα λανθασμένα θετικά αποτελέσματα εμφανίζονται μεταξύ όλων των διαθέσιμων αρνητικών δειγμάτων κατά τη διάρκεια της δοκιμής. Ο όρος ειδικότητα χρησιμοποιείται για να ορίσει την ιδιότητα της μεθόδου να ανιχνεύσει και να μετράει μόνο την ουσία που θέλει να μετρήσει.
- Η *θετική προβλεπόμενη τιμή (PPV*) ερμηνεύεται ως η πιθανότητα εμφάνισης θετικού περιστατικού μεταξύ όλων των θετικών προβλέψεων και η αρνητική προβλεπόμενη τιμή (NPV) ως η πιθανότητα εμφάνισης αρνητικού περιστατικού μεταξύ όλων των αρνητικών προβλέψεων.
- Ο *θετικός λόγος πιθανοφανειών (L)* εκφράζει πόσες φορές πιο συχνά εμφανίζεται το θετικό αποτέλεσμα στους πάσχοντες σε σχέση με τους μη πάσχοντες από το νόσημα που διερευνάται. Ο αρνητικός λόγος πιθανοφανειών (λ) εκφράζει πόσες φορές πιο συχνά εμφανίζεται το αρνητικό αποτέλεσμα στους μη πάσχοντες σε σχέση με τους πάσχοντες από το νόσημα που διερευνάται.

2.4.2 Σχεδιασμός καμπύλης ROC

Η καμπύλη ROC εκφράζει τη σχέση του ποσοστού των αληθώς θετικών ($\%\,\Theta\Alpha$) και ψευδώς θετικών ($\%\,\Psi\Theta = 1 - \%AA$) αποτελεσμάτων της διαγνωστικής δοκιμάσιας, καθώς μεταβάλλεται προοδευτικά προς μια κατεύθυνση το διαχωριστικό όριο (ΔO). Ένας χώρος ROC ορίζεται από το $\%\,\Psi\Theta$ στον x και το $\%\,A\Theta$ στον y άξονα αντίστοιχα και κάθε σημείο της εμπειρικής καμπύλης ROC προσδιορίζεται από ένα ορισμένο ζεύγος ($\%\,\Psi\Theta, \%\,A\Theta$).

Η ROC καμπύλη όπως είδαμε και πιο πάνω ορίζεται ως το μοναδιαίο τετράγωνο $[0,1] \times [0,1]$, το οποίο ξεκινά από το σημείο $(0,0)$ για να καταλήξει στο σημείο $(1,1)$. Η καλύτερη δυνατή μέθοδος πρόβλεψης θα απέφερε ένα σημείο στην επάνω αριστερή γωνιά ή τη συντεταγμένη $(0,1)$ του χώρου ROC, που αντιπροσωπεύει το 100% ευαισθησία (μηδέν ψευδώς αρνητικά) και 100% ειδικότητα (μηδέν ψευδώς θετικά). Μια εντελώς τυχαία εικασία θα έδινε ένα σημείο κατά μήκος μιας διαγώνιας γραμμής (*γραμμή της μη διάκρισης*), από την κάτω αριστερή προς την πάνω δεξιά γωνιά.



Γράφημα 7: Γραφική Απεικόνιση του τρόπου με τον οποίο αυξάνονται τα ποσοστά των ΑΘ, ΨΘ, ΨΑ, ΑΑ. Εμφανής επιλογή τριών ΔΟ.

2.4.2.1 *Περιοχές και σημεία που έχουν προβλεπτικές ικανότητες*

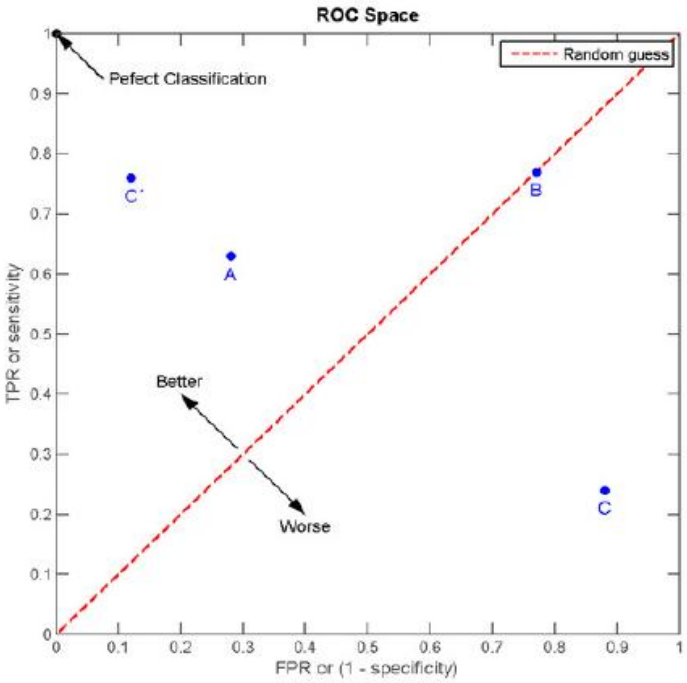
Σ' ένα επίπεδο όπου απεικονίζεται μια ROC καμπύλη, ορίζονται επιμέρους σημεία και περιοχές με ιδιαίτερη προβλεπτική ικανότητα. Ορισμένες εκδοχές θεωρούνται οι εξής :

- Το σημείο τομής της ROC καμπύλης με την κάθετη διαγώνιο

- Η περιοχή μεταξύ της καμπύλης ROC και της διαγωνίου
- Η περιοχή κάτω από την ROC καμπύλη (AUC)
- d', η απόσταση 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 δύο αυτές κατανομές είναι Κανονικές με την ίδια τυπική απόκλιση. Βάσει των υποθέσεων αυτών, μπορεί να αποδειχθεί ότι το σχήμα της ROC καμπύλης εξαρτάται μόνο από την d'.

2.4.2.2 *Η μέθοδος αντικατοπτισμού σημείου*

Η διαγώνιος χωρίζει το χώρο ROC. Τα σημεία πάνω από την διαγώνιο αντιπροσωπεύουν τα αποτελέσματα της καλής ταξινόμησης. Τα σημεία κάτω από τη διαγώνιο δίνουν πενιχρά αποτελέσματα. Παρατηρούμε, ωστόσο ότι ένα φτωχό μέσο πρόβλεψης μπορεί απλά να αποκτήσει σημεία πάνω από τη διαγώνιο, αν τα αποτελέσματα του πίνακα συνάφειας αντιστραφούν. Ας εξετάσουμε τα αποτελέσματα που προβλέπονται από τέσσερις διαδικασίες :



Γράφημα 8: Ο αντικατοπρισμός του σημείου C δίνει καλύτερα αποτελέσματα

Τα αποτελέσματα της μεθόδου Α έχει την καλύτερη προβλεπτική ικανότητα μεταξύ των Α, Β, C. Το αποτέλεσμα της Β σκουμπάει στη διαγώνιο, άρα η ακρίβεια της Β είναι 50% . Ωστόσο, όταν πάρουμε το συμμετρικό του σημείου C ως προς το κεντρικό σημείο $(0.5,0.5)$, η προκύπτουσα μέθοδος C' είναι ακόμη καλύτερη από την Α. Αυτή η μέθοδος αντιικατοπρισμού αντιστρέφει τις προβλέψεις οποιασδήποτε μεθόδου ή test παράγει ο πίνακας συνάφειας C. Παρόλο που η αρχική μέθοδος έχει αρνητική προβλεπτική ικανότητα, απλά αντιστρέφοντας τις αποφάσεις της οδηγούμαστα σε μια νέα διαγνωστική μέθοδο C' με θετική προβλεπτική ικανότητα. Στη περίπτωση αυτή η απόσταση σημείου από την διαγώνιο είναι ο καλύτερος δείκτης για την προβλεπτική δύναμη της μεθόδου. Αν το αποτέλεσμα είναι κάτω από τη συγκεκριμένη γραμμή, όλες οι προβλέψεις της μεθόδου πρέπει να αντιστραφούν, ώστε το αποτέλεσμα να βρεθεί πάνω από τη γραμμή και να αξιοποιηθεί η ισχύς της μεθόδου.

2.4.2.3 *Έννοια και χρήση του Εμβαδού κάτω από την καμπύλη ROC*

Η περιοχή κάτω από την καμπύλη ROC (AUC) ισούται με την πιθανότητα ένας ταξινομητής να κατατάξει ένα τυχαία επιλεγμένο θετικό παράδειγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό. Η AUC ερμηνεύεται αλλίως ως η πιθανότητα, η τιμή του test για έναν ασθενή $(N+)$ να είναι η μεγαλύτερη από την τιμή του test για ένα άτομο που δεν έχει την ασθένεια $(N-)$. Δηλαδή, $\mathrm{AUC} = \mathrm{P}(N+ > N-)$ και εκτιμάται από τον τύπο :

$$w = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} I(N_i^+, N_j^-)$$

όπου:

n_+ και n_- το πλήθος των ατόμων με ή χωρίς τη νόσο αντίστοιχα,

N_i^+ η τιμή του διαγνωστικού test για το i άτομο της ομάδας ασθενών,

N_j^- η τιμή του διαγνωστικού test για το j άτομο της ομάδας μη ασθενών,

$$I\left(N_i^+, N_j^-\right) = \begin{cases} 1, & N_i^+ > N_j^- \\ \frac{1}{2}, & N_i^+ = N_j^- \\ 0, & N_i^+ < N_j^- \end{cases}$$

Όταν το διαγνωστικό test συνδέεται αρνητικά με τη νόσο (μικρές τιμές του test υποδεικνύουν μεγάλη πιθανότητα εμφάνισης της νόσου) τότε υπολογίζουμε την ποσότητα $w' = 1 - w$ ή απλά μετασχηματίζουμε το test έτσι ώστε να συνδέεται θετικά με τη νόσο (π.χ πολλαπλασιάζουμε με το -1).

ΚΕΦΑΛΑΙΟ 3: ΕΦΑΡΜΟΓΗ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

3.1 Μεθοδολογία συλλογής δεδομένων

Τα έτη 2005-2006 πραγματοποιήθηκε μια μελέτη με θέμα «Επιδημιολογική Αναφορά και Διαχείριση Τραυματιών στην Ελλάδα» από την Ελληνική Εταιρεία Τραύματος και Επείγουσας Χειρουργικής (Ε.Ε.Τ&Ε.Χ). 'Ελαβαν μέρος 32 Εθνικοί Επίσημοι Αντιπροσώποι της Ε.Ε.Τ&Ε.Χ - πιστοποιημένοι χειρούργοι ελληνικών νοσοκομείων που περιλαμβάνουν τραυματίες – και έκαναν τρεις μεγάλες συναντήσεις με σκοπό την πλήρη κατάρτιση όσον αφορά τη φόρμα καταγραφής, τα κριτήρια εισαγωγής και τη διαδικασία διακίνησης των δεδομένων τραύματος. Για την ενιαία αναφορά δεδομένων τραύματος ακολουθήθηκαν οι βασικές διεθνείς συστάσεις.

Τα δεδομένα προήρθαν από ασθενείς οι οποίοι πληρούσαν τα παρακάτω κριτήρια εισαγωγής:

- Τραυματίες των οποίων κρίθηκε απαραίτητη η εισαγωγή στο νοσοκομείο
- Τραυματίες οι οποίοι προσκομίσθηκαν νεκροί ή απεβίωσαν στο Τμήμα Επειγόντων Περιστατικών (Τ.Ε.Π)
- Τραυματίες που μεταφέρθηκαν από ή προς άλλο υγειονομικό σχηματισμό

Οι φόρμες καταγραφής συμπληρώθηκαν προοπτικά με ευθύνη του κάθε Αντιπροσώπου από την άφιξη του ασθενούς στο Τ.Ε.Π μέχρι την έξοδο του από το νοσοκομείο, τη μεταφορά του σε άλλο νοσοκομείο ή το θάνατό του.

Στη συγκεκριμένη μελέτη καταγράφηκαν ακριβώς 8,862 ασθενείς οι οποίοι νοσηλεύτηκαν σε 32 νοσοκομεία της χώρας από τα οποία, 5 βρίσκονται στην Αττική και τα υπόλοιπα 27 νοσοκομεία στην περιφέρεια. Οι παράγοντες που καταγράφηκαν, οι οποίοι προέκυψαν μετά από βιβλιογραφική μελέτη και ιατρική εμπειρία, περιελάμβαναν δημογραφικά στοιχεία των ασθενών, στοιχεία που αφορούσαν στις συνθήκες του ατυχήματος καθώς και πληροφορίες για την ενδονοσοκομειακή αντιμετώπιση των ασθενών.

3.1.1 Εισαγωγή στο Πρόβλημα

Η βάση δεδομένων περιέχει 8.862 εγγραφές-ασθενείς και 92 επεξηγηματικές μεταβλητές.

Οι 92 επεξηγηματικές μεταβλητές καταγράφονται στο παρακάτω πίνακα :

X_1	**Βάρος** σε κιλά (kg)
X_2	**Ηλικία** σε έτη
X_3	**Κλίμακα γλασκώβης (G.C.S)*** (3-15)
X_4	**Σφύξεις** σε N/min*
X_5	**Αναπνοές** σε N/min*
X_6	**Συστολική αρτηριακή πίεση (Σ.Α.Π)** σε mmHg*
X_7	**Διαστολική αρτηριακή πίεση (Δ.Α.Π)** σε mmHg*
X_8	**Αιματοκρίτης (Ht)** σε (%)*
X_9	**Αιμοσφαιρίνη (Hb)** σε g/dl*
X_{10}	**Κορεσμός** σε (%)*
X_{11}	**Λευκά αιμοσφαίρια (/μl)***
X_{12}	**Αιμοπετάλια (/μl)***
X_{13}	**Νάτριο (Na)** σε mEq/L*
X_{14}	**Κάλιο (K)** σε mEq/L*
X_{15}	**Σάκχαρο** σε mg%*
X_{16}	**Κρεατινίνη** σε mg%*
X_{17}	**Ουρία** σε mg%*
X_{18}	**Αμυλάση***
X_{19}	**Εκτίμηση αναπηρίας** Αναμενόμενη μόνιμη μεγάλη; Αναμενόμενη μόνιμη μικρή; Αναμενόμενη προσωρινά μεγάλη; Αναμενόμενη προσωρινά μικρή; Καλή ανάνηψη
X_{20}	**Injury Security Score (I.S.S)*** (0-75)
X_{21}	**Revised Trauma Score (R.T.S)***
X_{22}	**Μηχανισμός κάκωσης** Αμβλεία; Διατιτραίνουσα; Έγκαυμα; Εισπνοή
X_{23}	**Αίτια κάκωσης** Πτώση ;Τροχαίο; Αθλητικό; Εργατικό; Εγκληματική ενέργεια; Άλλο
X_{79}	**Ημέρα εξόδου από το νοσοκομείο** Δευτέρα; Τρίτη; Τετάρτη; Πέμπτη; Παρασκευή; Σάββατο; Κυριακή
X_{80}	**Μήνας εξόδου από το νοσοκομείο** Ιανουάριος; Φεβρουάριος; Μάρτιος; Απρίλιος; Μάιος; Ιούνιος; Ιούλιος; Αύγουστος; Σεπτέμβριος; Οκτώβριος; Νοέμβριος; Δεκέμβριος

X_{24}	**Μέσο άφιξης στο Τ.Ε.Π*** Αερομεταφορά; ΕΚΑΒ; ΙΧ; Περιπατητικός
X_{25}	**ΕΚΑΒ στον τόπο ατυχήματος** (ναι ή όχι)
X_{26}	**Νοσοκομείο καταγραφής**
X_{27}	**Σύνολο στοιχείων υποδομής νοσοκομείου** Ορθοπεδικός; CT; Αγγειοχειρουργός; Νευροχειρουργός; Μ.Ε.Θ
X_{28}	**Συνοδές παθήσεις** (ναι ή όχι)
X_{29}	**Ημέρα συμβάντος** Δευτέρα; Τρίτη; Τετάρτη; Πέμπτη; Παρασκευή; Σάββατο; Κυριακή
X_{30}	**Μήνας συμβάντος** Ιανουάριος; Φεβρουάριος; Μάρτιος; Απρίλιος; Μάιος; Ιούνιος; Ιούλιος; Αύγουστος; Σεπτέμβριος; Οκτώβριος; Νοέμβριος; Δεκέμβριος
X_{31}	**Φύλο** (άνδρας ή γυναίκα)
X_{32}	**Διακομιδή από** (ναι ή όχι)
X_{33}	**Ημέρα άφιξης στο νοσοκομείο** Δευτέρα; Τρίτη; Τετάρτη; Πέμπτη; Παρασκευή; Σάββατο; Κυριακή
X_{34}	**Μήνας άφιξης στο νοσοκομείο** Ιανουάριος; Φεβρουάριος; Μάρτιος; Απρίλιος; Μάιος; Ιούνιος; Ιούλιος; Αύγουστος; Σεπτέμβριος; Οκτώβριος; Νοέμβριος; Δεκέμβριος
X_{35}	**Ειδικότητα ιατρού*** Αγγειχ/ος; Αγροτικός ιατρός; Γενικός ιατρός; Γενικός Χειρουργός; Γναθοχ/ος; Γυναικ/ος; Θωρακ/ος; Νευροχ/ος; Ορθοπεδικός; Ουρολόγος; Παθολόγος; Παιδίατρος; Παιδοχ/ος; Πλαστικός; Χειρούργος; Ωτοριν/ος
X_{36}	**Ειδικευόμενος Ιατρός*** (ναι ή όχι)
X_{37}	**Πιστοποίηση ιατρού με A.L.T.S*** (ναι ή όχι)
X_{38}	**Τριχοειδική επαναπλήρωση*** (ναι ή όχι)
X_{39}	**Ωχρός*** (ναι ή όχι)
X_{40}	**Εφίδρωση*** (ναι ή όχι)
X_{41}	**Ανησυχία*** (ναι ή όχι)
X_{42}	**Κεντρική κυάνωση*** (ναι ή όχι)
X_{43}	**Περιτοναικά σημεία*** (ναι ή όχι)
X_{44}	**Οξυγόνο*** (ναι ή όχι)
X_{45}	**Διασωλήνωση*** (ναι ή όχι)
X_{46}	**Μηχανικός αερισμός*** (ναι ή όχι)
X_{47}	**ΚΑΡΠΑ*** (ναι ή όχι)
X_{48}	**Παροχέτευση θώρακα*** (ναι ή όχι)
X_{49}	**Περικαρδιοκέντιση*** (ναι ή όχι)
X_{50}	**Καθετήρας κύστεως*** (ναι ή όχι)
X_{51}	**Ρινογαστρικός σωλήνας*** (ναι ή όχι)
X_{52}	**Κολάρο*** (ναι ή όχι)
X_{53}	**Ακινητοποίηση σπονδυλικής στήλης*** (ναι ή όχι)

X_{54}	**Ακινητοποίηση πυέλου*** (ναι ή όχι)
X_{55}	**Ακινητοποίηση άκρων*** (ναι ή όχι)
X_{56}	**Οροί*** (ναι ή όχι)
X_{57}	**Αίμα*** (ναι ή όχι)
X_{58}	**ICP monitoring*** (ναι ή όχι)
X_{59}	**Θωρακοτομή*** (ναι ή όχι)
X_{60}	**Αγγειογραφία*** (ναι ή όχι)
X_{61}	**Εμβολισμός*** (ναι ή όχι)
X_{62}	**Περιτοναική πλύση*** (ναι ή όχι)
X_{63}	**Αέρια*** (ναι ή όχι)
X_{64}	**Ακτινογραφία*** (ναι ή όχι)
X_{65}	**Αξονική τομογραφία (CT)*** (ναι ή όχι)
X_{66}	**Υπέρηχος (US)*** (ναι ή όχι)
X_{67}	**Γενική ούρων*** (ναι ή όχι)
X_{68}	**Τοξικολογικός έλεγχος*** (ναι ή όχι)
X_{69}	**Ημέρα εξόδου από το ΤΕΠ*** Δευτέρα; Τρίτη; Τετάρτη; Πέμπτη; Παρασκευή; Σάββατο; Κυριακή
X_{70}	**Μήνας εξόδου από το ΤΕΠ*** Ιανουάριος; Φεβρουάριος; Μάρτιος; Απρίλιος; Μάιος; Ιούνιος; Ιούλιος; Αύγουστος; Σεπτέμβριος; Οκτώβριος; Νοέμβριος; Δεκέμβριος
X_{71}	**Προορισμός μετά το ΤΕΠ** Άλλο νοσοκομείο; Κλινική; Μ.Α.Φ; Μ.Ε.Θ; Νεκροτομείο; Χειρουργείο
X_{72}	**Χειρουργική επέμβαση** (ναι ή όχι)
X_{73}	**Αξονική τομογραφία ενδονοσοκομειακά** (ναι ή όχι)
X_{74}	**Υπέρηχος (US) ενδονοσοκομειακά** (ναι ή όχι)
X_{75}	**M.R.I) ενδονοσοκομειακά** (ναι ή όχι)
X_{76}	**Αγγειογραφία) ενδονοσοκομειακά** (ναι ή όχι)
X_{77}	**Επιπλοκές** (ναι ή όχι)
X_{78}	**Παραμονή σε Μονάδα Εντατικής Θεραπείας** (ναι ή όχι)
X_{85}	**Ώρα συμβάντος (code 1) (ΩΩ:ΛΛ)** 00: 00-04: 00; 04 :01- 08 :00; 08 :01-12 :00 12 :01-16: 00; 16 :01-20: 00 ; 20 :01 -24 :00
X_{86}	**Ώρα άφιξης στο Τ.Ε.Π (code 1) (ΩΩ:ΛΛ)** 00: 00-04: 00; 04 :01- 08 :00; 08 :01-12 :00 ; 12 :01-16: 00; 16 :01-20: 00 ; 20 :01 -24 :00
X_{87}	**Ώρα εξόδου από το Τ.Ε.Π (code 1) (ΩΩ:ΛΛ)** 00: 00-04: 00; 04 :01- 08 :00; 08 :01-12 :00 ; 12 :01-16: 00; 16 :01-20: 00 ; 20 :01 -24 :00
X_{101}	**Κάκωση στο κεφάλι (ΚΕΚ) (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
X_{102}	**Κάκωση στο πρόσωπο (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
X_{103}	**Κάκωση στο λαιμό (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)

X_{104}	**Κάκωση στον θώρακα (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
X_{105}	**Κάκωση στην κοιλιά (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
X_{106}	**Κάκωση στην σπονδυλική στήλη (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
X_{107}	**Κάκωση στα άνω άκρα (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
X_{108}	**Κάκωση στα κάτω άκρα (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
X_{109}	**Εξωγενείς κακώσεις (code 2)** Καμία; Ελαφριά (AIS<=2); Σοβαρή (AIS>2)
y	**death**

3.2 Εφαρμογή στο CLEMENTINE

Για την εξαγωγή τω συμπερασμάτων όσο αφορά την εξέλιξη της ζωής των τραυματιών - ασθενών ακολουθήσαμε τα πιο κάτω βήματα στο πρόγραμμα Clementine SPSS, ενός λογισμικού εξόρυξης δεδομένων :

1. Ανοίγουμε το προγράμμα Clementine από το Windows Start menu και φορτώνουμε το αρχείο με τις 8.862 εγγραφές-ασθενείς και τις 92 επεξηγηματικές μεταβλητές με τη βοήθεια ενός variable file node το οποίο τοποθετούμε στο stream canvas.
2. Τοποθετούμε στο stream canvas ένα type node, έτσι ώστε να μπορούν να διαβαστούν οι τύποι των τιμών των πεδίων. Με αυτόν τον τρόπο καθορίζεται ο τύπος των δεδομένων (type) για κάθε πεδίο και ακόμη καθορίζεται η κατεύθυνση (direction) η οποία δείχνει τι ρόλο παίζει κάθε πεδίο στη μοντελοποίηση.

   Αφού έγινε το πρώτο βήμα και βρήκαμε ακριβώς τα δεδομένα που θα χρησιμοποιήσουμε στη συνέχεια ορίζουμε τον τύπο της πληροφορίας για κάθε πεδίο των δεδομένων. Ο τύπος πληροφορίας για κάθε πεδίο πρέπει να τεθεί πριν τα πεδία χρησιμοποιηθούν στα διάφορα modeling nodes. Στο πρόγραμμα Clementine υπάρχουν οι εξής τύποι δεδομένων :

- **<u>Εύρος (range)</u> :**

  Το range χρησιμοποιείται για να περιγράψει συνεχείς αριθμητικές τιμές, ένα σύνολο ή μια κλίμακα 0-100 ή 0.75-1.25. Μια τιμή range μπορεί να είναι ακέραιος, πραγματικός αριθμός ή ημερομηνία/ώρα.

- **<u>Διακριτοποίηση (discrete)</u> :**

  Το discrete χρησιμοποιείται για να περιγράψει αλφαριθμητικές τιμές όταν ένας ακριβής αριθμός διαφορετικών τιμών είναι άγνωστος.

  <u>π.χ</u> :1,5,8

- **<u>Δίτιμη παράμετρος-λογική παράμετρος τύπου Boolean (flag)</u> :**

  Το flag χρησιμοποιείται από δεδομένα με δύο μόνο τιμές yes/no ή 0/1 ή 1/2.

- **<u>Σύνολο (set)</u> :**

  Το set χρησιμοποιείται για να περιγράψει δεδομένα με πολλαπλές διακεκριμένες τιμές όπου η καθεμιά αντιμετωπίζεται σαν μονάδα ενός συνόλου ή διακεκριμένες κατηγορίες όπως small/medium/large.

- **<u>Ανένταχτος τύπος (typeless)</u> :**

  Το typeless χρησιμοποιείται για δεδομένα που δεν εντάσσονται σε καμία από τις παραπάνω κατηγορίες ή για δεδομένα τύπου set με πάρα πολλές διακεκριμένες τιμές. Η επιλογή του τύπου typeless ορίζει αυτόματα το πεδίο του direction σε none, δηλαδή το πεδίο που δεν μπορεί να χρησιμοποιηθεί σε μοντέλα.

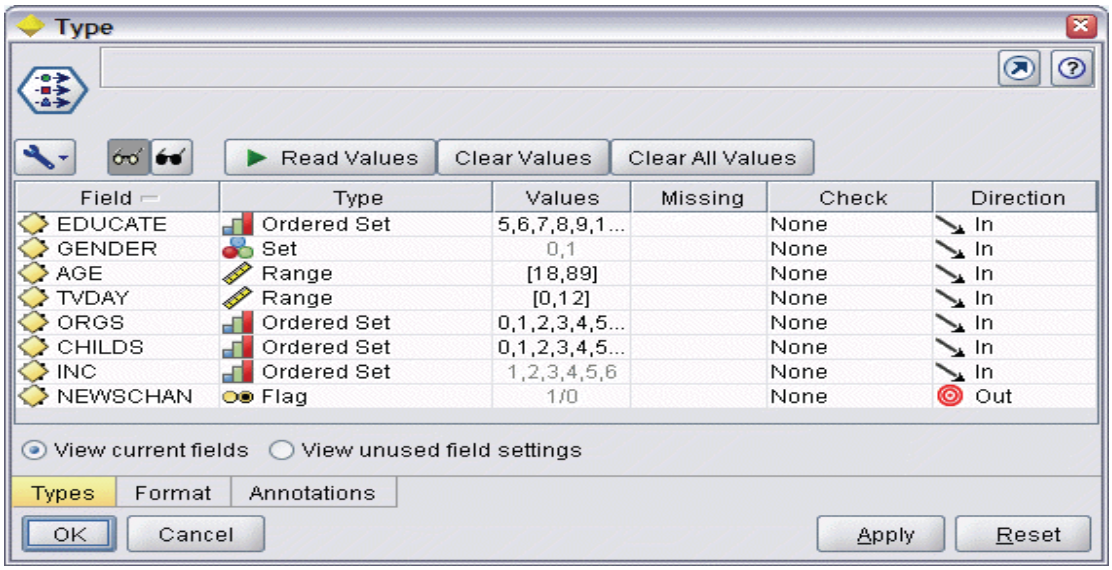
Τα δεδομένα εντάσσονται αρχικά σε μία από τις παραπάνω κατηγορίες μόλις τα εισαγάγουμε στο σύστημα. Για παράδειγμα, ο discrete τύπος δίνεται προσωρινά σε κατηγορικές μεταβλητές μέχρι να μπορεί να προσδιορισθεί αν πρόκειται για set ή flag τύπο και ο τύπος range δίνεται σε όλες τις αριθμητικές μεταβλητές.

Η τιμή του direction ενός πεδίου σχετίζεται μόνο με τη μοντελοποίηση. Υπάρχουν τέσσερις δυνατές κατευθύνσεις :

- **IN :** το πεδίο χρησιμοποιείται σαν input, δηλαδή είναι μία τιμή που θα βοηθήσει στη πρόβλεψη.
- **OUT:** το πεδίο χρησιμοποιείται σαν output-στόχος της τεχνικής μοντελοποίησης, είναι το πεδίο που θα προβλέψουμε.
- **BOTH:** το πεδίο επιτρέπεται να είναι και input και output σε κανόνα συσχέτισης (association rule). Όλες οι άλλες τεχνικές μοντελοποίησης αγνοούν αυτό το πεδίο.

  NONE: το πεδίο δεν χρησιμοποιείται στη μοντελοποίηση.

Dialog box :



Στην εφαρμογή μας στο dialog box του type node οι τύποι των 92 επεξηγηματικών μεταβλητών ήταν :

- **τύπου Range :** όταν το πεδίο έπαιρνε τιμές σε διάστημα,
- **τύπου Set :** όταν υπήρχαν διάφορες τιμές-κατηγορίες,
- **τύπου Flag :** η απόκριση y =death αποτελεί το στόχο πρόβλεψης, δηλαδή η έκβαση της υγείας του ασθενούς. Έτσι, η y είναι δίτιμη με τιμές :

$$y = \begin{cases} 0, & \text{ασθενής επιβίωσε} \\ 1, & \text{ασθενής απεβίωσε} \end{cases}$$

Στην εφαρμογή στην τιμή του *direction* έχουμε :

- ✓ *direction input* : για τις 92 επεξηγηματικές μεταβλητές, όπου το In δείχνει ότι θα χρησιμοποιηθούν σαν μεταβλητές πρόβλεψης
- ✓ *direction output*: για τη μεταβλητή y=death, όπου το Out δείχνει ότι πρόκειται για το πεδίο που θέλουμε να προβλέψουμε.

3. Τοποθετούμε στο stream canvas ένα Partition node έτσι ώστε να χωριστούν τα δεδομένα σε :

    - ✓ δεδομένα εκπαίδευσης(training set),
    - ✓ δεδομένα επαλήθευσης-επικύρωσης(quiz set-validation set),
    - ✓ δεδομένα ελέγχου-εξέτασης (test dataset).

Σε ένα τυπικό πρόβλημα του data mining, έχουμε ένα σύνολο δεδομένων εκπαίδευσης (training set) στο οποίο γνωρίζουμε την τιμή του αποτελέσματος και τις τιμές των χαρακτηριστικών που μας ενδιαφέρουν και προσπαθούμε με βάση αυτά τα δεδομένα να κατασκευάσουμε ένα μοντέλο πρόβλεψης. Το μοντέλο αυτό θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης (test set), στα οποία σύνολα είναι γνωστές οι τιμές τω χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης.

Στην περίπτωση τώρα, όπου ο αλγόριθμος που εφαρμόζουμε στηρίζεται σε κατασκευή και εκτίμηση μοντέλου, τα δεδομένα διαχωρίζονται σε τρία υποσύνολα:

- ➢ ***τα δεδομένα εκπαίδευσης (training data)*** τα οποία χρησιμοποιούνται για την προσαρμογή του μοντέλου,
- ➢ ***τα δεδομένα επικύρωσης (validation data)*** που χρησιμοποιούνται για την εκτίμηση του σφάλματος πρόβλεψης για την επιλογή του μοντέλου και
- ➢ ***τα δεδομένα ελέγχου (test data)*** που χρησιμοποιούνται για τον υπολογισμό της γενικευμένης τιμής σφάλματος του τελικά επιλεγμένου μοντέλου.

Καθένα από αυτά τα σύνολα θα πρέπει να επιλεγεί ανεξάρτητα. Το σύνολο επικύρωσης πρέπει να είναι διαφορετικό από το σύνολο εκπαίδευσης για να λαμβάνεται καλή απόδοση στη βελτιστοποίηση ή στο στάδιο της επιλογής και το σύνολο ελέγχου πρέπει να διαφορετικό και από τα δύο για να λαμβάνεται αξιόπιστη εκτίμηση του πραγματικού ποσοστού σφάλματος.

Καθώς το ποσοστό σφάλματος έχει προσδιοριστεί, τα δεδομένα ελέγχου μπορούν να χρησιμοποιηθούν πίσω στα δεδομένα εκπαίδευσης για την κατασκευή ενός άλλου, νέου, μοντέλου για πραγματική χρήση. Αυτός είναι ένας τρόπος να μεγιστοποιήσουμε τα δεδομένα που παράγουν το μοντέλο με το οποίο θα ασχοληθούμε στη πραγματικότητα. Αυτό που είναι σημαντικό, είναι ότι το ποσοστό σφάλματος δεν καθορίζεται με βάση κανένα από αυτά τα δεδομένα. Επίσης, μόλις τα δεδομένα επικύρωσης έχουν χρησιμοποιηθεί, τότε μπορούν να χρησιμοποιηθούν στα δεδομένα εκπαίδευσης για να γίνει επανεκπαίδευση, μεγιστοποιώντας τη χρήση των δεδομένων.

Γενικά, μπορούμε να πούμε ότι η ποιότητα του μοντέλου είναι ανάλογη του όγκου τω διαθέσιμων δεδομένων, αν και συχνά βαίνει φθίνουσα όταν ο όγκος του συνόλου εκπαίδευσης υπερβαίνει κάποιο όριο. Επίσης και η αξιοπιστία της εκτίμησης του σφάλματος είναι ανάλογη του όγκου των δεδομένων ελέγχου. Τα προβλήματα αρχίζουν όταν δεν υπάρχει επαρκής όγκος δεδομένων και επομένως περιορίζεται το ποσό των δεδομένων που μπορεί να χρησιμοποιηθεί ως σύνολο εκπαίδευσης, σύνολο επικύρωσης και σύνολο ελέγχου. Σε τέτοια σύνολα δεδομένων ένα μέρος των δεδομένων χρησιμοποιείται για τον έλεγχο και το υπόλοιπο για την εκπαίδευση. Αυτή η διαδικασία ονομάζεται *διαδικασία παρακράτησης* (holdout procedure) και το δίλημμα που προκύπτει τώρα είναι πώς διαχωρίσουμε το αρχικό σύνολο έτσι ώστε και τα δύο σύνολα να είναι μεγάλα.
Επανερχόμαστε τώρα στην περίπτωση που το σύνολο δεδομένων μας επαρκεί ώστε να κατασκευαστούν τα τρία σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Σε καθένα από αυτά τα σύνολα μπορούμε να διακρίνουμε τα αντίστοιχα σφάλματα και έτσι έχουμε το σφάλμα εκπαίδευσης (training error) που αναφέρεται στην προσαρμογή του μοντέλου, το σφάλμα επικύρωσης (validation error) που

αναφέρεται στην επιλογή του μοντέλου και το σφάλμα ελέγχου που αναφέρεται στην εκτίμηση του μοντέλου.

Είναι δύσκολο να δώσουμε ένα γενικό κανόνα σχετικά με το πώς επιλέγεται ο αριθμός των παρατηρήσεων που καταχωρείται σε καθένα από αυτά τα τρία σύνολα, καθώς εξαρτάται από το ποσοστό θορύβου στα δεδομένα και το μέγεθος του δείγματος εκπαίδευσης. Μία τυπική διάκριση που χρησιμοποιείται είναι το 50% στο σύνολο εκπαίδευσης και από 25% στα σύνολα επικύρωσης και ελέγχου αντίστοιχα.
Εφαρμόζουμε λοιπόν την παραπάνω διάκριση στις δικές μας 8.862 εγγραφές και προκύπτουν:

- Training set : $\approx 50\% \times 8.862 = 4417$ υποδείγματα
- Test set : $\approx 25\% \times 8.862 = 2231$ υποδείγματα
- Validation set : $\approx 25\% \times 8.862 = 2214$ υποδείγματα

4. Τοποθετούμε στο stream canvas ένα Feature selection node  το οποίο το συνδέουμε στο Partition node, έτσι ώστε να επιλεγούν για επίπεδο σημαντικότητας ***α=0.05*** οι σημαντικές μεταβλητές .

Στα περισσότερα προβλήματα εξόρυξης δεδομένων, εμπεριέχονται εκατοντάδες πεδία-μεταβλητές τα οποία είναι πιθανόν να χρησιμοποιηθούν με σκοπό την πρόβλεψη. Έτσι, άμεση συνέπεια είναι ότι απαιτείται πολύς χρόνος και προσπάθεια για να εξεταστεί ποια από αυτά τα πεδία πρέπει να συμπεριληφθούν στο μοντέλο. Για να μειώσουμε στο ελάχιστο τις πιθανές επιλογές, ο αλγόριθμος της επιλογής τω χαρακτηριστικών (Feature Selection Algorithm) μπορεί να χρησιμοποιηθεί για να προσδιορίσει τα πεδία εκείνα τα οποία είναι πιο σημαντικά για τη δεδομένη ανάλυση. Στην εφαρμογή προσπαθούμε να προβλέψουμε αποτελέσματα ασθενών ως προς την έκβαση της υγείας τους και έτσι πρέπει να βρούμε τους παράγοντες εκείνους που δείχνουν πιο σημαντικοί από τους υπόλοιπους.

Η επιλογή χαρακτηριστικών αποτελείται από τρία βήματα :

- **Screening (κρισάρισμα) :**

  Σε αυτό το βήμα απομακρύνονται οι μη σημαντικές και προβληματικές μεταβλητές πρόβλεψης καθώς και εγγραφές, όπως στην περίπτωση που έχουμε μεταβλητές με πολλές ελλειπούσες τιμές ή μεταβλητές με πολύ μεγάλη ή πολύ μικρή διακύμανση για να τις καθιστά χρήσιμες.

- **Ranking (Στοίχιση) :**

  Σε αυτό το βήμα ξεχωρίζονται οι εναπομείναντες μεταβλητές πρόβλεψης και καθορίζονται ranks βασισμένα στη σημαντικότητα.

- **Επιλογή :**

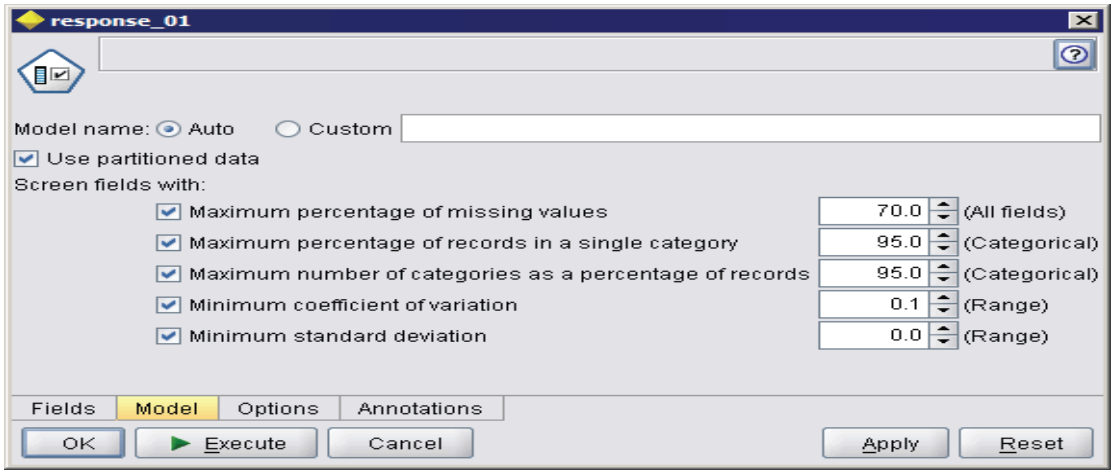
  Σε αυτό το βήμα αναγνωρίζεται το υποσύνολο τω χαρακτηριστικών που θα χρησιμοποιηθεί στα μοντέλα που ακολουθούν κρατώντας μόνο τις πιο σημαντικές μεταβλητές πρόβλεψης και φιλτράροντας ή αποκλείοντας όλες τις υπόλοιπες.

Τα πλεονεκτήματα από την επιλογή χαρακτηριστικών είναι ότι η διαδικασία της μοντελοποίησης απλοποιείται και φυσικά γίνεται ταχύτερη. Μειώνοντας τον αριθμό των πεδίων που χρησιμοποιούνται στο μοντέλο μειώνεται ο χρόνος αξιολόγησης του μοντέλου και επιπρόσθετα αποκτούμε απλούστερα και ακριβέστερα μοντέλα τα οποία μπορούν πολύ πιο εύκολα να εξηγηθούν .

Model tab-Options tab

Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα model tab το οποίο περιλαμβάνει βασικές επιλογές για το μοντέλο καθώς και ρυθμίσεις που επιτρέπουν την εύρεση κριτηρίων για το κρισάρισμα (Screening) τω μεταβλητών πρόβλεψης.

Model tab :



Model name : **(auto)** το όνομα του μοντέλου παράγεται αυτόματα.

Use partitioned data : **(√)** το επιλέγουμε γιατί με αυτό τον τρόπο εξασφαλίζουμε ότι χρησιμοποιηούμε μόνο τα δεδομένα από το το training set για την κατασκευή του μοντέλου.

Τα πεδία κρισάρονται(Screening) με τη βοήθεια των παρακάτω κριτηρίων :

- **Μέγιστο ποσοστό ελλειπουσών τιμών (Maximum percentage of missing values)**

  Κρισάρει τα πεδία με μεγάλο αριθμό ελλειπουσών τιμών που προσφέρουν ελάχιστη πληροφορία πρόβλεψης

- **Μέγιστο ποσοστό εγγραφών σε μια απλή κατηγορία (Maximum percentage of records in a single category)**

  Κρισάρει τα πεδία τα οποία έχουν πάρα πολλές εγγραφές να ανήκουν στην ίδια κατηγορία , για παράδειγμα το 95% των ασθενών στη βάση δεδομένων να πάσχουν από την ίδια ασθένεια, μιας και το να συμπεριληφθεί αυτή η πληροφορία δεν είναι χρήσιμη για να ξεχωρίσουμε τον ένα ασθενή από τον άλλο.

- **Μέγιστος αριθμός των κατηγοριών ως ποσοστό των εγγραφών (Maximum number of categories as a percentage of records)**

  Κρισάρει τα πεδία με πολλές κατηγορίες συγκριτικά με τον συνολικό αριθμό των εγγραφών δηλαδή εάν ένα μεγάλο ποσοστό των κατηγοριών περιέχει μόνο μία περίπτωση, το πεδίο δε μπορεί παρά να χρησιμοποιηθεί ελάχιστα.

- **<u>Ελάχιστος συντελεστής διακύμανσης (Minimum coefficient of variation)</u>**

  Κρισάρει τα πεδία με συντελεστή βάρους μικρότερο ή ίσο από το καθορισμένο ελάχιστο όριο. Εάν, η τιμή είναι κοντά στο 0 , δεν υπάρχει μεγάλη μεταβλητότητα στις τιμές της μεταβλητής.
- **<u>Ελάχιστη τυπική απόκλιση (Minimum standard deviation)</u>**

  Κρισάρει τα πεδία με τυπική απόκλιση μικρότερη ή ίση από το καθορισμένο ελάχιστο όριο.

Οι εγγραφές οι οποίες έχουν ελλείπουσες τιμές για το πεδίο στόχου ή ελλείπουσες τιμές για όλες τις μεταβλητές πρόβλεψης, αποκλείονται αυτόματα από όλους τους υπολογισμούς μέσα στα rankings.

Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα options tab το οποίο σου επιτρέπει να καθορίσεις τις default (εξ'ορισμού) ρυθμίσεις για την επιλογή ή τον αποκλεισμό των πεδίων πρόβλεψης του μοντέλου.

Σ' αυτό το βήμα θεωρείται μία μεταβλητή πρόβλεψης τη φορά για να εξεταστεί πόσο καλά κάθε μεταβλητή πρόβλεψης ξεχωριστά προβλέπει τη μεταβλητή στόχο. Οι μεταβλητές πρόβλεψης ιεραρχούνται σύμφωνα με το κριτήριο που καθορίζεται από το πειραματιστή.

Η τιμή σημαντικότητας κάθε μεταβλητής ή διαφορετικά ένα μέτρο το οποίο χρησιμοποιείται για να βάλει σε σειρά τα πεδία ή τα αποτελέσματα σε ποσοστιαία κλίμακα ορίζεται ως :

$$(1 - p)$$

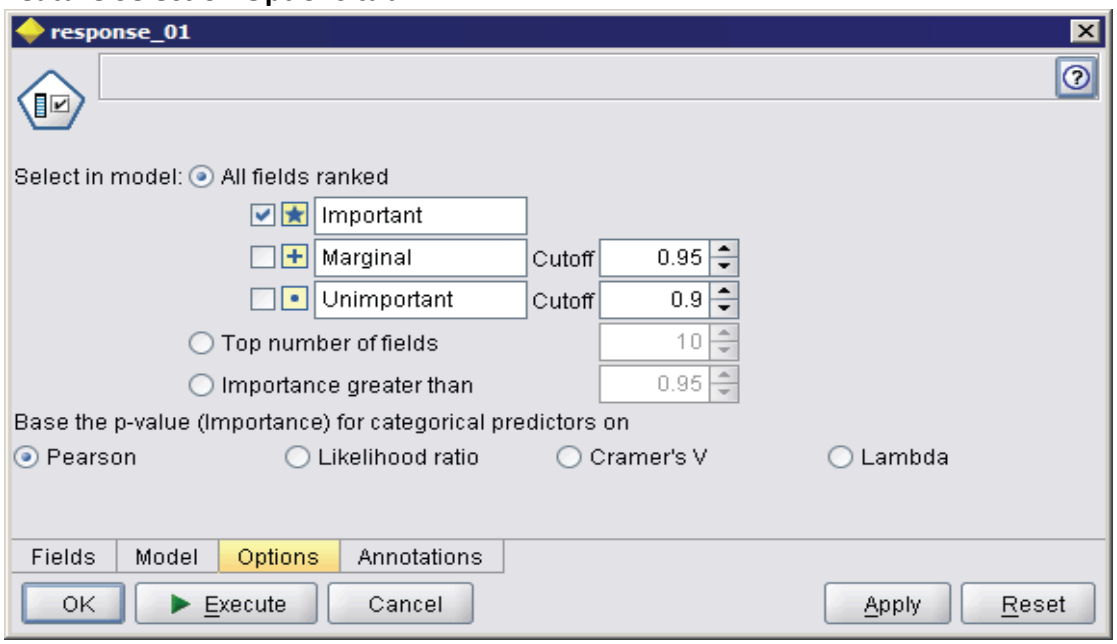
όπου:

p : είναι η τιμή $p\ value$ του κατάλληλου στατιστικού test της σχέσης μεταξύ της υποψήφιας μεταβλητής πρόβλεψης και της μεταβλητής στόχο .

Στην εφαρμογή χρησιμοποιήσαμε τιμή $p\ value$ βασισμένη στο στατιστικό του Pearson, το Pearson chi-square το οποίο εξετάζει την ανεξαρτησία του στόχου και

της μεταβλητής πρόβλεψης χωρίς να δείχνει τη δύναμη ή την κατεύθυνση οποιασδήποτε υπάρχουσας σχέσης .

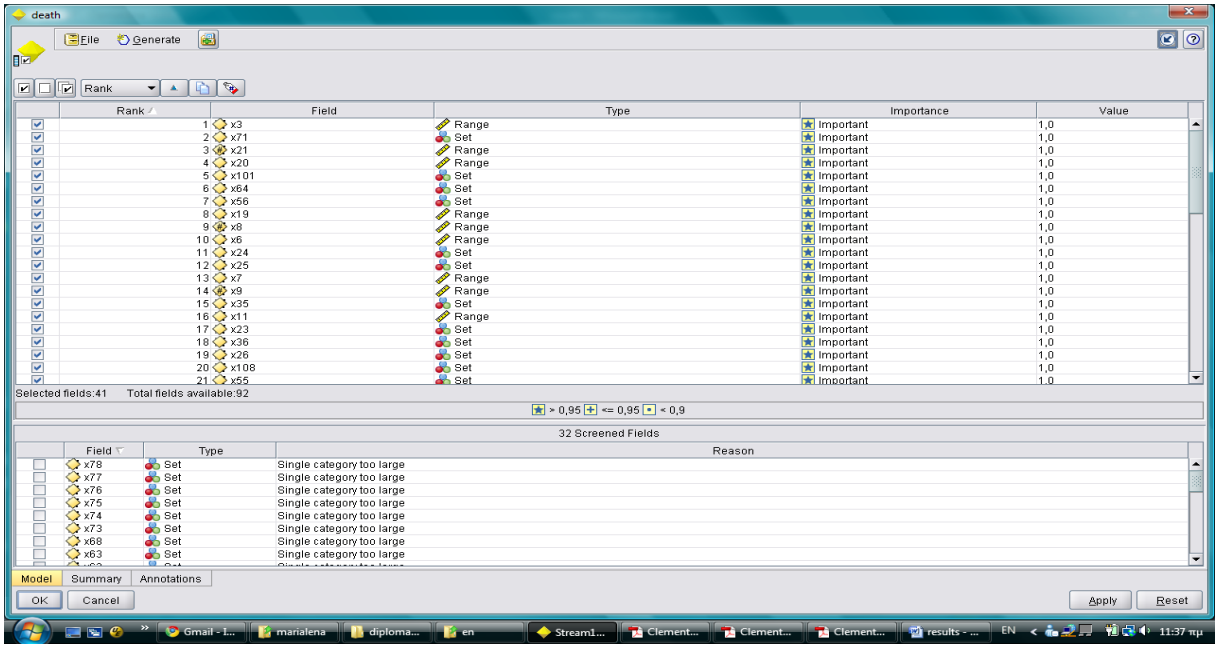
Feature Selection Options tab :

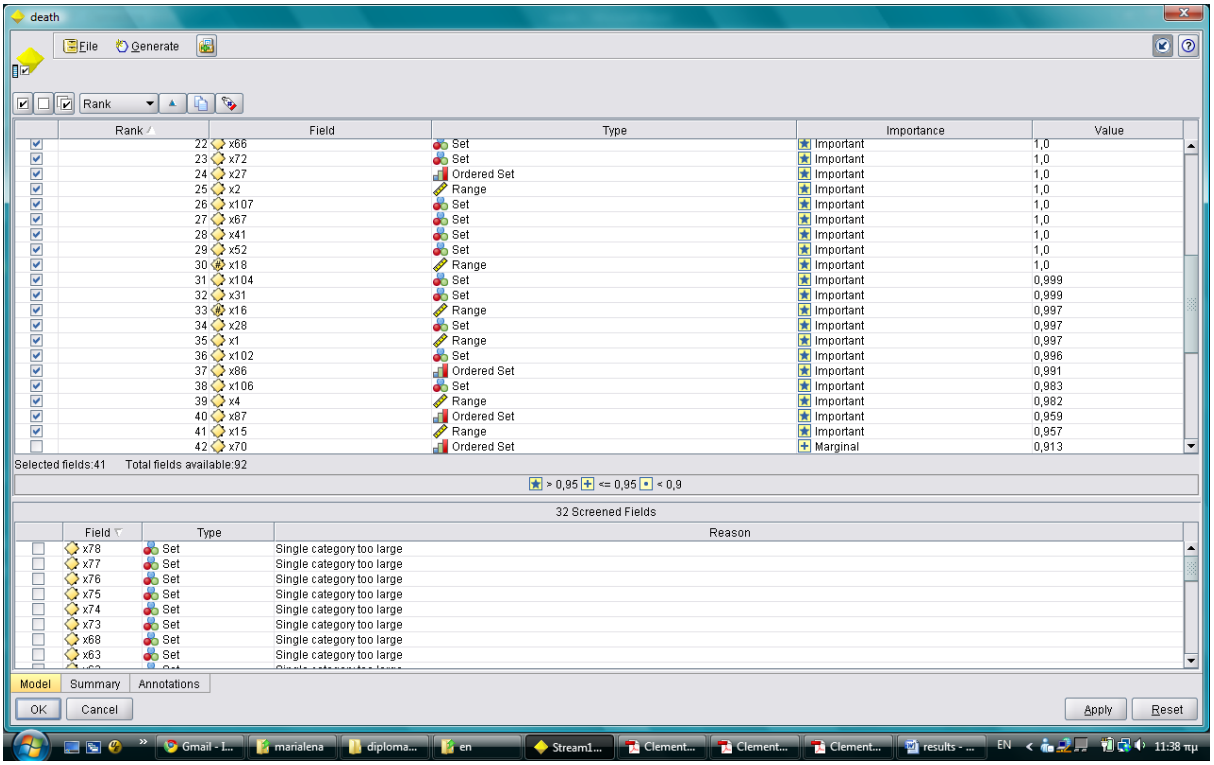


Μετά από όλη αυτή την διαδικασία επιλογής μεταβλητών παρατηρούμε ότι οι επεξηγηματικές μεταβλητές από 92 που ήταν αρχικά μειώνονται στις 41.

Για επίπεδο σημαντικότητας α=0.05 οι σημαντικές 41 αυτές μεταβλητές με $p\ value = 1{,}0$ παρουσιάζονται παρακάτω :

Feature Selection model results :





3.1.1.1 C&RT

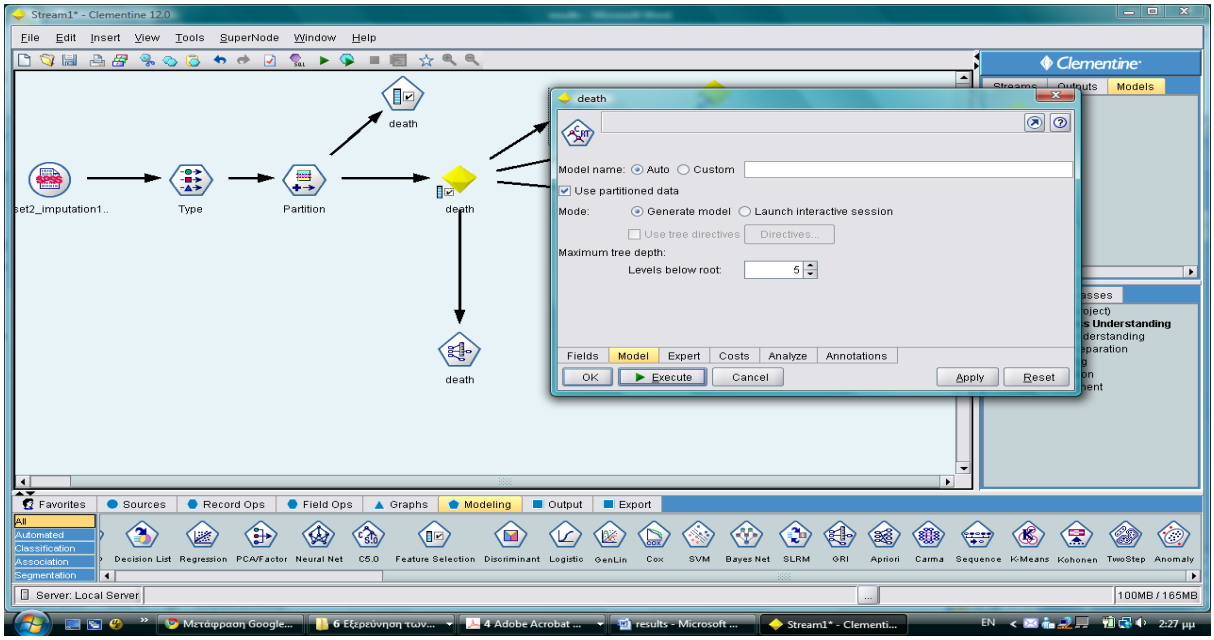
Έπειτα, από αυτά τα βήματα τοποθετούμε στο stream canvas ένα C&R Tree node 

.

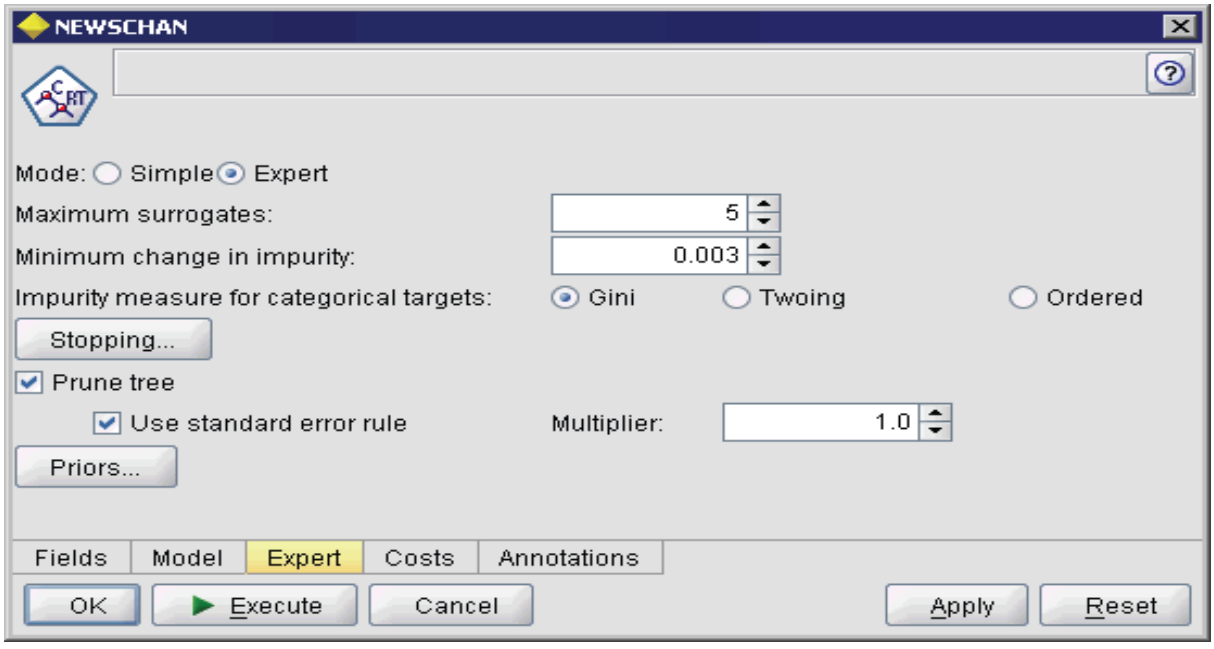
Συνδέουμε ένα C&R Tree node στο Feature selection model  . Ο Classification and Regression (C&R) Tree node παράγει ένα δέντρο απόφασης το οποίο μας επιτρέπει να προβλέψουμε ή να ταξινομήσουμε τις μελλοντικές παρατηρήσεις. Η μέθοδος χρησιμοποιεί επαναληπτικό διαμερισμό για να διασπάσει τις εγγραφές από το training set σε τμήματα ελαχιστοποιώντας τη μη καθαρότητα σε κάθε βήμα. Ένας κόμβος θεωρείται καθαρός εάν το 100% των περιπτώσεων στον κόμβο βρίσκονται μέσα σε μία συγκεκριμένη κατηγορία του πεδίου στόχος.

Στο Model tab επιλέγουμε το use paritioned data και το generate model. Επίσης, σαν μέγιστο βάθος δέντρου (max depth tree) αφήνουμε το default το οποίο είναι καθορισμένο στα 5 επίπεδα κάτω από τον αρχικό κόμβο ρίζα.

Ρύθμιση των επιλογών του μοντέλου :



Ρυθμίζοντας τις expert επιλογές :



Στο Expert tab στο δικό μας dialog box αλλάζουμε λίγο τις default εξ'ορισμού ρυθμίσεις και επιλέγουμε μόνο Expert mode και Prune tree. Ως maximum surrogates αφήνουμε το 5 (default). Η τιμή μη καθαρότητας είναι στο 0.003 (default). Εάν αυξήσουμε αυτή την τιμή τείνει να καταλήγει σ' ένα πιο απλό δέντρο, έχοντας σαν συνέπεια να αποτρέπονται διαχωρισμοί οι οποίοι έχουν σαν αποτέλεσμα μόνο μια μικρή βελτίωση.

Στη δική μας εφαρμογή, ρυθμίζουμε την τιμή μη καθαρότητας στο 0.0001. Επιλέγουμε σαν μέτρο μη καθαρότητας τη μία φορά το Gini κριτήριο και την άλλη φορά το Twoing κριτήριο και θα συγκρίνουμε τα αποτελέσματα που θα προκύψουν.

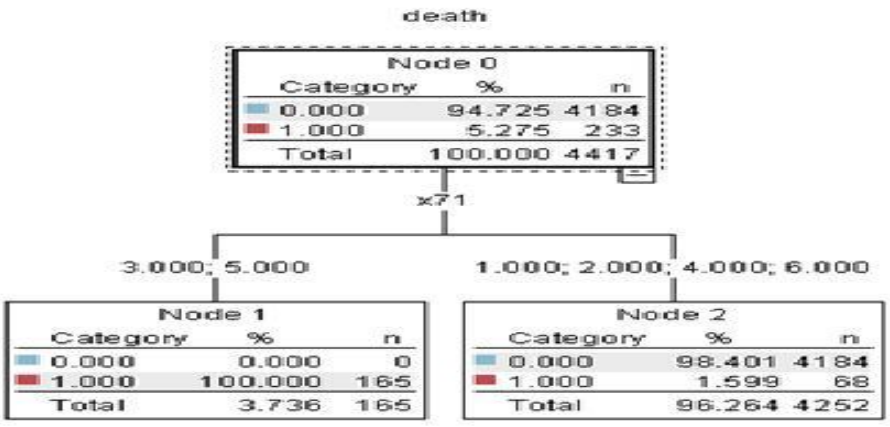
Για να καθορίσουμε τα κριτήρια διακοπής (stopping criteria) επιλέγουμε το Stopping στο Expert tab. Στο Stopping Criteria dialog επιλέγουμε use percentage αφήνοντας τις default ρυθμίσεις(2.1) :

- ελάχιστος αριθμός εγγραφών στο μητρικό κλαδί : 2
- ελάχιστος αριθμός εγγραφών στο θυγατρικό κλαδί : 1

Κριτήριο Gini

Ακολουθούν τα αποτελέσματα του C&RT χρησιμοποιώντας το Gini κριτήριο, με τις σημαντικές 41 μεταβλητές .

Προκύπτει το παρακάτω δέντρο ταξινόμησης χρησιμοποιώντας το Gini κριτήριο :



Σχήμα 14: Δέντρο Ταξινόμησης

Παρατηρούμε ότι :

- Για n=165 υποδείγματα το y=1 δηλαδή θάνατος όταν η X_{71} παίρνει τις τιμές 3; 5 όπου :

X_{71}	**Προορισμός μετά το ΤΕΠ** Άλλο νοσοκομείο; Κλινική; Μ.Α.Φ; Μ.Ε.Θ; Νεκροτομείο; Χειρουργείο

Άρα :

- Οι περισσότεροι από τους τραυματίες που οδηγούνται μετά το Τ.Ε.Π σε Μονάδα Αυξημένης Φροντίδας (Μ.Α.Φ), απεβιώνουν .
- Άμεσο επακόλουθο είναι ο θάνατος σε όσους τραυματίες οδηγούνται μετά το Τ.Ε.Π σε νεκροτομείο.

- Για n=4184 υποδείγματα το y=0 δηλαδή έζησαν όταν η X_{71} παίρνει τις τιμές 1; 2; 4 ;6 όπου :

X_{71}	**Προορισμός μετά το ΤΕΠ** Άλλο νοσοκομείο; Κλινική; Μ.Α.Φ; Μ.Ε.Θ; Νεκροτομείο; Χειρουργείο

Άρα:

- Οι περισσότεροι από τους τραυματίες που οδηγούνται μετά το Τ.Ε.Π σε κάποιο άλλο νοσοκομείο έχουν καλή έκβαση (επιβίωσαν)
- Οι περισσότεροι από τους τραυματίες που οδηγούνται μετά το Τ.Ε.Π σε κάποια κλινική νοσοκομείο έχουν καλή έκβαση (επιβίωσαν)
- Οι περισσότεροι από τους τραυματίες που οδηγούνται μετά το Τ.Ε.Π σε Μονάδα Εντατικής Θεραπείας (Μ.Ε.Θ) έχουν καλή έκβαση (επιβίωσαν)
- Οι περισσότεροι από τους τραυματίες που οδηγούνται μετά το Τ.Ε.Π στο χειρουργείο έχουν καλή έκβαση (επιβίωσαν)

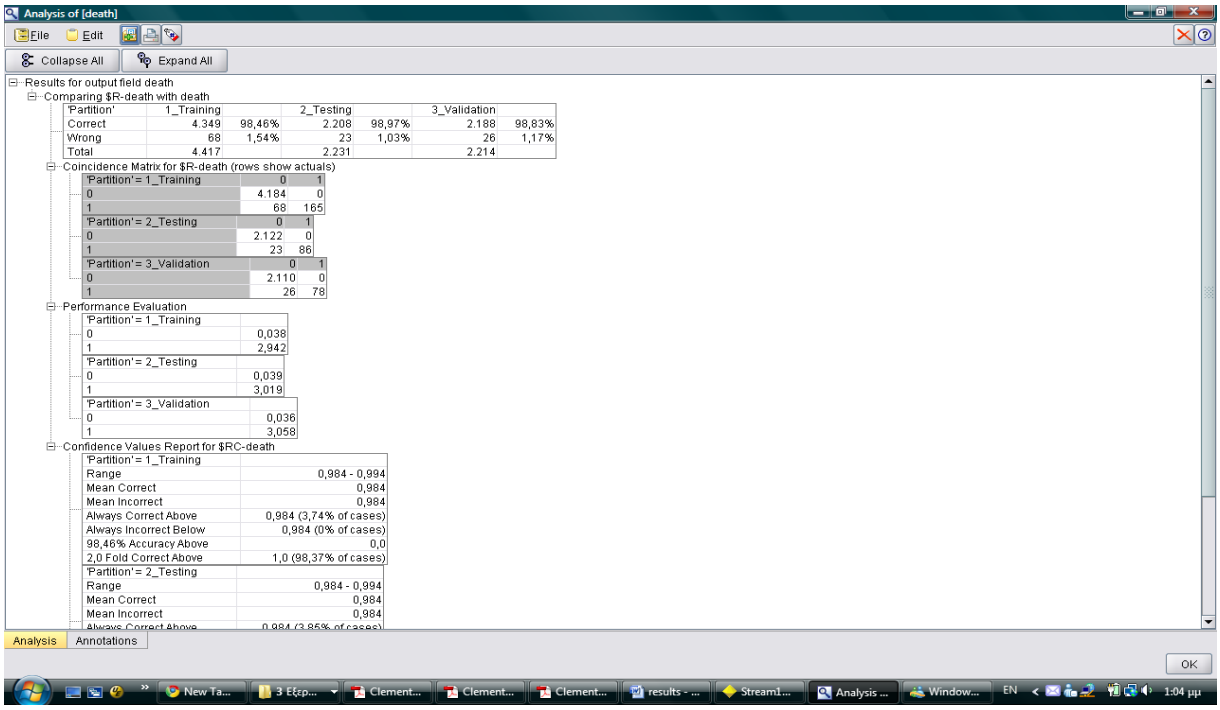
Αν συνδυάσουμε τους κανόνες ταξινόμησης συμπεραίνουμε ότι οι περισσότεροι από τους τραυματίες που οδηγούνται μετά το Τ.Ε.Π σε Μονάδα Αυξημένης Φροντίδας (Μ.Α.Φ), πεθαίνουν και αυτό το εύρημα είναι σημαντικό αν αναλογιστούμε ότι στην ανάλυση συμπεριλαμβάνονται και όσοι μεταφέρθηκαν σε

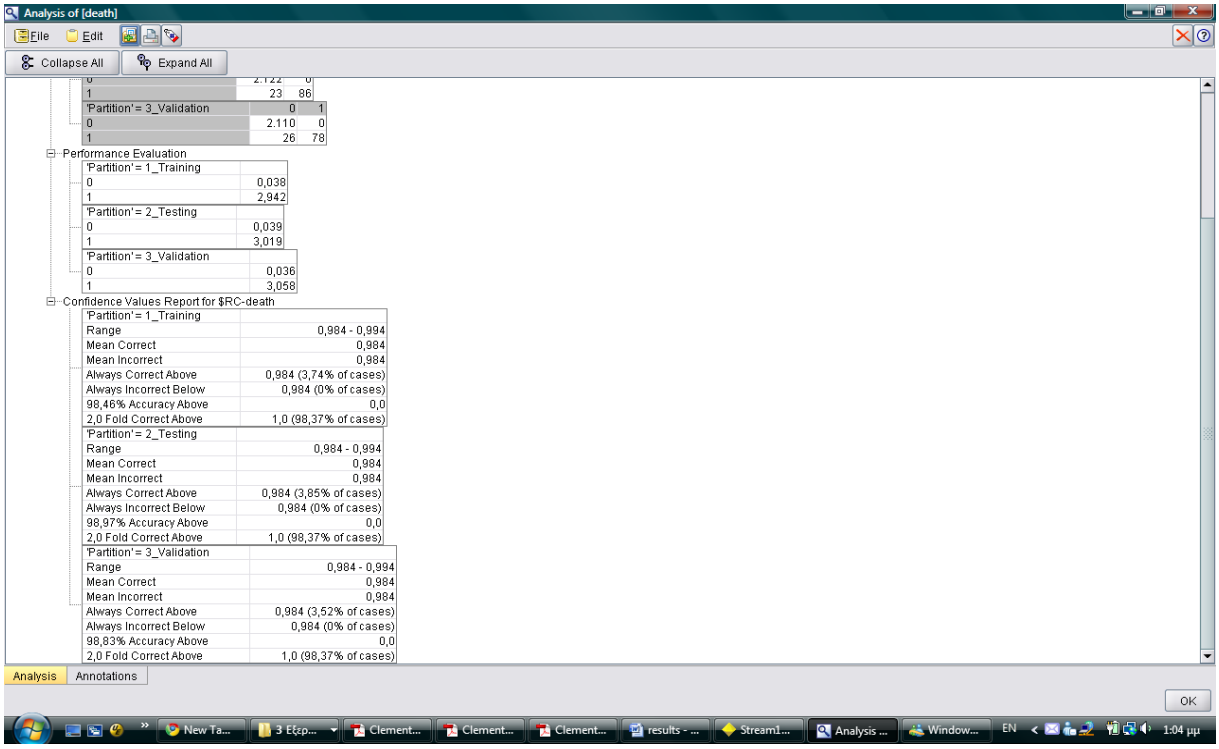
Μονάδα Εντατικής Θεραπείας (Μ.Ε.Θ) οι οποίοι όπως φαίνεται είχαν καλύτερη έκβαση (επιβίωσαν) καθώς είναι πιθανόν να έτυχαν καλύτερης περίθαλψης.

Παρατήρησεις:

- Αντίστοιχα αποτελέσματα με μόνη διαφορά τον αριθμό των υποδειγμάτων (για n=228 υποδείγματα το y=1 δηλαδή θάνατος όταν η X_{71} παίρνει τις τιμές 3; 5 για n=5875 υποδείγματα το y=0, δηλαδή έζησαν όταν η X_{71} παίρνει τις τιμές 1; 2; 4; 6) έδειξε η προηγούμενη μελέτη που πραγματοποιήθηκε στην ίδια βάση δεδομένων με τη διαφορά ότι εκεί χρησιμοποιήθηκε μόνο training και test set ενώ σ' αυτήν την εφαρμογή χρησιμοποιήσαμε και validation set.
- Παρατηρούμε ότι η μεταβλητή X_{71} δεν επηρεάζει την έκβαση όταν παίρνει τιμές 1; 2; 4; 6. Δηλαδή, στον αρχικό κόμβο-ρίζα είχαμε n=4184 υποδείγματα με y=0 (έζησαν) και τόσα παραμένουν και μετά την κατασκευή του δέντρου.

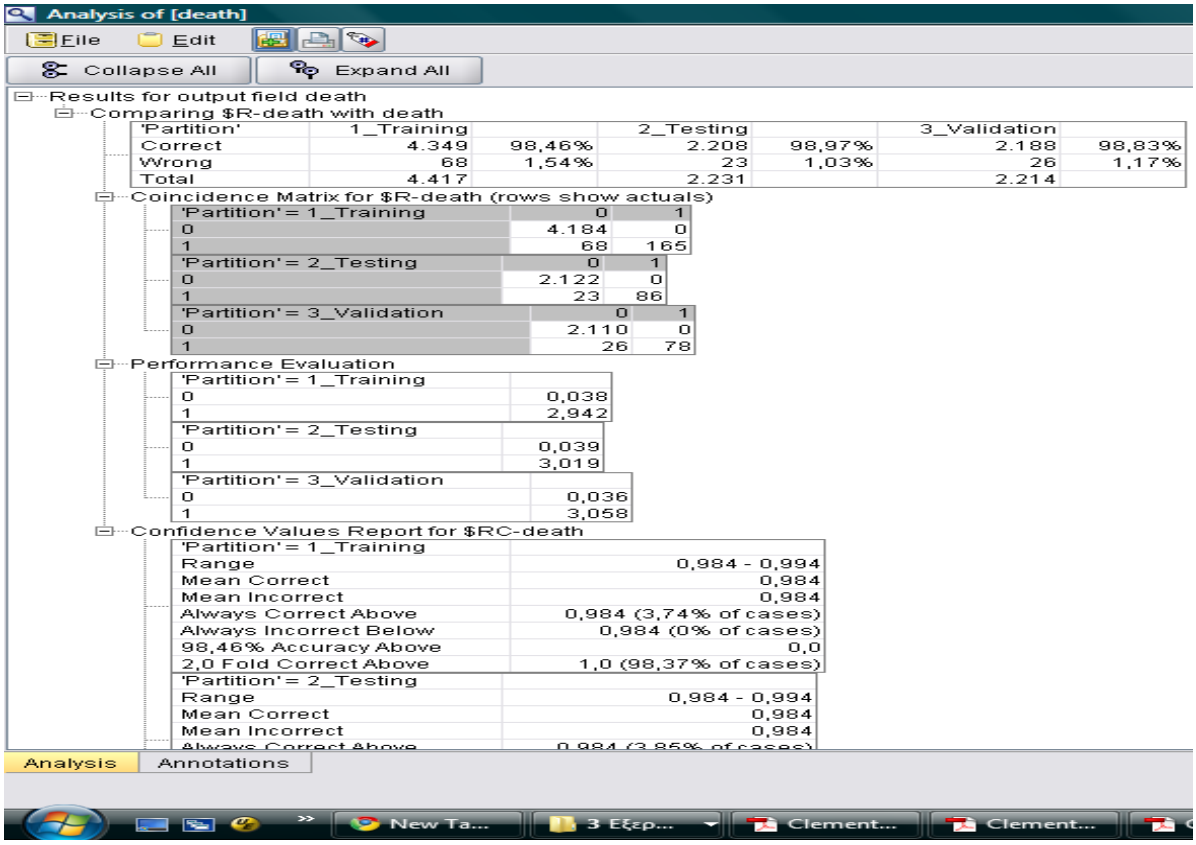
Για την αξιολόγηση του αλγόριθμου (με χρήση του Gini κριτηρίου) έχουμε τα πιο κάτω αποτελέσματα:





Παρατηρούμε ότι :

- συγκρίνοντας τα $R death με death δηλαδή τα προβλεπόμενα y με τα y από τα 4.417 υποδείγματα στο training set ($\approx 50\% * 8.862$) τα 4.349 είναι σωστά ταξινομημένα περίπου το 98.46%.
- συγκρίνοντας τα $R death με death δηλαδή τα προβλεπόμενα y με τα y από τα 2.231 υποδείγματα στο test set ($\approx 25\% * 8.862$) τα 2.208 είναι σωστά ταξινομημένα περίπου το 98.97%.
- συγκρίνοντας τα $R death με death δηλαδή τα προβλεπόμενα y με τα y από τα 2.214 υποδείγματα στο validation set ($\approx 25\% * 8.862$) τα 2.188 είναι σωστά ταξινομημένα περίπου το 98.83%.



Στην ιατρική για την εκτίμηση της λειτουργίας μιας διαγνωστικής διαδικασίας χρησιμοποιούνται :

- τα μέτρα ευαισθησία,
- ειδικότητα,
- θετική προγνωστική αξία (Θ.Π.Α),
- αρνητική προγνωστική αξία (Α.Π.Α)

Ωστόσο, σε πολλές εφαρμογές εξόρυξης δεδομένων το κριτήριο αξολόγησης είναι η ακρίβεια (accuracy), δηλαδή το ποσοστό των σωστά ταξινομημένων περιπτώσεων που προκύπτουν από τον αλγόριθμο όπως αναφερθήκαμε και παραπάνω.

Μέτρα αξιολόγησης διαγνωστικών test:

Διαγνωστικό test	**Αποτέλεσμα**	**Αποτέλεσμα**	
	Θάνατος (+)	Καλή έκβαση(-)	Σύνολο
Θετικό (+)	a (πραγματικά θετικά)	b (λανθασμένα θετικά)	a+b
Αρνητικό (-)	c (λανθασμένα αρνητικά)	d (πραγματικά αρνητικά)	c+d
Σύνολο	a+c	b+d	a+b+c+d

όπου :

- ευαισθησία = $\frac{a}{a+c}$
- ειδικότητα = $\frac{d}{b+d}$
- θετική προγνωστική αξία (Θ.Π.Α) = $\frac{a}{a+b}$
- αρνητική προγνωστική αξία (Α.Π.Α) = $\frac{d}{d+c}$
- ακρίβεια = $\frac{a+d}{a+c+b+d}$

Θα υπολογίσουμε τα μέτρα αυτά αξιολόγησης στα σύνολα εκπαίδευσης, επαλήθευσης και ελέγχου ξεχωριστά και θα τα απεικονίσουμε αναλυτικά σε μορφή ποσοστών .

Για το σύνολο εκπαίδευσης:

		0 (-)	1(+)
0	Καλή έκβαση(-)	4184	0
1	Θάνατος (+)	68	165

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 165
- b = (+ -) =λανθασμένα θετικά =68
- c =(- +) =λανθασμένα αρνητικά=0
- d = (- -) = πραγματικά αρνητικά=4.184

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 100 (%)
- ειδικότητα =98.4 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 70.8 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 100 (%)
- ακρίβεια = 98.46 (%)

Για το σύνολο ελέγχου:

		0 (-)	1(+)
0	Καλή έκβαση(-)	2122	0
1	Θάνατος (+)	23	86

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 86
- b = (+ -) =λανθασμένα θετικά =23
- c =(- +) =λανθασμένα αρνητικά=0
- d = (- -) = πραγματικά αρνητικά=2122

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία =100 (%)
- ειδικότητα =98.92 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 78.89 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) =100 (%)
- ακρίβεια = 98.97 (%)

Για το σύνολο επαλήθευσης :

		0 (-)	1(+)
0	Καλή έκβαση(-)	2110	0
1	Θάνατος (+)	26	78

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 78
- b = (+ -) =λανθασμένα θετικά =26
- c =(- +) =λανθασμένα αρνητικά=0
- d = (- -) = πραγματικά αρνητικά=2110

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία =100 (%)
- ειδικότητα =98.78 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 75 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) =100 (%)
- ακρίβεια = 98.83 (%)

Παρατήρηση :

Αντίστοιχα αποτελέσματα, με μικρές διαφορές στα ποσοστά των μέτρων αξιολόγησης διαγνωστικών tests , λάβαμε και από την προηγούμενη μελέτη που πραγματοποιήθηκε στην ίδια βάση δεδομένων όπου εκεί χρησιμοποιήθηκε μόνο training και test set ενώ στην δική μας εφαρμογή χρησιμοποιήθηκε και validation set.
Οι διαφορές στα ποσοστά τω μέτρων αξιολόγησης διαγνωστικών tests είναι της τάξης του 0.5% με 3% .

Κριτήριο Twoing:

Επιλέγουμε σαν μέτρο μη καθαρότητας αυτή τη φορά το Twoing κριτήριο και θα συγκρίνουμε τα αποτελέσματα που θα προκύψουν με τα αντίστοιχα όταν επιλέξαμε το Gini.

Παρατηρούμε ότι το δέντρο ταξινόμησης που προκύπτει με τη χρήση του Twoing κριτηρίου είναι ακριβώς το ίδιο με αυτό όταν χρησιμοποιήσαμε το Gini κριτήριο και φυσικά οι κανόνες ταξινόμησης που προκύπτουν είναι αντίστοιχα οι ίδιοι.
Συμπεραματικά, η αξιολόγηση του μοντέλου με το Twoing κριτήριο είναι ακριβώς η ίδια με την αξιολόγηση που πραγματοποίηθηκε παραπάνω με το Gini κριτήριο.
Φυσικά και τα μέτρα αξιολόγησης διαγνωστικών tests έχουν τις ίδιες ακριβώς τιμές.

Γενικές παρατηρήσεις :

- Παρατηρούμε ότι για τον C&RT αλγόριθμο τα κριτήρια αξιολόγησης για το test ελέγχου είναι υψηλότερα συγκριτικά με τα κριτήρια αξιολόγησης για το test επαλήθευσης και αυτά με της σειρά τους είναι υψηλότερα σε σχέση με αυτά του test εκπαίδευσης.

Στην προηγούμενη μελέτη που πραγματοποιήθηκε μόνο με σύνολα εκπαίδευσης και ελέγχου τα κριτήρια αξιολόγησης ήταν υψηλότερα για το test εκπαίδευσης για όλους τους αλγορίθμους .

- Παρατηρούμε ότι υπάρχει ταύτιση αποτελεσμάτων όσον αφορά την αξιολόγηση των ιατρικών μέτρων και τω μεθόδων εξόρυξης δεδομένων. Συγκεκριμένα, η ακρίβεια δηλαδή το ποσοστό των σωστά ταξινομημένων περιπτώσεων που προκύπτουν από τον αλγόριθμο στα σύνολα εκπαίδευσης, ελέγχου και επαλήθευσης είναι ίδιο με το ποσοστό ακρίβειας το οποίο χρησιμοποιείται ως μέτρο αξιολόγησης για την εκτίμηση της λειτουργίας μιας διαγνωστικής διαδικασίας στην Ιατρική.

- Παρατηρούμε ότι ενώ το δέντρο ταξινόμησης το οποίο προκύπτει τρέχοντας τον C&RT αλγόριθμο κάνοντας χρήση είτε του Gini κριτηρίου είτε του Twoing είναι απλό ο υπολογιστικός χρόνος που χρειάστηκε για να προκύψουν τα αποτελέσματα ήταν σχετικά μεγάλος($\approx 60'$ για κάθε C&RT αλγόριθμο μία φορά με Gini και την άλλη με Twoing).

3.1.1.2 Νευρωνικά Δίκτυα

Μέθοδοι MLP και RBFN

Με τον ίδιο ακριβώς τρόπο που έχουμε περιγράψει, δηλαδή αρχικά την κατασκευή και στη συνέχεια τα αποτελέσματα της ανάλυσης και αξιολόγησης ενός δέντρου απόφασης, συνεχίζουμε τώρα στην προσαρμογή ενός νευρωνικού δικτύου με τις μεθόδους MLP και RBFN στο Feature selection model. Στο τέλος, κατασκευάζουμε ένα διάγραμμα αξιολόγησης (evaluation chart) με τη βοήθεια των καμπύλων ROC, έτσι ώστε να μπορέσουμε να συγκρίνουμε τις επιδόσεις τω μοντέλων.

Στη μελέτη μας, το στρώμα εξόδου όλων των μοντέλων των Νευρωνικών Δικτύων αποτελείται από ένα νευρώνα, 1 για τη θετική ανταπόκριση ότι ο ασθενής

απεβίωσε και 0 για την αρνητική ότι ο ασθενής επιβίωσε. Χρησιμοποιήσαμε ένα κρυφό επίπεδο σε όλα τα νευρωνικά μοντέλα στην εφαρμογή μας. Ο αρχικός αριθμός των κρυφών μονάδων ορίστηκε σε 15 στα MLP και RBFN δίκτυα και ο μέγιστος αριθμός των εποχών εκπαίδευσης (training epochs) στις 600.

Η εκπαίδευση ενός πολυστρωματικού νευρωνικού δικτύου χρησιμοποιεί μια μέθοδο που ονομάζεται back propagation of error, με βάση το γενικεύμενο κανόνα Δέλτα.

Κάθε εγγραφή που παρουσιάζεται στο δίκτυο κατά τη διάρκεια της κατάρτισης (training), δίνει πληροφορίες (με τη μορφή των input fields) οι οποίες τροφοδοτούν το δίκτυο να έτσι ώστε να δημιουργήσει πρόβλεψη μέσω του στρώματος εξόδου. Συγκρίνουμε αυτή την πρόβλεψη με την πραγματική output value από το μητρώο εκπαίδευσης και η διαφορά μεταξύ της προβλεπόμενης και πραγματικής εξόδου αναπαράγεται προς τα πίσω μέσω του δικτύου έτσι ώστε να προσαρμόζουν τα βάρη σύνδεσης για τη βελτίωση της πρόβλεψη για παρόμοια εφαρμογές.

Όσον αφορά τον αριθμό των κρυφών μονάδων, η μέθοδος του κλαδέματος που χρησιμοποιήθηκε, εξαλείφει τα βάρη τα οποία είναι χαμηλότερα από το όριο, 0.05 στη δική μας εφαρμογή.
Τέλος, και στις δύο μεθόδους, MLP και RBFN, το δίκτυο είχε εκπαιδευτεί για το training set και η ακρίβεια υπολογίζεται με βάση το σύνολο δοκιμής (test set).

Θα υπολογίσουμε τα μέτρα αξιολόγησης στα σύνολα εκπαίδευσης, επαλήθευσης και ελέγχου ξεχωριστά και θα τα απεικονίσουμε αναλυτικά σε μορφή ποσοστών .

Για το σύνολο εκπαίδευσης:

		0 (-)	1(+)
0	Καλή έκβαση(-)	4289	5
1	Θάνατος (+)	31	38

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 38
- b = (+ -) =λανθασμένα θετικά =31

- c =(- +) =λανθασμένα αρνητικά=5
- d = (- -) = πραγματικά αρνητικά=4.289

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 88.37(%)
- ειδικότητα =99.28 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 55.07 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 99.88 (%)
- ακρίβεια = 99.17(%)

Για το σύνολο ελέγχου:

		0 (-)	1(+)
0	Καλή έκβαση(-)	2109	2
1	Θάνατος (+)	13	10

Έτσι έχουμε:

- a=(+ +)= πραγματικά θετικά = 10
- b = (+ -) =λανθασμένα θετικά =13
- c =(- +) =λανθασμένα αρνητικά=2
- d = (- -) = πραγματικά αρνητικά=2109

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία =83.33 (%)
- ειδικότητα =99.39 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 43.48 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) =99.90 (%)
- ακρίβεια = 99.29 (%)

Για το σύνολο επαλήθευσης :

		0 (-)	1(+)
0	Καλή έκβαση(-)	2304	2
1	Θάνατος (+)	44	15

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 15
- b = (+ -) =λανθασμένα θετικά =44
- c =(- +) =λανθασμένα αρνητικά=2
- d = (- -) = πραγματικά αρνητικά=2304

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία =88.24 (%)
- ειδικότητα =98.13 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 25.43 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) =99.91 (%)
- ακρίβεια = 98.05 (%)

Για την αξιολόγηση των MLP και RBFN μεθόδων (με χρήση του Gini κριτηρίου) έχουμε τα πιο κάτω αποτελέσματα:

Παρατηρούμε ότι :

MLP Μέθοδος:

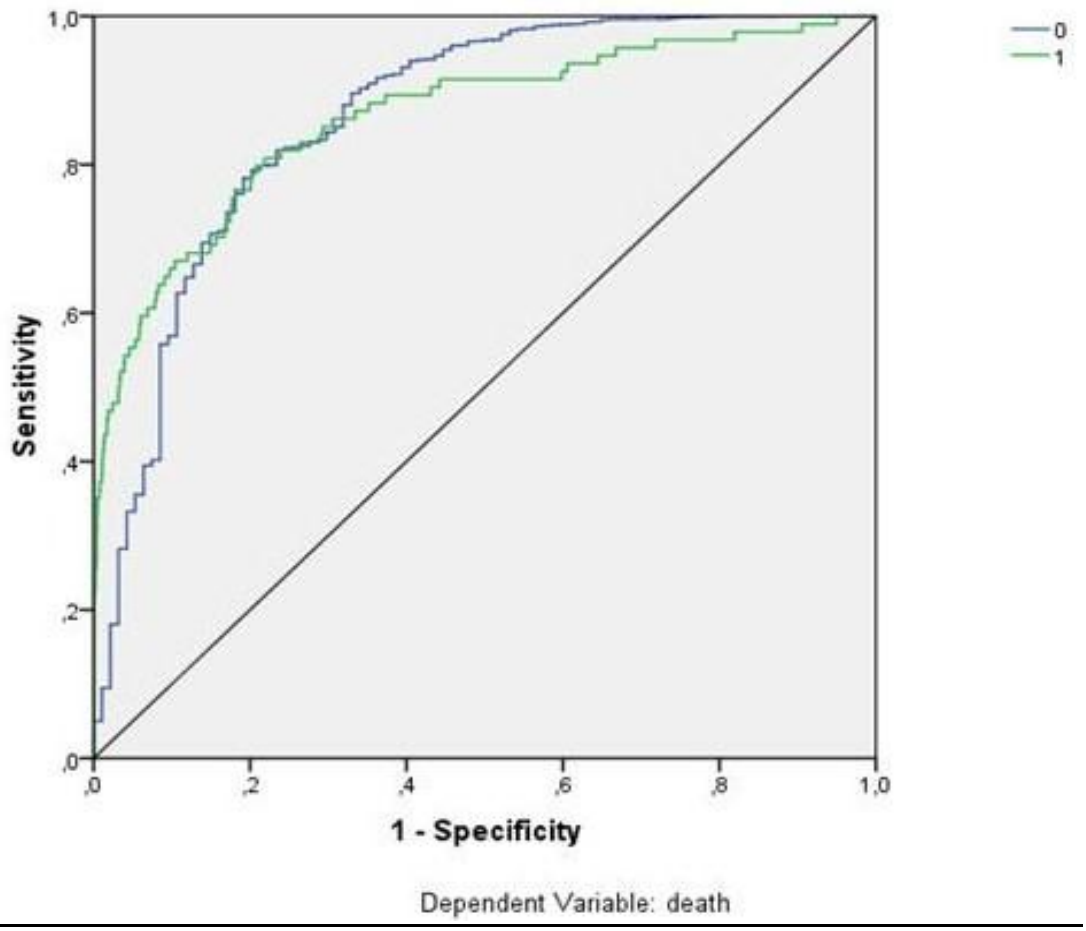
- στο training set $(\approx 50\% * 8.862)$, συγκρίνοντας τα $R death με death, δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 99.37%.
- στο test set $(\approx 25\% * 8.862)$, συγκρίνοντας τα $R death με death, δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 99.13%.
- στο validation set $(\approx 25\% * 8.862)$, συγκρίνοντας τα $R death με death δηλαδή, τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 99.45%.

RBFN Μέθοδος:

- στο training set $(\approx 50\% * 8.862)$, συγκρίνοντας τα $R death με death, δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 98.52%.

- στο test set $(\approx 25\% * 8.862)$, συγκρίνοντας τα $R death με death, δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 98.74%.
- στο validation set $(\approx 25\% * 8.862)$, συγκρίνοντας τα $R death με death δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 98.51%.

Ακολουθεί το διάγραμμα αξιολόγησης (evaluation chart) μέσω της μεθοδολογίας των καμπύλων ROC. Παρατηρούμε ότι και οι δύο μέθοδοι MLP και RBFN έχουν μεγάλες τιμές AUC (Area Under the Curve), περίπου στο 1 και αυτό φανερώνει εξαιρετική απόδοση των υποδειγμάτων και μια πολύ καλή διακριτική ικανότητα για την έκβαση του ασθενούς (ζωής ή θανάτου).



Γραφική 9: Average ROC Curve of NNs

3.1.1.3 Λογιστική Παλινδρόμηση

Μία τρίτη και τελευταία τεχνική που χρησιμοποιούμε στη συγκεκριμένη εφαρμογή με την ίδια διαδικασία όπως πριν, είναι αυτός της Λογιστικής παλινδρόμησης (Logistic Regression).

Θα υπολογίσουμε τα μέτρα αξιολόγησης στα σύνολα εκπαίδευσης, επαλήθευσης και ελέγχου ξεχωριστά και θα τα απεικονίσουμε αναλυτικά σε μορφή ποσοστών .

Για το σύνολο εκπαίδευσης:

		0 (-)	1(+)
0	Καλή έκβαση(-)	4204	12
1	Θάνατος (+)	38	45

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 45
- b = (+ -) =λανθασμένα θετικά =38
- c =(- +) =λανθασμένα αρνητικά=12
- d = (- -) = πραγματικά αρνητικά=4.204

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 78.95(%)
- ειδικότητα =99.10 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 54.21 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 99.72 (%)
- ακρίβεια = 98.84(%)

Για το σύνολο ελέγχου:

		0 (-)	1(+)
0	Καλή έκβαση(-)	2301	13
1	Θάνατος (+)	11	24

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 24
- b = (+ -) =λανθασμένα θετικά =11

- c =(- +) =λανθασμένα αρνητικά=13
- d = (- -) = πραγματικά αρνητικά=2.301

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία =64.86 (%)
- ειδικότητα =99.52 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 68.57 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) =99.44 (%)
- ακρίβεια = 98.98 (%)

Για το σύνολο επαλήθευσης :

		0 (-)	1(+)
0	Καλή έκβαση(-)	2191	2
1	Θάνατος (+)	10	11

Έτσι, έχουμε:

- a=(+ +)= πραγματικά θετικά = 11
- b = (+ -) =λανθασμένα θετικά =10
- c =(- +) =λανθασμένα αρνητικά=2
- d = (- -) = πραγματικά αρνητικά=2.191

και τα ποσοστά των μέτρων % είναι αναλυτικά :

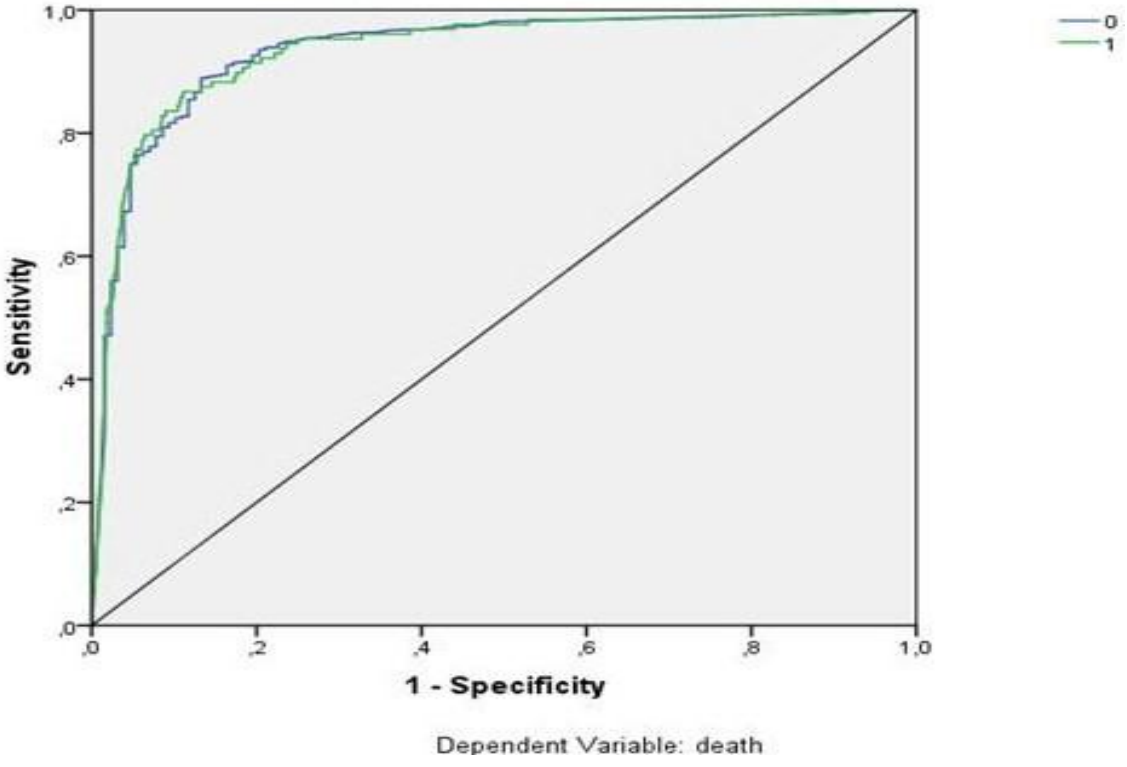
- ευαισθησία =84.62 (%)
- ειδικότητα =99.55 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 52.38 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) =99.91 (%)
- ακρίβεια = 99.46 (%)

Για την αξιολόγηση της λογιστικής παλινδρόμησης (με χρήση του Gini κριτηρίου) έχουμε τα πιο κάτω αποτελέσματα:

- στο training set $(\approx 50\% * 8.862)$, συγκρίνοντας τα $R death με death, δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 99%.

- στο test set $(\approx 25\% * 8.862)$, συγκρίνοντας τα $R death με death, δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 98.82%.
- στο validation set $(\approx 25\% * 8.862)$, συγκρίνοντας τα $R death με death δηλαδή τα προβλεπόμενα y με τα y, τα y υποδείγματα είναι σωστά ταξινομημένα περίπου το 99.05%.

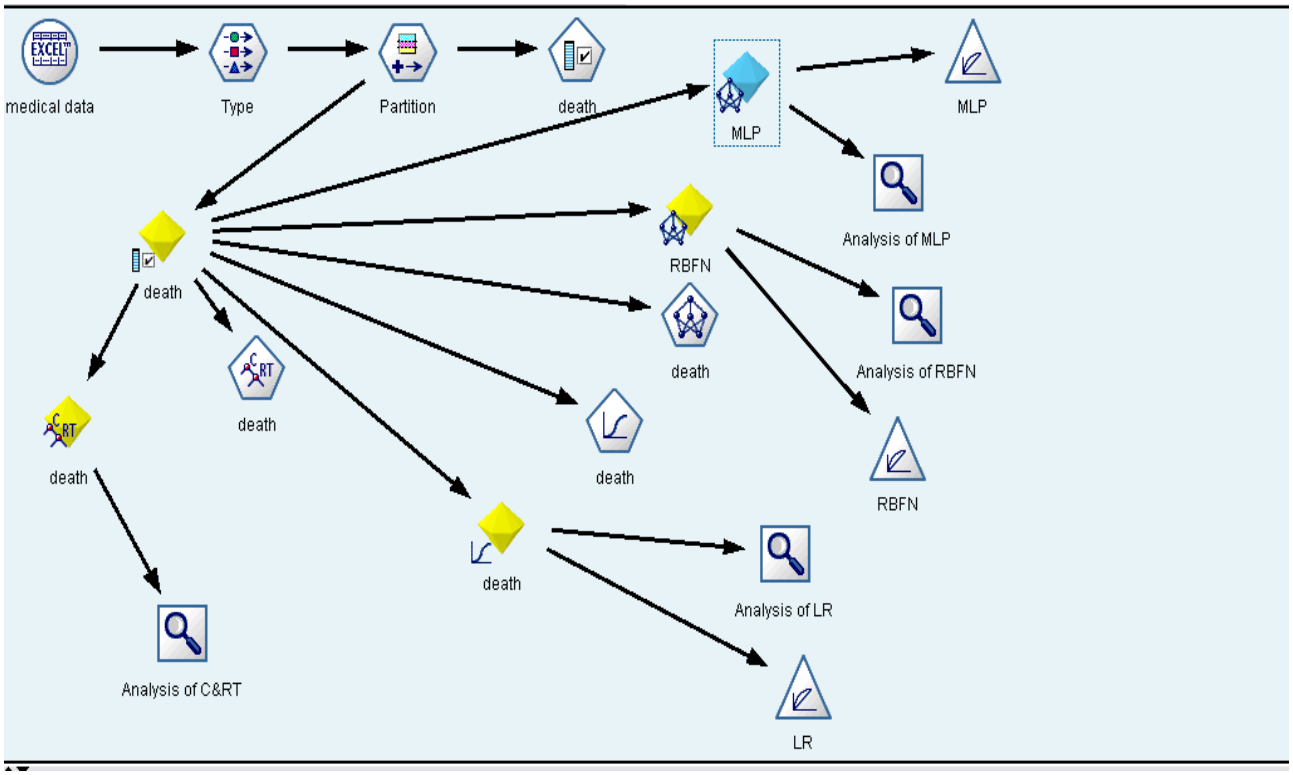
Ακολουθεί το διάγραμμα αξιολόγησης (evaluation chart) μέσω της μεθοδολογίας των καμπύλων ROC. Παρατηρούμε ότι η Λογιστική Παλινδρόμηση έχει μεγάλες τιμές AUC (Area Under the Curve), περίπου στο 1 και αυτό φανερώνει παρά πολύ καλή απόδοση των υποδειγμάτων και μια πολύ καλή διακριτική ικανότητα για την έκβαση του ασθενούς (ζωής ή θανάτου).



Γραφική 10: ROC Curve of LR

3.2 Γενικά Συμπεράσματα

Το τελικό Stream Canvas δείχνει όλη τη διαδικασία που ακολουθήσαμε στην εφαρμογή μας σε ιατρικά δεδομένα, όπως την περιγράψαμε αναλυτικά πιο πάνω.



Σχήμα 15 : Τελικό Stream Canvas

Μετά τη συλλογή και την ανάλυση των δεδομένων με σκοπό να προβλέψουμε την εξέλιξη της ζωής των ασθενών με τον αλγόριθμο C&RT, τις MLP και RBFN μεθόδους των Νευρωνικών Δικτύων και της Λογιστικής Παλινδρόμησης καταλήγουμε στα ακόλουθα συμπεράσματα:

- Η νευρωνική μέθοδος MLP είχε την υψηλότερη απόδοση σωστής ταξινόμησης των υποδειγμάτων-ασθενών και στα τρία σύνολα κατάρτισης, ελέγχου και επικύρωσης.
- Ως εκ τούτου μπορούμε να θεωρήσουμε το MLP μοντέλο ότι είναι το καλύτερο σε σύγκριση με τη μέθοδο RBFN, τον αλγόριθμο C&RT και το

μοντέλο της Λογιστικής Παλινδρόμησης τα οποία είχαν σχεδόν την ίδια συμπεριφορά.

- Το MLP Νευρωνικό Μοντέλο κατάφερε να ισορροπήσει με ακρίβεια μεταξύ των ψευδώς θετικών και των ψευδώς αρνητικά ποσοστών επιτυχίας (hit rates), δείχνοντας ότι ήταν το πιο ικανό για να καθορίσει τα σημαντικότερα χαρακτηριστικά ώστε να ταξινομήσει σωστά τους ασθενείς με αρνητικό αποτέλεσμα(απεβίωσαν) και εκείνων με θετική έκβαση(επιβίωσαν) στα τραύματα.
- Ο αλγόριθμος C&RT ξεπέρασε τα Νευρωνικά Δίκτυα και το μοντέλο της Λογιστικής Παλινδρόμησης όσον αφορά την ευαισθησία(sensitivity), τη θετική προγνωστική αξία (positive predictive value) και την αρνητική προγνωστική αξία (negative predictive value).
- Η ευαισθησία (sensitivity) και η specificity ratio, η θετική και αρνητική προγνωστική αξία καθώς και η συνολική ακρίβεια (overall accuracy) των Νευρωνικών Δικτύων ήταν υψηλότερες από τις αντίστοιχες τιμές στο μοντέλο της Λογιστικής Παλινδρόμησης.
- Τα διαγράμματα αξιολόγησης (evaluation charts) μέσω της μεθοδολογίας των καμπύλων ROC δείχνουν πολύ ξεκάθαρα ότι οι μεθόδοι MLP και RBFN τω Νευρωνικών Δικτύων και η Λογιστική Παλινδρόμηση έχουν μεγάλες τιμές AUC (Area Under the Curve), περίπου στο 1 και αυτό φανερώνει παρά πολύ καλή απόδοση των υποδειγμάτων και μια πολύ καλή διακριτική ικανότητα για την έκβαση του ασθενούς (ζωής ή θανάτου).
- Τα αποτελέσματα δείχνουν σαφώς ότι η ζωτικής σημασίας κατάσταση των ασθενών, για τους σκοπούς της πρόβλεψης, θα πρέπει να καθορίζεται από πολλές πτυχές και από ένα επαρκώς μεγάλο σύνολο μεταβλητών - εγραφών.
- Έτσι, πιθανή βελτίωση της επίδοσης του μοντέλου, μπορεί να είναι εφικτή από την αύξηση του αριθμού των προγνωστικών παραγόντων κινδύνου θανάτου οι οποίες περιλαμβάνονται στο dataset και ακόμη χρησιμοποιώντας άλλες εξελιγμένες μεθόδους, όπως οι support vector machines και άλλες ταξινομήσεις οι οποίες θα μπορούσαν να συνδυάστουν με τα νευρωνικά δίκτυα, προκειμένου να πετύχουμε ένα πιο επιτυχημένο μοντέλο πρόβλεψης της εξέλιξης της ζωής των ασθενών.

- Οι support vector machines, έχουν αποδείξει ότι προσφέρουν πολύ καλά αποτελέσματα ταξινόμησης, έτσι ένα από τα πιο ελπιδοφόρα θέματα για περαιτέρω μελέτη είναι η χρήση τους ως μια εναλλακτική μέθοδο για την υποστήριξη ιατρική ανακάλυψη της γνώσης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. P., Baker, B., O'Neil, W., Haddon and B., Long, (1974). The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care, *Journal of Trauma*, vol. 14 (3), pp. 187- 196.

2. C.M., Bishop, (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

3. G.A., Bors, (2001). *Introduction of the Radial Basis Function (RBF) Networks,* Department of Computer Science, University of York.

4. L., Breiman, J. H., Friedman, R. A., Olshen and C.J., Stone, (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

5. J., Buckley and I. James, (1979). Linear regression with censored data. *Biometrika*, Vol. 66, pp. 429-436.

6. R., Christensen, (1997). *Log-Linear Models and Logistic Regression*, New York, Springer-Verlac, second edition.

7. D., Collet, (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, Boca Raton.

8. D., Faraggi, B., Reiser, (2002). Estimation of the area under the ROC curve, *Statistics in Medicine*, Vol. 21, pp. 3093-3106.

9. D., Faraggi, and R., Simon, (1995). *A Neural Network Model for Survival Data. Statistics in medicine*, Vol. 14, pp. 73-82.

10. T., Fawcett: (2003). ROC Graphs: Notes and practical considerations for Data Mining researchers*, Intelligent Enterprise Technologies Laboratory*.

11. N., Fieller, (2007). *Medical Statistics: Survival Data, Course Booklet*. Department of Probability & Statistics, University of Sheffield.

12. M., Gönen, (2006). Receiver Operating Characteristic (ROC) Curves, SUGI 31 Proceedings, *Statistics and Data Analysis*, Paper 210-31.

13. K., Gurney, (1997). *An introduction to Neural Networks*, University of Sheffield.

14. T. Hastie, R. Tibshirani and J. Friedman, (2001*). The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, New York.

15. W.W., Hauck and A., Donner, (1997). Wald's test applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, Vol. 72, pp. 851-853.

16. S., Haykin, (1999). *Neural Networks, a comprehensive foundation*. McMaster University Hamilton, Ontario, Canada, second edition.

17. D.W., Hosmer, S., Lemeshow, (2000). *Applied Logistic Regression*, John Wiley and Sons, New York, second edition.

18. C., Kamath, (2009). *Scientific Data Mining, a Practical Perspective*. Philadelphia: Society for Industrial and Applied Mathematics.

19. K., Kawaguchi, (2000). *A multithreaded software model for backpropagation neural network applications*. [Chp. 2.4.4].

20. C., Koukouvinos, C., Parpoula and E.-M., Theodoraki, (2012). Classification methods and ROC analysis for outcome prediction of patients following injuries, *Int. J. Biomedical Engineering and Technology*, Vol. 8, No. 1, pp. 49-59.

21. D.T., Larose, (2005). *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley and Sons, Hoboken, New Jersey.

22. K., Liestol, P.K., Andersen, and U., Andersen, (1994). Survival analysis and neural nets*, Statistics in Medecine*, Vol. 13, pp. 1189-1200.

23. E., Marubini and G.M., Valecchi, (2004). *Analysing Survival Data from Clinical Trials and Observational Studies*.

24. S.W., Menard, *Applied Logistic Regression Analysis*. Second Edition, Sage Publications, London.

25. L., Noriega, (2005). *Multilayer Perceptron Tutorial*. School of Computing, Staffordshire University.

26. P.M., Pardalos, V.L., Boginski and A., Vazacopoulos, (2007). *Data Mining in Biomedicine*, Springer, New York.

27. R.L., Pearson, (1983). Karl Pearson and the chi-squared test, *International Statistical Review*, vol. 51, pp. 59-72.

28. R., Rojas, (1996). *Neural Networks*. Springer-Verlag, Berlin.

29. M., Suka, Shinichi, O., Ichimura, T., Yoshida, K. and Takezawa J. (2004). Comparison of Proportional Hazard Model and Neural Network Models in a real data set of intensive care unit patients. *Studies in Health Technology and Informatics*, Vol. 107, pp. 741-745.

30. T., Zhang, R., Ramakrishnan and M., Livny, (1996). An effcient data clustering method for very large databases, In Proc. of the ACM SIGMOD *International, Conference on Management of Data (SIGMOD)*, pp. 103-114.

31. X.-H., Zhou, N.A., Obuchowski, D.K., McClish, (2002). *Statistical Methods in Diagnostic Medicine*. Chapter 2, Wiley, New York.

32. Χ., Καρώνη, (2005). *Μοντέλα Αξιοπιστίας και Επιβίωσης*. Ε.Μ.Π.

33. Χ., Νάκας, (2002): *«Προσαρμογή καμπύλης, στατιστική συμπερασματολογία, επεκτάσεις και εφαρμογές στην ανάλυση των καμπυλών λειτουργικού χαρακτηριστικού δέκτη (ROC)».* Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

34. Γ., Ρίζος, (1996). *Τεχνητά νευρωνικά δίκτυα*. Αθήνα: Εκδόσεις Νέων Τεχνολογιών.