Διπλωματική Εργασία **:** *Πολυκριτήρια Αξιολόγηση Συστημάτων Ταξινόμησης*

Επιβλέπων καθηγητής : Δούμπος Μιχάλης
Εκπόνηση Διπλωματικής : Δαλάτσης Γεώργιος
Α.Μ : 1999010095

ΧΑΝΙΑ 2005

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ευκαιρία της ολοκλήρωσης της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω τον καθηγητή μου κύριο Μιχάλη Δούμπο για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο θέμα, καθώς και για την αμέριστη βοήθεια και γνώση που μου παρείχε κατά τη διάρκεια της συνεργασίας μας.

Επίσης, πολλές ευχαριστίες θα ήθελα να εκφράσω και στους φίλους που έκανα τα χρόνια αυτά και που σε όλες τις φοιτητικές μου στιγμές, ευχάριστες ή δυσάρεστες, στάθηκαν δίπλα μου.

Τέλος, θα ήθελα να απευθύνω ξεχωριστά ένα μεγάλο ευχαριστώ στην οικογένειά μου που με στήριξε ηθικά και υλικά καθ' όλη τη διάρκεια των σπουδών μου, και βοήθησε ώστε να γίνει πραγματικότητα ένα παιδικό μου όνειρο.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

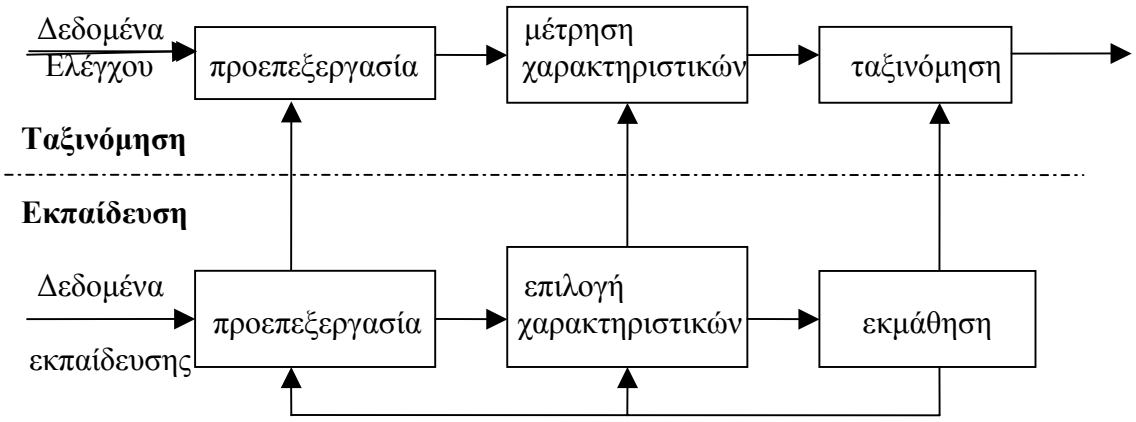
Τα προβλήματα λήψης διαφόρων αποφάσεων χαρακτηρίζονται συνήθως από την υψηλή πολυπλοκότητα που εμπεριέχουν λόγω του ότι επηρεάζονται από πλήθος παραγόντων, είτε ποσοτικούς είτε ποιοτικούς. Είναι λοιπόν επιβεβλημένη η οργάνωση των δεδομένων σε ομοιογενείς προκαθορισμένες ομάδες, μέσω των οποίων θα δίνεται η γενική περιγραφή και όλες τις απαραίτητες πληροφορίες όσων στοιχείων ανήκουν σε αυτές. Παράλληλα, κρίνεται απαραίτητη η ανάπτυξη ενός μοντέλου που εκτός από το παραπάνω, θα είναι σε θέση με βάση την εξέταση κάποιων συγκεκριμένων επιλεγμένων χαρακτηριστικών, να προβλέπει την προκαθορισμένη ομάδα στην οποία ανήκει κάθε αντικείμενο. Το γεγονός αυτό έχει επιφέρει τις τελευταίες δεκαετίες την ανάπτυξη πλήθους διαφόρων μεθοδολογιών και συστημάτων που έχουν ως στόχο να υλοποιήσουν την παραπάνω ταξινόμηση. Όμως, το να δημιουργηθεί ένα τέτοιο μοντέλο ταξινόμησης είναι στην πραγματικότητα το μισό του προβλήματος.

Δημιουργώντας μια τέτοια μέθοδο ταξινόμησης, προκύπτει αυτόματα η απορία κατά πόσο αυτή η διαδικασία είναι σε θέση να δώσει αξιόπιστες προβλέψεις και με τι ακρίβεια. Επίσης, δεδομένου ότι πλέον υπάρχουν αρκετά συστήματα ταξινόμησης, κρίνεται απαραίτητο να διερευνηθεί ποια από αυτά είναι καταλληλότερα σε σχέση με τα άλλα, για ταξινόμηση δεδομένων. Αναπτύχθηκαν λοιπόν, διάφορα εργαλεία αξιολόγησης της ποιότητας των συστημάτων ταξινόμησης, δημιουργώντας έτσι ένα νέο πεδίο μεγάλου ενδιαφέροντος και πρόκληση για όσους ασχολούνται με αυτό, την πολυκριτήρια αξιολόγηση συστημάτων ταξινόμησης. Κατά τη διαδικασία αυτή αποτιμώνται επιλεγμένα κριτήρια, κοινά για όλα τα συστήματα ταξινόμησης και ανάλογα με τη μέθοδο αξιολόγησης που έχει επιλεχθεί. Τα κριτήρια σε κάθε μέθοδο είναι συγκεκριμένα και επιλεγμένα έτσι ώστε να αποτελούν αντικειμενικά κριτήρια σύγκρισης και είναι τέτοια που να διασφαλίζουν ότι μας δίνουν αποτέλεσμα υπεροχής ενός συστήματος ταξινόμησης από ένα άλλο.

Η ανάγκη να αξιολογηθεί η ποιότητα των αποτελεσμάτων ενός συστήματος ταξινόμησης και να συγκριθεί με τα αποτελέσματα κάποιου ή κάποιων άλλων, γίνεται ολοένα πιο σημαντική και επιτακτική. Οι τομείς στους οποίους έχει εισέλθει η ταξινόμηση στην καθημερινότητα είναι πλέον πολλοί, έτσι ώστε η βελτίωση ακόμα και κατά ένα πολύ μικρό ποσοστό στην ακρίβεια τέτοιων συστημάτων, μπορεί να μεταφραστεί σε αξιοπρόσεκτα αποτελέσματα και προνόμια. Τα προβλήματα οικονομικής φύσεως , είναι ένα από τα πεδία στα οποία έχουν αναπτυχθεί πολλά συστήματα ταξινόμησης και συνεπώς και μέθοδοι αξιολόγησής τους. Ο έλεγχος του πιστωτικού κινδύνου , η ελάττωση του κόστους , η βελτίωση της παρακολούθησης των ήδη υπαρχόντων λογαριασμών , η ταχύτερη αξιολόγηση πιστωτών και η βελτίωση των ταμειακών ροών είναι κάποια από τα πεδία στα οποία οι τράπεζες και αρκετοί άλλοι οικονομικοί φορείς έχουν αναπτύξει μεθοδολογίες ταξινόμησης. Συνυπολογίζοντας ότι στο επίπεδο αυτό η αναφορά γίνεται σε επιχειρήσεις οι οποίες είναι και έχουν στόχο να παραμείνουν κερδοφόρες , θέλουν επίσης να προσελκύσουν νέους πελάτες διατηρώντας ταυτόχρονα και τους ήδη υπάρχοντες , ελαχιστοποιώντας ωστόσο το ρίσκο τους και διατηρώντας την κερδοφορία τους , διαπιστώνεται ότι η αξιοπιστία των συστημάτων ταξινόμησης που χρησιμοποιούν επιβάλλεται να είναι υψηλή έτσι ώστε να αποφεύγονται λάθη στην κατάτμηση της αγοράς και στη συνεργασία με ασύμφορους οικονομικά πελάτες.

Άλλος ένας τομέας στον οποίο γίνεται ευρεία χρήση των συστημάτων ταξινόμησης και συνεπώς είναι απαραίτητη η αξιολόγησή των αποτελεσμάτων τους , είναι και αυτός της ιατρικής. Επίσης, η ταξινόμηση μπορεί να βρει εφαρμογή και σε εκλογικές διαδικασίες με σκοπό τη σφυγμομέτρηση του εκλογικού σώματος και γενικότερα θα μπορούσε να ειπωθεί ότι η χρήση τέτοιων συστημάτων είναι απαραίτητη σε περιπτώσεις που είναι δυνατή η ταξινόμηση ατόμων ή αντικειμένων σε διακριτές κατηγορίες και αυτή διευκολύνει την εξαγωγή αποτελεσμάτων και τη λήψη σοβαρών και τεκμηριωμένων αποφάσεων. Αυτός είναι ένας ακόμα λόγος που καθιστά το πεδίο αυτό αρκετά χρήσιμο και την αξιολόγησή του αρκετά σημαντική. Μια λάθος ταξινόμηση είναι λογικό να επιφέρει κάποιο κόστος , το οποίο όμως διαφέρει από εφαρμογή σε εφαρμογή, όπως επίσης μπορεί να διαφέρει ακόμα και μέσα στην ίδια την εφαρμογή. Είναι λοιπόν αναγκαίος ο έλεγχος των διαφόρων μεθόδων ταξινόμησης ως προς την αξιοπιστία τους και η επιλογή σε κάθε περίπτωση του καταλληλότερου που θα αποφέρει τον ελάχιστο αριθμό σφαλμάτων.

Η φιλοσοφία της ανάπτυξης των συστημάτων ταξινόμησης και η αξιολόγησή τους στα γενικότερα πλαίσιά της, είναι κοινή για όλα . Λαμβάνεται ένα σύνολο δεδομένων το οποίο χωρίζεται σε δυο υποσύνολα , το σύνολο εκπαίδευσης (training pattern) του συστήματος ταξινόμησης και το σύνολο μέσα από το οποίο ελέγχεται η απόδοσή του και η ακρίβεια των αποτελεσμάτων του (test pattern). Τα δεδομένα αποτελούνται από διάφορες παρατηρήσεις οι οποίες περιγράφονται από διάφορα επιλεγμένα χαρακτηριστικά, τις τιμές τους και την ταξινόμησή τους , ώστε κατά την εκπαίδευσή του το σύστημα χρησιμοποιώντας το ανάλογο σύνολο δεδομένων, συσχετίζει αυτά τα χαρακτηριστικά ανάλογα με τις τιμές τους, με τις διάφορες κατηγορίες ταξινόμησης. Έτσι, με βάση αυτό και τη μέθοδο που κάθε φορά επιλέγεται αναπτύσσεται ο συνολικός κανόνας ταξινόμησης. Αφού αναπτυχθεί ο κανόνας ταξινόμησης , χρησιμοποιείται το δεύτερο υποσύνολο των δεδομένων, τις παρατηρήσεις του οποίου καλείται το σύστημα να ταξινομήσει αν και η κατηγορία ταξινόμησής τους είναι ήδη γνωστή. Έτσι, με τη βοήθεια κάποιων κριτηρίων και μεθόδων αξιολόγησης , ελέγχεται η απόδοση του ταξινομητή ως προς την ικανότητά του να διαχωρίζει νέες παρατηρήσεις. Η γενικότερη αυτή φιλοσοφία που μόλις περιγράφθηκε, μπορεί να αποδοθεί σχηματικά στην εικόνα 1 που ακολουθεί.



Εικόνα 1. Η γενική δομή ενός συστήματος ταξινόμησης

Στο κεφάλαιο 2 που ακολουθεί, στο πρώτο μέρος του γίνεται αναφορά σε αρκετές από τις μεθόδους ταξινόμησης που έχουν αναπτυχθεί, όπως διάφορες στατιστικές μέθοδοι , τα δένδρα απόφασης και τα νευρωνικά δίκτυα και εξετάζεται θεωρητικά ο τρόπος με τον οποίο λειτουργούν. Στη συνέχεια , αναπτύσσονται τα κριτήρια , οι διαδικασίες και οι τεχνικές που χρησιμοποιούνται ώστε να αξιολογηθούν οι μέθοδοι ταξινόμησης και ενδεικτικά κάποιες από αυτές είναι οι τεχνικές Bootstrap και Cross-Validation και τα κριτήρια του ποσοστού ακρίβειας , το στατιστικό μέγεθος Kolmogorov-Smirnov, ο δείκτης ακρίβειας και το τυπικό σφάλμα ακρίβειας. Τέλος , στο κεφάλαιο 3 γίνεται πειραματική ανάλυση και εφαρμογή όσων έχουν προηγουμένως αναπτυχθεί , μέσα από συγκεκριμένα παραδείγματα και δεδομένα πραγματικών προβλημάτων ταξινόμησης και συγκρίνονται τα αποτελέσματα των συστημάτων ταξινόμησης που εφαρμόζονται .

ΚΕΦΑΛΑΙΟ 2: ΣΥΣΤΗΜΑΤΑ ΤΑΞΙΝΟΜΗΣΗΣ

2.1 Περιγραφή προβλήματος – Εφαρμογές

Η αυτόματη αναγνώριση , η περιγραφή , η κατάταξη και η ομαδοποίηση δεδομένων είναι σημαντικά προβλήματα για τους μηχανικούς αλλά και σε άλλα επιστημονικά πεδία . Στους ανθρώπους , στην ηλικία των πέντε ετών ακόμα τα περισσότερα παιδιά είναι σε θέση να αναγνωρίσουν ψηφία και γράμματα . Λίγο αργότερα, είναι ικανός ο άνθρωπος εύκολα να ξεχωρίσει είτε τυπωμένους είτε γραμμένους στο χέρι χαρακτήρες , μικρούς ή μεγάλους , ακόμα και αντεστραμμένους. Το πρόβλημα που προκύπτει, είναι πώς αυτή η δεδομένη για τον άνθρωπο ικανότητα μπορεί να διδαχθεί και στις μηχανές έτσι ώστε να μπορούν να αναγνωρίζουν τα απαραίτητα δεδομένα και βασιζόμενες σε αυτά να παίρνουν λογικές αποφάσεις για τις κατηγορίες που ανήκουν.

Υπάρχουν πολλοί λόγοι που καθιστούν αναγκαία την ανάπτυξη μιας διαδικασίας ταξινόμησης και η λύση του παραπάνω προβλήματος γίνεται ακόμα πιο σημαντική αν αναλογιστεί κανείς τους τομείς στους οποίους η αναγνώριση δεδομένων και η ταξινόμηση μπορεί να βρει εφαρμογή , αλλά και τι κέρδος ή κόστος μπορεί να αποφέρει. Πιο συγκεκριμένα , στον έλεγχο πιστωτικού κινδύνου ενδιαφέρει άμεσα την τράπεζα η αξιολόγηση ενός πελάτη ως αξιόπιστο ή όχι. Κατατάσσοντας έναν αξιόπιστο πελάτη ορθά , αποκτά ταυτόχρονα έναν αναγκαίο δανειστή κάτι που σημαίνει κέρδος για την τράπεζα. Αντιθέτως, κατατάσσοντας τον ίδιο πελάτη ως αναξιόπιστο της αποφέρει κόστος ευκαιρίας του να χάσει έναν καλό πελάτη και πιθανόν κόστος στην ανάπτυξη της ανταγωνιστικότητάς της. Εύκολα μπορεί κανείς να αντιληφθεί το κόστος ή το κέρδος από λάθος ή σωστή αντίστοιχα ταξινόμηση , για ασθενείς και γιατρούς, όταν τέτοιες μεθοδολογίες χρησιμοποιούνται στο χώρο της ιατρικής με σκοπό τη διάγνωση και την ορθή αντιμετώπιση των ασθενειών.

Η αναγνώριση ενός δείγματος και η ταξινόμησή του επεκτείνεται επίσης σε τομείς όπως η βιολογία , η ψυχολογία , το μάρκετινγκ, η τεχνητή νοημοσύνη, η πρόβλεψη καιρικών φαινομένων, η συναλλαγή αποθεμάτων και οι επενδύσεις. Ταυτόχρονα όμως, τα συστήματα ταξινόμησης που αναπτύσσονται είναι απαραίτητο να συνδυάζουν και χαρακτηριστικά όπως ταχύτητα χωρίς ωστόσο να μειώνεται η ακρίβειά τους , ιδιαίτερα όταν εφαρμόζονται σε γραμμές παραγωγής στον αυτόματο εντοπισμό λαθών. Επίσης, επιβάλλεται να είναι κατανοητά έτσι ώστε να μπορούν εύκολα να ερμηνευτούν τα αποτελέσματά τους από τον άνθρωπο χωρίς παρανοήσεις που πιθανόν να οδηγήσουν σε λανθασμένες αποφάσεις.

Τέλος, σημαντικό ρόλο παίζει και ο χρόνος εκμάθησης που χρειάζεται ένα τέτοιο σύστημα ταξινόμησης. Σε ένα εργασιακό περιβάλλον που αλλάζει πολύ γρήγορα είναι απαραίτητα το σύστημα να εκπαιδεύεται εξίσου γρήγορα, συμπεριλαμβάνοντας στην έννοια γρήγορα και το γεγονός να μην απαιτείται μεγάλος αριθμός παρατηρήσεων ώστε να κατασκευαστεί ένας κανόνας ταξινόμησης. Σε όλα τα παραπάνω σημαντικό ρόλο παίζει η ικανότητα του συστήματος ταξινόμησης να εντοπίζει μέσα από τη διαδικασία της εκπαίδευσής του τα κατάλληλα χαρακτηριστικά, μέσα από ένα σύνολο πολλών, ώστε να αναπαρίστανται με το καταλληλότερο τρόπο δείγματα που χρησιμοποιούνται ως δεδομένα εισόδου.

2.2 Μεθοδολογίες

Η ταξινόμηση έχει δυο διακριτές έννοιες. Η μια προσέγγιση αφορά την κατηγοριοποίηση παρατηρήσεων , με δεδομένες όμως τις κατηγορίες στις οποίες είναι δυνατό να ταξινομηθεί. Στη δεύτερη περίπτωση, ζητείται από το σύστημα ταξινόμησης να κατατάξει τις παρατηρήσεις αφού όμως πρώτα μέσα από την διαδικασία εκπαίδευσής θα έχει κατασκευάσει εκτός από τον κανόνα ταξινόμησης και τις ίδιες τις κατηγορίες ταξινόμησης.

Τα παραπάνω, σε συνδυασμό με το ευρύ φάσμα στο οποίο μπορούν να βρουν εφαρμογή τα συστήματα ταξινόμησης οδήγησαν στην ανάπτυξη αρκετών μεθοδολογιών. Τέτοιες μεθοδολογίες είναι η γραμμική διακριτική ανάλυση (LDA), η λογιστική παλινδρόμηση (LR), τα πιθανοτικά νευρωνικά δίκτυα (PNN), η μέθοδος των πλησιέστερων γειτόνων (KNN), οι μηχανές διανυσμάτων υποστήριξης και τα δένδρα ταξινόμησης και παλινδρόμησης.

Οι παραπάνω τεχνικές μπορούν να διαχωριστούν μεταξύ τους σε δυο γενικές κατηγορίες, στις στατιστικές μεθόδους και στις μη-στατιστικές. Ειδικότερα, οι στατιστικές μέθοδοι μπορούν να διακριθούν περαιτέρω σε δυο γενικούς τύπους , τα διαγνωστικά παραδείγματα και τα δειγματοληπτικά. Οι μέθοδοι που ανήκουν στον πρώτο γενικό τύπο χρησιμοποιούν πληροφορίες από τα δεδομένα σχεδιασμού για να εκτιμήσουν την πιθανότητα μια παρατήρηση να ανήκει σε κάθε κατηγορία κατάταξης, βασιζόμενες σε μεταβλητές πρόβλεψης. Στις μεθόδους που ανήκουν στον δεύτερο γενικό τύπο, για κάθε δεδομένο που πρέπει να ταξινομηθεί γίνεται εκτίμηση της κατανομής των μεταβλητών πρόβλεψης ξεχωριστά για κάθε κατηγορία ταξινόμησης και συνδυάζεται με τις αρχικές πιθανότητες που είχε το δεδομένο αυτό να ανήκει σε κάποια από τις κατηγορίες αυτές. Έπειτα, με χρήση του θεωρήματος Bayes υπολογίζονται οι τελικές πιθανότητες κατάταξης των δεδομένων στις διάφορες κατηγορίες. Παρακάτω παρουσιάζονται αναλυτικότερα οι μεθοδολογίες αυτές.

2.2.1 Γραμμική Διακριτική Ανάλυση (LDA)

Η γραμμική διακριτική ανάλυση είναι στατιστική μέθοδος και είναι από τις πρώτες τεχνικές που αναπτύχθηκαν βασιζόμενη στις μεθόδους διάκρισης που προτάθηκαν από τον Fisher (1936). Σύμφωνα με τις διαδικασίες αυτής της μεθόδου, καθορίζεται η ταξινόμηση συνυπολογίζοντας τις μεταβλητές πρόβλεψης , αφού μεγιστοποιηθεί ο διαχωρισμός μεταξύ των κατηγοριών και υπολογισθεί η πιθανότητα κατάταξης σε κάθε κατηγορία . Αυτό μπορεί να γίνει με δυο τρόπους , με την μέθοδο των ελαχίστων τετραγώνων ή με τη μέθοδο της μέγιστης πιθανότητας .

Σύμφωνα με την πρώτη μέθοδο, καθορίζονται οι διαχωριστικές γραμμές ή τα διαχωριστικά επίπεδα αν μιλάμε για παραπάνω από δυο διαστάσεις του n-διάστατου χώρου των χαρακτηριστικών , μεταξύ των διαφόρων κατηγοριών ταξινόμησης. Έτσι , ανάλογα σε ποια πλευρά των διαχωριστικών αυτών θα βρεθεί κάποια παρατήρηση ταξινομείται και στην αντίστοιχη κατηγορία .

Στόχος λοιπόν της μεθόδου είναι η ανάπτυξη ενός υπερεπιπέδου ταξινόμησης,

$$g(\mathbf{x}) = \varepsilon \; a_j x_j$$

έτσι ώστε να μεγιστοποιηθεί η διακύμανση μεταξύ των στοιχείων διαφορετικών κατηγοριών σε σχέση με την διακύμανση εντός των κατηγοριών. Συμβολίζοντας ως $\bar{\mathbf{x}}$ το διάνυσμα με τις μέσες τιμές των χαρακτηριστικών για το σύνολο του δείγματος και ως $\bar{\mathbf{x}}_\mathbf{k}$ το αντίστοιχο διάνυσμα για την κατηγορία ω_k , το άθροισμα τετραγώνων μέσα σε μια κατηγορία ταξινόμησης ω_k είναι :

$$\varepsilon \; (g(\mathbf{x}) - g(\bar{\mathbf{x}}_\mathbf{k}))^2$$

Αν το συνολικό άθροισμα των παραπάνω αθροισμάτων τετραγώνων είναι u και

$$t = \varepsilon \; (g(\mathbf{x}) - g(\bar{\mathbf{x}}))^2$$

είναι το συνολικό άθροισμα τετραγώνων της $g(\boldsymbol{x})$, τότε $t - u$ είναι το συνολικό άθροισμα τετραγώνων μεταξύ των κατηγοριών ταξινόμησης.

Από τη σχέση

$$F = \frac{(t - u)}{u/(N - 2)}$$

φαίνεται καθαρά ότι μεγιστοποιώντας το F μεγιστοποιείται η σχέση t/u (άρα και η διακριτότητα μεταξύ των κατηγοριών ταξινόμησης) κάτι που επιτυγχάνεται χρησιμοποιώντας τους κατάλληλους συντελεστές $a_1, a_2, \ldots a_n$.

Σύμφωνα με τη δεύτερη μέθοδο, αυτή της μέγιστης πιθανότητας , ένα αντικείμενο ταξινομείται στην κατηγορία στην οποία έχει τη μεγαλύτερη τιμή συνάρτησης πυκνότητας πιθανότητας f_i. Συχνότερα , θεωρείται ότι η συνάρτηση πυκνότητας πιθανότητας είναι κανονική και συνεπώς είναι της μορφής :

$$\frac{1}{\sqrt{|2p\mathrm{S}|}}\exp(-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^{\mathrm{T}}\varepsilon^{-1}(\mathbf{x}-\bar{\mathbf{x}})$$

όπου $\bar{\mathbf{x}}$ είναι το n-διάστατο διάνυσμα που αναφέρεται στον μέσο μιας κατηγορίας ταξινόμησης και S ο nxn πίνακας διακύμανσης – συνδιακύμανσης.

Το όριο μεταξύ δυο κατηγοριών ταξινόμησης καθορίζεται από την εξίσωση :

$$\mathbf{x}^{\mathrm{T}}\varepsilon^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)-\frac{1}{2}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)^{\mathrm{T}}\varepsilon^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)$$

2.2.2 Λογιστική Παλινδρόμηση (LR)

Είναι μια από τις πρόσφατες διαγνωστικού τύπου μεθόδους. Σε περιπτώσεις ταξινόμησης σε δύο κατηγορίες , μέσω ενός γραμμικού συνδυασμού των μεταβλητών που περιγράφουν μια προς ταξινόμηση περίπτωση, μετασχηματίζονται τα αποτελέσματα έτσι ώστε να παίρνουν τιμές μεταξύ 0 και 1 και να μπορούν να εξισωθούν σε πιθανότητα. Έτσι, σε περιπτώσεις ταξινόμησης σε δύο κατηγορίες η λογιστική παλινδρόμηση θεωρείται ως μια από τις καταλληλότερες τεχνικές .

Αναλυτικότερα, η μέθοδος αυτή επιτρέπει την πρόβλεψη ενός διακριτού αποτελέσματος , όπως μέλος μιας ομάδας , από ένα σύνολο μεταβλητών που μπορεί να είναι συνεχές , διακριτό , διχοτομημένο ή και συνδυασμός όλων των παραπάνω. Συνήθως, στη λογιστική παλινδρόμηση η εξαρτημένη μεταβλητή είναι δυαδική (ή αλλιώς Bernoulli) γι'αυτό και μπορεί να πάρει τιμή 1 με πιθανότητα επιτυχίας P ή τιμή 0 με πιθανότητα αποτυχίας 1-P.

Έτσι , υπολογίζεται η παρακάτω συνάρτηση της πιθανότητας P, η οποία δείχνει την υπεροχή της μιας κατηγορίας κατάταξης σε σχέση με την άλλη :

$$odds = \frac{P}{1-P}$$

και κατόπιν καθορίζεται η εξαρτημένη μεταβλητή η οποία είναι ο λογάριθμος της παραπάνω σχέσης :

$$\log(odds) = \mathrm{logit}(P) = \ln(\frac{P}{1-P})$$

με $P = \frac{1}{1+e^{-(a+b\mathbf{x})}}$ και $\mathrm{logit}(P) = a + b\,\mathbf{x}$, όπου a , b οι παράμετροι του μοντέλου, με

την παράμετρο ***a*** να καθορίζει την πιθανότητα P όταν η μεταβλητή **x** είναι μηδέν και τις παραμέτροθς στο διάνυσμα ***b*** να ρυθμίζουν το πόσο γρήγορα μεταβάλλεται η πιθανότητα P καθώς αλλάζει μια μεταβλητή κατά μια μονάδα..

Καθορίζεται λοιπόν κατά αυτόν τον τρόπο το πόσο σχετίζεται μια μέτρηση με τα χαρακτηριστικά μιας ταξινόμησης (0) ή μιας άλλης (1) , αφού βέβαια έχει καθοριστεί πρώτα ένα κατώφλι σύμφωνα με το οποίο αν ξεπερνιέται ή όχι, αναλόγως κατατάσσεται το προς ταξινόμηση στοιχείο ως 0 ή 1. Συνυπολογίζοντας, ότι ένας λογιστικός ταξινομητής βασίζεται στην προσέγγιση της μέγιστης πιθανότητας και το σύστημα έχει ως στόχο τη μεγιστοποίηση :

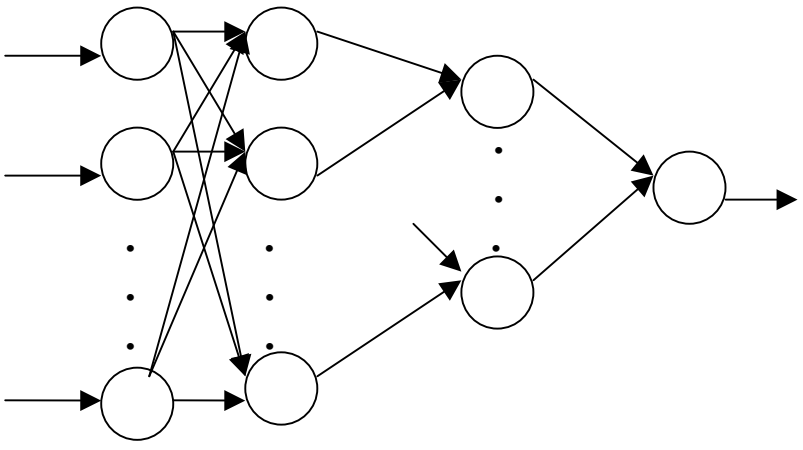
$$\max\{ \prod_{\mathbf{x_i} \Xi \omega_1} P(\mathbf{x_i}; q) \prod_{\mathbf{x_i} \Xi \omega_2} (1 - P(\mathbf{x_i}; q))\}$$

όπου θ το σύνολο των αγνώστων παραμέτρων(α και b) και ω_1, ω_2 οι κατηγορίες ταξινόμησης .

2.2.3 Πιθανοτικά Νευρωνικά Δίκτυα (PNN)

Μέσω των νευρωνικών δικτύων γίνεται μια προσπάθεια ώστε να μοντελοποιηθούν οι διαδικασίες του ανθρώπινου εγκεφάλου. Αν και γενικότερα οι μέθοδοι ταξινόμησης που στηρίζονται στα νευρωνικά δίκτυα κατατάσσονται στις μη στατιστικές , τα πιθανοτικά νευρωνικά δίκτυα οδηγούν σε μια λύση βασιζόμενα στην στατιστική και χρησιμοποιώντας τη θεωρία Bayes.

Παράδειγμα ενός πιθανοτικού νευρωνικού δικτύου φαίνεται στην εικόνα 2.1 παρακάτω, ενώ στη συνέχεια γίνεται μια περιγραφή στις διαδικασίες που ακολουθούν.



Εικόνα 2.1 Δομή ενός Πιθανοτικού Νευρωνικού Δικτύου

Το πλήθος των εισόδων στο δίκτυο καθορίζεται από το πλήθος n των χαρακτηριστικών $x_1,\ldots,x_n$. Ο αριθμός των κόμβων που αναπαριστούν τις μονάδες δείγματος (pattern units) είναι ίδιος με το πλήθος των αντικειμένων εκπαίδευσης και τα βάρη w_{ij}^P που τους αποδίδονται τα συνδέουν με τα δεδομένα εισόδου σύμφωνα με τη σχέση

$$w_{ij}^P = x_{ij}$$

όπου x_{ij} είναι η περιγραφή του αντικειμένου εκπαίδευσης i στο χαρακτηριστικό j.

Έπειτα, αποδίδονται βάρη w_{ik}^s μεταξύ του επιπέδου pattern units και του summation units ως εξής :

$$w_{ik}^{(s)} = \begin{cases} 1, & \text{αν } y_i = k \\ 0, & \text{αλλιώς} \end{cases}$$

όπου y_i είναι η ταξινόμηση του αντικειμένου εκπαίδευσης i.

Για να γίνει η ταξινόμηση ενός νέου αντικειμένου $\mathbf{x} = (x_1,\ldots,x_n)$, υπολογίζεται η είσοδος σε κάθε pattern unit i :

$$in_i^{(P)} = \sqrt{\sum_j (w_{ij}^{(P)} - \mathbf{x_j})^2}$$

και η έξοδος : $$P_i = \exp(-\frac{in_i^P}{2s^2})$$

όπου s είναι μια παράμετρος ομαλοποίησης του μοντέλου (pnn-smoothing).

Στη συνέχεια , στο επίπεδο summation units , όπου κάθε κόμβος αναπαριστά και μια κατηγορία ταξινόμησης, υπολογίζεται η έξοδος κάθε κόμβου:

$$S_k = \frac{1}{\sum_i w_{ik}^{(S)}} \sum_{i=1}^{n} w_{ik}^{(S)} P_i$$

Το αντικείμενο θα ταξινομηθεί στην κατηγορία στην οποία αντιστοιχεί η μεγαλύτερη έξοδος.

2.2.4 Πλησιέστεροι Γείτονες (KNN)

Η ταξινόμηση με τη μέθοδο των k-πλησιέστερων γειτόνων είναι μια από τις πιο δημοφιλείς τεχνικές , σε περιπτώσεις που είναι διαθέσιμη μεγάλη ποσότητα δεδομένων εκπαίδευσης του ταξινομητή. Οι KNN ταξινομητές κατηγοριοποιούν ένα άγνωστο δείγμα βασιζόμενοι στο ποια είναι η κυρίαρχη κατηγορία των k κοντινότερων γειτόνων του, θεωρώντας ως κοντινότερους γείτονες τα δείγματα που έχουν από αυτό την μικρότερη απόσταση.

Ένας τέτοιος αλγόριθμος μπορεί να αναλυθεί σε τρία κυρίως βήματα:

1. Υπολογισμός της απόστασης του προς ταξινόμηση αντικειμένου $\mathbf{x}$ με κάθε ένα από τα δείγματα εκπαίδευσης $\mathbf{x}_i$ χρησιμοποιώντας την Ευκλείδεια απόσταση.

2. Εντοπισμός των k αντικειμένων εκπαίδευσης με την μικρότερη απόσταση από το προς ταξινόμηση αντικείμενο και έλεγχος ποια κατηγορία ταξινόμησης κυριαρχεί σε αυτά .

3. Επανάληψη των βημάτων 1-2 για κάθε αντικείμενο.

2.2.5 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Η μέθοδος των μηχανών υποστήριξης διανυσμάτων αποτελεί αυτή τη στιγμή μια από τις πιο αποτελεσματικές μεθόδους στο χώρο της ταξινόμησης. Σκοπός της μεθόδου αυτής είναι να βρεθεί ένα υπερεπίπεδο τέτοιο ώστε το περιθώριο μεταξύ των κατηγοριών ταξινόμησης να είναι μέγιστο, αλλά ταυτόχρονα ο αριθμός των λάθος ταξινομημένων δειγμάτων να είναι ο ελάχιστος. Έτσι , το παραπάνω πρόβλημα μπορεί να διατυπωθεί ως εξής :

$$\min y(\mathbf{w}, x) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{m} x_i\right)$$

υπό :

$$y_i[(\mathbf{w} \cdot \mathbf{x}_i) + b] - 1 + x_i > 0, i = 1, .., m$$

όπου $(\mathbf{x}_i, y_i), i = 1, ..., m, \mathbf{x} \in R^n, y_i \in \{-1, +1\}$, είναι τα αντικείμενα που χρησιμοποιήθηκαν για την εκπαίδευση του ταξινομητή και C μια σταθερά για να ελέγχεται η σχέση μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης των λαθών.

Η συνάρτηση απόφασης είναι η :

$$f(\mathbf{x}) = \operatorname{sgn}\{(\mathbf{w} \cdot \mathbf{x}) + b\} = \operatorname{sgn}\left\{\sum_{i=1}^{m} a_i^* \cdot y_i (\mathbf{x}_i \cdot \mathbf{x}) + b^*\right\}$$

και το πρόβλημα βελτιστοποίησης λύνεται από το παρακάτω πρόβλημα :

$$\max Q(a) = \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m} a_i a_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

υπό : $$\sum_{i=1}^{m} y_i a_i = 0$$

και $$0 \le a_i \le C \text{ , i=1,…,m}$$

Στην τελική SVM συνάρτηση απόφασης μόνο ένα μικρό μέρος των μεταβλητών a_i είναι μη μηδενικό. Τα αντίστοιχα διανύσματα δεδομένων εκπαίδευσης είναι τα διανύσματα υποστήριξης , αφού υποστηρίζουν τα όρια της ταξινόμησης. Οι παραπάνω περιγραφή αφορά τις γραμμικές μηχανές υποστήριξης διανυσμάτων (LSVM).

Γενικά , οι μηχανές υποστήριξης διανυσμάτων χρησιμοποιούν την συνάρτηση :

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i a_i k(\mathbf{x}, \mathbf{x}_i) + b$$

όπου $k(\mathbf{x}, \mathbf{x}_i)$ είναι η συνάρτηση πυρήνα (kernel) και ανάλογα ποια είναι αυτή η συνάρτηση προκύπτουν διάφορες μηχανές υποστήριξης. Τέτοιες είναι οι RSVM στις οποίες χρησιμοποιείται πυρήνας RBF(Radial Basis Function). Σε αυτή την περίπτωση , η συνάρτηση kernel είναι η :

$$k(\mathbf{x}, \mathbf{x}_i) = \exp\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{2d^2}\right)$$

Μια άλλη διαδεδομένη επιλογή είναι ο πολυωνυμικός πυρήνας (Polynomial kernel)

$$k(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x} \cdot \mathbf{x}_i + 1)^P$$

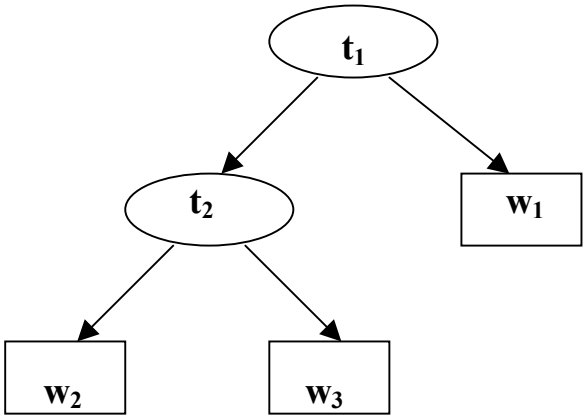
Στην ειδική περίπτωση όπου ο βαθμός του πολυωνύμου είναι p = 2, προκύπτουν οι μηχανές διανυσμάτων υποστήριξης τετραγωνικού πυρήνα (QSVM).

2.2.6 Δένδρα Ταξινόμησης και Παλινδρόμησης (CART)

Γενικά, ένα δένδρο απόφασης είναι μια αναπαράσταση, ανάλογα δομημένη, ενός προβλήματος απόφασης έτσι ώστε κάθε μη-φύλλο κόμβος να σχετίζεται με μια από τις μεταβλητές απόφασης , κάθε κλαδί από ένα μη-φύλλο κόμβο να σχετίζεται με ένα υποσύνολο τιμών των αντίστοιχων μεταβλητών απόφασης και κάθε κόμβος-φύλλο να σχετίζεται με μια τιμή της εξαρτημένης μεταβλητής.

Υπάρχουν δυο κύριοι τύποι δένδρων απόφασης : (1) τα δένδρα ταξινόμησης και (2) τα δένδρα παλινδρόμησης. Για τα δένδρα ταξινόμησης , οι εξαρτημένες μεταβλητές παίρνουν τιμές από μια διακριτή περιοχή και σε κάθε κόμβο-φύλλο το δένδρο απόφασης αντιστοιχίζει μια πιθανότητα για κάθε κατηγορία ταξινόμησης. Τελικά , η κατηγορία στην οποία ταξινομείται κάθε κόμβος προκύπτει με βάση το ποια έχει τη μεγαλύτερη πιθανότητα σε σχέση με τις άλλες. Τα δένδρα παλινδρόμησης , μπορούν να χρησιμοποιηθούν ως εναλλακτική προσέγγιση της ανάλυσης παλινδρόμησης , σύμφωνα με την οποία εκτιμάται η τιμή της εξαρτημένης μεταβλητής δίνοντας την τιμή κάθε ανεξάρτητης μεταβλητής.

Ειδικότερα , τα CART περιλαμβάνουν την αναγνώριση και την κατασκευή ενός δένδρου απόφασης χρησιμοποιώντας δείγματα ως δεδομένα εκπαίδευσης για τα οποία η σωστή κατηγορία ταξινόμησης τους είναι γνωστή. Σε κάθε κόμβο γίνεται ένας διαχωρισμός , με βάση κάποια κριτήρια, ώστε από αυτόν να προκύπτουν κάποιοι άλλοι υπό-κόμβοι και να γίνει έτσι η τελική εύρεση των κατηγοριών ταξινόμησης για κάθε στοιχείο. Παράδειγμα ενός δένδρου απόφασης φαίνεται στην εικόνα 2.2.



Εικόνα 2.2 Δυαδικό Δένδρο Απόφασης

Το δένδρο απόφασης ξεκινά από ένα κόμβο-ρίζα t που προκύπτει από όποια μεταβλητή του διαστήματος χαρακτηριστικών ελαχιστοποιεί την εντροπία των κόμβων που πρόκειται να προκύψουν από αυτόν. Το μέτρο της εντροπίας προκύπτει από την παρακάτω σχέση :

$$i(t) = - \sum_{j=1}^{k} p(w_k | t) \log p(w_k | t)$$

όπου $p(w_k | t)$ είναι η πιθανότητα ταξινόμησης ενός αντικειμένου σε κάποια από τις κατηγορίες $w_1, w_2, ..., w_k$ στον κόμβο t. Έπειτα, κάθε μη τελικός κόμβος χωρίζεται σε παραπέρα κόμβους , έστω δύο αν πρόκειται για δυαδικό δένδρο απόφασης , t_L και t_R με πιθανότητες p_L και p_R αντίστοιχα. Το καλύτερο τμήμα είναι αυτό που μεγιστοποιεί τη διαφορά που δίνεται από τον τύπο:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Το δένδρο απόφασης μεγαλώνει κατ'αυτόν τον τρόπο μέχρι να φτάσει σε ένα επίπεδο στο οποίο η προσθήκη ακόμα ενός τμήματος δεν προσφέρει ουσιαστικά τίποτα στην μείωση της εντροπίας. Μόλις επιτευχθεί αυτό το επίπεδο προκύπτουν και οι τελικοί κόμβοι αυτόματα . Η κατηγορία ταξινόμησης που συνδέεται με τον κάθε τελικό κόμβο είναι αυτή για την οποία η πιθανότητα $p(w_k | t)$ είναι μέγιστη.

2.3 Αξιολόγηση Συστημάτων Ταξινόμησης

Ένα από τα μεγαλύτερα προβλήματα που απασχόλει το πεδίο της ταξινόμησης πέρα από το πώς και μέσα από ποιες διαδικασίες θα αναπτυχθεί ένα σύστημα ταξινόμησης , είναι και αυτό της ακρίβειας με την οποία θα παρέχουν αποτελέσματα . Έχουν αναπτυχθεί λοιπόν αρκετές τεχνικές οι οποίες χρησιμοποιώντας κάποια συγκεκριμένα κριτήρια , στοχεύουν στην αξιολόγηση των συστημάτων ταξινόμησης. Βέβαια , σημαντικό ρόλο στην αξιολόγηση αλλά και σύγκριση τέτοιων συστημάτων μεταξύ τους παίζει και το γεγονός κάθε φορά να χρησιμοποιούνται τα κατάλληλα κριτήρια και η κατάλληλη μέθοδος ώστε η αξιολόγηση να αντικατοπτρίζει την πραγματικότητα. Επομένως, κρίνεται αναγκαίο πρωτίστως να αποσαφηνίζονται οι καταστάσεις κάτω από τις οποίες κάθε μέθοδος αξιολόγησης είναι η καταλληλότερη.

Σημαντικό επίσης θέμα που σχετίζεται άμεσα με την αποτίμηση της ταξινόμησης, είναι η σύγκριση της αξιολόγησης της ταξινόμησης μέσω μεμονωμένων μετρήσεων για κάθε κατηγορία ξεχωριστά , με την αξιολόγηση μέσω ενός πληθυσμού μετρήσεων όπου έτσι αποτιμάται συνολικά η απόδοση της ταξινόμησης στις διακριτές κατηγορίες . Τέτοιο είναι και το θέμα που αφορά την διαφορά της υποθετικής απόδοσης των συστημάτων ταξινόμησης με την μη υποθετική . Η πρώτη περίπτωση χρησιμοποιείται για την αποτίμηση της απόδοσης

ενός μοντέλου βασιζόμενη στα προκαθορισμένα δεδομένα με τα οποία αυτό θα αναπτυχθεί, ενώ η δεύτερη περίπτωση χρησιμοποιείται όταν σκοπός είναι να βγουν γενικότερα συμπεράσματα για το ποια μέθοδος ταξινόμησης είναι καλύτερη. Το τρίτο θέμα που σχετίζεται με την αξιολόγηση των μοντέλων ταξινόμησης είναι αυτό μεταξύ της απόλυτης και της συγκριτικής απόδοσης τους. Μέσω της απόλυτης αξιολόγησης κάθε σύστημα ταξινόμησης συγκρίνεται με κάποιο εξωγενές , ίδιο για όλα, μέτρο. Αντιθέτως, στη συγκριτική αξιολόγηση συγκρίνονται τα μοντέλα μεταξύ ώστε να βρεθεί που είναι καλύτερο το ένα από το άλλο, με την έννοια καλύτερα να καθορίζεται από τον αναλυτή ή αυτόν που παίρνει την απόφαση.

Ένα από τα πιο κοινά στοιχεία μέτρησης της απόδοσης ενός κανόνα ταξινόμησης , είναι η εκτίμηση του ποσοστού ακρίβειας (error rate) που αποδίδει την αναλογία των παρατηρήσεων που έχουν κατηγοριοποιηθεί λάθος. Υπάρχουν και διάφορα άλλα κριτήρια εκτίμησης της απόδοσης και αρκετές επίσης τεχνικές στις οποίες γίνεται χρήση κάποιων από αυτών, γι'αυτό και είναι σημαντικό να είναι ξεκάθαρο σε ποιο κριτήριο από όλα γίνεται αναφορά καθώς και μέσω ποιας μεθοδολογίας γίνεται η αξιολόγηση . Κριτήρια αξιολόγησης λοιπόν, είναι και ο δείκτης ακρίβειας (accuracy ratio AR), η Kolmogorov-Smirnov στατιστική (KS), η εκτίμηση Leave-one-out-bootstrap (LOO-B) του ποσοστού ακρίβειας , η εκτίμηση 0.632 και το τυπικό σφάλμα ακρίβειας. Παρακάτω, αναλύονται μαζί με τα κριτήρια αυτά οι διαδικασίες και οι τεχνικές μέσα από τις οποίες γίνεται η αξιολόγηση των συστημάτων ταξινόμησης.

2.3.1 Το ποσοστό ακρίβειας (error rate)

Το κριτήριο αυτό, αποτελεί ένα από τα ποιο δημοφιλή στοιχεία μέτρησης για την εκτίμηση της απόδοσης ενός κανόνα ταξινόμησης. Λόγω των πολλών διαφορετικών μεθόδων που αναπτύχθηκαν για την δημιουργία συστημάτων ταξινόμησης , παράλληλα αναπτύχθηκαν και διαφορετικοί τύποι για το ποσοστό ακρίβειας. Γενικά , αυτό ορίζεται ως η αναλογία των νέων αντικειμένων που έχουν ταξινομηθεί λανθασμένα από τον κανόνα ταξινόμησης που χρησιμοποιείται κάθε φορά και επιτρέπει μέσω του υπολογισμού του να γίνονται συγκρίσεις μεταξύ διαφορετικών ταξινομητών. Οι κυριότεροι τύποι του ποσοστού ακρίβειας είναι :

- **Bayes error rate (e_B).** Είναι το ελάχιστο-βέλτιστο πιθανό σφάλμα που μπορεί να επιτευχθεί από ένα σύστημα ταξινόμησης και για τον υπολογισμό του είναι απαραίτητη η γνώση της κατανομής της πιθανότητας κάθε κατηγορίας ταξινόμησης. Δίνεται από τον τύπο :

$$e_B = \int e(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int [1 - \max p(w_k | \mathbf{x})] f(\mathbf{x}) d\mathbf{x}$$

όπου $f(\mathbf{x})$ είναι η κατανομή όλου του πληθυσμού δεδομένων και $p(w_k|\mathbf{x})$ είναι οι τελικές τους πιθανότητες να ανήκουν στην κατηγορία i, δεδομένου του διανύσματος χαρακτηριστικών $\mathbf{x}$ του καθενός.

- **Actual error rate** e_c ενός ταξινομητή c. Εκφράζει την αναμενόμενη πιθανότητα ένας κανόνας ταξινόμησης να κατηγοριοποιήσει λανθασμένα μια καινούρια παρατήρηση. Οπότε ,

$$e_c = \int \hat{e}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} ,$$

όπου $\hat{e}(\mathbf{x})$ είναι το αποκαλούμενο true error rate ή διαφορετικά το υπό όρους σφάλμα λόγω της κατάστασης του δείγματος δεδομένων στο οποίο βασίζεται ο ταξινομητής.

- **Expected error rate** e_E **.** Εκφράζει την αναμενόμενη τιμή του e_c για διαφορετικά σύνολα m αντικειμένων εκπαίδευσης που χρησιμοποιούνται για τον σχεδιασμό του κανόνα ταξινόμησης. Συνεπώς $e_E = E(e_c)$, άρα

$$\int \cdots \int e_c(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod f(\mathbf{x}_i) d\mathbf{x}_1 \ldots d\mathbf{x}_m$$

2.3.2 Δείκτης Kolmogorov – Smirnov (KS-test)

Ένα πολύ διαδεδομένο κριτήριο αξιολόγησης συστημάτων ταξινόμησης ,ειδικά στην αξιολόγηση πιστωτικού κινδύνου, είναι ο δείκτης Kolmogorov – Smirnov ή αλλιώς KS – test. Το κριτήριο αυτό μετράει την απόσταση μεταξύ των διαγραμμάτων των συνολικών συναρτήσεων κατανομής των κατηγοριών ταξινόμησης, όπου κάθε σκορ ταξινόμησης έχει μετασχηματιστεί έτσι ώστε να κυμαίνεται μεταξύ 0 και 1. Εκείνη η τιμή για την οποία προκύπτει η μεγαλύτερη διαφορά μεταξύ των συναρτήσεων, θεωρείται ως η τιμή – όριο με βάση την οποία ένα αντικείμενο κατατάσσεται στην μια ή την άλλη κατηγορία . Το μοντέλο για το οποίο προκύπτει η μεγαλύτερη διαφορά μεταξύ των δυο κατανομών αξιολογείται ως το καταλληλότερο για την ταξινόμηση.

Ειδικότερα, συμβολίζοντας ως s_i τη βαθμολογία ενός αντικειμένου i από κάποιο μοντέλο ταξινόμησης και θεωρώντας ένα διαχωριστικό όριο S διαμορφώνεται η ακόλουθη συνάρτηση κατανομής:

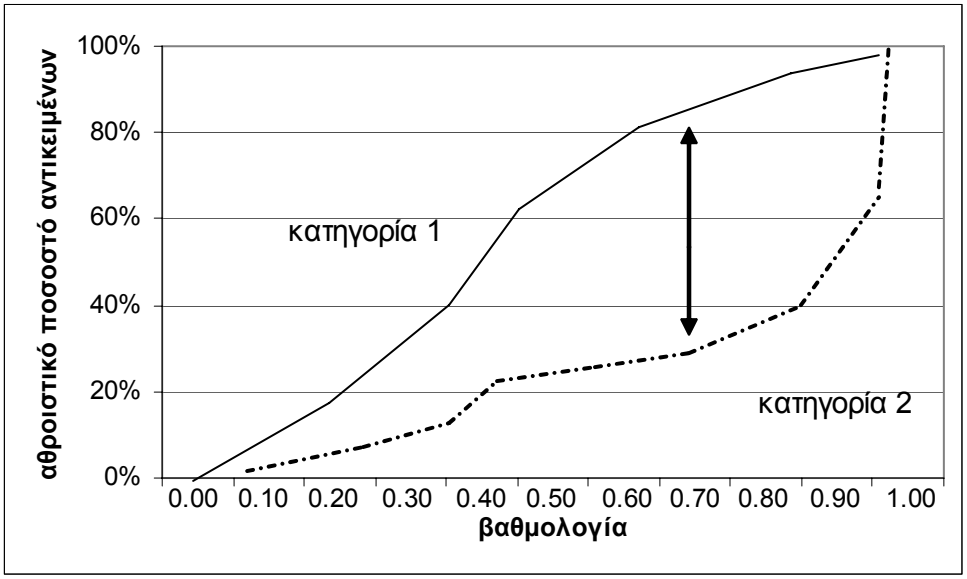
$$F(S) = \frac{1}{m} \sum_{i=1}^{m} I(s_i \leq S)$$

όπου $I(s_i \leq S) = 1$ εάν $s_i \leq S$ και $I(s_i \leq S) = 0$ εάν $s_i >$S.

Στην περίπτωση που η ταξινόμηση γίνεται σε δύο κατηγορίες (1 και 2), αυτή η συνάρτηση κατανομής προσδιορίζεται για κάθε κατηγορία και το μέγεθος της διαφοράς μεταξύ των δυο συναρτήσεων F_1 και F_2 προσδιορίζει το δείκτη KS ως εξής :

$$KS = \max_i \left| F_1(s_i) - F_2(s_i) \right| \text{ , με } i = 1, 2, \ldots, \text{m.}$$

Ένα γραφικό παράδειγμα της παραπάνω διαδικασίας φαίνεται παρακάτω στην εικόνα 2.4. Για κάποιο μοντέλο ταξινόμησης η μεγαλύτερη διαφορά των δυο συναρτήσεων κατανομής προκύπτει για την τιμή 0.7 και γι'αυτήν η μια συνάρτηση δίνει 80% και η άλλη 32%. Δηλαδή, το 80% των περιπτώσεων της κατηγορίας 1 έχουν βαθμολογία μικρότερη ή ίση με 0.7 ενώ το αντίστοιχο ποσοστό για την κατηγορία 2 είναι 32%. Άρα η max διαφορά $\left| F_1(s_i) - F_2(s_i) \right|$ είναι 80% - 32% = 48%. Αν για κάποιο άλλο μοντέλο προκύψει η αντίστοιχη διαφορά μεγαλύτερη του 48% τότε αυτό αξιολογείται ως καλύτερο.



Εικόνα 2.4 Απεικόνιση KS-test

Βασικό μειονέκτημα της μεθόδου αυτής είναι ότι λαμβάνει υπ'οψιν μόνο ένα είδος κόστους από λάθος ταξινόμηση, ενώ στην πραγματικότητα δεν είναι έτσι. Για παράδειγμα , στην περίπτωση της εκτίμησης πιστωτικού κινδύνου υπολογίζεται μόνο το κόστος της λάθος ταξινόμησης ενός αφερέγγυου πελάτη ως φερέγγυο, τη στιγμή που υπάρχει αλλά δεν λαμβάνεται υπ'όψιν και το κόστος ευκαιρίας που προκύπτει από την λάθος ταξινόμηση ενός φερέγγυου πελάτη ως αφερέγγυο. Κατά συνέπεια , δεν λαμβάνεται επίσης υπ'όψιν και ποιο από τα δύο κόστη είναι σημαντικότερο.

2.3.3 ROC Ανάλυση

Γενικά , αποτελεί μια κλασσική μεθοδολογία της θεωρίας εντοπισμού σήματος και πιο πρόσφατα της ιατρικής διάγνωσης. Η καμπύλη ROC (Receiver Operating Characteristic) βασίζεται στην αποσύνθεση της αποτελεσματικότητας σε δύο κατηγορίες , την πραγματική θετική (true positive) και την λάθος θετική (false positive).

Έστω λοιπόν ένα μοντέλο ταξινόμησης $g(X) \to Y$, όπου Y είναι η ταξινόμηση των αντικειμένων σε θετικές (p) και αρνητικές (n) περιπτώσεις. Η ανάλυση βασίζεται στην παρατήρηση τεσσάρων τύπων αποτελεσμάτων :

- true positives, όπου $(g(X) = p) \Omega (Y = p)$
- false positives, όπου $(g(X) = p) \Omega (Y = n)$
- true negatives, όπου $(g(X) = n) \Omega (Y = n)$
- false negatives, όπου $(g(X) = n) \Omega (Y = p)$

οπότε και προκύπτουν τα αποτελέσματα TP (true positives) και FP (false positives), τα οποία ορίζονται ως εξής :

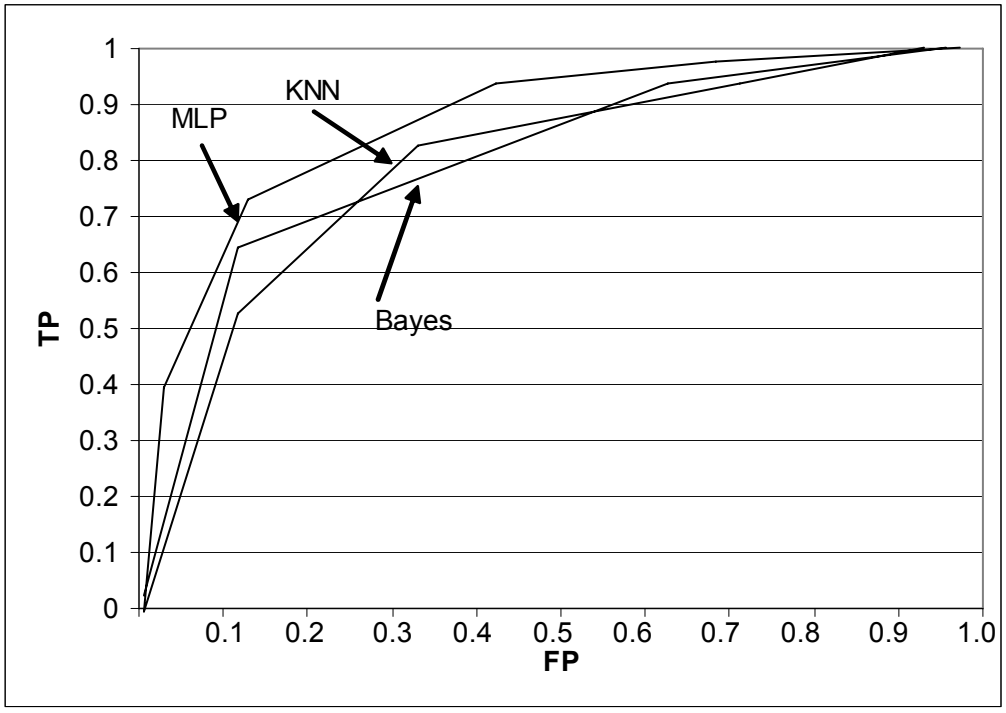
$$TP = \frac{true\ positives}{total\ positives}$$

$$= P(g(X) = p \mid Y = p)$$

$$FP = \frac{false\ positives}{total\ negatives}$$

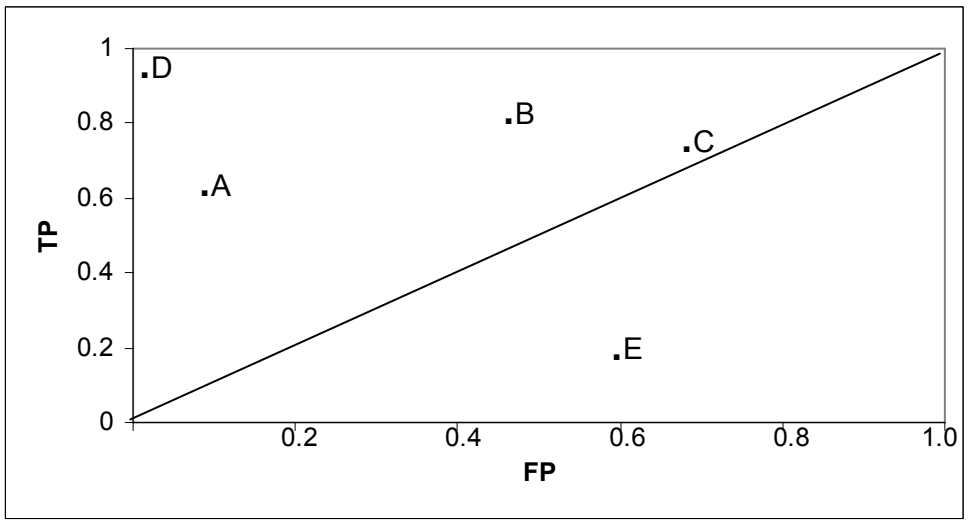
$$= P(g(X) = p \mid Y = n)$$

Σε μοντέλα λοιπόν που παράγουν συνεχή αποτελέσματα ,προκύπτουν ζευγάρια (TP, FP) , με τα TP να τοποθετούνται στον Y άξονα και τα FP στον X άξονα και από τις διάφορες τιμές τους προκύπτει για κάθε σύστημα ταξινόμησης η αντίστοιχη ROC καμπύλη που εκτείνεται από το (0,0), σημείο που δείχνει ότι ο ταξινομητής δεν διαπράττει false positives λάθη, μέχρι το (1,1), σημείο που έχει την ακριβώς αντίθετη έννοια με το προηγούμενο. Σημαντικό επίσης είναι και το σημείο (0,1) που αναπαριστά την τέλεια ταξινόμηση. Συγκρίνοντας τις καμπύλες που παράγονται για διάφορους ταξινομητές , μπορούν να εξαχθούν συμπεράσματα για το ποιο μοντέλο είναι καλύτερο από τα άλλα και σε ποια διαστήματα . Ένα σημείο είναι καλύτερο από ένα άλλο αν οι τιμές των TP και FP του, είναι μεγαλύτερη και μικρότερη αντίστοιχα από την τιμή του TP και του FP του άλλου σημείου. Στην περίπτωση μάλιστα που η καμπύλη ενός μοντέλου βρίσκεται σε όλο το ROC διάστημα πάνω από τις καμπύλες των άλλων, τότε το μοντέλο αυτό κυριαρχεί. Στην εικόνα 2.5 παρακάτω φαίνονται οι ROC καμπύλες διάφορων συστημάτων ταξινόμησης.



Εικόνα 2.5 ROC καμπύλες για τους ταξινομητές Bayes, K-NN και MLP

Στις περιπτώσεις των διακριτών ταξινομητών, όπου έχουν ως έξοδο μόνο μια ταμπέλα ταξινόμησης , η αναπαράστασή τους στο ROC διάστημα γίνεται από ένα μόνο σημείο – ζεύγος (TP, FP) , όπως φαίνεται στην εικόνα 2.6.



Εικόνα 2.6 Η γραφική ROC για πέντε διακριτούς ταξινομητές

Ισχύουν ότι και στα μοντέλα συνεχών αποτελεσμάτων και διαγώνια γραμμή $y = x$ βοηθά στο να καθορίζεται ποιο σύστημα ταξινόμησης έχει καλύτερη απόδοση και εκφράζει την στρατηγική της τυχαίας επιλογής μιας κατηγορίας ταξινόμησης . Έτσι όσοι ταξινομητές βρίσκονται στο τρίγωνο που ορίζεται κάτω και δεξιά της γραμμής αυτής , αποδίδουν χειρότερα ακόμα και από αυτή τη στρατηγική (ταξινομητής Ε στο παράδειγμα της εικόνας 2.6).

2.3.3.1 Η περιοχή κάτω από την ROC καμπύλη (AUC)

Ένα ακόμα κριτήριο μέτρησης της απόδοσης ενός συστήματος ταξινόμησης , το οποίο σχετίζεται άμεσα με την ROC καμπύλη, είναι και η περιοχή κάτω από αυτή την καμπύλη (AUC). Στον πίνακα 2.1 παρακάτω συνοψίζονται τα περισσότερα από τα στοιχεία που βοηθούν την διαδικασία αξιολόγησης. Τέτοια στοιχεία είναι οι τα αποτελέσματα TP (true positive), FP (false positive), FN (false negative), TN (true negative), όπως αυτά ορίστηκαν παραπάνω μέσα από την ROC ανάλυση.

Πίνακας 2.1 Περιγραφή των αποτελεσμάτων TP, FP, FN, TN

Πραγματική κατηγορία	**Κατηγορία που ταξινομήθηκε**	**Κατηγορία που ταξινομήθηκε**	
	1	**2**	
1	**TN**	**FP**	**Cn**
2	**FN**	**TP**	**Cp**
	Rn	**Rp**	

Οι μεταβλητές C_n, C_p εκφράζουν τον αριθμό των TN και των TP αντίστοιχα παραδειγμάτων και οι R_n, R_p οι συνολικοί αριθμοί των παραδειγμάτων που ταξινομήθηκαν ως negative και ως positive αντίστοιχα .

Τέλος, σημαντικά για τη διαδικασία ταξινόμησης στοιχεία μέτρησης είναι και τα ακόλουθα :

$$\text{ακρίβεια (1- error)} = \frac{T_p + T_n}{C_p + C_n} = P(C)$$

$$\text{Ευαισθησία (1-β)} = \frac{T_p}{C_p} = P(T_p)$$

$$\text{Ανάκληση (1-α)} = \frac{T_n}{C_n} = P(T_n)$$

$$\text{Αναλογία ταξινόμησης θετικών} = \frac{T_p}{R_p}$$

$$\text{Αναλογία ταξινόμησης αρνητικών} = \frac{T_n}{R_n}$$

Σκοπός είναι η ελαχιστοποίηση του κόστους λάθους ταξινόμησης , το οποίο υπολογίζεται ως εξής :

$$\cos t = FP \times C_{FP} + FN \times C_{FN}$$

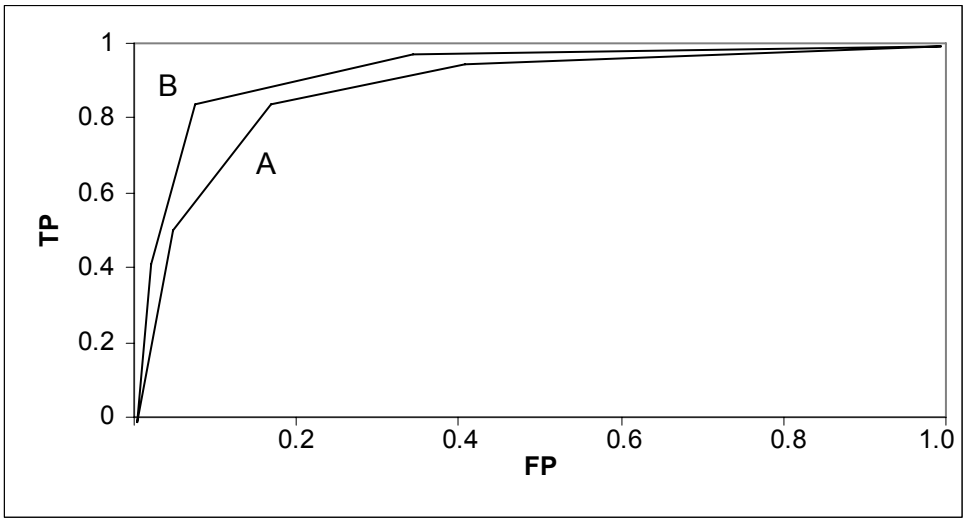
Ωστόσο, είναι αναγκαίο να βρεθεί ένας απλούστερος τρόπος αξιολόγησης και αυτό επιτυγχάνεται με τη χρήση της AUC ως κριτήριο μέτρησης. Ο απλούστερος τρόπος υπολογισμού του είναι η χρήση της τραπεζοειδούς ολοκλήρωσης

$$AUC = \sum_i \{(1- b_i) \times \Delta a) + \frac{1}{2}[\Delta(1- b) \times \Delta a\}$$

όπου $\Delta(1- b) = (1- b_i) - (1- b_{i-1})$

και $\Delta a = a_i - a_{i-1}$

Το σύστημα ταξινόμησης για το οποίο η περιοχή κάτω από την ROC καμπύλη του είναι μεγαλύτερη σε σχέση με των άλλων ταξινομητών, είναι αυτό που έχει την καλύτερη κατά μέσο όρο απόδοση. Στην εικόνα 2.7 απεικονίζεται η AUC για δυο ταξινομητές A , B.



Εικόνα 2.7 Η AUC για δυο ταξινομητές Α και Β

Όπως φαίνεται ο ταξινομητής Β υπερέχει σχεδόν εξολοκλήρου του ταξινομητή Α. Σχεδόν για όλες τις τιμές η ROC καμπύλη του Β βρίσκεται πάνω από την αντίστοιχη του Α. Φυσικά, υπάρχουν και περιπτώσεις που δεν συμβαίνει κάτι τέτοιο, δηλαδή σε κάποια υποδιαστήματα υπερέχει η καμπύλη του ενός συστήματος ταξινόμησης και σε άλλα η καμπύλη του άλλου.

2.3.3.2 Δείκτης Ακρίβειας (AR)

Με τη βοήθεια του εμβαδού κάτω από την ROC καμπύλη (AUC), έτσι όπως υπολογίζεται στην προηγούμενη παράγραφο, είναι ευκολο να υπολογιστεί και ο δείκτης ακρίβειας (AR). Ο δείκτης αυτός αποτελεί ακόμα ένα κριτήριο αξιολόγησης των συστημάτων ταξινόμησης και η τιμή του προκύπτει από τη σχέση που ακολουθεί:

$$AR = 2 * AUC - 1$$

Το μοντέλο ταξινόμησης είναι τόσο πιο αξιόπιστο όσο πιο κοντά βρίσκεται η τιμή του δείκτη ακρίβειας στη μονάδα.

2.3.4 K–fold Cross -Validation

Για τον υπολογισμό της αποτελεσματικότητας ενός μοντέλου ταξινόμησης απαιτούνται κατάλληλα δεδομένα ελέγχου τα οποία όμως δεν είναι πάντα εύκολα διαθέσιμα. Για την αντιμετώπιση του θέματος αυτού έχουν αναπτυχθεί κατάλληλες υπολογιστικές διαδικασίες. Μια από τις πιο διαδεδομένες τέτοιες διαδικασίες είναι το K-fold Cross-Validation. Σύμφωνα με τη διαδικασία αυτή το σύνολο των δεδομένων X χωρίζεται σε K υποσύνολα ίσου μεγέθους. Κάθε φορά , το k- οστό υποσύνολο σχηματίζει ένα σύνολο ελέγχου X_{val}, ενώ τα υπόλοιπα αποτελούν το σύνολο εκπαίδευσης X_{learn} που χρησιμοποιείται για την ανάπτυξη του μοντέλου. Το σφάλμα υπολογίζεται από τον τύπο

$$Err^{(cvk)} = \frac{\sum_{i=1}^{m/K} (g(\mathbf{x}_i^{val}) - y_i^{val})^2}{m/K}$$

με $(\mathbf{x}_i^{val}, y_i^{val})$ τα στοιχεία του X_{val} και $g(\mathbf{x}_i^{val})$ η προσέγγιση του y_i^{val} απ' το μοντέλο g.
Η παραπάνω διαδικασία επαναλαμβάνεται για όλα τα K υποσύνολα, δηλαδή από το 1 μέχρι το K. Το μέσο σφάλμα υπολογίζεται από τον ακόλουθο τύπο

$$Err^{cvk} = \frac{\sum_{k=1}^{K} Err^{cvk}}{K}$$

Μια πιο συγκεκριμένη περίπτωση της μεθόδου k-fold Cross-Validation, είναι και η **Leave-One-Out** όπου εκεί το K είναι ίσο με το m. Δηλαδή, κάθε φορά μένει έξω από το σύνολο εκπαίδευσης ένα μόνο σημείο, αυτό για το οποίο πρόκειται να γίνει η πρόβλεψη.

2.3.5 Η μέθοδος Bootstrap

Το bootstrap είναι μια εναλλακτική διαδικασία που μπορεί να χρησιμοποιηθεί αντί του Cross-Validation. Τα διαδοχικά βήματα που ακολουθεί αυτή η μέθοδος αξιολόγησης των συστημάτων ταξινόμησης είναι τα ακόλουθα :

1. Αρχικά ακολουθείται η διαδικασία της μετατροπής του ληφθέντος δείγματος δεδομένων (re-sampling). Σύμφωνα με αυτή, από το σύνολο των δεδομένων X επιλέγονται τυχαία m στοιχεία με επανατοποθέτηση (with replacement). Τα στοιχεία αυτά σχηματίζουν το σύνολο X_{learn} εκμάθησης του ταξινομητή. Το υπόλοιπο σύνολο αποτελεί πλέον το σύνολο εκτίμησης της εγκυρότητας του συστήματος ταξινόμησης X_{val}. Η διαδικασία αυτή επαναλαμβάνεται B φορές, οπότε διαμορφώνονται B δείγματα bootstrap X_{learn}^1, X_{learn}^2,..., X_{learn}^B που χρησιμοποιούνται για την ανάπτηξη

αντιστοίχων μοντέλων $g_1, g_2, ..., g_B$ και B δείγματα "out –of- the bootstrap" $X^1_{val,}$ $X^2_{val,...,}$ X^B_{val} που χρησιμοποιούνται για τον έλεγχο των μοντέλων.

2. Για κάθε δείγμα X^b_{learn} (b = 1, 2,…, B) αναπτύσσεται ένα μοντέλο g_b και υπολογίζονται τα σφάλματα Err^{learn}_k και Err^{val}_k από τους τύπους :

$$Err^{learn}_k = \frac{\varepsilon \quad (g_b(\mathbf{x}_i) - y_i)^2}{m}$$

και

$$Err^{val}_k = \frac{\varepsilon \quad (g_b(\mathbf{x}_i) - y_i)^2}{\left|x^b_{val}\right|}$$

Όπου $\left|x^b_{val}\right|$ είναι το πλήθος των περιπτώσεων στο δείγμα ελεγχου x^b_{val} .

3. Το αναμενόμενο σφάλμα του μοντέλου υπολογίζεται ως :

$$error_{OB} = \frac{1}{m}\varepsilon^{m}_{i=1}\frac{1}{B_i}\varepsilon \quad (g_b(\mathbf{x}_i - y_i)^2$$

Όπου B_i είναι το πλήθος των δειγμάτων ελέγχου $X^1_{val,}$ $X^2_{val,...,}$ X^B_{val}, που περιλαμβάνουν το αντικείμενο i .

Ειδικές περιπτώσεις της Bootstrap μεθόδου αποτελούν η μέθοδος **Bootstrap .632** και ο **Bootstrap εκτιμητής .632+,** καθώς και η μέθοδος **leave-one-out Bootstrap (LOO-B)**.

2.3.5.1 Bootstrap .632 και .632+

Στην μέθοδο **Bootstrap .632** η διαδικασία που ακολουθείται είναι ανάλογη της παραπάνω, με τη διαφορά ότι ο τύπος του γενικού σφάλματος προσαρμόζονται ως εξής :

$$Error_{gen} = .368 Error^{l} + .632 error_{OB}$$

Όπου $error^{l}$ είναι το σφάλμα του μοντέλου που αναπτύσσεται στο σύνολο των δεδομένων X**,** μετρούμενο βάσει των δεδομένων εκπαίδευσης.

Βελτίωση της παραπάνω μεθόδου αποτελεί η **Bootstrap .632+,** σύμφωνα με την οποία το σφάλμα υπολογίζεται από τον τύπο :

$$\overline{Error}_{gen} = (1- w)\,.368 Error^{l} + w\,.632 \overline{error}_{OB}$$

Όπως φαίνεται , αποδίδεται μεγαλύτερη βαρύτητα w στο σφάλμα $\overline{error}_{OB}$ και η τιμή του δίνεται από τη σχέση

$$w = \frac{.632}{1- .368r},$$

όπου

$$r = \frac{\overline{error}_{OB} - Error^{l}}{g - Error^{l}},$$

και

$$g = \sum_{i=1}^{K} p_i(1- q_i),$$

όπου p_i είναι η αναλογία των δειγμάτων της i-οστής κατηγορίας και q_i η αναλογία όσων τελικά ταξινομήθηκαν σε αυτήν από το μοντέλο.

2.3.5.2 Leave – One - Out Bootstrap (LOO-B)

Η μέθοδος αυτή αποτελεί εξομάλυνση του leave-one-out Cross-Validation μεθόδου μέσω του Bootstrap και προβλέπει το σφάλμα ταξινόμησης ενός στοιχείου $\mathbf{x}_i$ μέσα από τη χρήση ενός συνόλου δειγμάτων Bootstrap, το οποίο δεν περιλαμβάνει το $\mathbf{x}_i$. Πιο συγκεκριμένα, μέσα από τη διαδικασία λήψης δειγμάτων από το σύνολο των δεδομένων με αντικατάσταση (re-sampling) προκύπτουν B δείγματα μεγέθους m και δημιουργείται ο bootstrap κανόνας ταξινόμησης g_b για κάθε ένα από τα δείγματα Bootstrap. Τότε το σφάλμα εκτιμάται από την σχέση

$$Error = \sum_{i=1}^{m} E_i / m,$$

όπου

$$E_i = \frac{\sum_{b=1}^{B} I_{ib} Q_{ib}}{\sum_{b=1}^{B} I_{ib}},$$

με $I_{ib} = 1$, αν το $\mathbf{x}_i$ δεν περιλαμβάνεται στο b-οστό δείγμα Bootstrap , αλλιώς $I_{ib} = 0$ και $Q_{ib} = 1$ αν το μοντέλο g_b ταξινομεί λάθος το $\mathbf{x}_i$, αλλιώς $Q_{ib} = 0$.

2.3.6 Τυπικό Σφάλμα Ακρίβειας

Το Bootstrap εκτός από την εκτίμηση της αναμενόμενης αποτελεσματικότητας ενός μοντέλου μπορεί να χρησιμοποιηθεί και για τον υπολογισμό της αβεβαιότητας όσον αφορά την αποτελεσματικότητα του μοντέλου. Αυτό μπορεί να γίνει με τον υπολογισμό του τυπικού σφάλματος της ακρίβειας του μαντέλου:

$$se_B = \sqrt{\frac{1}{B-1}\sum_b (\hat{q}_i - \hat{q})^2}$$

όπου B ο αριθμός των δειγμάτων Bootstrap , $\hat{q}_i$ η i-οστή Bootstrap εκτίμηση και $\hat{q}$ η μέση τιμή των Bootstrap εκτιμήσεων.

ΚΕΦΑΛΑΙΟ 3: ΠΕΙΡΑΜΑΤΙΚΗ ΑΝΑΛΥΣΗ

3.1 Εφαρμογές – Δεδομένα

Στο κεφάλαιο αυτό γίνεται μια προσπάθεια εφαρμογής της πολυκριτήριας αξιολόγησης συστημάτων ταξινόμησης , μέσω της πειραματικής ανάλυσης δεδομένων που προέρχονται από προβλήματα ταξινόμησης πραγματικών εφαρμογών. Τα δεδομένα που χρησιμοποιούνται στην πειραματική ανάλυση προέρχονται από τη βάση UCI Machine Learning Repository. Συνολικά εξετάζονται 7 σύνολα δεδομένων σε καθένα από τα οποία εφαρμόζονται έξι μέθοδοι ταξινόμησης (Γραμμική Διακριτική Ανάλυση - LDA, Λογιστική Παλινδρόμηση - LR, Πιθανοτικά Νευρωνικά Δίκτυα - PNN, Κ-Πλησιέστεροι Γείτονες - KNN, Γραμμικές (LSVM) και μη γραμμικές μηχανές διανυσμάτων υποστήριξης με τετραγωνικό πυρήνα (QSVM) και πυρήνα RBF (RSVM), Δένδρα Ταξινόμησης και Παλινδρόμησης - CART). Οι μέθοδοι συγκρίνονται με τα κριτήρια αξιολόγησης που αναλύθηκαν στο προηγούμενο κεφάλαιο (Στατιστικό μέγεθος Kolmogorov-Smirnov, Δείκτης Ακρίβειας, εκτίμηση Leave-one-out-bootstrap του ποσοστού ακρίβειας, εκτίμηση 0.632, Τυπικό Σφάλμα Ακρίβειας). Όλα τα μεγέθη προέκυψαν ως μέσοι όροι 30 επαναλήψεων bootstrap. Παρακάτω παρουσιάζονται αναλυτικότερα και ξεχωριστά η καθεμία από τις εφαρμογές στις οποίες έγινε αναφορά και που χρησιμοποιήθηκαν για την πειραματική ανάλυση.

3.1.1 Διάγνωση καρκίνου του μαστού

Πρόκειται για μια εφαρμογή στη διάγνωση του καρκίνου του μαστού και στην αξιολόγηση της ταξινόμησης που προέκυψε από τη μελέτη του πανεπιστημίου Wisconsin. Οι ασθενείς ταξινομούνται σε δύο κατηγορίες , σε αυτούς που έχουν ήπιας μορφής καρκίνο (B = Benign) και σε αυτούς που βρίσκονται σε πολύ πιο προχωρημένο (θανάσιμο) στάδιο (M = Malignant). Η διαδικασία της ταξινόμησης βασίζεται στην μέτρηση 30 συγκεκριμένων χαρακτηριστικών σε κάθε ένα ασθενή ξεχωριστά . Τα χαρακτηριστικά αυτά υπολογίζονται με τη βοήθεια της ψηφιακής απεικόνισης μιας μάζας του στήθους και περιγράφουν τους κυτταρικούς πυρήνες που φαίνονται σε αυτήν την ψηφιακή απεικόνιση. Τα χαρακτηριστικά περιγραφής των κυτταρικών πυρήνων που υπολογίζονται είναι τα :

1. ακτίνα (ως η μέση απόσταση των σημείων της περιφέρειας από το κέντρο)
2. σύσταση (ως η απόκλιση από την ενδιάμεση αναλογία μεταξύ των τιμών)
3. περίμετρος
4. εμβαδόν
5. ομαλότητα (ως η τοπική απόκλιση σε διάφορα μήκη)
6. κατάληψη μικρού χώρου (από τον τύπο περίμετρος^2 / εμβαδόν -1)
7. κοιλότητα (ως η σοβαρότητα των κοίλων επιφανειών του περιγράμματος)
8. κοίλα σημεία (ως ο αριθμός των κοιλωμάτων του περιγράμματος)
9. συμμετρία
10. κλασματική διάσταση (ως “γραμμή προσέγγισης” – 1)

Για το κάθε ένα από αυτά τα δέκα χαρακτηριστικά υπολογίζονται τρεις τιμές , η μέση τιμή , το τυπικό σφάλμα και το χειρότερο ή μεγαλύτερο το οποίο υπολογίζεται από τη μέση τιμή των τριών μεγαλύτερων τιμών του κάθε χαρακτηριστικού. Έτσι, προκύπτουν συνολικά τα 30 χαρακτηριστικά με βάση τα οποία γίνεται η ταξινόμηση.

Στον πίνακα δεδομένων που χρησιμοποιήθηκε για την εφαρμογή αυτή, φαίνονται επίσης το ID του κάθε ασθενή στην πρώτη στήλη (A), η τελική ταξινόμηση κάθε περιστατικού στη δεύτερη στήλη (B) και στις υπόλοιπες (C – AF) οι τιμές των μετρήσεων για τα 30 χαρακτηριστικά σε 569 περιστατικά ταξινόμησης. Συνοπτικά, στον πίνακα 3.1 παρακάτω παρουσιάζεται η δομή του πίνακα δεδομένων.

Πίνακας 3.1 Δομή του πίνακα δεδομένων

Στήλη	**Χαρακτηριστικό**	**Τύπος**
A	ID του ασθενή	
B	**Ταξινόμηση**	**B(=Benign) ή M(=Malignant)**
C D E	1. ακτίνα	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
F G H	2. σύσταση	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
I J K	3. περίμετρος	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
L M N	4. εμβαδόν	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
O P Q	5. ομαλότητα	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
R S T	6. κατάληψη μικρού χώρου	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
U V W	7. κοιλότητα	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)

X Y Z	8. κοίλα σημεία	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
AA AB AC	9. συμμετρία	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)
AD AF AE	10. κλασματική διάσταση	Μέση τιμή (αριθμός) Τυπικό σφάλμα (αριθμός) μέσος των 3 μεγαλύτερων τιμών (αριθμός)

3.1.2 Αξιολόγηση πιστοληπτικής ικανότητας

Το παράδειγμα αυτό αναφέρεται στην έγκριση παροχής πίστωσης μέσα από την εξέταση αιτήσεων για πιστωτική κάρτα και πώς τελικά ταξινομούνται τα περιστατικά , αν εγκρίνονται ή όχι και η αξιολόγηση αφορά ακριβώς αυτή την ταξινόμηση. Τα δεδομένα παρέχονται κυρίως με τη μορφή συμβόλων, ώστε ουσιαστικά να μην αποκαλύπτεται η πραγματική σημασία των επιλεγμένων χαρακτηριστικών ταξινόμησης , προστατεύοντας έτσι το απόρρητο των δεδομένων.

Αρχικά, η βάση δεδομένων αφορά 690 περιστατικά, ενώ τελικά η αξιολόγηση γίνεται για την ταξινόμηση των 666 από αυτά αφού, όπως θα αναλυθεί παρακάτω στη μεθοδολογία που ακολουθείται, μερικά περιστατικά διαγράφονται από τα δεδομένα λόγω του ότι κάποια στοιχεία τους λείπουν. Αυτό το σύνολο των δεδομένων είναι ιδιαίτερα ενδιαφέρον σε ότι αφορά την ταξινόμηση γιατί αποτελεί ένα μείγμα ποσοτικών και ποιοτικών μεταβλητών. Όλες οι μεταβλητές κανονικοποιούνται κατά τέτοιο τρόπο ώστε τελικά να παίρνουν τιμές που κυμαίνονται από 0 έως 1.

Στον συγκεντρωτικό πίνακα όλων των αρχικών δεδομένων του συγκεκριμένου παραδείγματος, η δομή έχει ως εξής : οι πρώτες δεκαπέντε στήλες (A – O) αφορούν τα χαρακτηριστικά μέτρησης και τις τιμές που παίρνει σε αυτά κάθε περιστατικό προς ταξινόμηση και η τελευταία στήλη (P) δείχνει ποια είναι η τελική ταξινόμηση κάθε περίπτωσης.

Πιο συγκεκριμένα , το χαρακτηριστικό της πρώτης στήλης (A) μπορεί να χαρακτηρίζεται από τις μεταβλητές b ή a ανάλογα με την περίπτωση. Για τα χαρακτηριστικά της δεύτερης (B) και της τρίτης (C) στήλης , οι τιμές που μπορούν να πάρουν από περιστατικό σε περιστατικό είναι συνεχείς αριθμοί . Ομοίως και για τα χαρακτηριστικά της όγδοης (H), ενδέκατης (K), δέκατης τέταρτης (N) και δέκατης πέμπτης (O) στήλης. Κάθε περίπτωση

ταξινόμησης, με βάση το χαρακτηριστικό μέτρησης της τέταρτης (D) στήλης, συμβολίζεται με έναν από τους χαρακτήρες u, y, l, t, ενώ για το χαρακτηριστικό της πέμπτης (E) στήλης με έναν από τους g, p, gg. Για αυτό της έκτης (F) στήλης, έχουν επιλεχθεί οι χαρακτήρες c, d,

cc, i, j, k, m, r, q, w, x, e, aa, ff και για την έβδομη (G) οι v, h, bb, j, n, z, dd, ff, o. Για τα χαρακτηριστικά των στηλών I, J και L οι 690 περιπτώσεις χαρακτηρίζονται με ένα από τα γράμματα t, f και στην στήλη M το χαρακτηριστικό μέτρησης συμβολίζεται με ένα από τα g, p, s, ανάλογα με την περίπτωση. Τέλος, η ταξινόμηση γίνεται σε δύο κατηγορίες +, - και φαίνεται στην τελευταία στήλη (P) του πίνακα δεδομένων του παραδείγματος αυτού.

Συνολικά , παρατηρήθηκαν 37 περιπτώσεις (5%) στις οποίες έλειπαν κάποια στοιχεία τα οποία ανά χαρακτηριστικό (στήλη) κατανέμονται ως εξής :

Στήλη A : λείπουν 12 τιμές
Στήλη B : λείπουν 12 τιμές
Στήλη D : λείπουν 6 τιμές
Στήλη E : λείπουν 6 τιμές
Στήλη F : λείπουν 9 τιμές
Στήλη G : λείπουν 9 τιμές
Στήλη P : λείπουν 13 τιμές

Η κατανομή της τελικής ταξινόμησης είναι η παρακάτω :

\+ : 307 (44.5%)
\- : 383 (55.5%) .

Παρακάτω, στον πίνακα 3.2 φαίνονται συγκεντρωτικά πως κατανέμονται ανά στήλες στον πίνακα δεδομένων τα χαρακτηριστικά και οι πιθανές τιμές που μπορούν να πάρουν.

Πίνακας 3.2 Δομή του πίνακα δεδομένων

Στήλη	**Χαρακτηριστικό**	**Τύπος**
A	1.	Ποιοτικό
B	2.	Ποσοτικό-συνεχές
C	3.	Ποσοτικό-συνεχές
D	4.	Ποιοτικό
E	5.	Ποιοτικό
F	6.	Ποιοτικό
G	7.	Ποιοτικό
H	8.	Ποσοτικό-συνεχές
I	9.	Ποιοτικό
J	10.	Ποιοτικό
K	11.	Ποσοτικό-συνεχές
L	12.	Ποιοτικό
M	13.	Ποιοτικό
N	14.	Ποσοτικό-συνεχές
O	15.	Ποσοτικό-συνεχές
P	**ΤΑΞΙΝΌΜΗΣΗ**	+ ή -

3.1.3 Ταξινόμηση ηλεκτρονίων της Ιονόσφαιρας

Τα δεδομένα του παραδείγματος αυτού αφορούν στοιχεία που προέκυψαν από έρευνα της ομάδας φυσικής του διαστήματος του πανεπιστημίου John Hopkins. Μέσω ενός συστήματος ραντάρ που τοποθετήθηκε στο Goose Bay στο Labrador και που αποτελείται από μια διάταξη 16 κεραιών υψηλής συχνότητας με συνολική εκπεμπόμενη ισχύ 6.4 KW, στόχος ήταν η μελέτη των ελευθέρων ηλεκτρονίων της ιονόσφαιρας με απώτερο σκοπό την ταξινόμησή τους σε "καλά" και "κακά" βάση το σήμα που επιστρέφουν στο ραντάρ. Έτσι, "καλά" χαρακτηρίζονται εκείνα που επιστρέφουν ένδειξη ότι υπάρχει κάποια δομή στην ιονόσφαιρα και "κακά" εκείνα που δεν βοηθούν στην εξαγωγή κάποιου τέτοιου συμπεράσματος αφού δεν επιστρέφουν σήμα ότι υπάρχει δομή στην ιονόσφαιρα.

Τα σήματα που ελήφθησαν επεξεργάστηκαν μέσα από μια συνάρτηση αυτοσυσχέτισης της οποίας ορίσματα ήταν η στιγμή του ηχητικού παλμού και ο αριθμός αυτών. Στην περιοχή του Goose Bay παρατηρήθηκαν 17 είδη ηχητικών παλμών, ο κάθε ένας από τους οποίους περιγράφεται από 2 χαρακτηριστικά . Συνεπώς η ταξινόμηση των ελευθέρων ηλεκτρονίων γίνεται με βάση τις τιμές τους σε 34 χαρακτηριστικά .

Σύμφωνα λοιπόν με όλα τα παραπάνω, ο πίνακας που συγκεντρώνει όλα τα δεδομένα για την δημιουργία ενός συστήματος ταξινόμησης και την αξιολόγησή του, αποτελείται από 35 στήλες (A – AI). Οι 34 (A – AH) πρώτες είναι τα χαρακτηριστικά μέτρησης, τα οποία παίρνουν τιμές συνεχείς αριθμούς και η τελευταία στήλη (AI) δείχνει την ταξινόμηση σε δύο κατηγορίες του κάθε ηλεκτρονίου που μελετήθηκε, με βάση τις τιμές του στα 34 χαρακτηριστικά . Οι κατηγορίες ταξινόμησης εκφράζονται ως "good" και "bad" και ο συνολικός αριθμός των περιπτώσεων (ηλεκτρονίων) που έχουν ταξινομηθεί είναι 351.

3.1.4 Διάγνωση διαταραχών του ήπατος

Το παράδειγμα αυτό αποτελεί μια εφαρμογή της ταξινόμησης στο πεδίο της ιατρικής και σχετίζεται με διαταραχές του ήπατος και πως αυτές συνδέονται με κάποια χαρακτηριστικά η μέτρηση των οποίων στηρίζεται σε κάποιες εξετάσεις αίματος. Η πηγή των δεδομένων είναι η BUPA Medical Research Ltd. Οι 6 μεταβλητές που χρησιμοποιούνται είναι όλες τεστ αίματος που σχετίζουν τις διαταραχές με την υπερβολική κατανάλωση αλκοόλ. Έτσι, τα χαρακτηριστικά που υπολογίζονται είναι τα ακόλουθα :

1. mcv: μέσο μοριακό πλήθος
2. Alkphos: αλκαλική φωσφωτάση
3. Sgpt: alamine aminotransferase
4. Sgot: aspartate aminotransferase
5. Gammagt: gamma-glut amyl transpeptidase
6. Drinks: ο αριθμός των αλκοολικών ισοδύναμων ανά ημέρα

Στον πίνακα δεδομένων της εφαρμογής αυτής, φαίνονται αντίστοιχα οι τιμές των παραπάνω χαρακτηριστικών στις πρώτες 6 στήλες (A-F) , για 345 περιπτώσεις προς ταξινόμηση και στην τελευταία στήλη (G) παρουσιάζεται η τελική ταξινόμηση κάθε περίπτωσης. Οι κατηγορίες ταξινόμησης είναι δύο και αναφέρονται ως 1 αν δεν προκύπτουν διαταραχές στο συκώτι με βάση τις εξετάσεις αίματος ή 2 αν με βάση τα χαρακτηριστικά μέτρησης η πρόβλεψη είναι οτι θα υπάρξουν τέτοια προβλήματα .

Στον πίνακα 3.3 φαίνεται η γενική δομή του πίνακα δεδομένων.

Πίνακας 3.3 Δομή του πίνακα δεδομένων

Στήλη	**Χαρακτηριστικό**	**Τύπος**
A	mcv	Ποσοτικό-συνεχές
B	Alkphos	Ποσοτικό-συνεχές
C	Sgpt	Ποσοτικό-συνεχές
D	Sgot	Ποσοτικό-συνεχές
E	Gammagt	Ποσοτικό-συνεχές
F	Drinks	Ποσοτικό-συνεχές
G	**Ταξινόμηση**	1 ή 2

3.1.5 Διάγνωση διαβήτη

Ένα ακόμα πεδίο της ιατρικής στο οποίο βρίσκει εφαρμογή η ταξινόμηση και συνεπώς είναι απαραίτητη και η αξιολόγησή της , είναι και η πρόβλεψη αν κάποιος έχει συμπτώματα διαβήτη. Στο παράδειγμα που αναλύεται , τα δεδομένα έχουν προκύψει από την εξέταση ενός πλήθους ανθρώπων που ζει στο Φοίνιξ της Αριζόνα των Ηνωμένων Πολιτειών Αμερικής. Όλος ο πληθυσμός που εξετάστηκε αποτελείται από γυναίκες τουλάχιστον 21 ετών και καταγωγής Pima Indian και η ταξινόμησή του στηρίχθηκε στη μέτρηση οκτώ χαρακτηριστικών. Ο αριθμός των περιστατικών που ταξινομήθηκε ανέρχεται στα 768 και οι κατηγορίες ταξινόμησης είναι δύο.

Ο αρχικός συγκεντρωτικός πίνακας που δίνει τα δεδομένα και τα αποτελέσματα της ταξινόμησης, αποτελείται από εννέα στήλες (A – I), οι οκτώ πρώτες (A- H) για κάθε ένα από τα χαρακτηριστικά και η τελευταία (I) για την τελική ταξινόμηση η οποία δίδεται σε μορφή 0 ή 1 (0 αν το τεστ του κάθε περιστατικού είναι αρνητικό ως προς την εκδήλωση διαβήτη και 1 αν είναι θετικό). Συνολικά 768 περιστατικά ταξινομήθηκαν, εκ των οποίων 500 στην κατηγορία 0 και 268 στην κατηγορία 1. Επίσης, πιο συγκεκριμένα , τα χαρακτηριστικά στα οποία βασίστηκε το σύστημα ταξινόμησης ώστε να προκύψουν τα παραπάνω, με τη σειρά που φαίνονται στον πίνακα δεδομένων, είναι τα :

1. Πόσες φορές έμεινε έγκυος κάθε γυναίκα .
2. Η συγκέντρωση πλάσματος γλυκόζης σε 2 ώρες μέσα από τεστ αντοχής σε γλυκόζη.
3. Πίεση αίματος κατά τη διαστολή (mm Hg).
4. Πάχος δέρματος του τρικέφαλου (mm).
5. Δίωρος ορός ινσουλίνης (mu u/ml).
6. Δείκτης μυϊκής μάζας (βάρος σε kg/(ύψος σε m)^2).
7. Συνάρτηση κληρονομικότητας διαβήτη.
8. Ηλικία.

Το μέγεθος όλων των χαρακτηριστικών προσδιορίζεται μέσα από συνεχείς αριθμητικές τιμές και κάποια στατιστική ανάλυση σχετικά με τις τιμές που παίρνει η κάθε μία φαίνεται στον παρακάτω πίνακα (3.4) :

Πίνακας 3.4 Μέσος Όρος και Τυπική Απόκλιση των τιμών των χαρακτηριστικών

Αριθμός χαρακτηριστικού	**Μέσος όρος**	**Τυπική απόκλιση**
1.	3.8	3.4
2.	120.9	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

3.1.6 Πρόβλεψη Αποτελέσματος Παιγνίων

Το παράδειγμα αυτό αφορά το παιχνίδι τρίλιζα και τα δεδομένα που προέκυψαν ύστερα από μια προσπάθεια να αναλυθεί το παιχνίδι αυτό και όλες οι πιθανές καταστάσεις πάνω του πίνακα πάνω στον οποίο παίζεται . Για την μελέτη γίνεται αρχικά η υπόθεση ότι το "x" ξεκινά πρώτο το παιχνίδι και απώτερος σκοπός είναι να γίνει γνωστό το πότε κερδίζει το "x". Πρόκειται δηλαδή, για μια ταξινόμηση του τέλους του παιχνιδιού και όλων των δυνατών εκδοχών που μπορεί να προκύψουν για αυτό, σε δύο κατηγορίες :

- **Θετική**, όταν η κατάληξη του παιχνιδιού είναι νίκη του "x". Στον πίνακα των αρχικών δεδομένων και πιο συγκεκριμένα στην στήλη της ταξινόμησης (J) εκφράζεται από τον αριθμό 1.

- **Αρνητική**, όταν δεν κερδίζει το "x". Κάτι τέτοιο στον πίνακα των δεδομένων εκφράζεται από τον αριθμό 2.

Εξετάζονται συνολικά εννέα κριτήρια για την εξαγωγή της ταξινόμησης και όλα αφορούν τα τετράγωνα του πίνακα του παιχνιδιού και τι επιλογή έχει γίνει γι'αυτό. Αν δηλαδή έχει παίξει σ'αυτό ο παίκτης με το σύμβολο "x" ή ο παίκτης με το σύμβολο "o" ή αν ακόμα είναι κενό (b). Συνοπτικά λοιπόν τα κριτήρια και οι πιθανές τιμές που μπορούν να πάρουν παρουσιάζονται στον πίνακα 3.1.6 παρακάτω, όπου επίσης φαίνεται και σε ποιες στήλες του πίνακα δεδομένων αντιστοιχούν τα χαρακτηριστικά και η τελική ταξινόμηση.

Πίνακας 3.5 Δομή του πίνακα δεδομένων

Στήλη	χαρακτηριστικό	Πιθανές Τιμές
A	1. αριστερό τετράγωνο $1^{ης}$ γραμμής	{x,o,b}
B	2. μεσαίο τετράγωνο $1^{ης}$ γραμμής	{x,o,b}
C	3. δεξί τετράγωνο $1^{ης}$ γραμμής	{x,o,b}
D	4. αριστερό τετράγωνο μεσαίας γραμμής	{x,o,b}
E	5. μεσαίο τετράγωνο μεσαίας γραμμής	{x,o,b}
F	6. δεξί τετράγωνο μεσαίας γραμμής	{x,o,b}
G	7. αριστερό τετράγωνο τελευταίας γραμμής	{x,o,b}
H	8 μεσαίο τετράγωνο τελευταίας γραμμής	{x,o,b}
I	9. δεξί τετράγωνο τελευταίας γραμμής	{x,o,b}
J	**Ταξινόμηση**	1 ή 2

Συνολικά προέκυψαν 958 πιθανές περιπτώσεις , από τις οποίες το 65,3% κατέληξαν σε νίκη του "x" (θετικά ταξινομημένες περιπτώσεις). Στον πίνακα που παρουσιάζονται συνολικά τα δεδομένα και η τελική ταξινόμηση , οι εννέα πρώτες στήλες (A – I) αναφέρονται στα χαρακτηριστικά και η τελευταία στην τελική ταξινόμηση κάθε περίπτωσης. Επίσης , οι πιθανές επιλογές x, o, b έχουν αντικατασταθεί από τους αριθμούς 1, 2, 3 αντίστοιχα .

3.1.7 Μελέτη πολιτικής συμπεριφοράς

Η εφαρμογή αυτή αναλύει δεδομένα που προέκυψαν από την εκλογική διαδικασία του 1984 για την εκλογή του Κογκρέσου των Ηνωμένων Πολιτειών της Αμερικής. Στα δεδομένα περιλαμβάνονται ψήφοι από κάθε πολιτεία και με βάση 16 τύπους – κλειδιά ψήφων , έτσι όπως αναγνωρίστηκαν από το CQA (Congressional Quarterly Almanac). Το CQA, καθόρισε 9 τύπους ψήφων : voted for, paired for, announced for (οι 3 αυτοί τύποι ψήφων απλοποιούνται ως θετικοί), voted against, paired against, announced against (οι 3 αυτοί τύποι απλοποιούνται ως αρνητικοί), voted present, voted present to avoid conflict of interest και did not vote ή διαφορετικά make a position known (αυτοί οι 3 απλοποιούνται στην άγνωστη διάθεση).

Μέσα από τη μελέτη 435 ψηφοφόρων και 16 συγκεκριμένων χαρακτηριστικών, σκοπός ήταν η ταξινόμησή τους σε δύο κατηγορίες :

- Δημοκρατικοί (democrats)

και

- Ρεπουμπλικάνοι (republicans)

Στον αρχικό πίνακα δεδομένων που χρησιμοποιήθηκε σε αυτήν την εφαρμογή, εκτός από την ταξινόμηση που φαίνεται στην στήλη A, φαίνονται επίσης και οι τιμές για τα 16 χαρακτηριστικά σε κάθε μια από τις 435 περιπτώσεις. Τα χαρακτηριστικά που μετρώνται συνοψίζονται στον παρακάτω πίνακα (3.6).

Πίνακας 3.6 Τιμές των χαρακτηριστικών

Χαρακτηριστικό	**Πιθανές Τιμές**
1. handicapped-infants (στήλη B)	y (yes) ή n (no)
2. water-project-cost-sharing (στήλη C)	y (yes) ή n (no)
3. adoption-of-the-budget-resolution (στήλη D	y (yes) ή n (no)
4. physician-fee-freeze (στήλη E)	y (yes) ή n (no)
5. El-Salvador-aid (στήλη F)	y (yes) ή n (no)
6. religious-groups-in-schools (στήλη G)	y (yes) ή n (no)
7. anti-satellite-test-ban (στήλη H)	y (yes) ή n (no)
8. aid-to-Nicaraguan-contras (στήλη I)	y (yes) ή n (no)
9. mx-missile (στήλη J)	y (yes) ή n (no)
10. immigration (στήλη K)	y (yes) ή n (no)
11. synfuels-corporation-cutback (στήλη L)	y (yes) ή n (no)
12. education-spending (στήλη M)	y (yes) ή n (no)
13. superfund-right-to-sue (στήλη N)	y (yes) ή n (no)
14. crime (στήλη O)	y (yes) ή n (no)
15. duty-free-exports (στήλη P)	y (yes) ή n (no)
16. export-administration-act-south-Africa (στήλη Q)	y (yes) ή n (no)

Σε κάποιες από τις περιπτώσεις, σε μερικά χαρακτηριστικά η τιμή που παίρνουν δεν είναι καμία από αυτές που τους αναλογούν σύμφωνα με τον παραπάνω πίνακα , αλλά έχουν ένα ερωτηματικό (?). Αυτό δεν σημαίνει ότι η τιμή είναι άγνωστη αλλά ότι η απάντηση δεν είναι ούτε ναι (yes) ούτε όχι (no). Επίσης , στον συνολικό πίνακα δεδομένων κάποια από τα στοιχεία λείπουν και όπως αναφέρεται και παρακάτω στην περιγραφή της μεθοδολογίας που ακολουθείται, οι περιπτώσεις στις οποίες εμφανίζεται κάτι τέτοιο διαγράφονται και δεν λαμβάνουν μέρος στην διαδικασία της αξιολόγησης της ταξινόμησης.

Τελικά, από τους 435 ψηφοφόρους οι 168 (45.2%) έχουν ταξινομηθεί ως δημοκρατικοί και οι υπόλοιποι 267 (54.8%) ως ρεπουμπλικάνοι.

3.2 Μεθοδολογία

Για τις εφαρμογές που αναφέρθηκαν παραπάνω, ακολουθείται μια γενική μεθοδολογία με σκοπό την αξιολόγηση των συστημάτων ταξινόμησης που εφαρμόστηκαν. Οι μέθοδοι ταξινόμησης που αξιολογούνται είναι αυτές που αναλύονται στο κεφάλαιο 2, δηλαδή η Γραμμική Διακριτική Ανάλυση, η Λογιστική Παλινδρόμηση, τα Πιθανοτικά Νευρωνικά Δίκτυα, η μέθοδος των πλησιέστερων γειτόνων, οι μηχανές διανυσμάτων υποστήριξης και πιο συγκεκριμένα οι γραμμικές, ο πυρήνας RBF και ο τετραγωνικός πυρήνας και τέλος τα Δένδρα Ταξινόμησης. Τα παραπάνω ταξινομητικά συστήματα αξιολογούνται με την βοήθεια πέντε κριτηρίων τα οποία είναι :

1. Δείκτης ακρίβειας (accuracy ratio, AR)
2. Δείκτης Kolmogorov – Smirnov (KS)
3. Εκτίμηση Leave – one – out – bootstrap του ποσοστού ακρίβειας (LOO – B)
4. Εκτίμηση 0.632 bootstrap του ποσοστού ακρίβειας
5. Τυπικό Σφάλμα Ακρίβειας (SE)

Τα βήματα έτσι όπως ακολουθήθηκαν σε κάθε εφαρμογή είναι τα παρακάτω:

1. Μετατροπή όλων των στοιχείων των αρχικών πινάκων δεδομένων κατά τέτοιο τρόπο ώστε όλες οι τιμές για όλα τα χαρακτηριστικά και για όλα τα περιστατικά σε κάθε εφαρμογή κυμαίνονται μεταξύ των τιμών 0 και 1. Κάτι τέτοιο επιτυγχάνεται με τη χρήση του τύπου :

$$(\mathbf{a_{ij}} - \mathbf{min_j}) / (\mathbf{max_j} - \mathbf{min_j})$$

και εφαρμόζεται για όλες της στήλες των αρχικών πινάκων εκτός από αυτές που αναφέρονται στην τελική ταξινόμηση, για τις οποίες ακολουθείται η διαδικασία του βήματος 2.

Πιο αναλυτικά , για κάθε στήλη **j** του πίνακα δεδομένων βρίσκεται το στοιχείο που παίρνει την ελάχιστη τιμή ($\mathbf{min_j}$) και το στοιχείο με τη μέγιστη τιμή ($\mathbf{max_j}$) και για όλα τα στοιχεία της στήλης αυτής ($\mathbf{a_{ij}}$) εφαρμόζεται ο παραπάνω τύπος.

2. Οι στήλες που περιέχουν την περιγραφή της ταξινόμησης κάθε περιστατικού, συμπληρώνονται με τα στοιχεία 1 ή 2. Δηλαδή, αντικαθίστανται οι αρχικές κατηγορίες ταξινόμησης , όποια και αν ήταν η περιγραφή τους (σύμβολο, γράμμα ή αριθμός), με τον αριθμό 1 για την μια κατηγορία και με τον αριθμό 2 για την άλλη. Πρέπει εδώ να τονιστεί ότι όλες οι εφαρμογές που εξετάζονται είναι δυαδικής κατηγοριοποίησης , δηλαδή ταξινομούν τα περιστατικά σε δύο μόνο κατηγορίες.

3. Σε περιπτώσεις εφαρμογών όπου κάποια από τα χαρακτηριστικά που χρησιμοποιούνται για την ταξινόμηση είναι ποιοτικά και συμβολίζονται με κάποια γράμματα ανάλογα με το περιστατικό, τότε χρησιμοποιούνται δυαδικές ψευδομεταβλητές.

4. Σε κάποιες από τις εφαρμογές λείπουν κάποια στοιχεία σε ορισμένα περιστατικά . Σε αυτές τις περιπτώσεις , λόγω της έλλειψης δεδομένων , δεν είναι δυνατό να αξιολογηθούν τα συστήματα ταξινόμησης για τα συγκεκριμένα περιστατικά , συνεπώς αυτά διαγράφονται από το σύνολο των δεδομένων.

5. Αφού λοιπόν τα δεδομένα μετασχηματιστούν με βάση τα παραπάνω βήματα και οι πίνακες πάρουν την απαιτούμενη μορφή (όλα τα στοιχεία αριθμοί, οι τιμές των οποίων μεταξύ 0 και 1 εκτός της τελευταίας στήλης που είναι η ταξινόμηση και έχει τιμές ή 1 ή 2), εισάγονται σε ένα κατάλληλα διαμορφωμένο πρόγραμμα Matlab. Το πρόγραμμα αυτό, με την εισαγωγή των αρχείων των δεδομένων, δημιουργεί και εκπαιδεύει τα συστήματα ταξινόμησης που έχουν επιλεχθεί , χρησιμοποιώντας ένα μέρος των δεδομένων και με βάση αυτή προχωράει στην πρόβλεψη της ταξινόμησης για τα υπόλοιπα δεδομένα .

6. Τα αποτελέσματα των μεθόδων σε κάθε σύνολο δεδομένων αξιολογούνται με τη μέθοδο **Promethee.** Πρόκειται για το τελευταίο και σημαντικότερο βήμα, αφού μέσα από αυτό προκύπτει η τελική αξιολόγηση των συστημάτων ταξινόμησης και χρήσιμα στατιστικά στοιχεία μεταξύ τους.

   Η μέθοδος Promethee είναι μια πολυκριτήρια μέθοδος αξιολόγησης. Η μέθοδος αυτή ανήκει στη θεωρία των σχέσεων υπεροχής και προτάθηκε για πρώτη φορά το 1982 από τον Brans. Οι βασικές αρχές που την διέπουν είναι οι ακόλουθες :

   - Επέκταση στην έννοια των κριτηρίων. Προτείνονται στον αποφασίζοντα κάποιες νέες συναρτήσεις κριτηρίων, όπως κριτήριο τελείως αυστηρό, κριτήριο λιγότερο αυστηρό (περιέχει δηλαδή και περιοχή αδιαφορίας), κριτήριο με γραμμική προτίμηση, κριτήριο με περιοχές προτίμησης κ.λ.π.

   - Εκτιμώμενη σχέση υπεροχής. Σε μικρές τροποποιήσεις , η μέθοδος δεν είναι πολύ ευαίσθητη και συνεπώς τα αποτελέσματα που δίνει μπορούν εύκολα να ερμηνευτούν.

   - Εκμετάλλευση της σχέσης υπεροχής. Πραγματοποιείται όταν οι εναλλακτικές λύσεις πρέπει να ταξινομηθούν από την καλύτερη προς τη χειρότερη.

Σύμφωνα με τη μέθοδο αυτή, ο πολυκριτήριος δείκτης προτίμησης που προσδιορίζει την εκτιμώμενη σχέση υπεροχής υπολογίζεται από τη σχέση :

$$P(a,b) = \frac{\sum_{i=1}^{k} p_i P_i(a,b)}{\sum_{i=1}^{k} p_i}$$

Όπου π_i είναι το βάρος κάθε κριτηρίου , $P_i(a,b)$ η συνάρτηση προτίμησης και Π(α, b) αντιπροσωπεύει την ένταση προτίμησης του αποφασίζοντα για την εναλλακτική λύση α έναντι της b. Παίρνει τιμές μεταξύ 0 και 1 με όσο πιο κοντά την τιμή στο 0 να σημαίνει τόσο πιο αδύνατη προτίμηση της λύσης α έναντι της b και όσο πιο κοντά στην τιμή 1, τόσο πιο ισχυρή προτίμηση της α από την b.

Σε ότι αφορά την εκμετάλλευση της σχέσης υπεροχής, ορίζονται δύο ροές:

- Η εξερχόμενη $f^{+}(a) = \sum_{b \Xi K} P(a,b)$, K= το σύνολο των εναλλακτικών λύσεων.
- Η εισερχόμενη $f^{-}(a) = \sum_{b \Xi K} P(b,a)$

Οι δύο αυτές ροές υποδηλώνουν μια πρώτη ταξινόμηση για κάθε εναλλακτική λύση. Η μεγαλύτερη εξερχόμενη ροή δηλώνει ότι μια εναλλακτική υπερέχει των άλλων και η μικρότερη εισερχόμενη ροή ότι μια εναλλακτική λύση κυριαρχείται από τις άλλες.

Στη συγκεκριμένη ανάλυση εξετάστηκαν 50 διαφορετικά τυχαία σενάρια βαρών για τα 5 κριτήρια αξιολόγησης. Τα βάρη αυτά προέκυψαν από μια γεννήτρια τυχαίων αριθμών ως ομοιόμορφα κατανεμημένες τυχαίες μεταβλητές στο διάστημα [0 100]και παρουσιάζονται στον πίνακα 3.7 που ακολουθεί.

Πίνακας 3.7 Τιμές των βαρών των κριτηρίων αξιολόγησης

AR	KS	LOO-B	0.632-B	SE
71	54	58	29	31
78	2	77	82	71
5	42	87	80	38
97	88	6	95	37
53	77	6	60	47
30	63	65	27	28
83	83	59	99	92
23	70	99	25	54
11	100	68	2	58
11	11	80	29	5
30	39	31	95	98
41	28	17	17	65
42	42	72	33	64
21	19	59	9	46
91	27	79	38	29
92	64	63	43	10
57	70	92	84	3
55	92	44	68	51
52	47	36	41	27
6	25	98	7	40
37	49	16	48	26
63	55	16	94	66
51	40	11	79	46
76	60	84	2	22
8	11	34	13	1
54	66	55	83	9
20	68	46	36	15
71	93	54	9	76
41	47	50	21	33
10	59	17	93	10
45	28	88	76	28
68	26	9	4	33
80	30	24	49	26
35	5	49	21	87
59	76	93	34	55
9	64	42	97	12
93	63	35	15	48
22	100	14	3	35
55	93	54	41	85
83	68	73	100	34
50	42	70	18	43
55	82	55	43	51
23	62	49	69	89
38	31	30	16	53
23	59	37	88	48
20	69	75	62	79
17	81	21	96	7
7	80	38	47	12
12	18	5	72	54
57	22	47	75	76

Στη συνέχεια, για κάθε ένα από τα παραπάνω βάρη εφαρμόζονται 10 διαφορετικά σενάρια και συνολικά προκύπτουν 500 ροές για κάθε σύστημα ταξινόμησης. Τα σενάρια που εξετάζονται χαρακτηρίζονται από ελαστικά έως αυστηρά. Ο χαρακτηρισμός αυτός αναφέρεται στην ευκολία με την οποία μπορεί να εξαχθεί συμπέρασμα για το ποιο σύστημα ταξινόμησης είναι καλύτερο από κάποιο άλλο. Ένα ελαστικό σενάριο δεν απαιτεί η απόκλιση μεταξύ των αντίστοιχων ροών των σεναρίων που εξετάζονται να είναι μεγάλη. Αντιθέτως, τα πιο αυστηρά απαιτούν η υπεροχή ενός συστήματος ταξινόμησης από ένα άλλο να είναι ξεκάθαρη.

Από τις τιμές των ροών για κάθε βάρος σε κάθε σενάριο και για τα οκτώ συστήματα που αξιολογούνται και τη σύγκριση μεταξύ τους , προκύπτει κάθε φορά ποιος από τους ταξινομητές θεωρείται ως καλύτερος.

3.3 Αποτελέσματα

3.3.1 Διάγνωση καρκίνου του μαστού

Με την πραγματοποίηση των τεσσάρων πρώτων βημάτων, τα δεδομένα μετασχηματίζονται στην κατάλληλη μορφή που απαιτείται ώστε να γίνει εφαρμογή όλων των μεθόδων ταξινόμησης. Ο πίνακας 3.8 παρακάτω δείχνει ποιες είναι οι τιμές των κριτηρίων όπως αυτές προέκυψαν μετά το πέμπτο βήμα της διαδικασίας και για κάθε ένα από τα οκτώ συστήματα ταξινόμησης που πρόκειται να αξιολογηθούν.

Πίνακας 3.8 Τιμές των κριτηρίων για κάθε σύστημα ταξινόμησης για την εφαρμογή διάγνωσης του καρκίνου του μαστού

Ταξινομητής	AR	KS	LOO-B	0.632-B	SE
LDA	0,9777	0,90838	0,9577	0,96162	0,0074445
LR	0,9464	0,88022	0,94432	0,96093	0,0073051
PNN	0,95029	0,92433	0,96936	0,9787	0,0061637
KNN	0,97667	0,9179	0,95915	0,96384	0,0069476
LSVM	0,97765	0,91554	0,9633	0,96516	0,0073535
RSVM	0,98232	0,93223	0,96297	0,96495	0,0071837
QSVM	0,97995	0,92544	0,95583	0,96433	0,0074124
CART	0,86832	0,81382	0,91124	0,91545	0,008439

Οι τιμές των παραμέτρων pnn_smoothing, svm_smoothing, nearest_neighbor, cart_splitmin και bs_samples για τις οποίες προέκυψε η μέγιστη απόδοση στις τιμές των κριτηρίων για κάθε ταξινομητή, φαίνονται στον παρακάτω πίνακα (3.9):

Πίνακας 3.9 Τιμές των παραμέτρων των ταξινομητών για την εφαρμογή διάγνωσης του καρκίνου του μαστού

Παράμετροι	**Τιμή**
pnn_smoothing	0.2
svm_smoothing	0.1
nearest_neighbor	21
cart_splitmin	0.1

Η αντιστοιχία των παραπάνω παράμετρων με τη μορφή που εμφανίζονται στους τύπους του κάθε συστήματος ταξινόμησης, όπως αυτοί παρουσιάστηκαν στο Κεφάλαιο 2 είναι η εξής :

pnn_smoothing = σ
svm_smoothing = δ
nearest_neighbor = k

και είναι ίδια για όλες τις εφαρμογές που αναλύονται.

Χρησιμοποιώντας ως δεδομένα για τη μέθοδο PROMETHEE τα αποτελέσματα του πίνακα 3.8 προκύπτουν κάποια αποτελέσματα που συγκεντρώνονται σε ένα πίνακα στον οποίο φαίνονται οι τιμές όλων των ροών για κάθε βάρος των κριτηρίων, σε κάθε σενάριο και για κάθε σύστημα ταξινόμησης. Στον πίνακα 3.10 που ακολουθεί και με βάση τα αποτελέσματα που παρουσιάζονται στον συγκεντρωτικό πίνακα των ροών, διακρίνονται χρήσιμα στατιστικά στοιχεία για την αποτελεσματικότητα των συστημάτων ταξινόμησης όταν αυτά συγκρίνονται μεταξύ τους. Έτσι, φαίνεται η κατάταξή τους με βάση το μέσο όρο των ροών τους (RANK(av)), η κατάταξη τους στην μέγιστη ροή τους (RANK(at max) και η αντίστοιχη στην ελάχιστη (RANK(at min)) όπως επίσης και η κατάταξή που προκύπτει γι'αυτά από τη σύγκριση των μέγιστων (RANK(max)) και των ελάχιστων τιμών όλων (RANK(min)).

Από τον συγκεντρωτικό πίνακα των ροών μπορεί επίσης εύκολα να προκύψει και πίνακας που να παρουσιάζει αναλυτικότερα την κατάταξη των συστημάτων ταξινόμησης για κάθε σενάριο και κάθε βάρος. Με βάση αυτά τα αποτελέσματα, ο πίνακας 3.11 παρουσιάζει συνοπτικά ποια είναι η συνήθης κατάταξη κάθε ταξινομητή στην συγκεκριμένη εφαρμογή.

Πίνακας 3.10 Κατάταξη των συστημάτων ταξινόμησης με βάση χαρακτηριστικά στατιστικά μεγέθη

ταξινομητής	AVERAGE	RANK (av)	MAX	MIN	RANK (max)	RANK (min)	RANK (at max)	RANK (at min)
LDA	0.40142	6	1.0176	-0.7848	6	6	6	6
LR	-0.69872	7	0.1144	-4.1642	7	7	7	7
PNN	1.83026	1	5.6337	0.45085	1	1	1	1
KNN	0.90179	3	2.6788	0.32728	3	2	2	2
LSVM	0.72564	4	2.0450	0.20494	4	4	3	4
RSVM	1.02958	2	3.3858	0.27869	2	3	1	3
QSVM	0.62178	5	1.7115	-0.23660	5	5	4	5
CART	-4.81178	8	-2.0452	-6.99999	8	8	8	8

Πίνακας 3.11 Κατανομή κατατάξεων των ταξινομητών για όλες τις ροές

κατάταξη/πόσες φορές	LDA	LR	PNN	KNN	LSVM	RBF SVM	QSVM	CART
1η	0	0	487	0	0	13	0	0
2η	0	0	10	112	0	378	0	0
3η	0	0	2	309	64	109	16	0
4η	0	0	0	76	358	0	66	0
5η	4	0	1	3	78	0	414	0
6η	496	0	0	0	0	0	4	0
7η	0	500	0	0	0	0	0	0
8η	0	0	0	0	0	0	0	500
χειρότερη κατάταξη	6	7	5	5	5	3	6	8
καλύτερη κατάταξη	5	7	1	2	3	1	3	8

Όπως φαίνεται, για τη συγκεκριμένη εφαρμογή τα πιθανοτικά νευρωνικά δίκτυα είναι η καταλληλότερη μέθοδος ταξινόμησης. Περίπου στο 97% των σεναρίων στα οποία γίνεται η σύγκριση με τα άλλα συστήματα ταξινόμησης, τα PNN κατατάσσονται στην $1^{η}$ θέση από την πολυκριτήρια μέθοδο PROMETHEE. Η χειρότερη θέση που κατατάχτηκε ο συγκεκριμένος ταξινομητής είναι η $5^{η}$ θέση, αλλά αυτό δεν έγινε παρά μόνο σε μια περίπτωση από σύνολο 500.

Τα πιθανοτικά νευρωνικά δίκτυα τα ακολουθούν σε αποτελεσματικότητα οι μηχανές διανυσμάτων υποστήριξης με πυρήνα RBF και το σύστημα Κ πλησιέστερων γειτόνων (KNN). Το πρώτο σύστημα αποδίδει σχετικά καλύτερα από το δεύτερο και αυτό διαπιστώνεται από το γεγονός ότι στις 378 από τις 500 συνολικά ροές που συγκρίνονται για κάθε ταξινομητή, η RSVM κατατάσσεται στη $2^{η}$ θέση και σε 109 στην $3^{η}$, ενώ για τον KNN συμβαίνει το αντίθετο. Επίσης, η μέθοδος ταξινόμησης RSVM στη χειρότερη των περιπτώσεων έχει καταταχθεί στην $3^{η}$ θέση και στην καλύτερη στην $1^{η}$, ενώ η μέθοδος των Κ πλησιέστερων γειτόνων στην καλύτερη θέση που έχει αξιολογηθεί είναι η $2^{η}$ και η χειρότερη η $5^{η}$.

Οι δύο μέθοδοι με βάση τα αποτελέσματα , που είναι καλύτερο να μην χρησιμοποιηθούν για την πρόβλεψη ταξινόμησης στην εφαρμογή αυτή , είναι τα δένδρα ταξινόμησης (CART) που και στα 500 σενάρια σύγκρισης καταλαμβάνουν την τελευταία θέση ($8^{η}$), η λογιστική παλινδρόμηση (LR) που συνεχώς αξιολογείται στην $7^{η}$ θέση και η γραμμική διακριτική ανάλυση (LDA) που κυρίως βρίσκεται στην $6^{η}$ θέση. Για τα υπόλοιπα συστήματα ταξινόμησης (LSVM και QSVM) η αποτελεσματικότητά τους κυμαίνεται σε μέτρια επίπεδα , δηλαδή μεταξύ $3^{ης}$ και $6^{ης}$ θέσης. Στον πίνακα 3.12 παρουσιάζεται μια γενική κατάταξη των συστημάτων ταξινόμησης για την εφαρμογή της διάγνωσης του καρκίνου του μαστού.

Πίνακας 3.12 Γενική κατάταξη των ταξινομητών για την διάγνωση του καρκίνου του μαστού

Θέση	**Σύστημα ταξινόμησης**
$1^{η}$	PNN
$2^{η}$	RSVM
$3^{η}$	KNN
$4^{η}$	LSVM
$5^{η}$	QSVM
$6^{η}$	LDA
$7^{η}$	LR
$8^{η}$	CART

3.3.2 Αξιολόγηση πιστοληπτικής ικανότητας

Αφού πρώτα εφαρμοστούν όλα τα απαραίτητα βήματα ώστε τα αρχικά δεδομένα να μετασχηματιστούν στους κατάλληλους πίνακες συγκεκριμένης μορφής , εφαρμόζονται όλες οι εξεταζόμενοι μέθοδοι ταξινόμησης. Τα αποτελέσματα που προέκυψαν για τη συγκεκριμένη εφαρμογή παρουσιάζονται στον πίνακα 3.13 παρακάτω. Στον πίνακα 3.14 που ακολουθεί φαίνονται ποιες τιμές επιλέχθηκαν για τις παραμέτρους που χρησιμοποιεί κάθε σύστημα ταξινόμησης.

Πίνακας 3.13 Τιμές των κριτηρίων για κάθε σύστημα ταξινόμησης για την αξιολόγηση της πιστοληπτικής ικανότητας

ταξινομητής	AR	KS	LOO- B	0.632-B	se
LDA	0,83625	0,74056	0,86302	0,8648	0,012813
LR	0,81672	0,71213	0,84962	0,86296	0,011906
PNN	0,83824	0,74628	0,85999	0,86454	0,012528
KNN	0,78333	0,66828	0,84731	0,85266	0,012157
LSVM	0,83549	0,74462	0,86398	0,86375	0,013144
RSVM	0,82383	0,74081	0,86408	0,86824	0,012555
QSVM	0,79469	0,70454	0,84908	0,87368	0,011534
CART	0,7976	0,71404	0,84874	0,85468	0,012232

Πίνακας 3.14 Τιμές των παραμέτρων των ταξινομητών για την αξιολόγηση της πιστοληπτικής ικανότητας

Παράμετροι	**Τιμή**
pnn_smoothing	1.2
svm_smoothing	0.1
nearest_neighbor	31
cart_splitmin	0.2

Με την εισαγωγή των δεδομένων του πίνακα 3.13 στην πολυκριτήρια μέθοδο PROMETHEE, προκύπτουν όλες οι τιμές των ροών που συγκεντρώνονται σε ένα πίνακα όπως ακριβώς και στην 1ης εφαρμογή. Τα βάρη και τα σενάρια παραμένουν τα ίδια . Κατά ανάλογο τρόπο προκύπτουν και τα στατιστικά δεδομένα που αφορούν την κατάταξη των συστημάτων ταξινόμησης για διάφορες χαρακτηριστικές τιμές των ροών τους αλλά και αυτών που δείχνουν γενικά σε ποιες θέσεις κατατάσσεται τις περισσότερες φορές το κάθε μοντέλο. Οι πίνακες που παρουσιάζουν αυτά τα αποτελέσματα φαίνονται ακριβώς παρακάτω και είναι οι 3.15 και 3.16 αντίστοιχα .

Πίνακας 3.15 Κατάταξη των συστημάτων ταξινόμησης με βάση χαρακτηριστικά στατιστικά μεγέθη

ταξινομητής	AVERAGE	RANK (av)	MAX	RANK (max)	MIN	RANK (min)	RANK (at max)	RANK (at min)
LDA	1.014	3	3.500	4	-0.342	3	2	5
LR	-0.220	6	1.020	6	-2.324	5	3	6
PNN	1.140	2	3.202	5	0.138	2	3	4
KNN	-2.632	8	-0.494	8	-6.206	8	8	8
LSVM	0.799	4	3.607	3	-1.345	4	2	6
RSVM	1.251	1	4.101	2	0.197	1	1	3
QSVM	0.121	5	4.579	1	-2.885	6	1	6
CART	-1.474	7	-0.173	7	-4.018	7	7	7

Πίνακας 3.16 Κατανομή κατατάξεων των ταξινομητών για όλες τις ροές

κατάταξη/πόσες φορές	LDA	LR	PNN	KNN	LSVM	RBF SVM	QSVM	CART
1η	28	0	146	0	1	262	63	0
2η	101	0	179	0	43	147	60	0
3η	265	11	126	0	18	66	13	0
4η	66	29	47	0	291	25	41	0
5η	40	149	0	0	75	0	235	0
6η	0	311	0	0	72	0	118	0
7η	0	0	0	0	0	0	0	500
8η	0	0	0	500	0	0	0	0
χειρότερη κατάταξη	5	6	4	8	6	4	6	7
καλύτερη κατάταξη	1	3	1	8	1	1	1	7

Όπως φαίνεται από τους παραπάνω πίνακες, το σύστημα ταξινόμησης που δείχνει να είναι πιο αξιόπιστο στην εφαρμογή αυτή είναι οι μηχανές διανυσμάτων υποστήριξης με πυρήνα RBF (RSVM). Σε αυτό συνηγορεί και το γεγονός ότι στις περισσότερες των περιπτώσεων (πάνω από τις μισές) το συγκεκριμένο σύστημα αξιολογείται ως το καλύτερο ($1^{η}$ θέση) σε σύγκριση με τα υπόλοιπα , ενώ σε πάνω από το 80% των περιπτώσεων κατατάσσεται ως το καλύτερο ή το δεύτερο καλύτερο σύστημα ταξινόμησης από σύνολο οκτώ συστημάτων. Ταυτόχρονα, η χειρότερη κατάταξη στην οποία βρέθηκε είναι η $4^{η}$ θέση και αυτό συνέβη μόλις σε 25 περιπτώσεις (5%) σε σύνολο 500 ροών που εξετάστηκαν.

Συνεχίζοντας, αμέσως μετά ακολουθούν τα πιθανοτικά νευρωνικά δίκτυα (PNN) τα οποία επίσης κατηγοριοποιούνται σε όλες τις περιπτώσεις μεταξύ της $1^{ης}$ και της $4^{ης}$ θέσης. Χαρακτηριστικό βέβαια είναι ότι όπως φαίνεται στον πίνακα 3.15, στη σύγκριση της μέγιστης τιμής των ροών του συστήματος αυτού με αυτές των υπολοίπων επτά, τα πιθανοτικά νευρωνικά δίκτυα βρίσκονται στην πέμπτη θέση. Παρόλα αυτά, για όλες τις άλλες περιπτώσεις που εξετάζονται (RANK(av), RANK(at max), (RANK(at min) και (RANK(min)) αυτή η μέθοδος ταξινόμησης δεν έπεσε κάτω από την $4^{η}$ θέση κατάταξης.

Τα συστήματα ταξινόμησης που σε αυτή την εφαρμογή δεν απέδωσαν καλά, είναι τα δένδρα ταξινόμησης (CART) και η μέθοδος των Κ πλησιέστερων γειτόνων. Και στις 500 περιπτώσεις που προέκυψαν από τα σενάρια που αναλύθηκαν, η κατάταξή τους που προέκυπτε από την σύγκριση των ροών τους με τα άλλα συστήματα ήταν η $7^{η}$ και η $8^{η}$ αντίστοιχα. Οι απόδοση των υπολοίπων μεθόδων (LDA, LR, LSVM, QSVM) κινήθηκε σε μέτρια επίπεδα, με το σύστημα της γραμμικής διακριτικής ανάλυσης να ξεχωρίζει λίγο από τα άλλα, αφού στο 73% των περιπτώσεων κατατάχθηκε στη $2^{η}$ ή $3^{η}$ θέση προτίμησης.

Στον πίνακα 3.17 που ακολουθεί παρουσιάζεται συνοπτικά η αξιολόγηση των ταξινομητών που χρησιμοποιήθηκαν στην εφαρμογή αυτή.

Πίνακας 3.17 Γενική κατάταξη των ταξινομητών για την Αξιολόγηση πιστοληπτικής ικανότητας

Θέση	**Σύστημα ταξινόμησης**
$1^{η}$	RSVM
$2^{η}$	PNN
$3^{η}$	LDA
$4^{η}$	LSVM
$5^{η}$	QSVM
$6^{η}$	LR
$7^{η}$	CART
$8^{η}$	KNN

3.3.3 Ταξινόμηση ηλεκτρονίων της Ιονόσφαιρας

Τα αποτελέσματα της εφαρμογής αυτής, όπως προέκυψαν από τη γενική μεθοδολογία που ακολουθήθηκε, παρουσιάζονται συνοπτικά και κατά ανάλογο τρόπο με τις δύο προηγούμενες εφαρμογές, στους πίνακες 3.18, 3.19, 3.20 και 3.21 παρακάτω. Ο πίνακας 3.18 παρουσιάζει τις τιμές των κριτηρίων AR, KS, LOO-B, εκτίμηση 0.632-B και SE για κάθε σύστημα ταξινόμησης που εφαρμόστηκε στα δεδομένα της εφαρμογής αυτής. Στον πίνακα 3.19 φαίνονται οι τιμές των των παραμέτρων των μεθόδων που χρησιμοποιήθηκαν και προέκυψαν τα παραπάνω αποτελέσματα, ενώ οι πίνακες 3.20 και 3.21 περιέχουν διάφορα στατιστικά στοιχεία σε ότι αφορά την κατάταξη των συστημάτων ταξινόμησης. Τέλος, αξίζει να σημειωθεί ότι χρησιμοποιήθηκε και πίνακας που περιείχε συγκεντρωτικά όλες τις τιμές των ροών για κάθε σενάριο και βάρος όλων των ταξινομητών, έτσι όπως προέκυψαν από τη μέθοδο πολυκριτήριας αξιολόγησης PROMETHEE. Τα αποτελέσματα των πινάκων 3.20 και 3.21 βασίστηκαν στα δεδομένα που παρείχε ο πίνακας αυτός.

Πίνακας 3.18 Τιμές των κριτηρίων για κάθε σύστημα ταξινόμησης για την ταξινόμηση των ηλεκτρονίων της Ιονόσφαιρας

Ταξινομητής	AR	KS	LOO- B	0.632-B	se
LDA	0,75209	0,67285	0,86588	0,88168	-0,015268
LR	0,70089	0,68717	0,86562	0,88991	-0,014027
PNN	0,91957	0,85209	0,8688	0,90031	-0,016866
KNN	0,82288	0,78814	0,8607	0,86793	-0,017377
LSVM	0,75955	0,65266	0,85629	0,87248	-0,016524
RSVM	0,93061	0,80788	0,90359	0,91181	-0,014264
QSVM	0,76063	0,68207	0,86082	0,89211	-0,014293
CART	0,80086	0,75134	0,86617	0,89235	-0,014331

Πίνακας 3.19 Τιμές των παραμέτρων των ταξινομητών για την ταξινόμηση των ηλεκτρονίων της Ιονόσφαιρας

Παράμετροι	**Τιμή**
pnn_smoothing	0.5
svm_smoothing	0.1
nearest_neighbor	11
cart_splitmin	0.1

Πίνακας 3.20 Κατάταξη των συστημάτων ταξινόμησης με βάση χαρακτηριστικά στατιστικά μεγέθη

Ταξινομητής	**AVERAGE**	**RANK (av)**	**MAX**	**MIN**	**RANK (max)**	**RANK (min)**	**RANK (at max)**	**RANK (at min)**
LDA	-0.974089	6	-0.1424	-2.9028	7	6	6	7
LR	-0.522097	5	0.97938	-2.72435	5	5	5	7
PNN	1.2778244	2	4.98378	-0.57015	2	3	2	6
KNN	-1.010038	7	0.70615	-4.55908	6	7	4	7
LSVM	-2.063548	8	-0.5429	-5.28823	8	8	8	8
RSVM	3.2302773	1	6.59364	0.847897	1	1	1	1
QSVM	-0.402626	4	1.14672	-2.27009	4	4	4	7
CART	0.4642952	3	2.1573	-0.17727	3	2	2	3

Πίνακας 3.21 Κατανομή κατατάξεων των ταξινομητών για όλες τις ροές

κατάταξη/πόσες φορές	**LDA**	**LR**	**PNN**	**KNN**	**LSVM**	**RBF SVM**	**QSVM**	**CART**
1η	0	0	0	0	0	500	0	0
2η	0	0	430	0	0	0	0	70
3η	0	1	56	0	0	0	13	430
4η	0	98	4	81	0	0	317	0
5η	10	276	7	65	0	0	142	0
6η	273	85	3	113	0	0	26	0
7η	217	40	0	213	0	0	2	0
8η	0	0	0	28	500	0	0	0
χειρότερη κατάταξη	7	7	6	8	8	1	7	3
καλύτερη κατάταξη	5	3	2	4	7	1	3	2

Όπως φαίνεται από τους παραπάνω πίνακες, η καλύτερη μέθοδος πρόβλεψης της ταξινόμησης των ελευθέρων ηλεκτρονίων της ιονόσφαιρας σε αυτά που βοηθούν στην πρόβλεψη ύπαρξης δομής στην ιονόσφαιρα και σε αυτά που δεν βοηθούν στην εξαγωγή τέτοιων συμπερασμάτων, είναι η χρήση των μηχανών διανυσμάτων υποστήριξης πυρήνα RBF. Σε κάθε σενάριο, αυτό το σύστημα ταξινόμησης υπερέχει και η μέθοδος αξιολόγησης του (PROMETHEE) το κατατάσσει πάντα στην $1^{η}$ θέση.

Επίσης, αρκετά καλά φαίνεται ότι αποδίδει και το σύστημα ταξινόμησης που έχει ως βάση τα πιθανοτικά νευρωνικά δίκτυα . Αν και δεν έχει αξιολογηθεί για κανένα σενάριο ως η καλύτερη μέθοδος σε σχέση με τις υπόλοιπες, στο 86% των περιπτώσεων αλλά και για τα περισσότερα των επιμέρους στατιστικά στοιχεία του πίνακα 3.20 βρίσκεται στη $2^{η}$ θέση. Ποσοστό ιδιαίτερα υψηλό που αναδεικνύει το επίπεδο αξιοπιστίας του. Άλλωστε, η χειρότερη κατάταξη που κατέλαβε είναι η $6^{η}$ θέση και αυτό συνέβη σε ελάχιστες περιπτώσεις, όπως έγινε και με τις θέσεις 3, 4 και 5 για τη σειρά προτίμησης του σε σχέση με τα άλλα συστήματα ταξινόμησης. Περίπου ίδια είναι και τα αποτελέσματα για το παράδειγμα αυτό και για τα δένδρα ταξινόμησης, με μοναδικές διαφορές ότι στο 86% των ροών το σύστημα αυτό κατατάσσεται στην $3^{η}$ θέση που είναι και η χαμηλότερη θέση που κατέλαβε. Η υψηλότερη θέση που βρέθηκε είναι η $2^{η}$, αν και αυτό συνέβη σε λίγες περιπτώσεις.

Συνεχίζοντας, σε μέτρια επίπεδα αποτελεσματικότητας, δηλαδή κατατάσσονται κυρίως μεταξύ των θέσεων 4 και 6, κινήθηκαν οι ταξινομητές LR και QSVM. Πιο κάτω ακόμα βρίσκονται οι LDA και KNN, ενώ η χειρότερη μέθοδος βρέθηκε η LSVM καθώς και στις 500 από τις ροές της συγκρινόμενη με τις αντίστοιχες των άλλων μεθόδων κατέλαβε την $8^{η}$ και τελευταία θέση. Ακολουθεί ο πίνακας 3.22, που συγκεντρώνει τα παραπάνω συμπεράσματα .

Πίνακας 3.22 Γενική κατάταξη των ταξινομητών για την ταξινόμηση των ηλεκτρονίων της ιονόσφαιρας

Θέση	**Σύστημα ταξινόμησης**
$1^{η}$	RSVM
$2^{η}$	PNN
$3^{η}$	CART
$4^{η}$	QSVM
$5^{η}$	LR
$6^{η}$	KNN
$7^{η}$	LDA
$8^{η}$	LSVM

3.3.4 Διάγνωση διαταραχών του ήπατος

Σε έναν ακόμα ιατρικό τομέα όπως αυτός που ασχολείται με τις παθήσεις του ήπατος και πως επηρεάζεται η λειτουργία του από το αλκοόλ, η εφαρμογή συστημάτων ταξινόμησης για την πρόβλεψη μελλοντικών ασθενών σε τέτοια προβλήματα είναι αρκετά σημαντική. Εξίσου σημαντική όμως είναι και η επιλογή τα κατάλληλου συστήματος ταξινόμησης γεγονός που κάνει τα αποτελέσματα της πολυκριτήριας αξιολόγησής τους ιδιαίτερα χρήσιμα .

Από την εφαρμογή του βήματος 5 της γενικότερης μεθοδολογίας που περιγράφθηκε στην παράγραφο 3.2 και αφού τα αρχικά δεδομένα μετασχηματίσθηκαν στην κατάλληλη μορφή σύμφωνα με τα βήματα 1 – 4, τα αποτελέσματα που προέκυψαν συνοψίζονται στον πίνακα 3.23 που ακολουθεί, ενώ οι τιμές των παραμέτρων φαίνονται στον πίνακα 3.24 παρακάτω.

Πίνακας 3.23 Τιμές των κριτηρίων για κάθε σύστημα ταξινόμησης για την διάγνωση διαταραχών του ήπατος

ταξινομητής	AR	KS	LOO- B	0.632-B	se
LDA	0,37229	0,28194	0,63094	0,63342	0,02199
LR	0,39828	0,31224	0,6479	0,65587	0,022202
PNN	0,29353	0,22233	0,58437	0,70425	0,018831
KNN	0,24246	0,17796	0,58703	0,6174	0,01797
LSVM	0,25172	0,20347	0,57872	0,58868	0,02071
RSVM	0,33949	0,26191	0,62739	0,63544	0,021209
QSVM	0,42343	0,35224	0,66945	0,67803	0,021498
CART	0,28991	0,25034	0,63063	0,69722	0,014282

Πίνακας 3.24 Τιμές των παραμέτρων των ταξινομητών για την διάγνωση διαταραχών του ήπατος

Παράμετροι	**Τιμή**
pnn_smoothing	0.08
svm_smoothing	0.3
nearest_neighbor	21
cart_splitmin	0.08

Σύμφωνα με τη διαδικασία που ακολουθείται στο βήμα 6 της μεθοδολογίας, τα παραπάνω δεδομένα επεξεργάζονται με τη βοήθεια της πολυκριτήριας αξιολόγησης PROMETHEE ώστε να προκύψουν τα βάρη (50 τιμές) για κάθε κριτήριο (πίνακας 3.7). Επίσης, για αυτά τα βάρη και για τα όλα τα σενάρια υπολογίζονται οι ροές (σύνολο 500) μέσω των οποίων γίνεται η σύγκριση και η τελική αξιολόγηση των συστημάτων ταξινόμησης. Σε συγκεντρωτικό πίνακα βρίσκονται όλες οι τιμές των ροών για τους επιλεχθέντες προς σύγκριση και αξιολόγηση ταξινομητές.

Με βάση τα στοιχεία αυτού του πίνακα προκύπτουν και τα αποτελέσματα που παρουσιάζονται στους πίνακες 3.25 και 3.26 παρακάτω. Οι συγκεκριμένοι πίνακες περιέχουν πληροφορίες για την κατάταξη των συστημάτων ταξινόμησης με βάση συγκεκριμένες τιμές των ροών τους (πίνακας 3.25) και συνολικά σε πόσες περιπτώσεις κατηγοριοποιούνται ανά θέση (πίνακας 3.26).

Πίνακας 3.25 Κατάταξη των συστημάτων ταξινόμησης με βάση χαρακτηριστικά στατιστικά μεγέθη

ταξινομητής	AVERAGE	RANK (av)	MAX	RANK (max)	MIN	RANK (min)	RANK (at max)	RANK(at min)
LDA	0.1411	4	1.7503	5	-1.8874	5	3	7
LR	1.0446	3	3.8668	3	-0.2156	3	2	4
PNN	-0.2851	6	3.2002	4	-2.7646	6	2	6
KNN	-1.7655	7	-0.2364	7	-5.1662	7	7	7
LSVM	-2.3747	8	-0.6239	8	-5.7180	8	7	8
RSVM	-0.1556	5	0.6822	6	-1.4679	4	4	5
QSVM	2.1629	1	5.9825	1	0.2873	1	1	3
CART	1.2324	2	4.6699	2	0.1080	2	1	4

Πίνακας 3.26 Κατανομή κατατάξεων των ταξινομητών για όλες τις ροές

κατάταξη/πόσες φορές	LDA	LR	PNN	KNN	LSVM	RBF SVM	QSVM	CART
1η	0	0	0	0	0	0	402	98
2η	0	219	10	0	0	0	88	183
3η	30	257	14	0	0	0	10	189
4η	298	24	146	0	0	6	0	26
5η	157	0	46	0	0	293	0	4
6η	12	0	284	3	0	201	0	0
7η	3	0	0	491	6	0	0	0
8η	0	0	0	6	494	0	0	0
χειρότερη κατάταξη	7	4	6	8	8	6	3	5
καλύτερη κατάταξη	3	2	2	6	7	4	1	1

Με βάση τα αποτελέσματα που παρουσιάζονται στους δύο τελευταίους πίνακες, φαίνεται αρκετά καθαρά ότι στην συγκεκριμένη εφαρμογή το σύστημα ταξινόμησης που συμπεριφέρεται καλύτερα από όλα στην πρόβλεψη της κατηγορίας των ασθενών, είναι οι μηχανές διανυσμάτων υποστήριξης τετραγωνικού πυρήνα (QSVM). Η μέγιστή του τιμή είναι η μεγαλύτερη από όλες τις αντίστοιχες για τα άλλα συστήματα ταξινόμησης, καθώς επίσης αυτό συμβαίνει και για την μέση τιμή όλων των ροών του. Γι'αυτό και σε αυτές τις περιπτώσεις κατατάσσεται στη $1^{η}$ θέση, όπως και για 402 τιμές ροών (περίπου στο 80% των περιπτώσεων). Στη χειρότερη περίπτωση το συγκεκριμένο σύστημα ταξινόμησης καταλαμβάνει την $3^{η}$ θέση σε ότι αφορά την αποτελεσματικότητα του στην πρόβλεψη της κατηγορίας ταξινόμησης και αυτό δεν συμβαίνει παρά μόνο στο 2% του συνόλου των περιπτώσεων.

Λιγότερο, αλλά αρκετά αξιόπιστα παρουσιάζονται στη συγκεκριμένη εφαρμογή και οι ταξινομητές Λογιστικής Παλινδρόμησης (LR) και δένδρα ταξινόμησης (CART). Η λογιστική παλινδρόμηση σε ποσοστό 95% περίπου βρίσκεται σε πολύ ψηλές θέσης κατάταξης ($2^{η}$ και $3^{η}$) και παρόλο που ποτέ δεν ήταν το καλύτερο μοντέλο ταξινόμησης , σε καμία των περιπτώσεων δεν βρέθηκε κάτω από την $4^{η}$ θέση. Ανάλογα είναι τα αποτελέσματα για τα CART με την διαφορά ότι το σύστημα αυτό στο 20% των περιπτώσεων έχει αξιολογηθεί ως το καλύτερο από όλα .

Τα μοντέλα των Κ πλησιέστερων γειτόνων (KNN) και των γραμμικών μηχανών διανυσμάτων υποστήριξης (LSVM), στο πλήθος των σεναρίων που αναλύονται, παρουσιάζονται να έχουν αρκετά χαμηλή αποτελεσματικότητα σε σχέση με τα άλλα . Τα υπόλοιπα συστήματα ταξινόμησης που αξιολογήθηκαν, όπως φαίνεται και από τους πίνακες παρουσιάζονται να έχουν μια μέτρια αποτελεσματικότητα σε όλες τις περιπτώσεις. Ο πίνακας 3.27 παρουσιάζει την γενική κατάταξη των συστημάτων ταξινόμησης γι'αυτήν την εφαρμογή, με βάση τα αποτελέσματα του πίνακα 3.26.

Πίνακας 3.27 Γενική κατάταξη των ταξινομητών για την διάγνωση διαταραχών του ήπατος

Θέση	**Σύστημα ταξινόμησης**
$1^{η}$	QSVM
$2^{η}$	CART
$3^{η}$	LR
$4^{η}$	LDA
$5^{η}$	PNN
$6^{η}$	RSVM
$7^{η}$	KNN
$8^{η}$	LSVM

3.3.5 Διάγνωση διαβήτη

Επαναλαμβάνοντας την ίδια ακριβώς διαδικασία με τις προηγούμενες εφαρμογές, προκύπτουν οι συνολικοί πίνακες αποτελεσμάτων και για το συγκεκριμένο παράδειγμα Έτσι, οι πίνακες των αποτελεσμάτων (3.28 και 3.29) παρακάτω παρουσιάζουν τις τιμές των πέντε κριτηρίων που επιλέχθηκαν ώστε βάσει αυτών να γίνει η αξιολόγηση των συστημάτων ταξινόμησης. Επίσης , φαίνονται οι τελικές τιμές των παραμέτρων pnn_smoothing, svm_smoothing, nearest_neighbour, cart_splitmin και bs_samples οι οποίες χρησιμοποιήθηκαν στη συγκεκριμένη εφαρμογή.

Πίνακας 3.28 Τιμές των κριτηρίων για κάθε σύστημα ταξινόμησης για την διάγνωση διαβήτη

ταξινομητής	AR	KS	LOO- B	0.632-B	se
LDA	0,65723	0,508	0,75417	0,75934	0,014193
LR	0,65508	0,50054	0,74945	0,75348	0,01432
PNN	0,59655	0,45928	0,72202	0,72273	0,015225
KNN	0,51145	0,39322	0,70964	0,72785	0,013113
LSVM	0,61809	0,49524	0,73009	0,73454	0,014971
RSVM	0,65942	0,50911	0,75344	0,75553	0,014443
QSVM	0,67093	0,51467	0,7609	0,76791	0,013575
CART	0,56579	0,43121	0,70562	0,73058	0,011765

Πίνακας 3.29 Τιμές των παραμέτρων των ταξινομητών για την διάγνωση διαβήτη

Παράμετροι	**Τιμή**
pnn_smoothing	1.6
svm_smoothing	0.1
nearest_neighbor	11
cart_splitmin	0.08

Με βάση αυτά τα αποτελέσματα εφαρμόστηκε, όπως και προηγούμενα, η μέθοδος PROMETHEE ώστε να αξιολογηθούν τα αποτελέσματα των μεθόδων. Τα αναλυτικά αποτελέσματα (ροές) περιλαμβάνονται σε συγκεντρωτικό πίνακα όπως και στις προηγούμενες εφαρμογές και βάσει αυτών υπολογίζονται τα τελικά στατιστικά στοιχεία που αφορούν την αξιολόγηση των συστημάτων ταξινόμησης. Οι πίνακες 3.30 και 3.31 που ακολουθούν, παρουσιάζουν συνοπτικά τα συγκεκριμένα στοιχεία .

Πίνακας 3.30 Κατάταξη των συστημάτων ταξινόμησης με βάση χαρακτηριστικά στατιστικά μεγέθη

ταξινομητής	AVERAGE	RANK	MAX	MIN	RANK (max)	RANK (min)	RANK (at max)	RANK (at min)
LDA	1.3905120	2	3.69074	0.210311	2	2	2	3
LR	1.0504555	4	2.87761	0.112495	4	4	4	5
PNN	-1.730852	7	-0.3139	-5.38167	7	7	6	8
KNN	-2.338520	8	-0.3876	-5.82400	8	8	7	8
LSVM	-0.571254	5	0.28243	-3.13343	6	5	5	7
RSVM	1.2253381	3	3.43855	0.13105	3	3	3	4
QSVM	2.1323432	1	5.57917	0.50174	1	1	1	1
CART	-1.158021	6	0.79401	-5.04581	5	6	2	7

Πίνακας 3.31 Κατανομή κατατάξεων των ταξινομητών για όλες τις ροές

κατάταξη/πόσες φορές	LDA	LR	PNN	KNN	LSVM	RBF SVM	QSVM	CART
1η	0	0	0	0	0	0	500	0
2η	490	0	0	0	0	0	0	10
3η	10	0	0	0	0	483	0	7
4η	0	480	0	0	0	17	0	3
5η	0	20	0	0	364	0	0	116
6η	0	0	120	11	125	0	0	244
7η	0	0	305	64	11	0	0	120
8η	0	0	75	425	0	0	0	0
χειρότερη κατάταξη	3	5	8	8	7	4	1	7
καλύτερη κατάταξη	2	4	6	6	5	3	1	2

Όπως φαίνεται από τα στατιστικά στοιχεία παραπάνω, το QSVM ως σύστημα ταξινόμησης των δεδομένων του συγκεκριμένου προβλήματος, αποδίδει πολύ καλύτερα σε σχέση με οποιοδήποτε άλλο από τα υπόλοιπα επτά συστήματα . Για όλες τις τιμές των ροών του κατατάσσεται στην $1^{η}$ θέση, ενώ και η μέγιστή τιμή τους είναι μεγαλύτερη και από όλες τις μέγιστες τιμές των ροών των υπολοίπων ταξινομητών.

Αμέσως μετά ακολουθεί η γραμμική διακριτική ανάλυση. Στο 98% των σεναρίων για κάθε βάρος που εξετάζονται, αξιολογείται ως το δεύτερο καλύτερο σύστημα ταξινόμησης ενώ στη χειρότερη περίπτωση κατατάσσεται στην $3^{η}$ θέση. Κάτι τέτοιο συμβαίνει στην μικρότερη από τις τιμές των ροών του και σε ελάχιστες ακόμα περιπτώσεις.

Ακολουθούν οι μηχανές υποστήριξης διανυσμάτων με πυρήνα RBF, με αποτελέσματα που είναι περίπου ίδια αυτά της LDA (96% των σεναρίων στην $3^{η}$ θέση και 4% στην $4^{η}$ θέση) με τη διαφορά ότι εδώ η αναφορά γίνεται για την $3^{η}$ και $4^{η}$ θέση (την οποία καταλαμβάνει στην ελάχιστη τιμή των ροών του) αντίστοιχα . Σε επίσης μέτρια επίπεδα κινούνται και τα συστήματα LR και LSVM με το πρώτο να αξιολογείται κυρίως στην $4^{η}$ θέση και στην χειρότερη περίπτωση στην $5^{η}$, ενώ οι γραμμικές μηχανές διανυσμάτων υποστήριξης η καλύτερη θέση στην οποία έχουν αξιολογηθεί είναι η $5^{η}$ και η χειρότερη η $7^{η}$.

Στις τελευταίες θέσεις και ως οι λιγότερο αποδοτικοί ταξινομητές για την εφαρμογή αυτή, βρίσκονται οι μέθοδοι CART, PNN και KNN. Τα δένδρα ταξινόμησης, αν και σε κάποιες από τις περιπτώσεις έχουν αξιολογηθεί μέχρι και στη $2^{η}$ θέση (για 10 περιπτώσεις), κινούνται κυρίως στην $6^{η}$ θέση, ενώ πολύ χαμηλότερα από όλα, στην $8^{η}$ θέση, αξιολογούνται τα πιθανοτικά νευρωνικά δίκτυα . Ο πίνακας (3.32) παρακάτω δείχνει αναλυτικά την τελική κατάταξη των συστημάτων ταξινόμησης για την εφαρμογή τους στη διάγνωση διαβήτη.

Πίνακας 3.32 Γενική κατάταξη των ταξινομητών για την διάγνωση διαβήτη

Θέση	**Σύστημα ταξινόμησης**
$1^{η}$	QSVM
$2^{η}$	LDA
$3^{η}$	RSVM
$4^{η}$	LR
$5^{η}$	LSVM
$6^{η}$	CART
$7^{η}$	PNN
$8^{η}$	KNN

3.3.6 Πρόβλεψη αποτελέσματος παιγνίων

Λόγω του ενδιαφέροντος που παρουσιάζει η δυνατότητα πρόβλεψης της εξέλιξης ενός παιχνιδιού και της τελικής του έκβασης (ποιος παίκτης κερδίζει σε κάθε πιθανή περίπτωση), αναλύθηκε πειραματικά ένα από αυτά , η τρίλιζα . Η συμπεριφορά των οκτώ επιλεγμένων μεθόδων ταξινόμησης παρουσιάζεται παρακάτω μέσα από συγκεντρωτικούς πίνακες οι οποίοι περιέχουν τα αποτελέσματα των δύο τελευταίων και κύριων βημάτων της μεθοδολογίας της πειραματικής ανάλυσης, δηλαδή τα αποτελέσματα των εξεταζομένων μεθόδων στα πέντε κριτήρια (πίνακες 3.33) και την αξιολόγηση των αποτελεσμάτων από την μέθοδο PROMETHEE (πίνακες 3.35, 3.36, 3.37)

Πίνακας 3.33 Τιμές των κριτηρίων για κάθε σύστημα ταξινόμησης για την πρόβλεψη αποτελέσματος παιγνίων

ταξινομητής	AR	KS	LOO- B	0.632-B	se
LDA	0,96087	0,95586	0,9833	0,9833	0,0041425
LR	0,98765	0,95312	0,96623	0,97251	0,0046346
PNN	0,98139	0,89122	0,94052	0,95934	0,0054733
KNN	0,94703	0,82482	0,90333	0,91278	0,0060199
LSVM	0,82294	0,63798	0,81441	0,82086	0,010662
RSVM	0,89364	0,73789	0,65344	0,69954	0,015383
QSVM	0,99768	0,98122	0,98691	0,99019	0,0033301
CART	0,88631	0,75156	0,87024	0,89072	0,0057618

Στον πίνακα 3.34 που ακολουθεί φαίνονται οι τιμές των μεταβλητών των συστημάτων ταξινόμησης που επιλέχθηκαν για τη συγκεκριμένη εφαρμογή.

Πίνακας 3.34 Τιμές των παραμέτρων των ταξινομητών πρόβλεψη αποτελέσματος παιγνίων

Παράμετροι	**Τιμή**
pnn_smoothing	1
svm_smoothing	2
nearest_neighbor	31
cart_splitmin	0.05

Τα αποτελέσματα που παρουσιάζονται στους πίνακες 3.35 και 3.36 παρακάτω και δείχνουν στατιστικά στοιχεία για την τελική αξιολόγηση των μεθόδων ταξινόμησης, εξάγονται από τα δεδομένα ενός συγκεντρωτικού πίνακα που περιέχει τις τιμές όλων των ροών ανά βάρος και κάθε σενάριο, για όλους τους ταξινομητές που χρησιμοποιήθηκαν.

Πίνακας 3.35 Κατάταξη των συστημάτων ταξινόμησης με βάση χαρακτηριστικά στατιστικά μεγέθη

ταξινομητής	AVERAGE	RANK	MAX	MIN	RANK (max)	RANK (min)	RANK (at max)	RANK (at min)
LDA	1.714691	3	4.3121	0.4788	2	3	2	3
LR	1.740070	2	4.1346	0.5064	3	2	2	3
PNN	1.198434	4	2.6597	0.3598	4	4	4	4
KNN	0.196172	5	0.8082	-1.0109	5	5	5	5
LSVM	-2.753536	7	-0.5830	-6.3612	7	7	7	8
RSVM	-3.445311	8	-0.9441	-6.5099	8	8	7	8
QSVM	2.150596	1	5.2918	0.6453	1	1	1	1
CART	-0.801117	6	0.0508	-3.1953	6	6	6	6

Πίνακας 3.36 Κατανομή κατατάξεων των ταξινομητών για όλες τις ροές

κατάταξη/πόσες φορές	LDA	LR	PNN	KNN	LSVM	RBF SVM	QSVM	CART
1η	0	0	0	0	0	0	500	0
2η	210	290	0	0	0	0	0	0
3η	290	210	0	0	0	0	0	0
4η	0	0	500	0	0	0	0	0
5η	0	0	0	500	0	0	0	0
6η	0	0	0	0	0	0	0	500
7η	0	0	0	0	434	66	0	0
8η	0	0	0	0	66	434	0	0
χειρότερη κατάταξη	3	3	4	5	8	8	1	6
καλύτερη κατάταξη	2	2	4	5	7	7	1	6

Η μέθοδος της γραμμικής διακριτικής ανάλυσης στην εφαρμογή αυτή, δείχνει να έχει καλή αποτελεσματικότητα. Στις 290 από τις 500 περιπτώσεις η μέθοδος PROMETHEE κατατάσσει το συγκεκριμένο σύστημα ταξινόμησης τρίτο, κάτι που συμβαίνει και στη σύγκριση των μέσων τιμών όλων των ροών από τον κάθε ταξινομητή. Στις υπόλοιπες περιπτώσεις (42%) βρίσκεται στην $2^{η}$ θέση. Ακριβώς αντίθετα για τις θέσεις 2 και 3, είναι τα αποτελέσματα για τη μέθοδο LR η οποία συνήθως κατατάσσεται στην $3^{η}$ θέση απόδοσης.

Τα πιθανοτικά νευρωνικά δίκτυα αποδίδουν μέτρια αποτελέσματα στις διάφορες περιπτώσεις για το παιχνίδι της τρίλιζας, κάτι που φαίνεται και από το γεγονός ότι και στις 500 των περιπτώσεων αλλά και σε όλα τα στατιστικά αποτελέσματα (RANK(av), RANK(max), RANK(min), RANK(at max), RANK(at min)) αξιολογούνται ως τέταρτα σε απόδοση. Επίσης μέτρια είναι και τα αποτελέσματα της μεθόδου των Κ πλησιέστερων γειτόνων, η οποία συνεχώς κατατάσσεται από την πολυκριτήρια μέθοδο αξιολόγησης στην $5^{η}$ θέση.

Πιο χαμηλή αποτελεσματικότητα έχουν στην συγκεκριμένη εφαρμογή τα συστήματα LSVM και RSVM που εναλλάσσονται μεταξύ τους στις θέσεις 7 και 8, με τα πρώτα να βρίσκονται κυρίως στην $7^{η}$ θέση και τα RSVM κυρίως στην $8^{η}$. Ελάχιστα καλύτερα φαίνεται να αποδίδουν τα δένδρα ταξινόμησης , τα οποία στο 100% των σεναρίων αλλά και σε όλα τα στατιστικά στοιχεία που παρουσιάζονται στον πίνακα 3.36 και με βάση τα οποία συγκρίνονται μεταξύ τους οι ταξινομητές, καταλαμβάνουν την $6^{η}$ θέση.

Τέλος, και σε αυτή την εφαρμογή προέκυψε από την αξιολόγηση ότι την καλύτερη αποτελεσματικότητα την είχαν οι μηχανές διανυσμάτων υποστήριξης τετραγωνικού πυρήνα (για το 100% των τιμών των ροών βρίσκονται στην $1^{η}$ θέση). Ο πίνακας 3.37 που ακολουθεί παρουσιάζει συνολικά τα συμπεράσματα για την κατάταξη των συστημάτων ταξινόμησης, που αναλύθηκαν παραπάνω.

Πίνακας 3.37 Γενική κατάταξη των ταξινομητών για την πρόβλεψη αποτελέσματος παιγνίων

Θέση	**Σύστημα ταξινόμησης**
$1^{η}$	QSVM
$2^{η}$	LR
$3^{η}$	LDA
$4^{η}$	PNN
$5^{η}$	KNN
$6^{η}$	CART
$7^{η}$	LSVM
$8^{η}$	RSVM

3.3.7 Μελέτη πολιτικής συμπεριφοράς

Η εφαρμογή της ταξινόμησης και κατά συνέπεια και της αξιολόγησης αυτής, δεν θα μπορούσε να λείπει από των χώρο των εκλογών. Αποτελεί άλλωστε σημαντικό κομμάτι για την οργάνωση και την πορεία της προεκλογικής εκστρατείας, αφού συμβάλλει στη δημιουργία γενικότερης εικόνας που είναι ανταγωνιστικός και που λιγότερο ανταγωνιστικός ο κάθε υποψήφιος. Τα αποτελέσματα για την συγκεκριμένη εφαρμογή από την επεξεργασία των αρχικών δεδομένων παρουσιάζονται στον πίνακα 3.38 που ακολουθεί.

Πίνακας 3.38 Τιμές των κριτηρίων για κάθε σύστημα ταξινόμησης για την μελέτη πολιτικής συμπεριφοράς

ταξινομητής	AR	KS	LOO- B	0.632-B	se
LDA	0,95182	0,91285	0,95546	0,95663	0,0096978
LR	0,94258	0,87653	0,93977	0,96109	0,0083784
PNN	0,94454	0,83355	0,90193	0,90503	0,014009
KNN	0,95412	0,87163	0,91639	0,9167	0,012486
LSVM	0,98705	0,91725	0,95623	0,95796	0,0095106
RSVM	0,9828	0,91282	0,95984	0,96531	0,0079842
QSVM	0,96414	0,91092	0,95431	0,96351	0,0083513
CART	0,89918	0,90768	0,95317	0,95433	0,009707

Οι τιμές των παραμέτρων που χρησιμοποιούνται στη συγκεκριμένη εφαρμογή και για τις οποίες προέκυψαν τα παραπάνω αποτελέσματα, συγκεντρώνονται στον πίνακα 3.39 που ακολουθεί.

Πίνακας 3.39 Τιμές των παραμέτρων των ταξινομητών για την μελέτη πολιτικής συμπεριφοράς

Παράμετροι	**Τιμή**
pnn_smoothing	1.6
svm_smoothing	0.1
nearest_neighbor	21
cart_splitmin	0.1

Από τη σύγκριση των τιμών των ροών υπολογίζονται διάφορα στατιστικά στοιχεία όπως η μέγιστη, η ελάχιστη τιμή τους και ο μέσος όρος τους. Τα στοιχεία αυτά συμβάλλουν στην εξαγωγή συμπερασμάτων σε ότι αφορά την ειδικότερη αλλά και γενικότερη κατάταξη των συστημάτων ταξινόμησης. Τα αποτελέσματα αυτά συγκεντρώνονται στους πίνακες 3.40 και 3.41.

Πίνακας 3.40 Κατάταξη των συστημάτων ταξινόμησης με βάση χαρακτηριστικά στατιστικά μεγέθη

ταξινομητής	AVERAGE	RANK (av)	MAX	MIN	RANK (MAX)	RANK (min)	RANK (at max)	RANK (at min)
LDA	0,87325	4	2,1714	0,1445	4	4	4	4
LR	0,02581	5	1,8037	-2,4125	5	5	3	6
PNN	-3,54947	8	-0,9317	-6,7957	8	8	8	8
KNN	-2,11681	7	-0,3246	-5,0364	7	7	6	7
LSVM	1,51916	2	3,8690	0,4082	2	2	2	3
RSVM	1,87062	1	4,9740	0,5571	1	1	1	1
QSVM	1,37946	3	3,4829	0,3957	3	3	2	3
CART	-0,00203	6	1,6328	-3,0276	6	6	5	7

Πίνακας 3.41 Κατανομή κατατάξεων των ταξινομητών για όλες τις ροές

κατάταξη/πόσες φορές	LDA	LR	PNN	KNN	LSVM	RBF SVM	QSVM	CART
1η	0	0	0	0	8	492	0	0
2η	0	0	0	0	387	8	105	0
3η	0	1	0	0	104	0	395	0
4η	487	12	0	0	1	0	0	0
5η	13	288	0	0	0	0	0	199
6η	0	199	0	9	0	0	0	292
7η	0	0	0	491	0	0	0	9
8η	0	0	500	0	0	0	0	0
χειρότερη κατάταξη	5	6	8	7	4	2	3	7
καλύτερη κατάταξη	4	4	8	6	1	1	2	5

Από τα αποτελέσματα που παρουσιάζονται παραπάνω, είναι εύκολο να διακριθεί ότι οι μηχανές διανυσμάτων υποστήριξης με πυρήνα RBF προβλέπουν την ταξινόμηση των περιστατικών της εφαρμογής αυτής πολύ καλύτερα από τις υπόλοιπες μεθόδους που χρησιμοποιήθηκαν. Πιο συγκεκριμένα, στο 98% περίπου των εξεταζόμενων ροών από τα διάφορα σενάρια η συγκεκριμένη μέθοδος αξιολογείται ως η καλύτερη από όλες , ενώ μόλις σε 8 περιπτώσεις κατατάσσεται στη $2^{η}$ θέση η οποία είναι και η χειρότερη θέση που καταλαμβάνει.

Τις RSVM ακολουθούν οι γραμμικές μηχανές διανυσμάτων υποστήριξης. Αν και η χειρότερη κατάταξη που προκύπτει για το σύστημα αυτό είναι η $4^{η}$ θέση, αυτό δεν συμβαίνει παρά σε μια μόλις περίπτωση. Ελάχιστες βέβαια είναι και οι περιπτώσεις (μόλις 8 στις 500) που ο συγκεκριμένος ταξινομητής κατατάσσεται στην $1^{η}$ θέση και κυρίως βρίσκεται στην $2^{η}$ θέση κάτι που συμβαίνει και για την κατάταξη με βάση τον μέσο όρο των τιμών των ροών του. Λίγο πιο κάτω στην $3^{η}$ θέση κατατάσσονται οι μηχανές διανυσμάτων υποστήριξης τετραγωνικού πυρήνα.

Σε μέτρια επίπεδα κινείται η αποτελεσματικότητα των ταξινομητών LDA και LR που βρίσκονται κυρίως στην $4^{η}$ και $5^{η}$ θέση αντίστοιχα, στο μεγαλύτερο μέρος του συνόλου των περιπτώσεων. Ενώ η μέθοδος της γραμμικής διακριτικής ανάλυσης κινείται σταθερά στην $4^{η}$ θέση πλην ελαχίστων περιπτώσεων (13), η μέθοδος της λογιστικής παλινδρόμησης σε μια περίπτωση καταλαμβάνει την $3^{η}$ θέση που είναι και η καλύτερη. Στην χειρότερη περίπτωση καταλαμβάνει την $6^{η}$ θέση, κάτι που όμως συμβαίνει για αρκετές περιπτώσεις και άρα σε αρκετά από τα σενάρια (199) στα οποία συγκρίνονται όλα τα συστήματα ταξινόμησης.

Στην προτελευταία θέση αξιολόγησης βρίσκεται η μέθοδος ταξινόμησης KNN. Σε ελάχιστες περιπτώσεις (μόλις 9) το σύστημα αυτό αξιολογείται λίγο πιο πάνω στην $6^{η}$ θέση ενώ στο 98,2% των περιπτώσεων βρίσκεται στην $7^{η}$ θέση. Κάτι τέτοιο άλλωστε μπορεί να διαπιστωθεί και από την κατάταξη που καταλαμβάνει όταν η σύγκριση με τα άλλα συστήματα ταξινόμησης γίνεται με βάση την μέση τιμή των ροών τους.

Τέλος, χειρότερη όλων των μεθόδων για την πρόβλεψη της κατηγορίας ταξινόμησης στην εφαρμογή αυτή είναι τα πιθανοτικά νευρωνικά δίκτυα. Σε όλα τα σενάρια και για όλα τα βάρη η θέση που αξιολογούνται είναι η $8^{η}$ και τελευταία. Το ίδιο βέβαια συμβαίνει και για όλες τις επιμέρους στατιστικές τιμές (average, max, min) που εξετάζονται στον πίνακα 3.40.

Ο πίνακας 3.42 που ακολουθεί συγκεντρώνει όλα τα παραπάνω συμπεράσματα για την συνολική τελική κατάταξη των μεθόδων ταξινόμησης, για την εφαρμογή αυτή.

Πίνακας 3.42 Γενική κατάταξη των ταξινομητών για την μελέτη πολιτικής συμπεριφοράς

Θέση	**Σύστημα ταξινόμησης**
$1^{η}$	RSVM
$2^{η}$	LSVM
$3^{η}$	QSVM
$4^{η}$	LDA
$5^{η}$	LR
$6^{η}$	CART
$7^{η}$	KNN
$8^{η}$	PNN

ΚΕΦΑΛΑΙΟ 4 : ΣΥΜΠΕΡΑΣΜΑΤΑ

Όπως έχει ήδη αναφερθεί, η επιστήμη της ταξινόμησης δεδομένων σε κατηγορίες βρίσκει εφαρμογή σε πολλούς τομείς της καθημερινότητας και σε πολλά επιστημονικά πεδία. Η δυνατότητα της πρόβλεψης που δίνει, αποτελεί ένα ισχυρό εργαλείο στα χέρια επιστημόνων είτε αυτοί προέρχονται από τον χώρο της ιατρικής, είτε είναι μηχανικοί, είτε οικονομολόγοι ή από οποιοδήποτε επιστημονικό χώρο. Παρόλ'αυτά , επειδή είναι πολύ δύσκολο έως αδύνατο προς το παρών να γίνεται 100% ακριβής πρόβλεψη της ταξινόμησης , η επιλογή των καταλληλότερων μεθόδων ταξινόμησης και κυρίως η αξιολόγηση των υπαρχόντων συστημάτων πριν την χρήση τους παρουσιάζει ιδιαίτερο ενδιαφέρον και αποτελεί αναγκαίο βήμα στο πεδίο αυτό.

Από την πειραματική ανάλυση που έγινε στο κεφάλαιο 3, εύκολα μπορεί να διαπιστώσει κάποιος πως δεν υπάρχει πάντα κάποια μέθοδος ταξινόμησης που να υπερέχει φανερά από τις υπόλοιπες. Από εφαρμογή σε εφαρμογή η αποτελεσματικότητά τους μεταβάλλεται και μια μέθοδος από πολύ καλή σε μια εφαρμογή μπορεί να γίνει πολύ κακή σε κάποια άλλη. Βέβαια αυτό δεν μπορεί να επιτευχθεί χωρίς κάποιες άλλες παραχωρήσεις, όπως υψηλότερο υπολογιστικό κόστος και πιο χρονοβόρα διαδικασία . Επίσης, κάποιες μέθοδοι απαιτούν μεγάλη βάση δεδομένων είτε για τη δημιουργία του μοντέλου είτε για την εκπαίδευσή του ώστε να αποδώσουν σε υψηλό βαθμό. Ακόμα, άλλα συστήματα απαιτούν πολλά χαρακτηριστικά μέτρησης. Είναι λοιπόν φανερό ότι σε μεγάλο βαθμό η επιλογή ταξινομητή από εφαρμογή σε εφαρμογή εξαρτάται και από τον αποφασίζοντα και τις παραχωρήσεις που είναι διατεθειμένος να κάνει.

Το γεγονός λοιπόν ότι δεν υπάρχει κάποιο κυρίαρχο μοντέλο, καθιστά ακόμα πιο δύσκολη και αναγκαία την πολυκριτήρια αξιολόγησή τους. Αναπτύσσονται έτσι, γι'αυτόν τον σκοπό πολλά χρήσιμα εργαλεία όπως η ROC ανάλυση και τεχνικές επαναληπτικής δειγματοληψίας όπως το Bootstrap, οι οποίες και αναπτύχθηκαν στο $2^{ο}$ κεφάλαιο. Απαιτείται όμως για τη χρήση τους πλήρη γνώση των χαρακτηριστικών και των περιορισμών τους, ώστε να εφαρμοστούν με τον κατάλληλο τρόπο και τα αποτελέσματά τους να μεταφραστούν σωστά. Άλλωστε υπάρχει και η παράμετρος κόστος (θετικό ή αρνητικό) που σε συνδυασμό με τα πεδία στα οποία μπορεί να εφαρμοστεί η ταξινόμηση, αποκτά ιδιαίτερη σημασία για τον αποφασίζοντα στην τελική επιλογή.

Γενικά λοιπόν, μπορεί να πει κανείς ότι όλες οι μέθοδοι ταξινόμησης που χρησιμοποιήθηκαν πρέπει να θεωρηθούν ως μέθοδοι που δίνουν μια γενικότερη ιδέα του πως συμπεριφέρονται οι ταξινομητές και όχι την ακριβή τους συμπεριφορά κάθε φορά που αντιμετωπίζουν προβλήματα ταξινόμησης ίδιας ή έστω παρόμοιας φύσης με αυτά που αναλύθηκαν στα Κεφάλαια 2 και 3. Επίσης δεν μπορεί να εξαχθεί συμπέρασμα ότι κάποιος ταξινομητής είναι γενικότερα καλύτερος ή χειρότερος σε σύγκριση με τους άλλους. Κάτι τέτοιο άλλωστε διαπιστώνεται και από τον πίνακα 4.1 παρακάτω, που συγκεντρώνει τις θέσεις που κατατάχθηκε κάθε σύστημα ταξινόμησης ανά εφαρμογή.

Πίνακας 4.1 Γενική κατάταξη των ταξινομητών για όλες τις εφαρμογές

Ταξινομητής	**1η εφαρ-μογή**	**2η εφαρ-μογή**	**3η εφαρ-μογή**	**4η εφαρ-μογή**	**5η εφαρμογή**	**6η εφαρμογή**	**7η εφαρμογή**
LDA	6η	3η	7η	4η	2η	3η	4η
LR	7η	6η	5η	3η	4η	2η	5η
PNN	1η	2η	2η	5η	7η	4η	8η
KNN	3η	8η	6η	7η	8η	5η	7η
LSVM	4η	4η	8η	8η	5η	7η	2η
RSVM	2η	1η	1η	6η	3η	8η	1η
QSVM	5η	5η	4η	1η	1η	1η	3η
CART	8η	7η	3η	2η	6η	6η	6η

Κανένα από τα συστήματα ταξινόμησης δεν βρίσκεται και στις 7 από τις εφαρμογές που αναλύθηκαν στις πρώτες θέσεις της κατάταξης. Για όλα ισχύει το ίδιο συμπέρασμα ότι δηλαδή σε κάποια από τα παραδείγματα αποδίδουν πάρα πολύ καλά, σε κάποια άλλα κυμαίνονται σε μέτρια απόδοση και στα υπόλοιπα δεν είναι προτιμότερα από τα υπόλοιπα αφού η ακρίβεια της πρόβλεψης ταξινόμησης που κάνουν δεν είναι από τις υψηλότερες. Παρόλα αυτά, αν θεωρηθούν οι LSVM, RSVM ΚΑΙ QSVM ως μια μόνο μέθοδος ταξινόμησης, δηλαδή ως μηχανές διανυσμάτων υποστήριξης (SVM), που συγκρίνεται με τις υπόλοιπες τότε παρατηρείται ότι καταλαμβάνουν την πρώτη θέση στις 6 από τις 7 εφαρμογές που εξετάζονται.

Το πεδίο της ταξινόμησης τις τελευταίες δεκαετίες γνώρισε ιδιαίτερη ανάπτυξη κάτι που συνεχίζει να συμβαίνει. Πολλές είναι οι μελέτες και οι εφαρμογές για την βελτίωση και για την αναλυτικότερη γνώση του αντικειμένου. Με την βοήθεια και της διαρκούς ανάπτυξης της τεχνολογίας των υπολογιστών οι διαδικασίες ταξινόμησης αλλά και αξιολόγησης αυτής συνεχώς βελτιώνονται ή και αναπτύσσονται νέες, με σκοπό βέβαια την παροχή της καλύτερης δυνατής βοήθειας στην εξαγωγή συμπερασμάτων και λήψης αποφάσεων σε πολλούς τομείς της καθημερινής ζωής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] *N.M. Adams, D.J. Hand (2000) – Improving The Practice of Classifier Performance Assessment, Neural Computation, vol. 12, 305 – 311.*

[2] *S. Balakrishnama, A. Ganapathiraju – Linear Discriminant Analysis: A Brief Tutorial, Mississipi State University Institute Signaland Information Processing, Internal Publications, 1-8.*

[3] *H.R. Bittencourt, R.T. Clarke – Feature Selection by Using Classification And Regression Trees (CART), XX ISPRS Congress, 12-13, July 2004, Instanbul Turkey.*

[4] *A.P. Bradley (1996) – The Use of The Area Under The ROC Curve In The Evaluation of Machine Learning Algorithms, Pattern Recognition, vol. 30, No 7, 1145-1159.*

[5] *R. Wehrens, H. Putter, L.M.C. Buydens (2000) – The Bootstrap: A Tutorial, Chemometrics and Intelligent Laboratory Systems, vol.54, 35-52.*

[6] *T.G. Dietterch (1998) – Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Neural Computation, 10, 1895-1923.*

[7] *R.P.W. Duin, A.K. Jain, J. Mao (2000) – Statistical Pattern Recognition: A Review, transactions on Pattern Analysis and Machine Intelligence, vol.22, No.1, 4-37, IEEE.*

[8] *B. Engelmann, E. Hayden, D. Tasche (2002) – Measuring The Discriminative Power of Rating Systems, Deutsche Bundesbank Banking Supervision Discussion Papers 200301.*

[9] *T. Fawcett (2003) – ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, Technical Report HPL – 2003-04, HP Labs.*

[10] *D.J. Hand, R.A. Schiavo (2000) – Ten More Years of Error Rate Research, International Statistical Review, vol. 68, No. 3, 295-310.*

[11] *D.J. Hand (2001) – Measuring Diagnostic Accuracy of Statistical Prediction Rules, Statistica Neerlandica, Vol. 55, No 1, 3-16.*

[12] *D.W. Hosmer, S. Lemeshow (1998) – Applied Logistic Regression,* John Wiley & Sons, New York .

[13] *C.J. Huang (2002) – A Performance Analysis of Cancer Classification Using Feature Extraction And Probablistic Neural Networks, 7th Conference on Artificial Intelligence and Applications.*

[14] *A. Lendasse, M.l Verteysen, V. Wertz (2003) – Model Selection With Cross-Validations And Bootstraps-Application to Time Series Prediction With RBFN Models, O. Kaynak et. Al (Eds).:ICANN/ICONIP 2003, LNCS 2714, 573-580.*

[15] *D. Michie, D.J. Spiegelhalter, C.C. Taylor (1994) – Machine Learning, Neural And Statistical Classification, Ellis Horwood Publications.*

[16] *S. Nargundkar, J.L. Priestley – Assessment of Evaluation Methods for Binary Classification Modeling, Proceedings of the 2003 Decision Scienes Institute National Conference, Washington DC, 1-6.*

[17] *F. Portera, A. Sperduti (2004) – A Generalized Quadratic Loss for Support Vector Machines, ECAI 2004, 628-634.*

[18] *N.A. Thacker (1998) – Tutorial: Supervised Neural Networks In Machine Vision.*
http://www.Tina-vision.net/docs/memos/1997-003.pdf

[19] *K.M. Ting, G.I. Webb (2005) – On The Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions, Machine Learning, vol. 58, no 1, 25-32.*

[20] *W.H. Wong, X. Zhang (2001) – Recursive Sample Classification And Gene Selection Based on SVM: Method And Software Description, Dept of Biostatistics, Harvard School of Puplic Health, Boston.*