ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. «ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»
ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

Διπλωματική εργασία

«Ανάλυση συνόλου ιατρικών δεδομένων MIMIC με ανάπτυξη εφαρμογής μηχανικής μάθησης, στο πλαίσιο πληροφοριακού συστήματος υγείας»

Ελπίς Μεζίλη
A.M.: ΜΕ1720

Επιβλέπων:

Καθηγητής Γεώργιος Βασιλακόπουλος

Πειραιάς, Φεβρουάριος 2020

Περιεχόμενα

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του Προγράμματος Μεταπτυχιακών Σπουδών «Ψηφιακά συστήματα και υπηρεσίες» του Πανεπιστημίου Πειραιώς και πιο συγκεκριμένα στην κατεύθυνση «Μεγάλα δεδομένα και αναλυτική». Η εργασία πραγματοποιήθηκε υπό την επίβλεψη του κ. Γεώργιου Βασιλακόπουλου, Καθηγητή του Τμήματος Πληροφοριακών Συστημάτων.

Αντικείμενο της εργασίας αποτελεί η ανάλυση συνόλου ιατρικών δεδομένων MIMIC με ανάπτυξη εφαρμογής μηχανικής μάθησης, στο πλαίσιο πληροφοριακού συστήματος υγείας. Συγκεκριμένα, αποτελείται από τρία τμήματα. Αρχικά δημιουργήθηκε ένα πληροφοριακό σύστημα επικεντρωμένο στην πνευμονολογική κλινική του νοσοκομείου, στη συνέχεια μελετήθηκε το περιεχόμενο της βάσης δεδομένων MIMIC και αντλήθηκε το σχετικό με την πνευμονολογική κλινική μέρος των δεδομένων, και στο τέλος έγινε η ανάλυση τους, με την ανάπτυξη εφαρμογής μηχανικής μάθησης.

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω όλους αυτούς που συνέβαλαν στην εκπόνηση της διπλωματικής μου εργασίας. Οφείλω να εκφράσω τις θερμές μου ευχαριστίες, προς τον επιβλέποντα της εργασίας, Καθηγητή Γεώργιο Βασιλακόπουλο, για την καθοδήγησή του, και την πολύτιμη βοήθεια που προσέφερε σε κάθε στάδιο εκπόνησής της. Επίσης, την Δρ. Βασιλική Κουφή, μεταδιδακτορική ερευνήτρια στο Τμήμα Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και μέλος Ε.Τ.Ε.Π. του Πανεπιστημίου Πειραιώς, για την καθοδήγηση της σχετικά με το τεχνικό μέρος του πληροφοριακού συστήματος. Χωρίς τη συμπαράσταση και συνεχή βοήθειά τους, η ολοκλήρωση αυτής της εργασίας δεν θα ήταν δυνατή. Τέλος, ευχαριστώ θερμά την οικογένεια και τους φίλους μου, για την κατανόηση και συμπαράσταση που έδειξαν ολόκληρη την περίοδο εκπόνησης της εργασίας αυτής.

1. Εισαγωγή

Η σωστή και αξιόπιστη πληροφόρηση αποτελεί τον θεμέλιο λίθο της λήψης αποφάσεων σε κάθε οργανισμό. Ειδικότερα, η ορθή καταγραφή δεδομένων στις μονάδες υγείας, είναι ουσιαστικής σημασίας, διότι επηρεάζονται άμεσα τόσο η βέλτιστη παροχή υπηρεσιών υγείας στους ασθενείς, η επιστημονική έρευνα, η εκπαίδευση και κατάρτιση, αλλά και η ανάπτυξη ανθρώπινων πόρων, όσο και ο σχεδιασμός και η εφαρμογή των πολιτικών και της χρηματοδότησης για το σύστημα υγείας.

Ένα πληροφοριακό σύστημα υγείας έχει τέσσερεις βασικές λειτουργίες: παραγωγή δεδομένων, σύνταξη, ανάλυση και σύνθεση, επικοινωνία και χρήση. Συλλέγει δεδομένα από τον τομέα της υγείας και άλλους συναφείς τομείς, αναλύει τα δεδομένα και εξασφαλίζει τη συνολική τους ποιότητα, τη συνάφεια και την έγκαιρη ενημέρωσή τους. Παράλληλα, μετατρέπει δεδομένα σε πληροφορίες για τη λήψη αποφάσεων που σχετίζονται με την υγεία.[1]

Στο επίπεδο των μονάδων υγείας το πληροφοριακό σύστημα ενώ συνήθως είναι ισοδύναμο με την παρακολούθηση και την αξιολόγηση, ταυτόχρονα εξυπηρετεί ευρύτερους σκοπούς, όπως η διαχείριση των ασθενών και της υγειονομικής μονάδας, η δυνατότητα συναγερμού και έγκαιρης προειδοποίησης –ιδιαίτερα σε μεγαλύτερη κλίμακα, π.χ. εθνικό πληροφοριακό σύστημα υγείας–, αλλά και την υποστήριξη της έρευνας, επιτρέποντας την ανάλυση των τάσεων. Η πληροφορία είναι μικρής αξίας, αν δεν είναι διαθέσιμη σε μορφές που ανταποκρίνονται στις ανάγκες πολλών χρηστών-υπεύθυνων για την χάραξη πολιτικής, διαχειριστών, παρόχων υπηρεσιών υγείας, κοινότητες και μεμονωμένα άτομα. Ως εκ τούτου, η διάδοση και η επικοινωνία είναι βασικές ιδιότητες του συστήματος πληροφοριών για την υγεία.

Οι παραπάνω χρήστες χρειάζονται διαφορετικά είδη πληροφοριών, όπως:

- παράγοντες καθοριστικούς για την υγεία (κοινωνικο-οικονομικοί, περιβαλλοντικοί συμπεριφοριστικοί, γενετικοί παράγοντες) και συναφή περιβάλλοντα εντός των οποίων λειτουργεί το σύστημα υγείας
- εισαγωγές στο σύστημα υγείας και σχετικές διαδικασίες (συμπεριλαμβανομένων των πολιτικών, οργανωσιακών και υγειονομικών υποδομών), εγκαταστάσεις και εξοπλισμός, κόστος, ανθρώπινους και οικονομικούς πόρους, πληροφορίες για το ίδιο το πληροφοριακό σύστημα
- τις επιδόσεις ή τις εξόδους του συστήματος υγείας, όπως διαθεσιμότητα, προσβασιμότητα, ποιότητα και χρήση των πληροφοριών και των υπηρεσιών υγείας, ανταπόκριση του συστήματος στις ανάγκες των χρηστών και προστασία από χρηματοοικονομικό κίνδυνο
- αποτελέσματα υγείας (θνησιμότητα, νοσηρότητα, εκδηλώσεις ασθενειών, κατάσταση υγείας, αναπηρία, ευεξία)
- ανισότητες στον τομέα της υγείας όσον αφορά τους καθοριστικούς παράγοντες, την κάλυψη της χρήσης των υπηρεσιών και τα αποτελέσματα της υγείας, και συμπεριλαμβανομένων των βασικών στρωματοποιητών όπως φύλο, κοινωνικοοικονομική κατάσταση, εθνοτική ομάδα, γεωγραφική θέση κ.λπ.

Ένα καλό πληροφοριακό σύστημα για την υγεία συγκεντρώνει όλους τους σχετικούς εταίρους για να εξασφαλίσει ότι οι χρήστες των πληροφοριών υγείας έχουν πρόσβαση σε αξιόπιστα, έγκυρα, χρήσιμα, κατανοητά συγκριτικά δεδομένα.

Επιπλέον, η διαρκώς αναπτυσσόμενη πίεση για τελειοποίηση και καινοτομία προκάλεσαν τη σύλληψη της ιδέας, πως μία διεργασία (process) θα μπορεί να ιδωθεί ως μονάδα ανάλυσης. Έτσι, ως μέρος της ποιότητας διαχείρισης, ο σημαντικός ρόλος της ποιότητας της διεργασίας οδήγησε σε μία πληθώρα τεχνικών ανάλυσης διεργασιών με κορύφωση τις μεθόδους Six Sigma (6σ), που αποτελούν ίσως τις σημαντικότερες μεθοδολογικές πρακτικές για την βελτίωση των επιχειρηματικών διεργασιών.

Χαρακτηριστικό παράδειγμα αποτελεί η μεθοδολογία DMAIC που πηγάζει από τις πρακτικές 6σ και αποτελείται από πέντε επίπεδα:

- Καθορισμός του προβλήματος και των στόχων (Define)
- Λεπτομερειακή μέτρηση των διαφόρων πτυχών της παρούσας διεργασίας (Measure)
- Ανάλυση των δεδομένων και, μεταξύ άλλων, εύρεση των ελαττωμάτων της διεργασίας (Analyse)
- Βελτίωση της διεργασίας (Improve)
- Έλεγχος του τρόπου εκτέλεσης της διεργασίας στο μέλλον (Control)

2. Ενσωμάτωση ευφυίας σε πληροφοριακό σύστημα

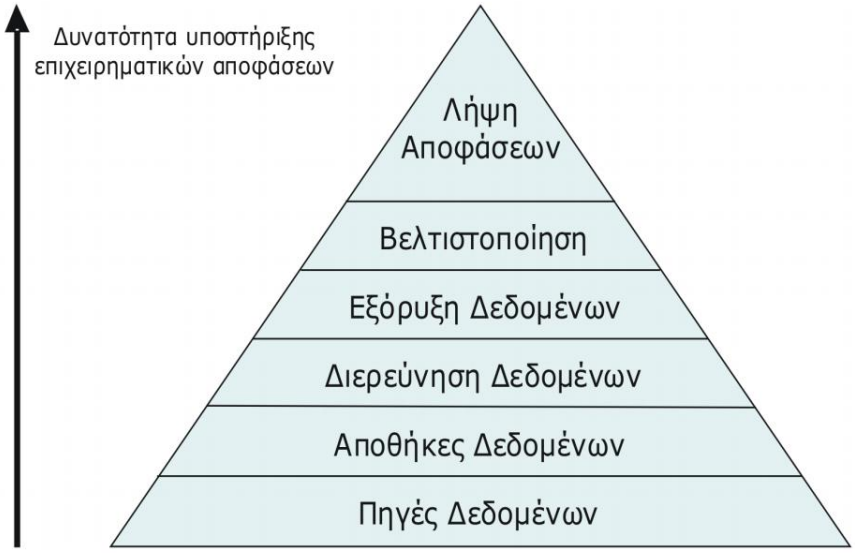
Ορίζουμε την Επιχειρηματική Ευφυΐα ως ένα σύνολο από μεθόδους ανάλυσης, τεχνολογίες, ικανότητες και στρατηγικές, οι οποίες στόχο έχουν την επεξεργασία των διαθέσιμων δεδομένων και την εξαγωγή χρήσιμης πληροφορίας από αυτά, για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Ένας άλλος συγγενής, αν και όχι ταυτόσημος όρος, ο οποίος γνωρίζει ιδιαίτερη διάδοση τον τελευταίο καιρό είναι «Αναλυτική των Επιχειρήσεων» (Business Analytics). Η Επιχειρηματική Ευφυΐα επιτρέπει σε έναν οργανισμό να μαθαίνει, να αντιλαμβάνεται καταστάσεις και συμβάντα, να σκέφτεται αφαιρετικά, να προβλέπει τάσεις και μελλοντικά συμβάντα, να σχεδιάζει και να καινοτομεί. Η παραγόμενη πληροφορία μετουσιώνεται σε γνώση που αξιοποιείται από τα διοικητικά στελέχη, ώστε να δρομολογήσουν κατάλληλες δράσεις, που θα οδηγήσουν στον καθορισμό και την επίτευξη επιχειρηματικών στόχων, με τρόπο αποτελεσματικό και αποδοτικό.

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι εξειδικευμένα πληροφοριακά συστήματα, τα οποία προσφέρουν ποιοτική πληροφορία. Η πληροφορία βασίζεται σε ποιοτικά και συγκεντρωτικά δεδομένα, τα οποία συνδυάζονται με λογισμικό ικανό να διεξάγει κατάλληλες αναλύσεις. Η βελτίωση της ποιότητας της πληροφορίας οφείλεται στις δυνατότητες αυτών των συστημάτων, τα οποία επιτρέπουν την ταχύτερη πρόσβαση στην πληροφορία, την ευκολότερη υποβολή ερωτημάτων στο σύστημα και τη σύνταξη αναφορών, την προχωρημένη ανάλυση των δεδομένων, καθώς και τη βελτίωση της ποιότητας των δεδομένων. Οι τελικοί αποδέκτες του προϊόντος των συστημάτων Επιχειρηματικής Ευφυΐας, οι οποίοι πολλές φορές αναφέρονται στη βιβλιογραφία ως «εργάτες γνώσης», τροφοδοτούνται έγκαιρα με γνώση που χρησιμοποιούν για τη λήψη αποφάσεων.

Στη σημερινή εποχή ένας νέος κλάδος της Πληροφορικής, η Εξόρυξη Δεδομένων, έρχεται να δώσει νέα ώθηση στην Επιχειρηματική Ευφυΐα. Η Εξόρυξη Δεδομένων (Data Mining) ή Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) στοχεύει στην ανακάλυψη γνώσης που είναι κρυμμένη σε μεγάλους όγκους δεδομένων. Οι τεχνικές Εξόρυξης Δεδομένων δεν απαιτούν τον προκαθορισμό μοντέλων. Αντιθέτως, τα μοντέλα προκύπτουν από την επεξεργασία των δεδομένων. Επίσης, τα μοντέλα μπορούν να χρησιμοποιηθούν για τη διατύπωση προβλέψεων.

2.1. Δομικά Επίπεδα Συστημάτων Επιχειρηματικής Ευφυΐας

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι δομημένα σε μια σειρά από επάλληλα επίπεδα, τα οποία συγκροτούν μια πυραμίδα. Στη βάση της πυραμίδας βρίσκονται τα αρχικά ακατέργαστα δεδομένα, ενώ στην κορυφή της βρίσκεται η λήψη των τελικών αποφάσεων. Κάθε μετάβαση από ένα επίπεδο σε κάποιο ανώτερο, αυξάνει τη δυνατότητα υποστήριξης επιχειρηματικών αποφάσεων. Η πυραμίδα Συστημάτων Επιχειρηματικής Ευφυΐας παρουσιάζεται στην Εικόνα 2.1.



Σχήμα 2.1. Η πυραμίδα συστημάτων Επιχειρηματικής Ευφυΐας [2]

Πηγές Δεδομένων

Στη βάση της πυραμίδας βρίσκονται οι πηγές των αρχικών δεδομένων. Τα δεδομένα αυτά προέρχονται κυρίως από συστήματα παρακολούθησης συναλλαγών και από εταιρικές βάσεις δεδομένων. Άλλες πρόσθετες πηγές δεδομένων είναι οι εταιρικοί δικτυακοί servers, εσωτερικά έγγραφα ή και εξωτερικές πηγές. Τα δεδομένα αυτά μπορεί να είναι σημαντικά για την καθημερινή λειτουργία της επιχείρησης, είναι όμως ακατάλληλα για τη λήψη αποφάσεων. Η πληροφορία ότι το ταμείο No. 3 ενός υποκαταστήματος super market εξέδωσε απόδειξη για την πώληση ενός κουτιού καφέ, μια συγκεκριμένη ημέρα και ώρα, είναι σημαντική για το λογιστήριο και την αποθήκη, είναι όμως αδιάφορη για τη διοίκηση. Αυτό που ενδιαφέρει τη διοίκηση είναι οι συγκεντρωτικές πωλήσεις καφέ, σε μια γεωγραφική περιοχή και σε μια χρονική περίοδο. Τα λειτουργικά δεδομένα είναι υπερβολικά αναλυτικά και για τον λόγο αυτό, ακατάλληλα για επεξεργασία και εξαγωγή συμπερασμάτων. Επίσης, τα δεδομένα αυτά είναι διάσπαρτα σε διάφορες πηγές και πρέπει να ενοποιηθούν. Τέλος, τα δεδομένα μπορεί να έχουν διαφόρων ειδών προβλήματα, τα οποία πρέπει να αντιμετωπιστούν.

Αποθήκες Δεδομένων

Το επόμενο επίπεδο είναι αυτό των Αποθηκών Δεδομένων. Πρόκειται για βάσεις δεδομένων που περιέχουν τα ενοποιημένα, συγκεντρωτικά και καθαρά δεδομένα. Αυτά τα δεδομένα θα χρησιμοποιηθούν για την ανάλυση και την εξαγωγή συμπερασμάτων. Οι εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης των δεδομένων στις Αποθήκες, γνωστές και ως εργασίες ETL (Extract, Transform, Load), εκτελούνται σε τακτά χρονικά διαστήματα. Στο πλαίσιο των εργασιών αυτών, επιλέγονται καταρχήν τα λειτουργικά δεδομένα που είναι σχετικά με την ανάλυση που πρέπει να πραγματοποιηθεί. Οι Αποθήκες Δεδομένων είναι θεματικά προσανατολισμένες, επικεντρώνονται δηλαδή σε θεματικές περιοχές, όπως π.χ. πελάτες ή προμηθευτές. Για τον λόγο αυτό, πρέπει να περιληφθούν τα σχετικά δεδομένα και να αποκλειστούν τα μη σχετικά. Επίσης τα δεδομένα πρέπει να συνολικοποιηθούν σύμφωνα με θέματα που ενδιαφέρουν τη διοίκηση, όπως π.χ. πωλήσεις ανά περιοχή ή ανά χρονική περίοδο ή ανά κατηγορία προϊόντος, καθώς επίσης και να οριστεί ο βαθμός λεπτομέρειας ή γενίκευσης, όπως π.χ. πωλήσεις ανά εβδομάδα ή ανά μήνα ή ανά τρίμηνο.

Διερεύνηση Δεδομένων

Το τρίτο επίπεδο περιλαμβάνει εργασίες αρχικής επεξεργασίας των δεδομένων. Στο στάδιο αυτό ο χρήστης υποβάλλει ερωτήματα (queries) στη βάση δεδομένων, λαμβάνει απαντήσεις και συντάσσει αναφορές. Στις αναφορές μπορεί να περιλαμβάνονται αριθμητικές τιμές αλλά και πίνακες και γραφήματα. Τα γραφήματα μπορούν να αποδώσουν με πιο παραστατικό και ευχάριστο τρόπο την πληροφορία. Γενικώς οι μέθοδοι οπτικοποίησης βοηθούν στην καλύτερη παράθεση και κατανόηση των δεδομένων. Στο στάδιο αυτό μπορεί να γίνει και μια αρχική στατιστική επεξεργασία των δεδομένων. Μπορούν για παράδειγμα να υπολογίζονται μέσοι όροι, τυπικές αποκλίσεις κ.λπ. Χαρακτηριστικό αυτού του επιπέδου είναι ότι ο χρήστης, σύμφωνα με το σκεπτικό του, αναπτύσσει εκ των προτέρων υποθέσεις και στη συνέχεια χρησιμοποιεί τα εργαλεία ανάλυσης για να επιβεβαιώσει ότι οι υποθέσεις του υποστηρίζονται από τα δεδομένα.

Εξόρυξη Δεδομένων

Στο τέταρτο στάδιο εκτελείται υψηλού επιπέδου ανάλυση των δεδομένων, με τη χρήση των πιο εξελιγμένων τεχνικών. Χρησιμοποιούνται προχωρημένες στατιστικές μέθοδοι, αλλά και μέθοδοι που προέρχονται από την Τεχνητή Νοημοσύνη και τη Μηχανική Μάθηση. Οι μέθοδοι κατηγοριοποίησης (classification) επιτρέπουν την πρόβλεψη της κατηγορίας στην οποία ανήκει ένα αντικείμενο με βάση τα χαρακτηριστικά του. Η πρόβλεψη χρεοκοπίας και η εκτίμηση πιστοληπτικής ικανότητας είναι χαρακτηριστικά παραδείγματα εφαρμογής τεχνικών κατηγοριοποίησης. Μέθοδοι ανάλυσης συστάδων (cluster analysis) επιτρέπουν τον εντοπισμό ομάδων ομοειδών αντικειμένων. Ένα χαρακτηριστικό που συναντάται συχνά στις μεθόδους αυτού του επιπέδου είναι ότι ο χρήστης δεν χρειάζεται να διατυπώσει δικές του αρχικές υποθέσεις. Οι αλγόριθμοι επεξεργάζονται τα δεδομένα και εξάγουν την πληροφορία απευθείας από αυτά. Συχνά το αποτέλεσμα είναι ένα μοντέλο. Για παράδειγμα ένα δένδρο απόφασης μπορεί να περιγράφει τα χαρακτηριστικά των αγοραστών μιας κατηγορίας προϊόντων, π.χ. τετρακίνητων αυτοκινήτων. Ο αλγόριθμος θα διαβάσει τα στοιχεία των πωλήσεων, θα εντοπίσει τα κοινά χαρακτηριστικά των καταναλωτών του συγκεκριμένου προϊόντος και θα κατασκευάσει ένα μοντέλο από κανόνες της μορφής εάν-τότε, οι οποίοι θα περιγράφουν ποιοι αγοράζουν το προϊόν και με ποια πιθανότητα. Ο χρήστης δεν χρειάζεται να διατυπώσει καμία αρχική υπόθεση.

Βελτιστοποίηση

Η λήψη αποφάσεων είναι μια διαδικασία επιλογής. Οι αναλύσεις που πραγματοποιήθηκαν στα χαμηλότερα επίπεδα αποφέρουν μια σειρά ενδεχόμενων λύσεων. Ο υπεύθυνος για τη λήψη της απόφασης καλείται να επιλέξει μια από τις πολλές εναλλακτικές λύσεις. Ως προς το πλήθος των πιθανών λύσεων, τα προβλήματα χωρίζονται σε τρεις κατηγορίες. Τα δυαδικά προβλήματα μπορούν να έχουν δύο δυνατές λύσεις, π.χ. έγκριση του δανείου ή απόρριψη της αίτησης. Τα προβλήματα πολλαπλών λύσεων μπορούν να έχουν έναν περιορισμένο αριθμό ενδεχόμενων λύσεων. Η επιλογή ενός προμηθευτή μέσα από ένα σύνολο υποψήφιων προμηθευτών είναι τέτοιου είδους πρόβλημα. Τέλος, υπάρχουν προβλήματα απεριόριστου αριθμού ενδεχόμενων λύσεων. Αντικείμενο των εργασιών αυτού του επιπέδου είναι ο εντοπισμός της βέλτιστης λύσης.

Λήψη απόφασης

Στο κορυφαίο επίπεδο της πυραμίδας γίνεται η λήψη της οριστικής απόφασης. Στο σημείο αυτό, είναι σημαντικό να τονιστεί ότι όλες οι μέθοδοι και τα συστήματα που αναφέρονται παραπάνω, έχουν στόχο την υποβοήθηση ενός ανθρώπου στη λήψη της απόφασης και όχι την αυτοματοποιημένη λήψη απόφασης από

έναν υπολογιστή. Πρόκειται ουσιαστικά για εργαλεία ανάλυσης δεδομένων και παραγωγής πληροφοριών. Η τελική απόφαση λαμβάνεται από άνθρωπο, ο οποίος φέρει και την ευθύνη για αυτήν την απόφαση. Ο άνθρωπος, όταν λαμβάνει μια απόφαση, διευκολύνεται στην εργασία του εάν χρησιμοποιήσει περίτεχνα εργαλεία, τα οποία θα του προσφέρουν κατάλληλη πληροφόρηση. Την πληροφόρηση αυτή θα τη χρησιμοποιήσει σε συνδυασμό με τη δική του λογική, τη γνώση και τις ικανότητες του. Πέρα όμως από αυτά, ο άνθρωπος διαθέτει και άλλες ικανότητες και ιδιότητες, τις οποίες μπορεί να επιστρατεύσει. Τέτοιες είναι η φαντασία, το ένστικτο, η διαίσθηση καθώς και πλευρές του χαρακτήρα του.

2.2. Οφέλη και Περιορισμοί της Επιχειρηματικής Ευφυΐας

Τα Συστήματα Επιχειρηματικής Ευφυΐας αξιοποιούν τεχνολογίες της Πληροφορικής για να επεξεργαστούν δεδομένα, να παράξουν πληροφορία και να συνδράμουν τη διοίκηση στον έλεγχο και την καλύτερη λειτουργία ενός οργανισμού. Όπως κάθε τεχνολογική λύση, μπορούν να προσφέρουν πολλά οφέλη, ταυτόχρονα όμως υπόκεινται σε περιορισμούς.

Τα βασικά οφέλη που προσφέρουν τα συστήματα Επιχειρηματικής Ευφυΐας είναι τα ακόλουθα:

- Καλύτερη κατανόηση πελατών, αγορών, ανταγωνιστών, προμηθειών και πόρων. Η κατάλληλη οργάνωση των δεδομένων και τα εξελιγμένα εργαλεία πληροφορικής δίνουν πρωτόγνωρες δυνατότητες στην εμβάθυνση όλων των παραπάνω ζητημάτων.
- Τροφοδότηση της διοίκησης με τη σωστή πληροφόρηση, την κατάλληλη στιγμή και με τον κατάλληλο τρόπο. Τα συστήματα της Επιχειρηματικής Ευφυΐας μπορούν να αναδείξουν την ουσιαστική πληροφορία. Ταυτόχρονο και βασικό μέλημα όμως είναι και η έγκαιρη πληροφόρηση.
- Βελτίωση της ποιότητας των αποφάσεων. Η αναβαθμισμένη και έγκαιρη πληροφόρηση επιτρέπει στη διοίκηση του οργανισμού να λάβει βελτιωμένες αποφάσεις.
- Συμβολή στη διαμόρφωση των στρατηγικών στόχων. Τα συστήματα Επιχειρηματικής Ευφυΐας απευθύνονται κυρίως στα υψηλά ή και κορυφαία στελέχη των επιχειρήσεων. Στο επίπεδο αυτό λαμβάνονται οι στρατηγικές αποφάσεις. Η διοίκηση αξιοποιεί τα συστήματα αυτά για την άντληση ποιοτικής πληροφόρησης και τον καθορισμό των στρατηγικών στόχων.
- Επίτευξη συγκριτικού πλεονεκτήματος. Η εξασφάλιση συγκριτικού πλεονεκτήματος αποτελεί μόνιμη επιδίωξη κάθε επιχείρησης. Η βελτίωση των αποφάσεων και μέσω αυτού η αύξηση της αποτελεσματικότητας και αποδοτικότητας της διοίκησης, καθώς και ο καθορισμός σωστών στρατηγικών στόχων, μπορούν να αποτελέσουν το συγκριτικό πλεονέκτημα και να οδηγήσουν σε αυξημένη ανταγωνιστικότητα.
- Δυνατότητες αύξησης της κερδοφορίας, μείωσης του κόστους και βελτίωσης της αποδοτικότητας. Η βελτίωση της πληροφόρησης σχετικά με τη διαχείριση της εφοδιαστικής αλυσίδας μπορεί να βοηθήσει στη συμπίεση του κόστους, ενώ η κατανόηση των αγορών μπορεί να αυξήσει τις πωλήσεις και τα κέρδη. Γενικώς, επιτυχημένα συστήματα Επιχειρηματικής Ευφυΐας συμβάλλουν στην αύξηση των επιδόσεων και της κερδοφορίας.
- Αύξηση της πιθανότητας πρόβλεψης συμβάντων και επιχειρηματικών ευκαιριών. Η βαθύτερη κατανόηση της αγοράς επιτρέπει τον εντοπισμό επιχειρηματικών ευκαιριών. Επιπλέον, οι μέθοδοι προγνωστικής ανάλυσης (predictive analytics) επεξεργάζονται ιστορικά δεδομένα και επιτρέπουν τη διατύπωση προβλέψεων.
- Μεγαλύτερη αξιοποίηση των δεδομένων και αύξηση της απόδοσης της επένδυσης σε τεχνολογίες πληροφορικής. Οι σημερινές επιχειρήσεις έχουν επενδύσει εκατομμύρια ευρώ σε πληροφοριακά

συστήματα. Τα δεδομένα αυτών των συστημάτων μπορούν να αποδειχθούν πολύτιμη πηγή πρόσθετης, μη συμβατικής πληροφόρησης, εάν αξιοποιηθούν με τη χρήση της Επιχειρηματικής Ευφυΐας. Με τον τρόπο αυτό, οι επενδύσεις πληροφορικής αποδίδουν πρόσθετους καρπούς.

Η ανάπτυξη συστημάτων Επιχειρηματικής Ευφυΐας έχει να αντιμετωπίσει διάφορους ανασχετικούς παράγοντες, προβλήματα και ενδεχόμενους κινδύνους:

- Κόστος απόκτησης και λειτουργίας Αποθηκών Δεδομένων και συστημάτων Επιχειρηματικής Ευφυΐας. Απαιτούνται επενδύσεις σε υλικό, λογισμικό και τεχνογνωσία. Επίσης οι εργασίες ETL είναι χρονοβόρες, δύσκολες και δαπανηρές. Όλα τα παραπάνω επιφέρουν ένα όχι ευκαταφρόνητο κόστος, το οποίο πρέπει να αναλάβει η επιχείρηση.
- Χαμηλή ποιότητα δεδομένων. Το πρόβλημα αυτό είναι ένα από τα σημαντικότερα στην ανάπτυξη συστημάτων Επιχειρηματικής Ευφυΐας. Τα αρχικά δεδομένα είναι διάσπαρτα, ανομοιογενή, ελλιπή και πιθανώς λανθασμένα ή αντιφατικά. Τροφοδότηση του συστήματος με προβληματικά δεδομένα θα οδηγήσει σε εσφαλμένη πληροφόρηση. Όπως χαρακτηριστικά λέγεται «garbage in, garbage out».
- Ζητήματα συμβατότητας με τα υπάρχοντα συστήματα. Τα συστήματα Επιχειρηματικής Ευφυΐας λειτουργούν επί δεδομένων άλλων συστημάτων. Τα συστήματα αυτά μπορεί να είναι πολλά, διαφορετικά, και πιθανότατα δεν έχει ληφθεί εκ των προτέρων καμία πρόνοια για ενοποίηση των δεδομένων τους. Μπορεί να εμφανιστούν προβλήματα συμβατότητας, τόσο μεταξύ των βασικών συστημάτων όσο και μεταξύ αυτών και του συστήματος Επιχειρηματικής Ευφυΐας.
- Πιθανή ύπαρξη επιφυλάξεων, δυσπιστίας και μη συνεργασίας από την πλευρά των στελεχών. Η ανάπτυξη συστημάτων Επιχειρηματικής Ευφυΐας επιφέρει αλλαγές σε λειτουργίες των οργανισμών. Έχει παρατηρηθεί ότι τέτοιες αλλαγές μπορεί να προκαλέσουν τις επιφυλάξεις και τη δυσπιστία των εμπλεκόμενων στελεχών. Είναι πολύ σημαντικό, τα ανώτατα στελέχη της διοίκησης να εφαρμόσουν πολιτικές διαχείρισης της αλλαγής (change management) και να επιληφθούν τέτοιων προβλημάτων.
- Προβλήματα επικοινωνίας και συνεννόησης μεταξύ των στελεχών και των ειδικών πληροφορικής. Τα στελέχη της επιχείρησης και οι ειδικοί της πληροφορικής έχουν ο καθένας τη δική του οπτική γωνία. Τα στελέχη επικεντρώνονται στα επιχειρησιακά ζητήματα, ενώ οι ειδικοί πληροφορικής στα τεχνικά. Αυτό μπορεί να προκαλέσει προβλήματα συνεννόησης. Ειδικά στα συστήματα Επιχειρηματικής Ευφυΐας, όπου τα επιχειρησιακά ζητήματα παίζουν βαρύνοντα ρόλο, το πρόβλημα αυτό μπορεί να ενταθεί.
- Ανάγκη ειδικά εκπαιδευμένου προσωπικού. Πρέπει να προσληφθεί νέο προσωπικό, αλλά κυρίως πρέπει τα στελέχη να μάθουν να χρησιμοποιούν, με τον βέλτιστο τρόπο, τα νέα αυτά συστήματα.
- Κίνδυνος υπερβολικής και άκριτης εμπιστοσύνης στο σύστημα Επιχειρηματικής Ευφυΐας και συνακόλουθης επανάπαυσης. Έχει ήδη τονιστεί ότι ο τελικός υπεύθυνος για τη λήψη των αποφάσεων είναι ο άνθρωπος. Συστήματα ευφυούς ανάλυσης των δεδομένων και κυρίως συστήματα ικανά να διατυπώνουν προβλέψεις, μπορεί μετά από κάποιον χρόνο να εμπνεύσουν υπερβολική εμπιστοσύνη στους χρήστες τους. Τα στελέχη δεν πρέπει να επαναπαύονται στις προβλέψεις του συστήματος, και πρέπει να αντιμετωπίζουν την πληροφόρηση στη βάση της δικής τους υποκειμενικής κρίσης.
- Πολλές περιπτώσεις αποτυχίας σε έργα Επιχειρηματικής Ευφυΐας. Τα έργα Επιχειρηματικής Ευφυΐας έχουν να αντιμετωπίσουν πολλές προκλήσεις. Ως αποτέλεσμα αυτού του γεγονότος καταγράφεται μεγάλο ποσοστό αποτυχίας έργων επιχειρηματικής ευφυΐας. Σύμφωνα με τον Saran (2012), ο οποίος επικαλείται πηγές του οίκου Gartner, λιγότερο από το 30% των έργων Επιχειρηματικής Ευφυΐας επιτυγχάνει τους σκοπούς του. [2]

2.3. Βασικές πτυχές της αναλυτικής επιχειρησιακών διεργασιών

Παρόλο που δεν είναι εύκολο να ταξινομήσει κανείς τις λειτουργίες της Αναλυτικής Επιχειρησιακών Διεργασιών υπάρχει μία ποικιλομορφία βασικών εργασιών ανάλυσης και μια αναλυτική διεργασία συχνά συνδυάζει πολλές από αυτές. Μερικά από τα βασικά είδη είναι:

- Παρακολούθηση για εξαιρέσεις: ο έλεγχος του αν τα πράγματα πηγαίνουν όπως θα έπρεπε.
- Παρακολούθηση και αξιολόγηση τάσεων: η μέτρηση και η αποτίμηση της αλλαγής.
- Σχεδίαση, προϋπολογισμός, πρόβλεψη: πολλές από τις σημαντικότερες αποφάσεις σε έναν οργανισμό αφορούν την κατανομή των παραγωγικών συντελεστών εντός του οργανισμού, δηλαδή τον κατάλληλο προϋπολογισμό. Ο προϋπολογισμός απαιτεί σχεδίαση και η σχεδίαση, σχεδόν πάντα, απαιτεί πρόβλεψη.
- Επενδυτική ανάλυση: σε κάθε οργανισμό οι κάθε είδους επενδύσεις απαιτείται να αναλύονται λεπτομερώς.
- Στατιστική ανάλυση: υπάρχει ένα ολόκληρο πεδίο εξελιγμένης στατιστικής ανάλυσης κυρίως στον τομέα της διασφάλισης ποιότητας και του marketing.

Βελτιστοποίηση: συνεχής βελτιστοποίηση των επιμέρους διεργασιών αλλά και ολόκληρου του οργανισμού ως σύνολο. [3]

2.4. Σχεδιασμός Μοντέλου Διεργασίας

Θεωρούμε την Πνευμονολογική Κλινική ως ένα σύστημα αποτελούμενο από επιχειρησιακές διεργασίες που παράγουν, ενημερώνουν ή διακινούν δεδομένα, οπότε οι μεθοδολογίες καθοδηγούν την ανάπτυξη συστημάτων με βάση αυτήν την άποψη.

Στην παρούσα εργασία θα γίνει χρήση του Oracle BPM [4]. Αυτό αποτελεί μια Σουίτα εργαλείων επιχειρησιακής μοντελοποίησης, βελτιστοποίησης, παρακολούθησης δραστηριοτήτων και διεργασιών και εξαγωγής χρήσιμων αναλυτικών δεδομένων.

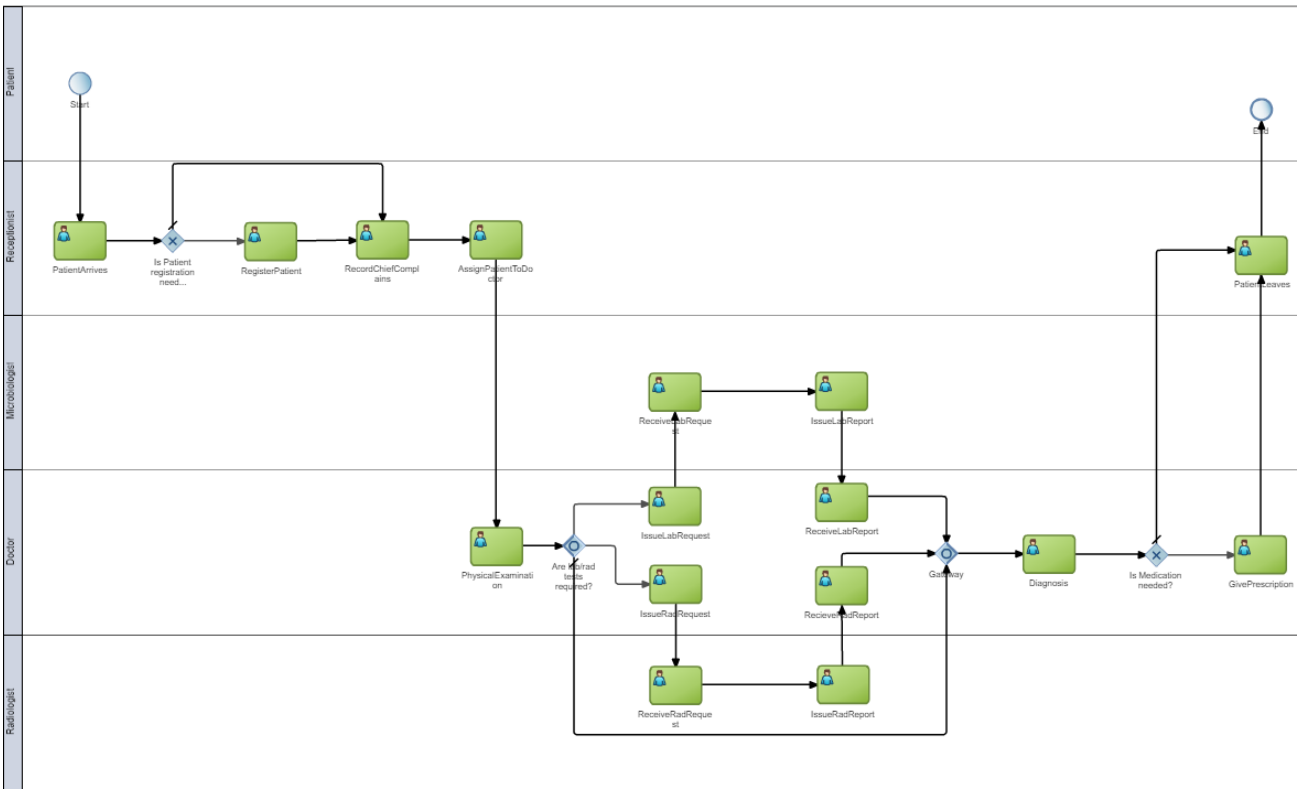
Στο Μοντέλο διεργασίας έχουμε δημιουργήσει 5 ρόλους: Patient, Secretary, Doctor, Microbiologist και Radiologist. Σε αυτούς τους ρόλους έχουν ανατεθεί συνολικά τα παρακάτω 16 δραστηριότητες (user tasks): Patient Arrives, Register Patient, Record Chief Complains, Assign Patient to Doctor, Physical Examination, Issue Lab Request, Issue Rad Request, Receive Lab Request, Receive Rad Request, Issue Lab Report, Issue Rad Report, Receive Lab Report, Receive Rad Report, Diagnosis, Give Prescription, Patient Leaves.

Μόλις ο Ασθενής (Patient) εισέλθει στην πνευμονολογική κλινική (PatientArrives), περνάει από την Υποδοχή (Receptionist), όπου ελέγχεται εάν τα στοιχεία του είναι ήδη καταχωρημένα στη βάση δεδομένων. Εάν όχι γίνεται η εισαγωγή των στοιχείων (RegisterPatient) και στη συνέχεια γίνεται μια αρχική καταγραφή της κατάστασης του ασθενή (RecordChiefComplains). Όταν ολοκληρωθεί η καταγραφή, γίνεται ανάθεση του Ασθενή στον αρμόδιο Γιατρό (AssignPatientToDoctor).

Στην εξέταση που ακολουθεί (PhysicalExamination) ο Γιατρός καταγράφει τα συμπτώματα και κρίνει εάν είναι απαραίτητες μικροβιολογικές ή/και ακτινολογικές εξετάσεις. Σε αυτή την περίπτωση συντάσσει τα αντίστοιχα παραπεμπτικά (IssueLabRequest/IssueRadRequest). Ο Ακτινολόγος ή/και ο Μικροβιολόγος αφού πραγματοποιήσουν τις συνιστώμενες εξετάσεις συντάσσουν τη σχετική αναφορά, στην οποία θα παραλάβει ο Γιατρός.

Μετά την ολοκλήρωση της διαδικασίας των εξετάσεων –εφόσον αυτές έχουν θεωρηθεί απαραίτητες–, ο Γιατρός πραγματοποιεί τη διάγνωση (Diagnosis) του περιστατικού και κρίνει εάν χρειάζεται η χορήγηση κάποιας φαρμακευτικής αγωγής (GivePrescription). Πριν αποχωρήσει ο ασθενής από την κλινική περνάει μια τελευταία φορά από την Υποδοχή.

Πιο αναλυτικά το μοντέλο διεργασίας παρουσιάζεται στο Σχήμα 2.2.

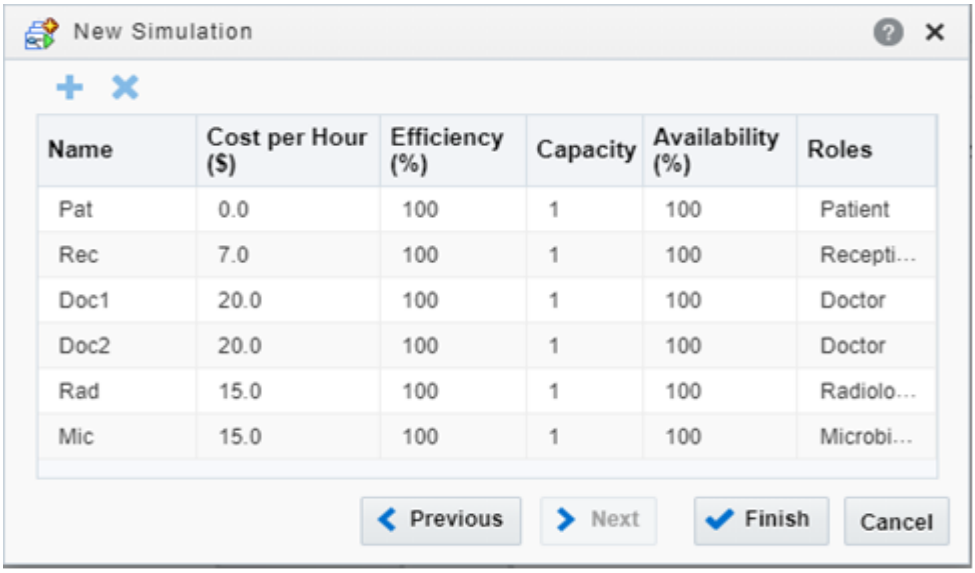


Σχήμα 2.2. Μοντέλο διεργασίας της πνευμονολογικής κλινικής

2.5. Προσομοίωση διεργασίας

Την παραπάνω διεργασία την υποβάλουμε σε μια διαδικασία προσομοίωσης. Στην προσομοίωση αυτή έχουμε θεωρήσει ότι η πνευμονολογική κλινική είναι μεσαίας κλίμακας και ότι έχουμε κατά μέσο όρο 5 περιστατικά την ώρα. Επίσης, έχουμε ορίσει το κόστος του προσωπικού να είναι σχετικά συμβατό με τα σημερινά επίπεδα. Αυτές οι υποθέσεις φαίνονται στο Σχήμα 2.3. και στον Πίνακα 2.4.

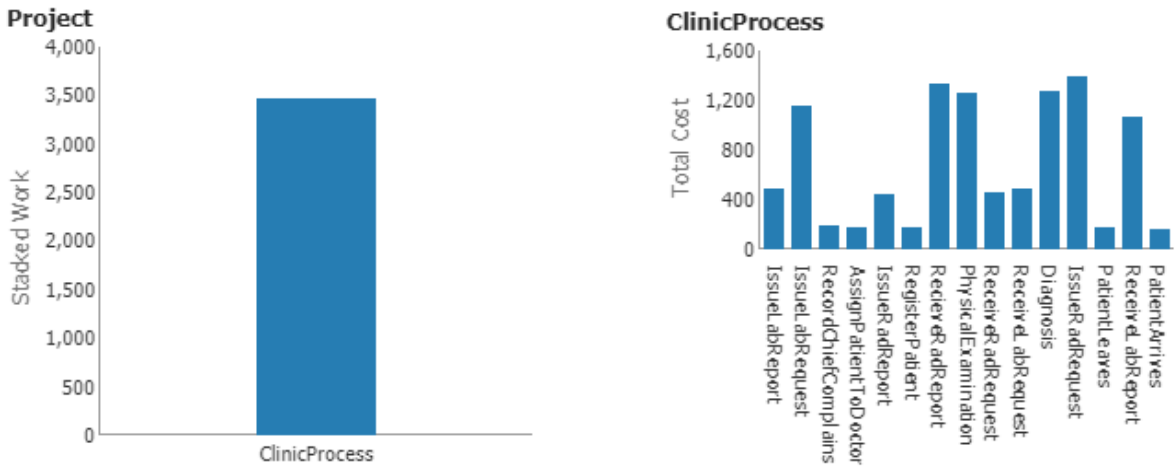
Με βάση τις παραπάνω υποθέσεις προκύπτουν τα Σχήματα 2.5 έως 2.13.



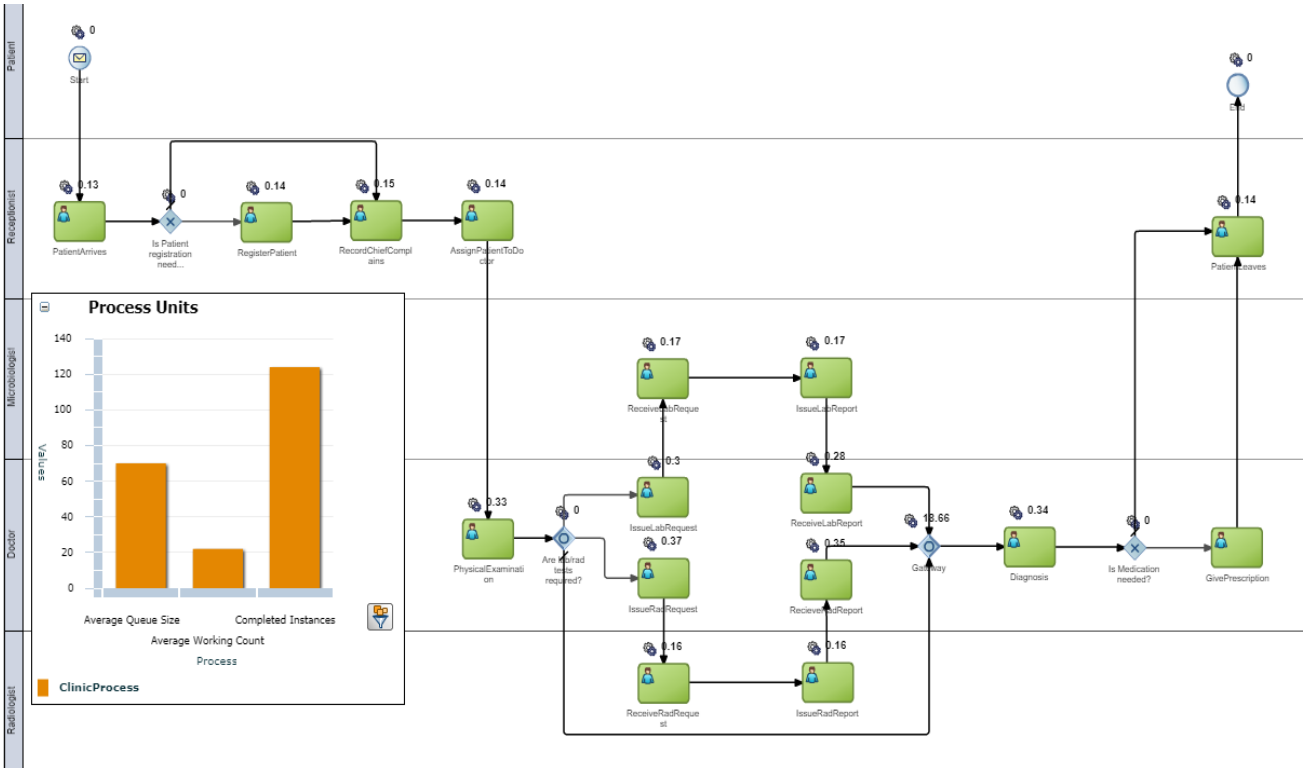
Σχήμα 2.3. Προσομοίωση ρόλων διαδικασίας

Activity Name	*Duration (Distribution Type, Frequency)*	*Resources*	*Outgoing Flows*
Start Event	Exponential, 5 instances/h		
PatientArrives	Exponential, 5 instances/h	As defined in organization resources	N/A
Is Patient registration needed?			RegisterPatient – 70% RecordChiefComplains – 30%
RegisterPatient	Exponential, 4 instances/h	As defined in organization resources	N/A
RecordChiefComplains	Exponential, 5 instances/h	As defined in organization resources	N/A
AssignPatientToDoctor	Exponential, 5 instances/h	As defined in organization resources	N/A
PhysicalExamination	Exponential, 2 instances/h	As defined in organization resources	N/A
Are lab/rad tests required?			IssueLabRequest – 30% IssueRadRequest – 30% Diagnosis – 40%
IssueLabRequest	Exponential, 2 instances/h	As defined in organization resources	N/A
IssueRadRequest	Exponential, 2 instances/h	As defined in organization resources	N/A
ReceiveLabRequest	Exponential, 4 instances/h	As defined in organization resources	N/A
ReceiveRadRequest	Exponential, 4 instances/h	As defined in organization resources	N/A
IssueLabReport	Exponential, 4 instances/h	As defined in organization resources	N/A
IssueRadReport	Exponential, 4 instances/h	As defined in organization resources	N/A
ReceiveLabReport	Exponential, 2 instances/h	As defined in organization resources	N/A
ReceiveRadReport	Exponential, 2 instances/h	As defined in organization resources	N/A
Diagnosis	Exponential, 2 instances/h	As defined in organization resources	N/A
GivePrescription	Exponential, 2 instances/h	As defined in organization resources	N/A
PatientLeaves	Exponential, 5 instances/h	As defined in organization resources	N/A

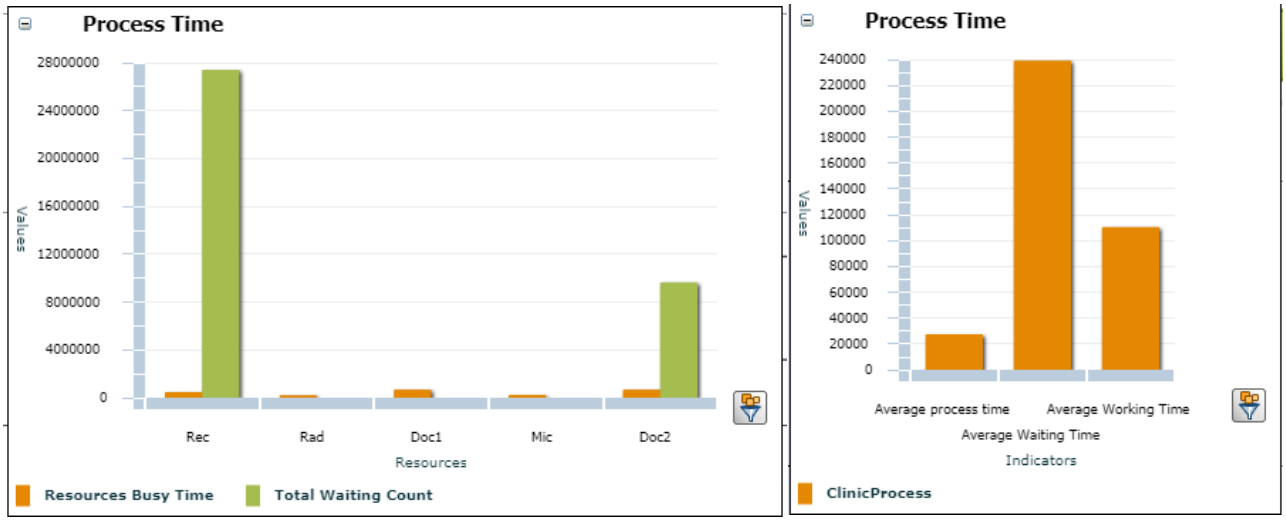
Πίνακας 2.4. Συχνότητα περιστατικών ανά δραστηριότητα



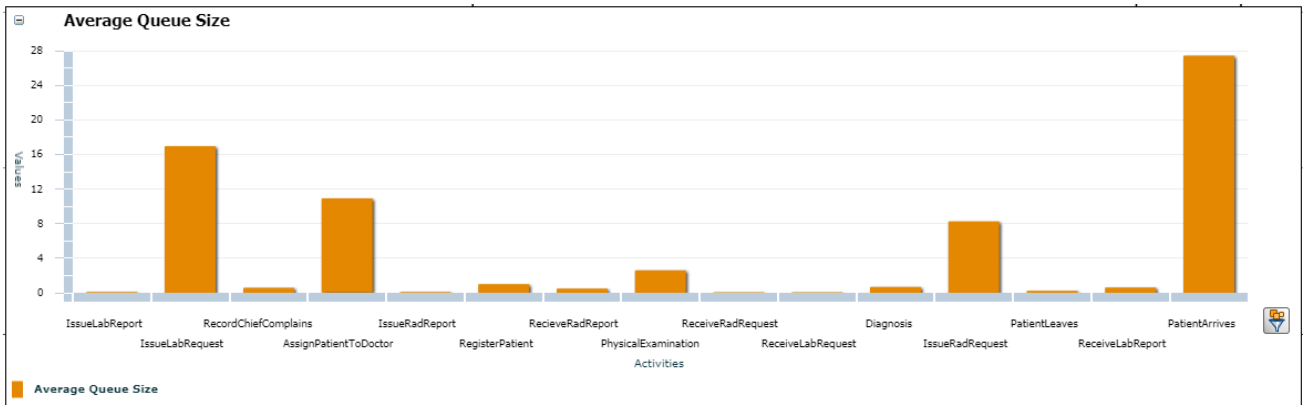
Σχήμα 2.5. Συνολικό κόστος της διεργασίας αλλά και κάθε επιμέρους δραστηριότητας Όπως είναι αναμενόμενο οι πιο «ακριβές» δραστηριότητες είναι αυτές που εμπεριέχουν γιατρό, ακτινολόγο και μικροβιολόγο.



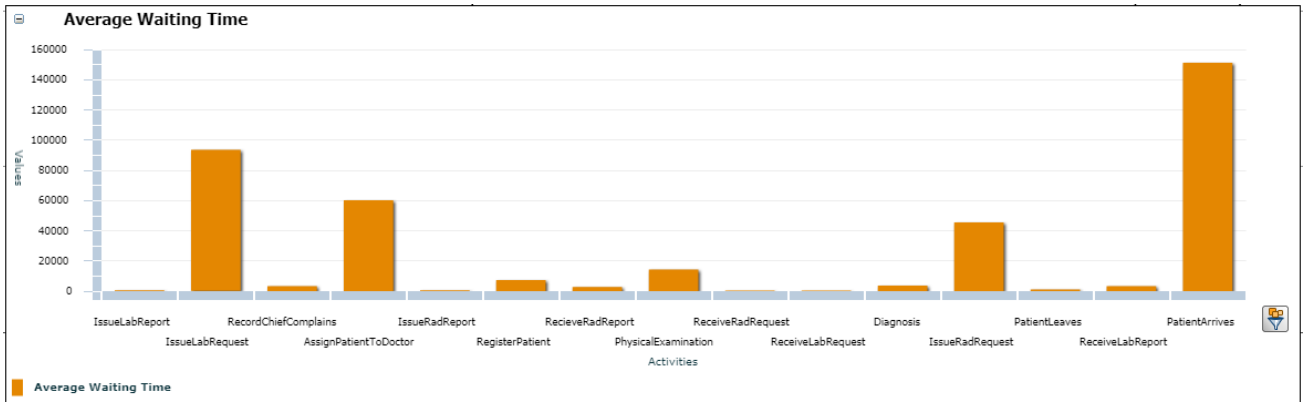
Σχήμα 2.6. Στην αριστερή στήλη βλέπουμε τη μέση ουρά αναμονής, στη μεσαία τον μέσο αριθμό εργασιών και στα δεξιά τα ολοκληρωμένα στιγμιότυπα-περιστατικά



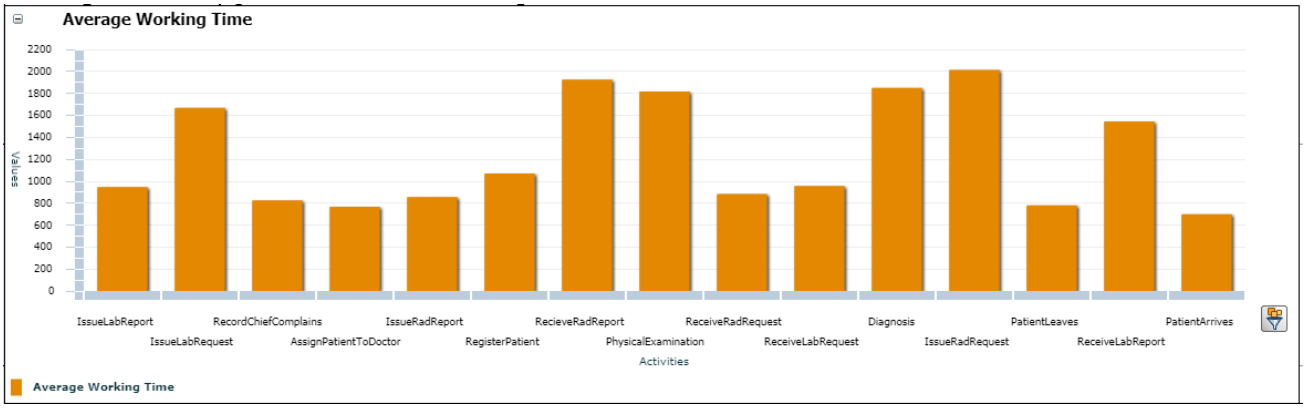
Σχήμα 2.7. Αριστερά βλέπουμε τον συνολικό χρόνο αναμονής για κάθε εργαζόμενο σε σύγκριση με τον συνολικό χρόνο που ήταν απασχολημένος. Δεξιά απεικονίζεται ο μέσος χρόνος της διεργασίας σε σύγκριση με τον μέσο χρόνο αναμονής και τον μέσο χρόνο εργασίας για τους εργαζόμενους.



Σχήμα 2.8. Η μέση ουρά των ασθενών για κάθε δραστηριότητα της διεργασίας



Σχήμα 2.9. Ο μέσος χρόνος αναμονής για κάθε δραστηριότητα της διεργασίας



Σχήμα 2.10. Ο μέσος χρόνος εργασίας για κάθε δραστηριότητα της διεργασίας

Αναλύοντας τα παραπάνω διαγράμματα και με δεδομένο ότι η συχνότητα των περιστατικών δεν θα αλλάξει, είναι προφανές ότι δημιουργείται bottleneck στην αρχή της διαδικασίας, στην υποδοχή, διότι σύμφωνα με το σχήμα 2.8 η ουρά πλησιάζει τα 28 άτομα. Προσλαμβάνοντας λοιπόν άλλον έναν υπάλληλο σε αυτή τη θέση η εμπειρία του ασθενή θα ήταν πολύ πιο ευχάριστη και οι χρόνοι αναμονής τόσο του ασθενή όσο και του υπόλοιπου προσωπικού θα μειώνονταν.

2.6. Υλοποίηση διεργασίας

2.6.1. Ορισμός των Data Objects

Κατά τη μοντελοποίηση μιας διεργασίας ορίζονται τα Data Objects, μέσα στα οποία αποθηκεύονται τα σχετικά με τη διεργασία δεδομένα. Οι τιμές τους μεταβάλλονται κατά τη διάρκεια της εκτέλεσης ενός στιγμιότυπου της διεργασίας και ενδέχεται να επηρεάσουν τη ροή εκτέλεσης. Στο Oracle Bpm Studio μπορούν να οριστούν δύο τύποι: τα Basic και τα Complex Data objects. Στα Basic αποθηκεύονται βασικοί τύποι δεδομένων όπως Integer, Boolean, String, Time κ.λπ. Ενώ τα Complex χρησιμοποιούνται για την ανάθεση ομαδοποιημένων δεδομένων.

Για την υλοποίηση της διεργασίας δημιουργήθηκαν 20 Basic data objects και 3 Complex data objects τα PatientBO, DoctorBO και PrescriptionBO, που υλοποιούνται ως business objects και αποτελούνται από τα παρακάτω επιμέρους βασικά data objects.

PatientBO: Αφορά τα προσωπικά δεδομένα του εκάστοτε ασθενή και πιο συγκεκριμένα τον μοναδικό του ΑΜΚΑ, όνομα, επίθετο, ημερομηνία γέννησης, φύλο, διεύθυνση κατοικίας, πόλη, ταχυδρομικό κώδικα, τηλέφωνο επικοινωνίας, και τρόπο πληρωμής.

- patAMKA (String)
- patName (String)
- patSurname (String)
- patDOB (Date)
- patGender (String)
- patStreet (String)
- patCity (String)
- patZipC (String)
- patPhone (String)

- patPaymentMethod (String)

DoctorBO: Αφορά βασικές πληροφορίες των γιατρών με τους οποίους συνεργάζεται η κλινική και πιο συγκεκριμένα τον μοναδικό τους ΑΜΚΑ, όνομα, επίθετο και ειδικότητα.

- docAMKA (String)
- docName (String)
- docSurname (String)
- docSpeciality (String)

PrescriptionBO: Αφορά τις πληροφορίες της συνταγογραφούμενης θεραπείας που προτείνεται στον ασθενή και πιο συγκεκριμένα τον κωδικό ICD της δραστικής ουσίας του φαρμάκου, το σκεύασμα, τον τύπο του φαρμάκου, τις ημερομηνίες έναρξης και λήξης της θεραπείας και την δόση του φαρμάκου.

- medICD (String)
- medication (String)
- drugType (String)
- startDate (Date)
- endDate (Date)
- dose (String)

Αντίστοιχα τα Process Data Objects είναι τα παρακάτω:

admTime (DateTime): Η ώρα που ο ασθενής έρχεται στην υποδοχή

patient: PatientBO

registered (Boolean): Αποθηκεύεται στο σύστημα προκειμένου να γίνει εφικτή η χρήση του κόμβου απόφασης που σχετίζεται με το αν ο ασθενής είναι ήδη καταχωρημένος στο σύστημα και μπορεί να προχωρήσει ο υπάλληλος στο βήμα της καταγραφής των προβλημάτων του ασθενή ή χρειάζεται να περαστούν πρώτα τα στοιχεία του στη βάση δεδομένων

caseID (String): Ο μοναδικός κωδικός που δίνεται στο κάθε περιστατικό. Σε αυτό τον κωδικό αντιστοιχίζονται όλες οι ενέργειες που θα γίνουν μέχρι ο ασθενής να πάρει εξιτήριο και να αποχωρίσει από την κλινική.

chiefComplains (String): Η καταγραφή των προβλημάτων που αναφέρει ο ασθενής

clinicalSigns (String): Τα ευρήματα που εντοπίζει ο γιατρός κατά τη διάρκεια της εξέτασης του ασθενή

doctor: DoctorBO

labExams (Boolean): Αποθηκεύεται στο σύστημα προκειμένου να γίνει εφικτή η χρήση του κόμβου απόφασης που σχετίζεται με το αν ο ασθενής θα παραπεμφθεί για αιματολογικές εξετάσεις

radExams (Boolean): Αποθηκεύεται στο σύστημα προκειμένου να γίνει εφικτή η χρήση του κόμβου απόφασης που σχετίζεται με το αν ο ασθενής θα παραπεμφθεί για ακτινολογικές εξετάσεις

labRequest (String): Το παραπεμπτικό για τις αιματολογικές εξετάσεις, όπου περιγράφονται ποιες τιμές πρέπει να ελεγχθούν

radRequest (String): Το παραπεμπτικό για τις ακτινολογικές εξετάσεις, όπου περιγράφεται το είδος τους

labReport (String): Τα αποτελέσματα των αιματολογικών εξετάσεων

radReport (String): Τα αποτελέσματα των ακτινολογικών εξετάσεων

labDate (Date): Η ημερομηνία διεξαγωγής των αιματολογικών εξετάσεων

radDate (Date): Η ημερομηνία διεξαγωγής των ακτινολογικών εξετάσεων

diagnosis (String): Η περιγραφή της διάγνωσης του γιατρού

diagnosisICD9 (String): Ο κωδικός ICD 9 της διάγνωσης

needMedication (Boolean): Αποθηκεύεται στο σύστημα προκειμένου να γίνει εφικτή η χρήση του κόμβου απόφασης που σχετίζεται με το αν στον ασθενή θα χορηγηθεί φαρμακευτική αγωγή

prescription: PrescriptionBO

dischTime (DateTime): Η ώρα που ο ασθενής παίρνει εξιτήριο από την υποδοχή

2.6.2. Human Tasks

Τα Human Tasks είναι βήματα που απαιτούν αλληλεπίδραση με τον εκάστοτε χρήστη και συνεπώς εκτελούνται από τον ίδιο και όχι αυτόματα από το σύστημα. Μέσω αυτών υλοποιούνται τα User Tasks που χρησιμοποιούνται στη διεργασία. Περιλαμβάνουν δε το λεγόμενο Payload, τα απαραίτητα δεδομένα δηλαδή που μεταφέρονται κάθε φορά μέσα από τη ροή της διεργασίας.

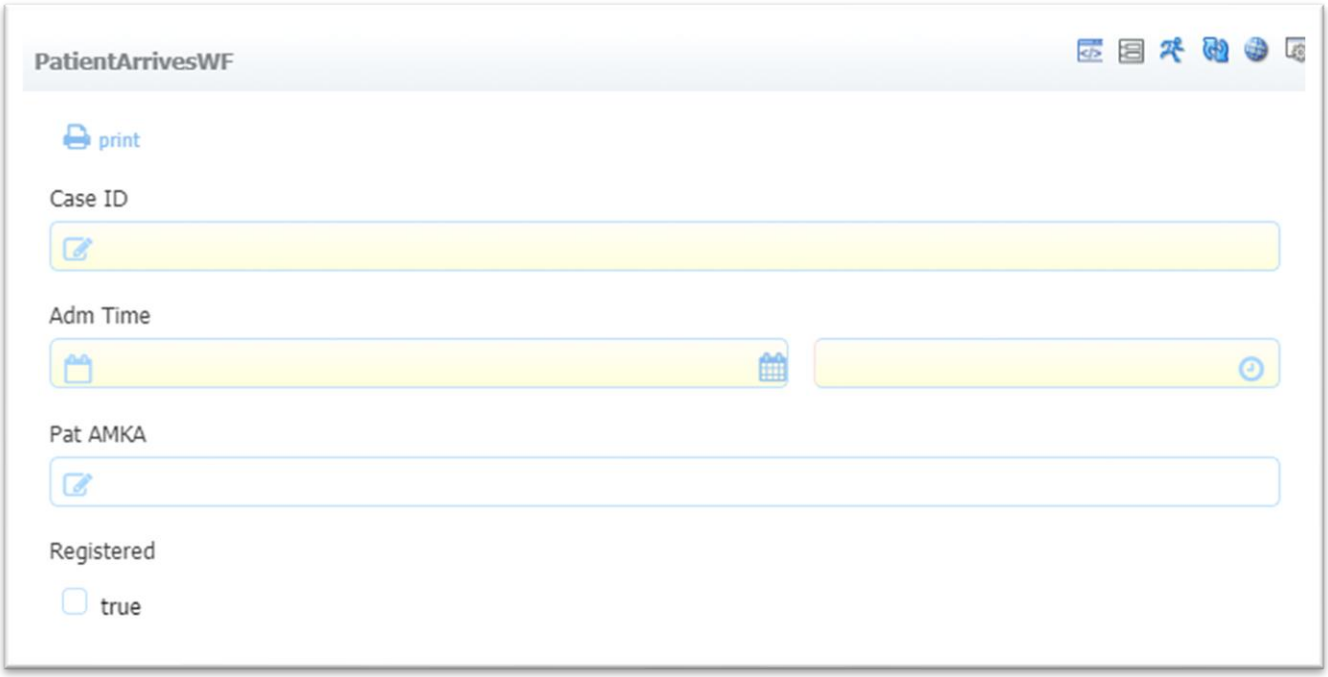
Παρακάτω ακολουθούν τα Human Tasks έτσι όπως ορίστηκαν καθώς και η απαραίτητη επεξήγηση αναφορικά με τη λειτουργία που επιτελεί το καθένα.

- **PatientArrivesHT:** Αποτελεί το σημείο εκκίνησης της διεργασίας και αφορά την είσοδο του ασθενή στην κλινική. Είναι η πρώτη επαφή με την Υποδοχή (Receptionist), όπου ελέγχεται εάν τα στοιχεία του είναι ήδη καταχωρημένα στη βάση δεδομένων.
- **RegisterPatientHT:** Έχει να κάνει με την καταχώρηση των στοιχείων του ασθενή στην βάση δεδομένων.
- **RecordChiefComplainsHT:** Αφορά την αρχική καταγραφή της κατάστασης του ασθενή.
- **AssignPatientToDoctorHT:** Έχει να κάνει με την ανάθεση του Ασθενή στον αρμόδιο Γιατρό.
- **PhysicalExaminationHT:** Είναι η πρώτη εξέταση του ασθενή από τον γιατρό, όπου καταγράφονται τα συμπτώματα και κρίνεται εάν είναι απαραίτητες μικροβιολογικές ή/και ακτινολογικές εξετάσεις.
- **IssueLabRequestHT** και **IssueRadRequestHT:** Εδώ συντάσσονται τα αντίστοιχα παραπεμπτικά.
- **ReceiveLabRequestHT**: Ο Μικροβιολόγος παραλαμβάνει το παραπεμπτικό και πραγματοποιεί τις συνιστώμενες εξετάσεις.
- **ReceiveRadRequestHT:** Ο Ακτινολόγος παραλαμβάνει το παραπεμπτικό και πραγματοποιεί τις συνιστώμενες εξετάσεις.

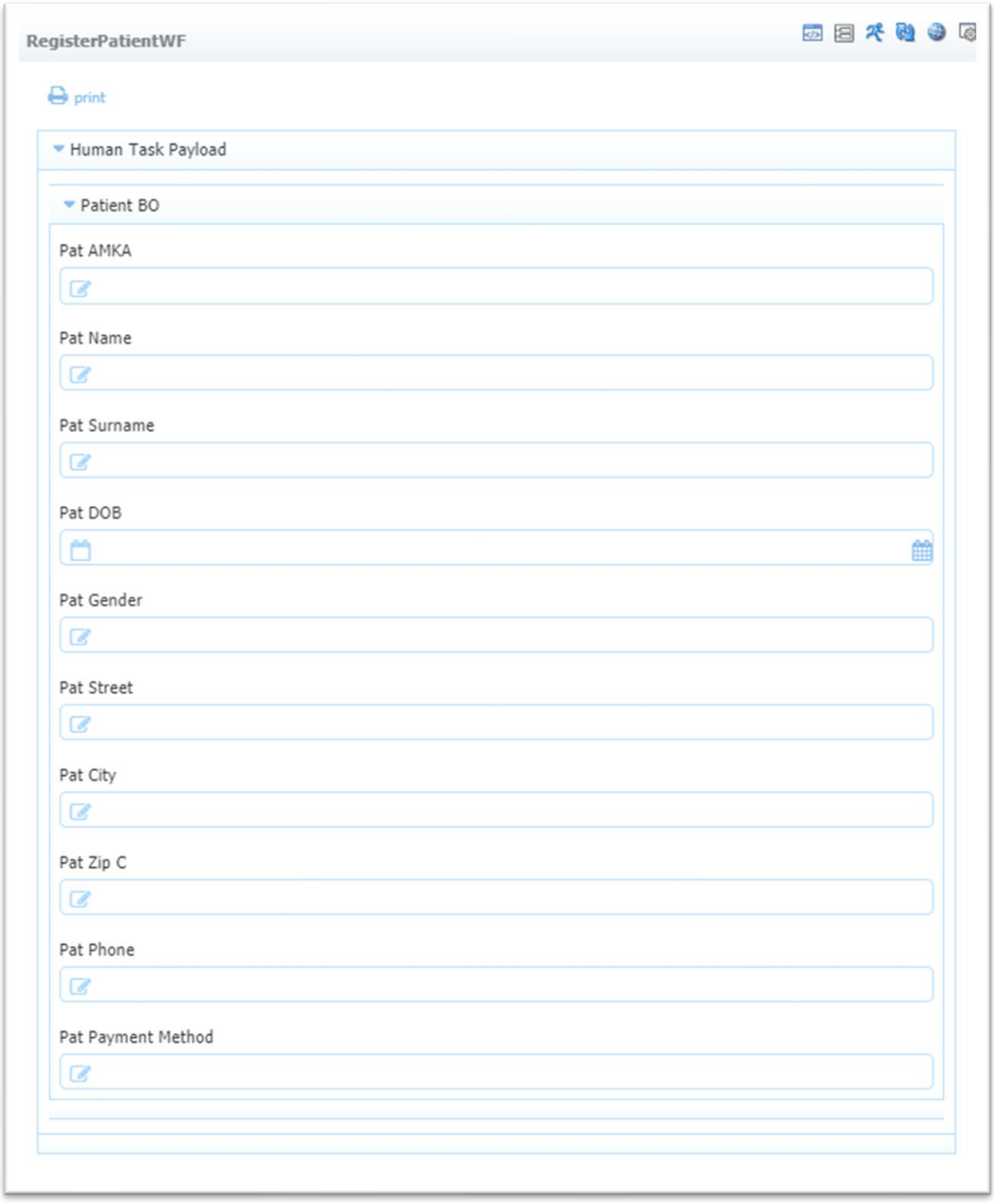
- **IssueLabRequestHT:** Ο Μικροβιολόγος συντάσσει τη σχετική αναφορά, στην οποία θα παραλάβει ο Γιατρός
- **IssueRadRequestHT**: Ο Ακτινολόγος συντάσσει τη σχετική αναφορά, στην οποία θα παραλάβει ο Γιατρός
- **DiagnosisHT:** Μετά την ολοκλήρωση της διαδικασίας των εξετάσεων, εφόσον αυτές έχουν θεωρηθεί απαραίτητες, ο Γιατρός πραγματοποιεί τη διάγνωση του περιστατικού και κρίνει εάν χρειάζεται η χορήγηση κάποιας φαρμακευτικής αγωγής
- **GivePrescriptionHT:** Χορήγηση φαρμακευτικής αγωγής.
- **PatientLeavesHT:** Πριν αποχωρήσει ο ασθενής από την κλινική περνάει μια τελευταία φορά από την Υποδοχή

2.6.3. Webforms

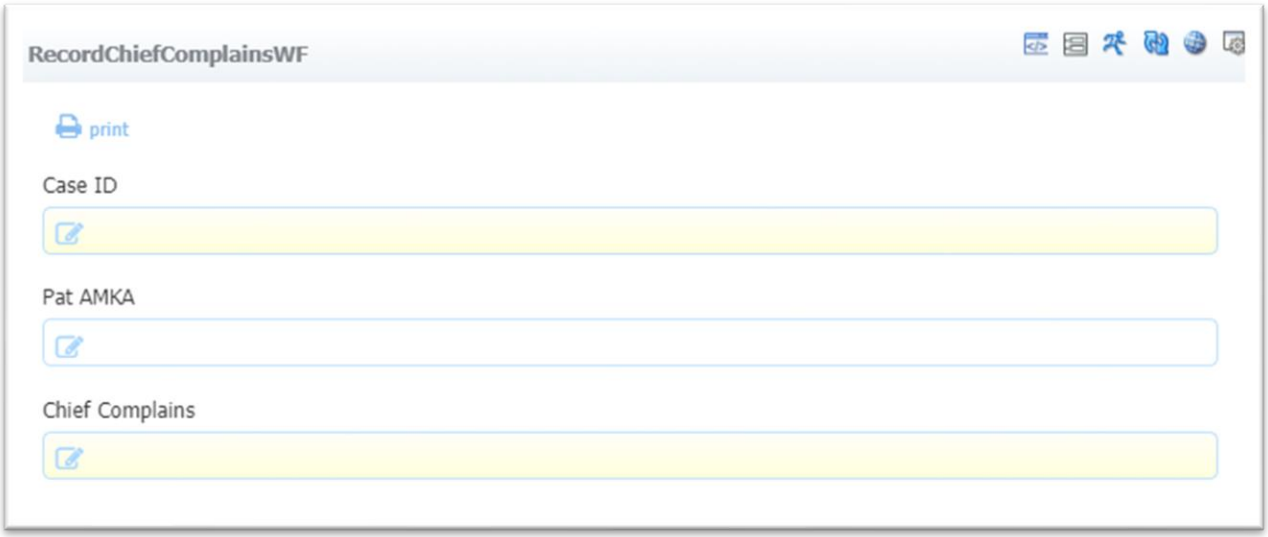
Απαραίτητο στοιχείο για να μπορέσει ο χρήστης να αλληλεπιδράσει με το πληροφοριακό μας σύστημα, είναι η κατασκευή των Webforms που αντιστοιχούν σε κάθε Human Task. Παρακάτω απεικονίζονται όλα τα παράθυρα που θα εμφανιστούν, καθ' όλη τη διάρκεια της διεργασίας, ανάλογα με τη δραστηριότητα, σε κάθε χρήστη. Στο επάνω αριστερά μέρος της εικόνας φαίνεται το Human Task στο οποίο αντιστοιχούν.



Σχήμα 2.11. Patient Arrives Web Form
Το πρώτο παράθυρο που καλούνται να συμπληρώσουν οι υπάλληλοι της υποδοχής, μόλις ο ασθενής εισέλθει στην κλινική. Εδώ δίνεται και ο αριθμός του περιστατικού.

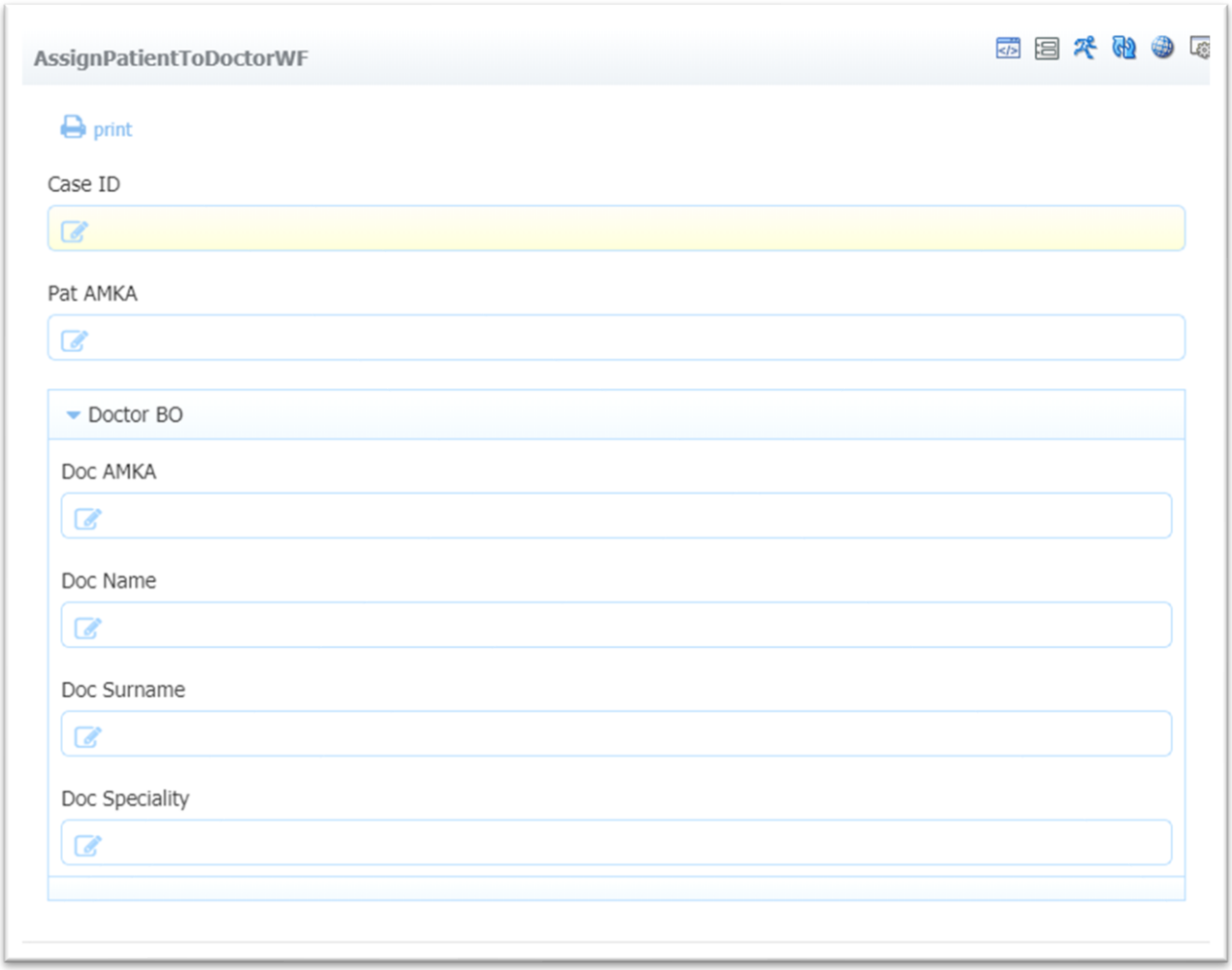


Σχήμα 2.12. Register Patient Web Form
Σε περίπτωση που ο ασθενής δεν είναι ήδη καταχωρημένος,
περνούν στο σύστημα τα στοιχεία του μέσω αυτού του παραθύρου



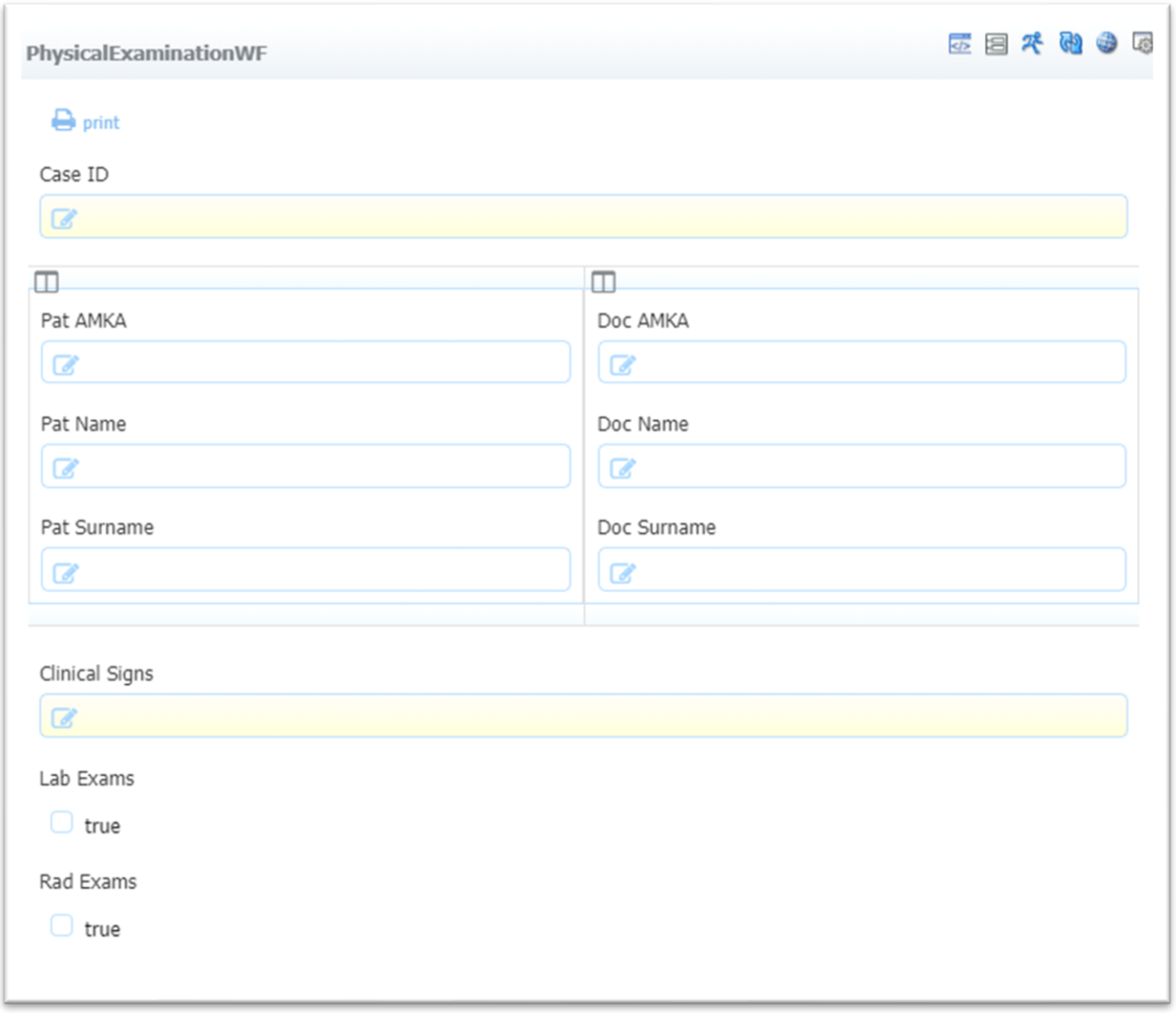
Σχήμα 2.13. Record Chief Complains Web Form

Μέσω αυτό του παραθύρου οι υπάλληλοι της υποδοχής καταγράφουν τα κύρια προβλήματα του Ασθενή



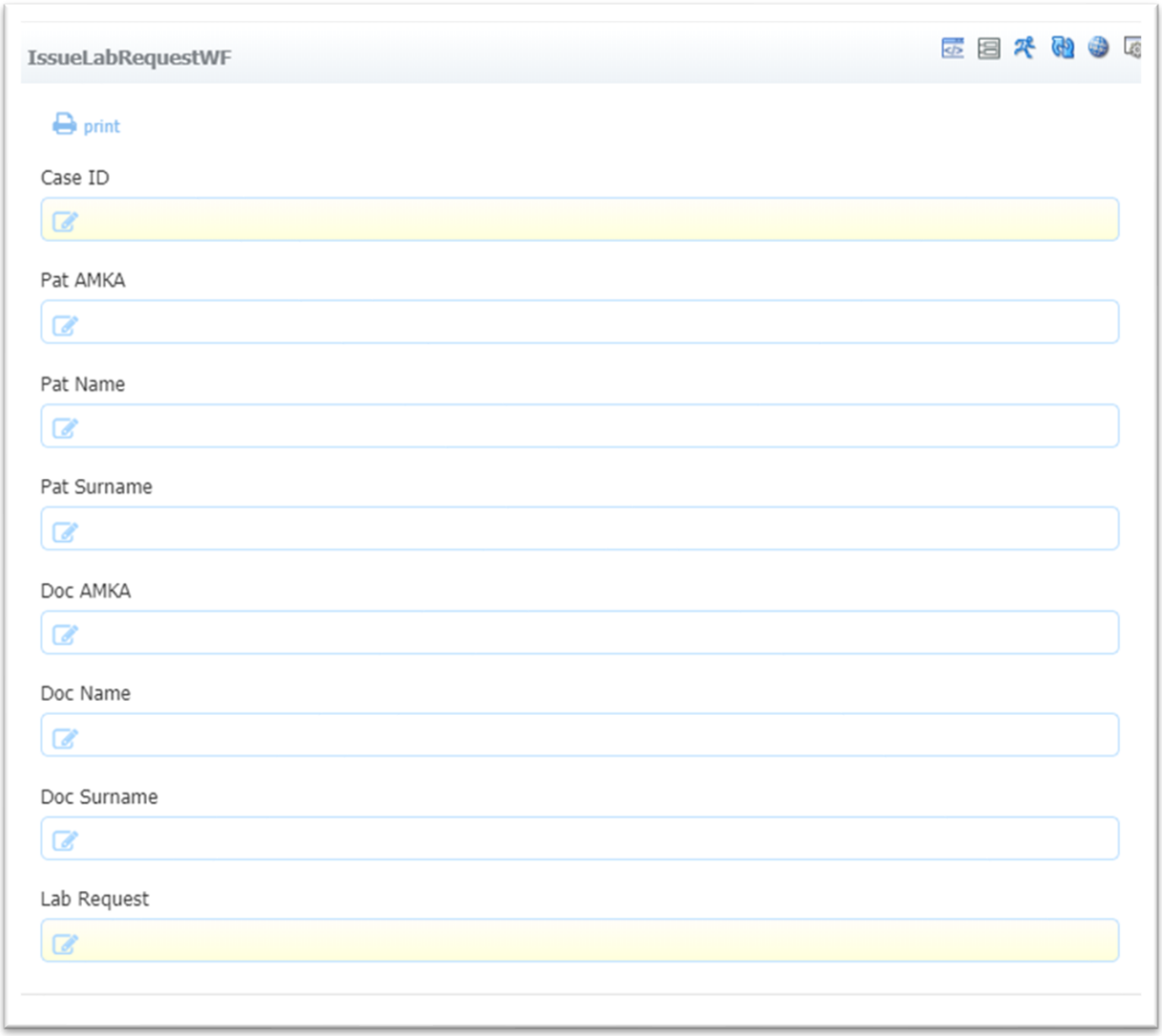
Σχήμα 2.14. Assign Patient to Doctor Web Form

Οι υπάλληλοι της υποδοχής καλούνται να αναθέσουν το περιστατικό στον κατάλληλο γιατρό.



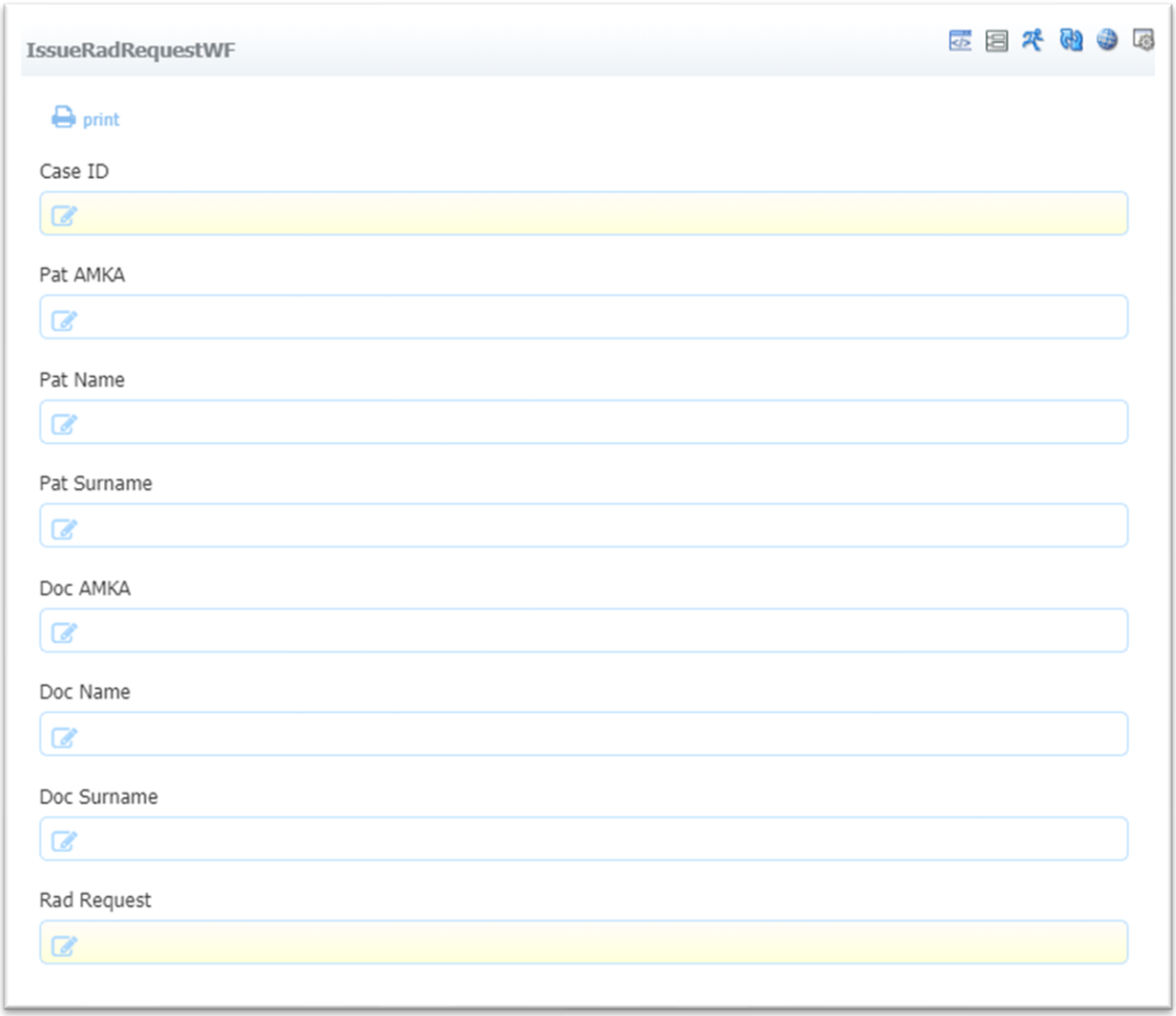
Σχήμα 2.15. Physical Examination Web Form

Το πρώτο παράθυρο που εμφανίζεται στον Γιατρό. Ο αριθμός του περιστατικού, τα στοιχεία του ασθενή και τα δικά του είναι προσυμπληρωμένα. Καλείται να συμπληρώσει τα ευρήματα της εξέτασης και να επιλέξει αν χρειάζονται Αιματολογικές ή/και ακτινολογικές εξετάσεις.



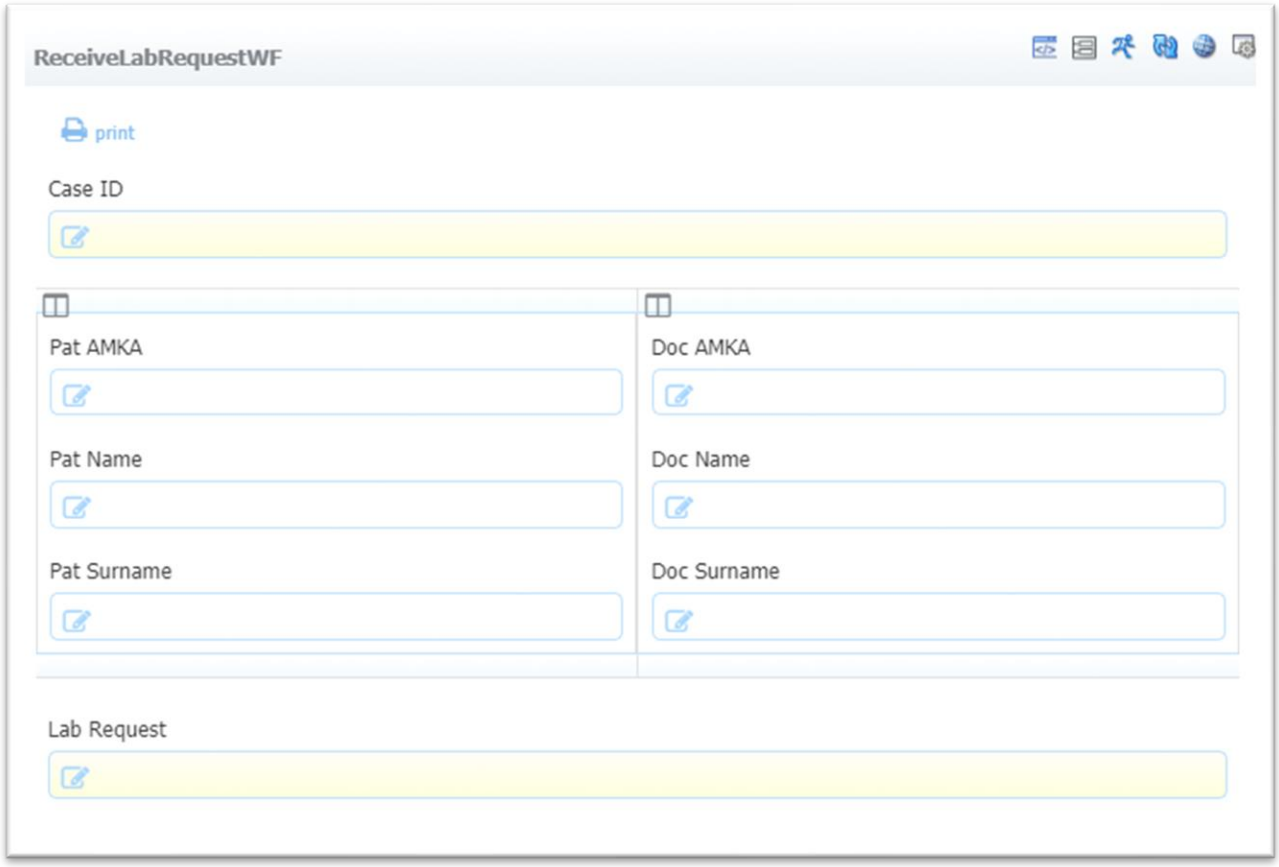
Σχήμα 2.16. Issue Lab Request Web Form

Σε αυτό το παράθυρο ο γιατρός συμπληρώνει μόνο το αίτημα για τις αιματολογικές εξετάσεις που χρειάζεται, ούτως ώστε να έχει μια σαφέστερη εικόνα του περιστατικού.



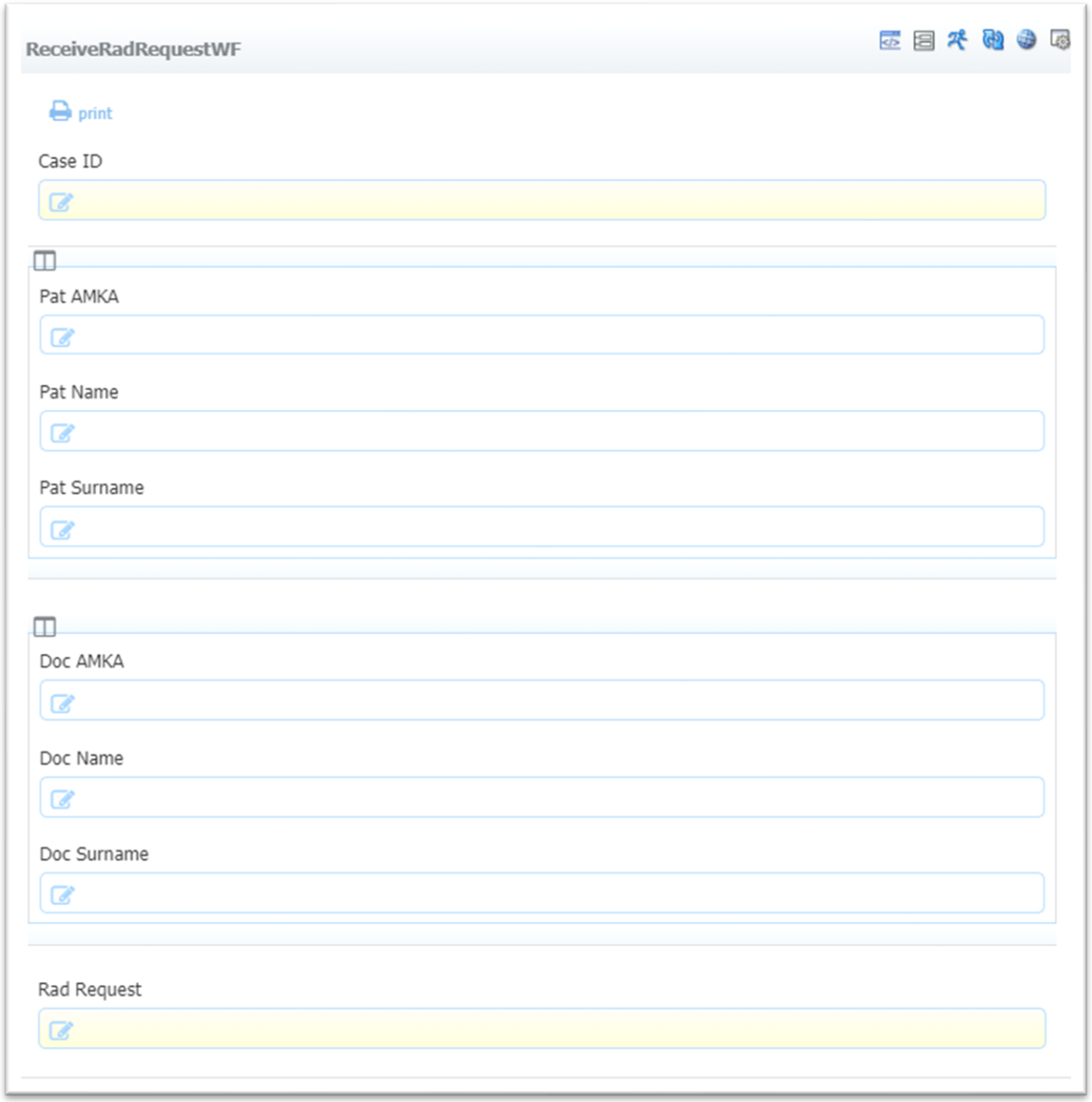
Σχήμα 2.17. Issue Rad Request Web Form

Σε αυτό το παράθυρο ο γιατρός συμπληρώνει μόνο το αίτημα για τις ακτινολογικές εξετάσεις που χρειάζεται, ούτως ώστε να έχει μια σαφέστερη εικόνα του περιστατικού.



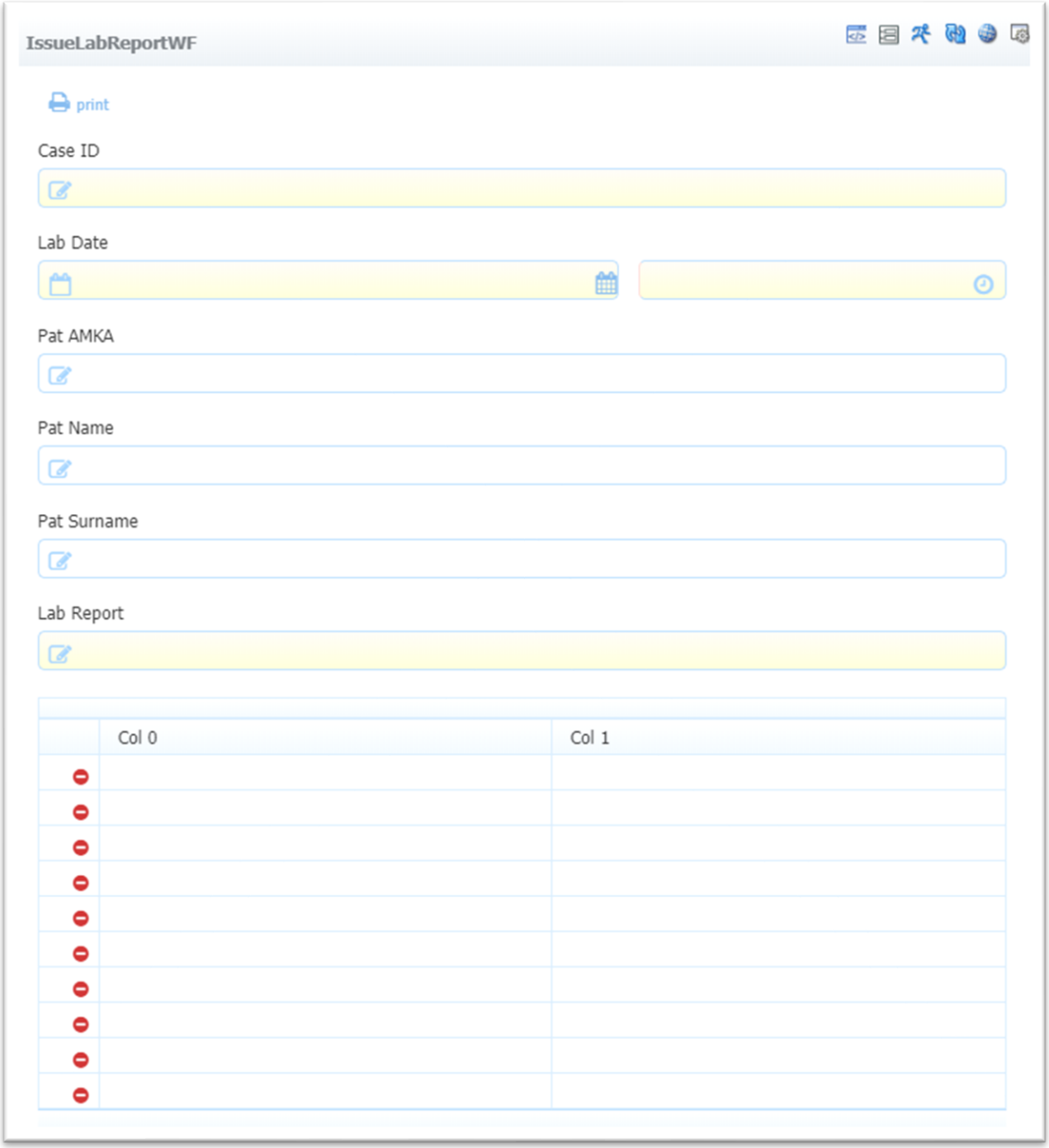
Σχήμα 2.18. Receive Lab Request Web Form

Αυτό το παράθυρο εμφανίζεται στον αιματολόγο, όπου εκτός από τον αριθμό του περιστατικού, τα στοιχεία του ασθενή και του γιατρού του, αναφέρονται και οι εξετάσεις που πρέπει να γίνουν στον ασθενή.



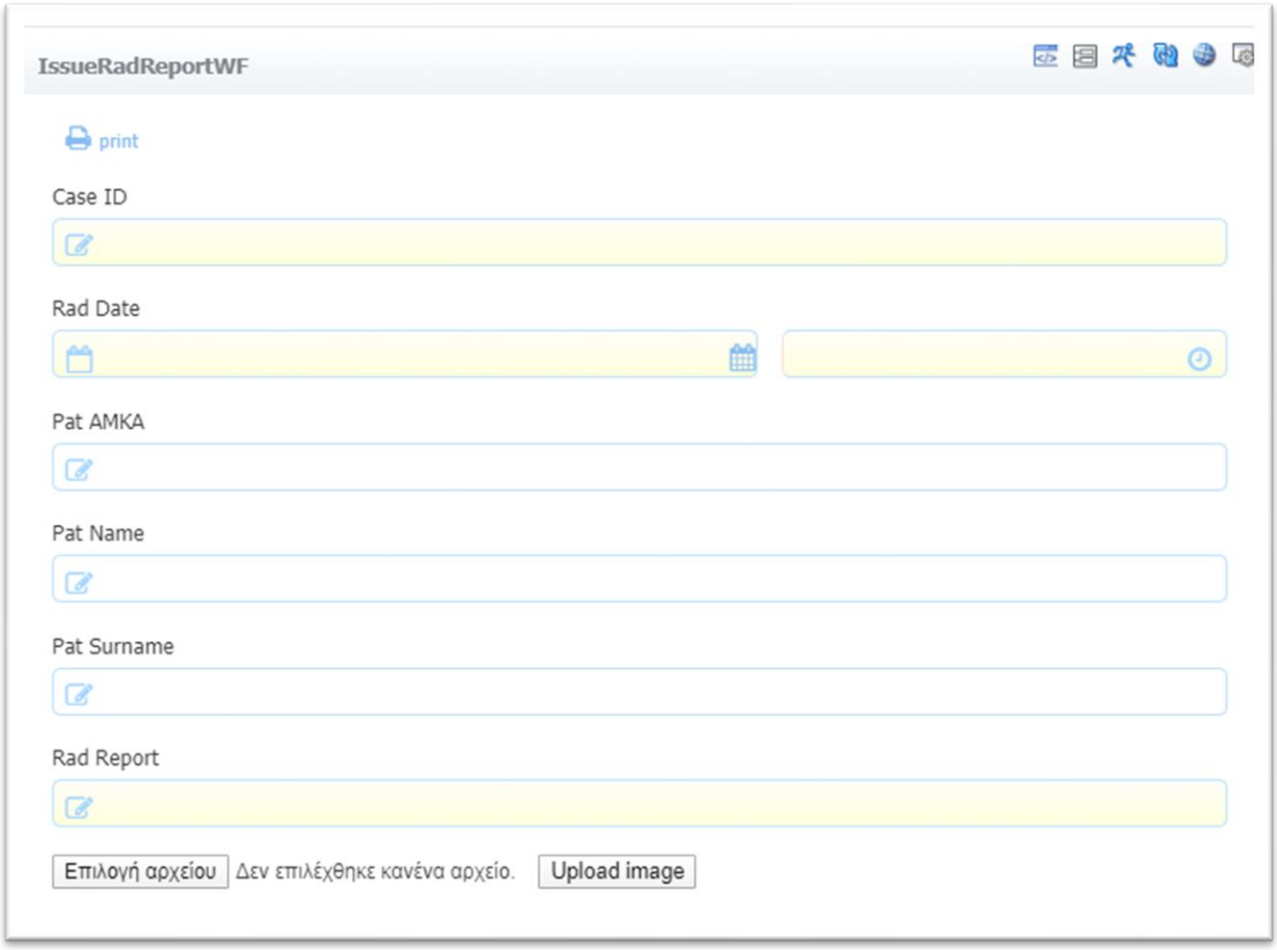
Σχήμα 2.19. Receive Rad Request Web Form

Αυτό το παράθυρο εμφανίζεται στον ακτινολόγο, όπου εκτός από τον αριθμό του περιστατικού, τα στοιχεία του ασθενή και του γιατρού του, αναφέρονται και οι εξετάσεις που πρέπει να γίνουν στον ασθενή.



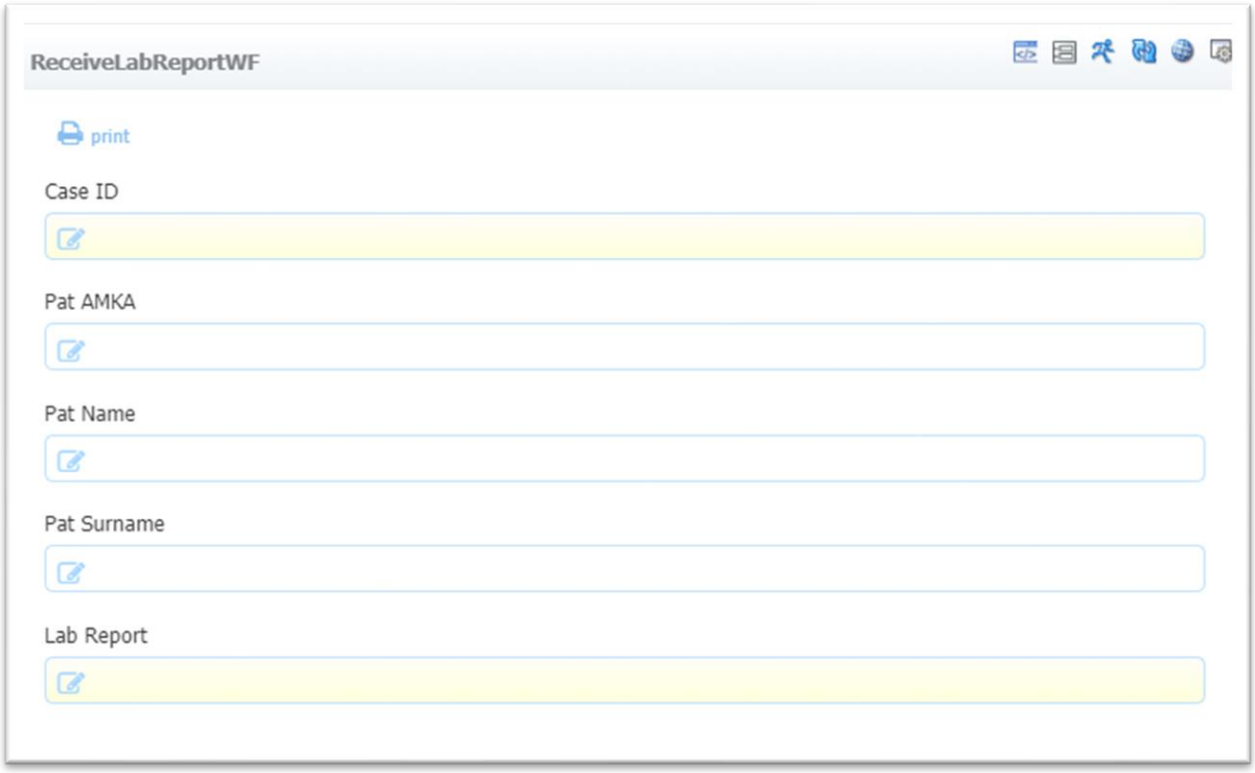
Σχήμα 2.20. Issue Lab Report Web Form

Όταν βγουν τα αποτελέσματα των εξετάσεων, ο αιματολόγος καλείται να συμπληρώσει μέσω του παραπάνω παραθύρου την ημερομηνία και την ώρα, και τις τιμές των εξετάσεων με τα όποια σχόλιά του

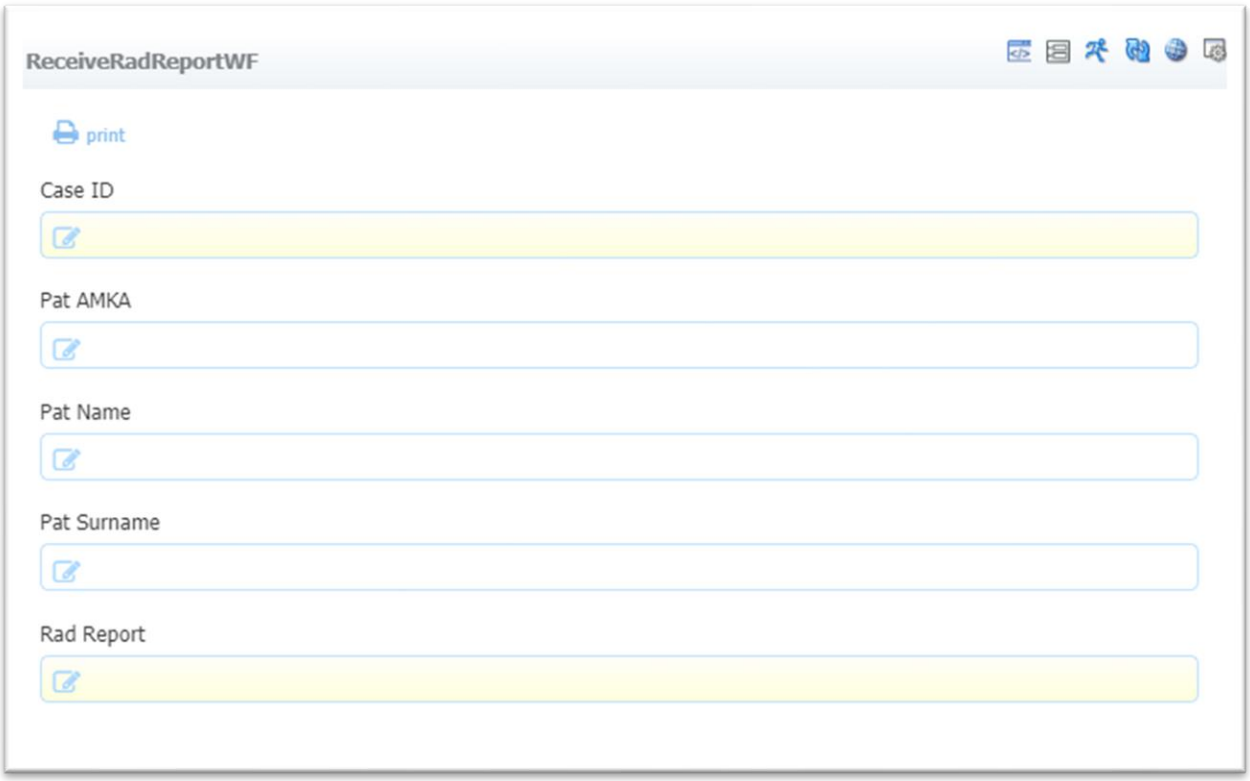


Σχήμα 2.21. Issue Rad Report Web Form

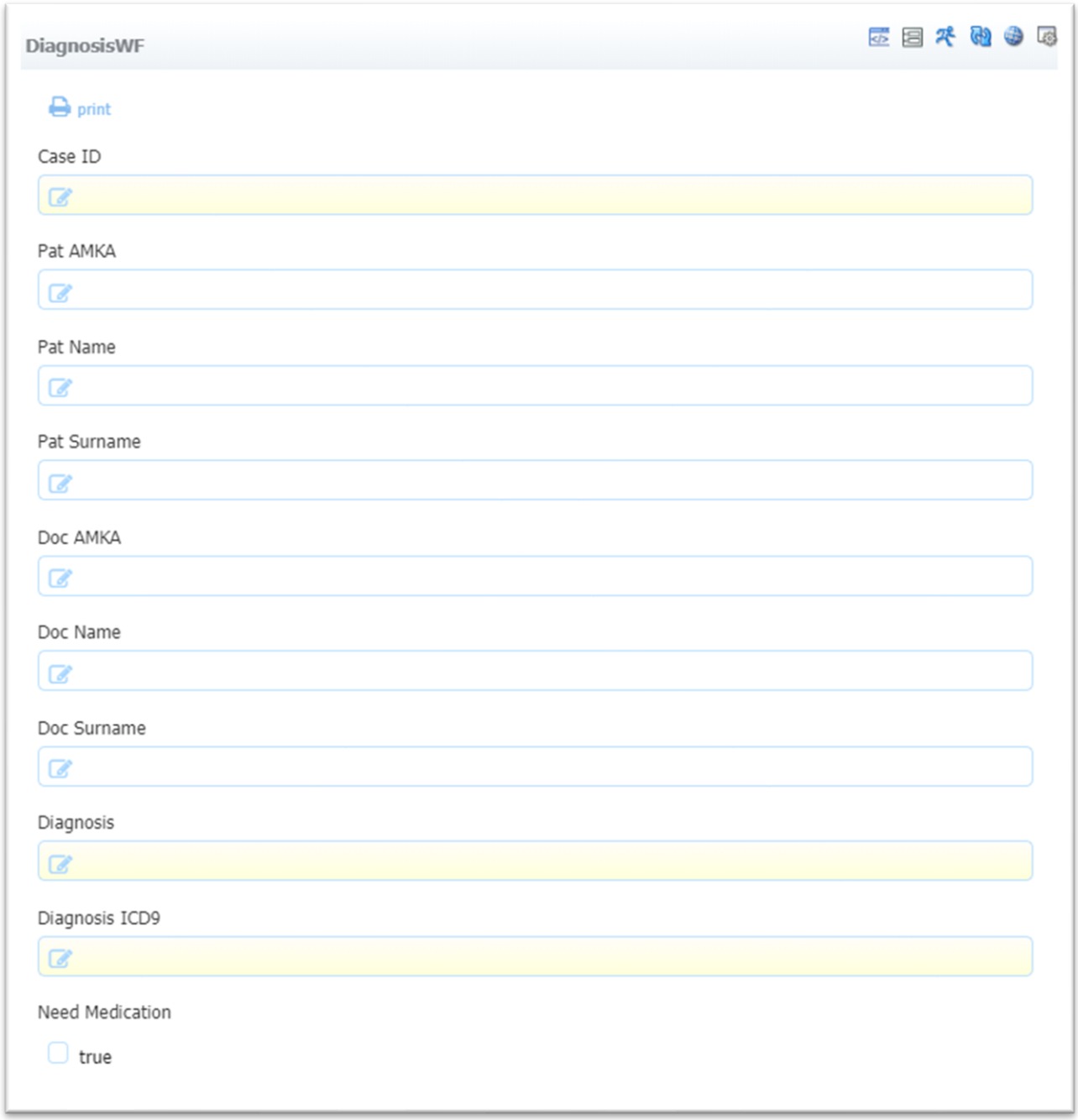
Όταν βγουν τα αποτελέσματα των εξετάσεων, ο ακτινολόγος καλείται συμπληρώσει την ημερομηνία και την ώρα, και να αναφέρει τα όποια σχόλιά του σχετικά με την εικόνα του ασθενή και να επισυνάψει μέσω του παραπάνω παραθύρου τις σχετικές εικόνες.



Σχήμα 2.22. Receive Lab Report Web Form
Η αναφορά του αιματολόγου εμφανίζεται στον γιατρό.

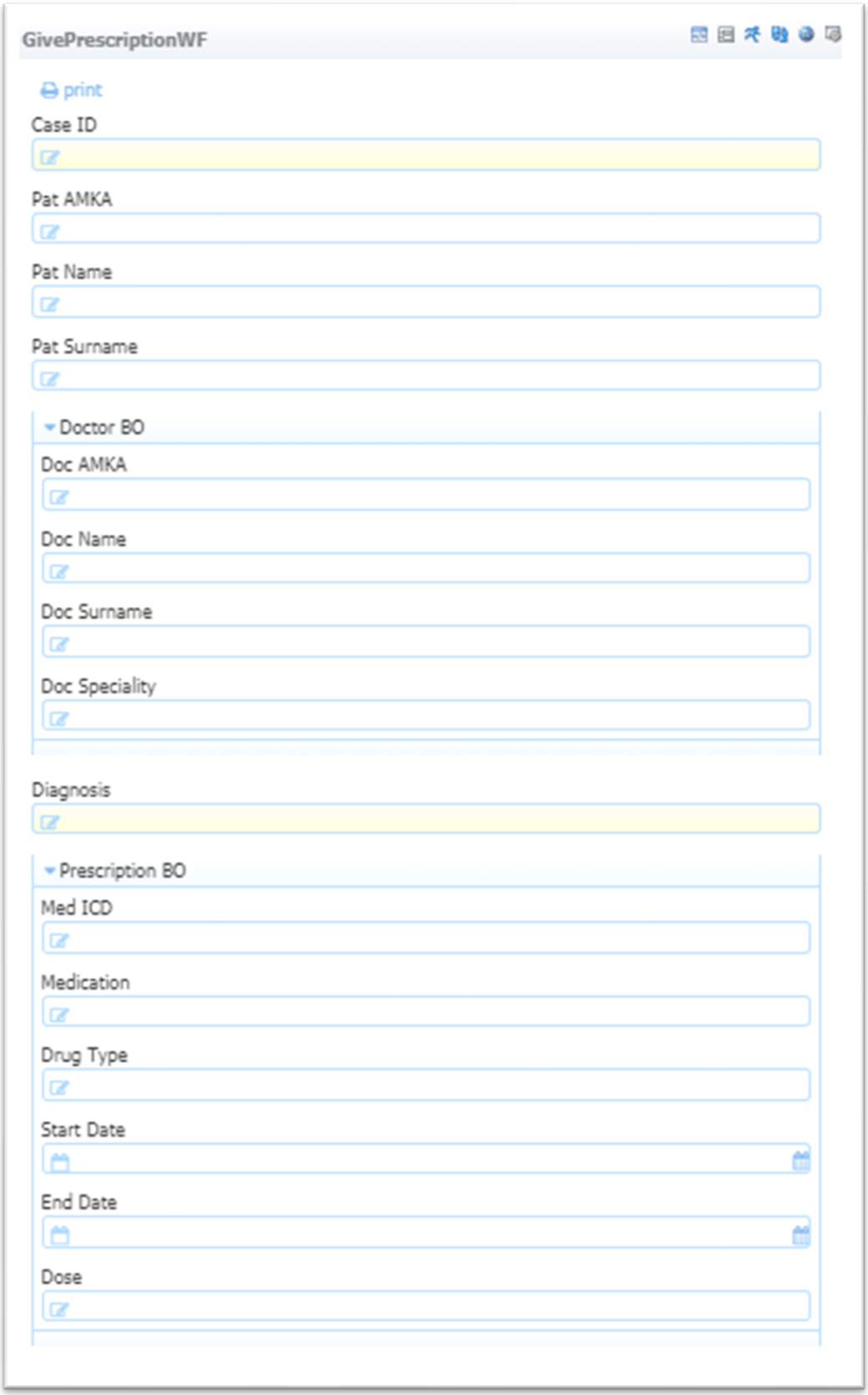


Σχήμα 2.23. Receive Rad Report Web Form
Η αναφορά του ακτινολόγου εμφανίζεται στον γιατρό.



Σχήμα 2.24. Diagnosis Web Form

Σε αυτό το παράθυρο ο γιατρός συντάσσει την διάγνωση του για το περιστατικό και επιλέγει αν χρειάζεται φαρμακευτική θεραπεία.



Σχήμα 2.25. Give Prescription Web Form

Σε αυτό το παράθυρο ο γιατρός περιγράφει την φαρμακευτική θεραπεία που θα χρειαστεί να λάβει ο ασθενής, συμπληρώνοντας τα απαραίτητα πεδία.

3. Βάση δεδομένων MIMIC III

Για την διαδικασία της ανάλυσης έγινε χρήση της βάσης δεδομένων MIMIC-III (**M**edical **I**nformation **M**art for **I**ntensive **C**are III), η οποία είναι μια μεγάλη σχεσιακή βάση ιατρικών δεδομένων, με ελεύθερη πρόσβαση σε μη ταυτοποιήσιμα με τα φυσικά πρόσωπα δεδομένα, με περισσότερους από 40000 ασθενείς. Οι ασθενείς αυτοί νοσηλεύθηκαν στις μονάδες εντατικής θεραπείας του Beth Israel Deaconess Medical Center ανάμεσα στο έτος 2001 και το έτος 2012. [5], [6]

Η βάση δεδομένων περιλαμβάνει πληροφορίες, όπως δημογραφικά στοιχεία, ζωτικές μετρήσεις που έγιναν στο κρεβάτι του ασθενή ανά μια ώρα περίπου, αποτελέσματα εργαστηριακών εξετάσεων, δεδομένα εσωτερικών διαδικασιών, φαρμακευτική αγωγή, σημειώσεις θεραπόντων ιατρών, καθώς και δεδομένα θνησιμότητας εντός και εκτός νοσοκομείου.

Η MIMIC υποστηρίζει ένα ευρύ φάσμα αναλυτικών σπουδών σχετικά με την επιδημιολογία, την βελτίωση των διεργασιών σε ιατρικές μονάδες και την ανάπτυξη ηλεκτρονικών εργαλείων. Είναι διακεκριμένη για τρεις παράγοντες:

- Είναι διαθέσιμη, ελεύθερα, σε ερευνητές παγκοσμίως.
- Συμπεριλαμβάνει έναν ποικιλόμορφο και πολύ μεγάλο αριθμό ασθενών που νοσηλεύθηκαν σε Μονάδα Εντατικής Θεραπείας.
- Περιέχει δεδομένα υψηλής χρονικής ανάλυσης συμπεριλαμβανομένων εργαστηριακών τιμών, ηλεκτρονικής τεκμηρίωσης, και δεδομένα από τις τάσεις και τις κυματομορφές στην οθόνη του κρεβατιού του ασθενή.

Η βάση δεδομένων MIMIC αποτελείται από 26 πίνακες.
Οι ακόλουθοι πίνακες χρησιμοποιούνται για τον ορισμό και την παρακολούθηση παραμονής ασθενών:

- **ADMISSIONS**
  Οι μοναδικές νοσηλείες για κάθε ασθενή στη βάση δεδομένων (ορίζεται από το HADM_ID) Οι διαθέσιμες πληροφορίες περιλαμβάνουν χρονικά δεδομένα για την εισαγωγή και το εξιτήριο, δημογραφικές πληροφορίες, πηγή εισόδου και ούτω καθεξής.
- **CALLOUT**
  Πληροφορίες σχετικά με το πότε ένας ασθενής μεταφέρθηκε από τη Μονάδα Εντατικής Θεραπείας και πότε ο ασθενής πήρε εξιτήριο.
- **ICUSTAYS**
  Κάθε μοναδική παραμονή στην ΜΕΘ που βρίσκεται στη βάση δεδομένων (καθορίζει το ICUSTAY_ID).
- **PATIENTS**
  Κάθε μοναδικός ασθενής στη βάση δεδομένων (καθορίζει το SUBJECT_ID).
- **SERVICES**
  Οι κλινικές υπηρεσίες τις οποίες έλαβε ένας ασθενής.

- **TRANSFERS**
  Η μετακίνηση των ασθενών από κρεβάτι σε κρεβάτι εντός του νοσοκομείου, συμπεριλαμβανομένης της μεταφοράς από και προς την ΜΕΘ.

Οι παρακάτω πίνακες περιέχουν δεδομένα που συλλέχθηκαν από την μονάδα αυξημένης φροντίδας (critical care unit):

- **CAREGIVERS**
  Κάθε μέλος του νοσηλευτικού προσωπικού (καθορίζει το CGID).

- **CHARTEVENTS**
  Περιέχει όλα τα δεδομένα του ηλεκτρονικού πίνακα που είναι διαθέσιμα για έναν ασθενή κατά τη διάρκεια της παραμονής τους στη ΜΕΘ. Ο ηλεκτρονικός πίνακας εμφανίζει τα ζωτικά σημάδια ρουτίνας των ασθενών και κάθε πρόσθετη πληροφορία σχετική με τη φροντίδα τους: ρυθμίσεις του αναπνευστήρα, εργαστηριακές τιμές, κατάσταση κωδικού, νοητική κατάσταση κ.λπ. Ως αποτέλεσμα, ο όγκος των πληροφοριών σχετικά με τη διαμονή ενός ασθενούς περιέχεται στα CHARTEVENTS. Επιπλέον, παρόλο που οι εργαστηριακές τιμές αναφέρονται αλλού (LABEVENTS), επαναλαμβάνονται συχνά στα CHARTEVENTS. Αυτό συμβαίνει επειδή είναι επιθυμητό να εμφανίζονται οι εργαστηριακές τιμές στο ηλεκτρονικό διάγραμμα του ασθενούς και έτσι οι τιμές αντιγράφονται από την βάση δεδομένων αποθήκευσης εργαστηριακών τιμών στη βάση δεδομένων που αποθηκεύει τα CHARTEVENTS.

- **DATETIMEEVENTS**
  Όλες οι καταγεγραμμένες παρατηρήσεις που είναι ημερομηνίες, π.χ. χρόνος αιμοκάθαρσης. Καθώς όλες οι ημερομηνίες στο MIMIC-III είναι ανώνυμες για την προστασία του απορρήτου του ασθενή, όλες οι ημερομηνίες σε αυτόν τον πίνακα έχουν μεταβληθεί. Παρ' όλα αυτά, σημειώνεται ότι η χρονική αναδρομή για έναν μεμονωμένο ασθενή δεν έχει επηρεαστεί και ποσότητες όπως η διαφορά μεταξύ δύο ημερομηνιών παραμένουν αληθείς στην πραγματικότητα.

- **INPUTEVENTS_CV**
  Η πρόσληψη για ασθενείς που παρακολουθούνται χρησιμοποιώντας το σύστημα Philips CareVue ενώ βρίσκονται στη ΜΕΘ.

- **INPUTEVENTS_MV**
  Η πρόσληψη για ασθενείς που παρακολουθούνται χρησιμοποιώντας το σύστημα iMDSoft Metavision ενώ βρίσκεται στη ΜΕΘ.

- **NOTEEVENTS**
  Μη ταυτοποιήσιμες σημειώσεις, συμπεριλαμβανομένων σημειώσεων νοσηλευτών και γιατρών, εκθέσεων ECG, εκθέσεων απεικόνισης και περιλήψεων απαλλαγής.

- **OUTPUTEVENTS**
  Πληροφορίες παραγωγής για τους ασθενείς ενώ βρίσκονται στη ΜΕΘ.

- **PROCEDUREEVENTS_MV**
  Διαδικασίες ασθενών για το υποσύνολο αυτών που νοσηλεύθηκαν στη ΜΕΘ χρησιμοποιώντας το σύστημα iMDSoft MetaVision.

Οι παρακάτω πίνακες περιέχουν δεδομένα που συλλέχθηκαν στο σύστημα αρχείων του νοσοκομείου:

- **CPTEVENTS**
  Διαδικασίες που καταγράφονται ως κωδικοί τρέχουσας διαδικαστικής ορολογίας Current Procedural Terminology (CPT).

- **DIAGNOSES_ICD**
  Οι νοσοκομειακές διαγνώσεις, κωδικοποιημένες με τη χρήση του συστήματος διεθνούς στατιστικής ταξινόμησης ασθενειών και συναφών προβλημάτων υγείας (International Statistical Classification of Diseases and Related Health Problems-ICD).

- **DRGCODES**
  Συγγενείς ομάδες διάγνωσης (Diagnosis Related Groups-DRG), οι οποίες χρησιμοποιούνται από το νοσοκομείο για σκοπούς χρέωσης.

- **LABEVENTS**
  Εργαστηριακές μετρήσεις για ασθενείς τόσο στο νοσοκομείο όσο και στις εξωτερικές κλινικές ασθενών.

- **MICROBIOLOGYEVENTS**
  Μικροβιολογικές μετρήσεις και ευαισθησίες από τη βάση δεδομένων νοσοκομείων.

- **PRESCRIPTIONS**
  Φάρμακα που έχουν παραγγελθεί, και όχι απαραίτητα χορηγούμενα, για έναν συγκεκριμένο ασθενή.

- **PROCEDURES_ICD**
  Επεμβάσεις ασθενών, που κωδικοποιούνται με τη χρήση του συστήματος διεθνούς στατιστικής ταξινόμησης ασθενειών και συναφών προβλημάτων υγείας (International Statistical Classification of Diseases and Related Health Problems-ICD).

Οι παρακάτω πίνακες λειτουργούν σαν λεξικά:

- **D_CPT**
  Λεξικό υψηλού επιπέδου των κωδικών της τρέχουσας διαδικαστικής ορολογίας (Current Procedural Terminology-CPT).

- **D_ICD_DIAGNOSES**
  Λεξικό του συστήματος διεθνούς στατιστικής ταξινόμησης ασθενειών και συναφών προβλημάτων υγείας (International Statistical Classification of Diseases and Related Health Problems-ICD) codes σχετιζόμενο με τις διαγνώσεις.

- **D_ICD_PROCEDURES**
  Λεξικό του συστήματος διεθνούς στατιστικής ταξινόμησης ασθενειών και συναφών προβλημάτων υγείας (International Statistical Classification of Diseases and Related Health Problems-ICD) codes σχετιζόμενο με τις επεμβάσεις.

- **D_ITEMS**
  Λεξικό των ITEMID που εμφανίζονται στην βάση δεδομένων MIMIC database, εκτός από αυτά που σχετίζονται με τις εργαστηριακές εξετάσεις.

- **D_LABITEMS:** Λεξικό των ITEMID που σχετίζονται με τις εργαστηριακές εξετάσεις.

Συμπληρωματικοί πίνακες

Η βάση δεδομένων MIMIC-II περιλαμβάνει μια ποικιλία συμπληρωματικών πινάκων που απλοποιούν τη χρήση της βάσης δεδομένων. Για παράδειγμα, ένας κοινώς χρησιμοποιούμενος πίνακας ήταν ο πίνακας ICUSTAY_DETAIL, ο οποίος παρέχει πρόσθετες πληροφορίες που συνοψίζουν τη διαμονή τη ΜΕΘ ενός ασθενή. http://github.com/MIT-lcp/mimic-code

3.1. Επιλογή δεδομένων για ανάλυση

Από τη βάση δεδομένων MIMIC III έγινε χρήση μόνο των δεδομένων που θα μπορούσαν να συσχετιστούν με μια Πνευμονολογική Κλινική. Οι πίνακες που χρησιμοποιήθηκαν είναι οι: ADMISSIONS, PATIENTS, CAREGIVERS, LABEVENTS, D_LABEVENTS, DIAGNOSES_ICD και D_ICD_DIAGNOSES. Αυτοί οι πίνακες θα παρουσιαστούν και θα αναλυθούν παρακάτω.

- Οι στήλες του πίνακα ADMISSIONS είναι οι εξής:

Name	**Postgres data type**
ROW_ID	INT
SUBJECT_ID	INT
HADM_ID	INT
ADMITTIME	TIMESTAMP(0)
DISCHTIME	TIMESTAMP(0)
DEATHTIME	TIMESTAMP(0)
ADMISSION_TYPE	VARCHAR(50)
ADMISSION_LOCATION	VARCHAR(50)
DISCHARGE_LOCATION	VARCHAR(50)
INSURANCE	VARCHAR(255)
LANGUAGE	VARCHAR(10)
RELIGION	VARCHAR(50)
MARITAL_STATUS	VARCHAR(50)
ETHNICITY	VARCHAR(200)
EDREGTIME	TIMESTAMP(0)
EDOUTTIME	TIMESTAMP(0)
DIAGNOSIS	VARCHAR(300)
HOSPITAL_EXPIRE_FLAG	TINYINT
HAS_CHARTEVENTS_DATA	TINYINT

Όπου,

SUBJECT_ID, HADM_ID: Κάθε γραμμή του πίνακα περιέχει ένα μοναδικό HADM_ID, το οποίο αντιπροσωπεύει μια μοναδική εισαγωγή του ασθενή στο νοσοκομείο. Το εύρος του HADM_ID είναι 1000000 - 1999999. Είναι πιθανό στον πίνακα αυτόν να εμφανιστεί περισσότερες από μια φορά το

ίδιο SUBJECT_ID, που αντιστοιχεί σε κάθε ένα μοναδικό ασθενή, καθώς ενδέχεται να έχει εισαχθεί πολλές φορές στο νοσοκομείο. Ο πίνακας ADMISSIONS συνδέεται μέσω του SUBJECT_ID με τον πίνακα PATIENTS.

ADMITTIME, DISCHTIME, DEATHTIME: Το ADMITTIME παρέχει την ημερομηνία και την ώρα που ο ασθενής εισήχθη στο νοσοκομείο, ενώ το DISCHTIME παρέχει την ημερομηνία και την ώρα που πήρε εξιτήριο. Εάν παρέχεται, το DEATHTIME αντιστοιχεί την ημερομηνία και την ώρα του θανάτου του ασθενή μέσα στο νοσοκομείο και σημειώνεται ότι είναι συμπληρωμένο μόνο τότε, και είναι σχεδόν πάντα ίδιο με το DISCHTIME.

ADMISSION_TYPE: Περιγράφει το είδος της εισαγωγής: "ELECTIVE", "URGENT", "NEWBORN" ή "EMERGENCY". Emergency/urgent υποδεικνύει μη προγραμματισμένη ιατρική φροντίδα. Το Elective αντιστοιχεί σε προγραμματισμένη εισαγωγή και το Newborn υποδεικνύει ότι το περιστατικό είναι γέννα.

ADMISSION_LOCATION: Παρέχει πληροφορίες για την τοποθεσία του ασθενή προτού φτάσει στο νοσοκομείο. Οι πιθανές τιμές του είναι οι παρακάτω 9:EMERGENCY ROOM ADMIT, TRANSFER FROM HOSP/EXTRAM, TRANSFER FROM OTHER HEALT, CLINIC REFERRAL/PREMATURE, ** INFO NOT AVAILABLE **, TRANSFER FROM SKILLED NUR, TRSF WITHIN THIS FACILITY, HMO REFERRAL/SICK, PHYS REFERRAL/NORMAL DELI.

INSURANCE, LANGUAGE, RELIGION, MARITAL_STATUS, ETHNICITY: Οι αυτές περιγράφουν τα δημογραφικά στοιχεία του ασθενή.

EDREGTIME, EDOUTTIME: Η ώρα που έγινε η εγγραφή του ασθενή και η έξοδός του στα επείγοντα περιστατικά

DIAGNOSIS: Η στήλη παρέχει μια προκαταρκτική περιγραφή, ελευθέρου κειμένου κατά τη διάρκεια της εισαγωγής του ασθενή. Η περιγραφή αυτή συνήθως συμπληρώνεται από τον κλινικό ιατρό και δεν χρησιμοποιεί συστηματική οντολογία και μπορεί να είναι αρκετά αναλυτική όπως chronic kidney failure (χρόνια νεφρική ανεπάρκεια), μέχρι ιδιαίτερα ασαφής όπως weakness(αδυναμία). Η τελική διάγνωση για κάθε περιστατικό είναι κωδικοποιημένη στην στήλη discharge του πίνακα DIAGNOSES_ICD.

HOSPITAL_EXPIRE_FLAG: Υποδεικνύει αν ο ασθενής απεβίωσε κατά τη διάρκεια της συγκεκριμένης νοσηλείας. Το "1" αντιστοιχεί σε θάνατο μέσα στο νοσοκομείο και το "0" σε επιβίωση και εξιτήριο.

- Οι στήλες του πίνακα PATIENTS είναι:

Name	Postgres data type
ROW_ID	INT
SUBJECT_ID	INT
GENDER	VARCHAR(5)
DOB	TIMESTAMP(0)
DOD	TIMESTAMP(0)
DOD_HOSP	TIMESTAMP(0)
DOD_SSN	TIMESTAMP(0)
EXPIRE_FLAG	VARCHAR(5)

Όπου,

SUBJECT_ID: Ένα μοναδικό αναγνωριστικό που προσδιορίζει έναν συγκεκριμένο ασθενή και αποτελεί υποψήφιο κλειδί για τον πίνακα, οπότε είναι μοναδικό για κάθε σειρά. Οι πληροφορίες που είναι συνεπείς για τη διάρκεια ζωής ενός ασθενούς αποθηκεύονται σε αυτόν τον πίνακα.

GENDER: Το φύλο είναι το γονοτυπικό φύλο του ασθενούς.

DOB: Η ημερομηνία γέννησης του συγκεκριμένου ασθενούς. Οι ασθενείς ηλικίας άνω των 89 ετών ανά πάσα στιγμή στη βάση δεδομένων έχουν μετατοπιστεί η ημερομηνία γέννησής της για να αποκρύψουν την ηλικία της και να συμμορφωθούν με την HIPAA. Η διαδικασία της μετατόπισης ήταν η εξής: καθορίστηκε η ηλικία του ασθενούς κατά την πρώτη εισαγωγή. Η ημερομηνία γέννησης ορίστηκε τότε ακριβώς 300 χρόνια πριν από την πρώτη εισαγωγή.

DOD, DOD_HOSP, DOD_SSN: Το DOD είναι η ημερομηνία θανάτου για τον συγκεκριμένο ασθενή. Το DOD_HOSP είναι η ημερομηνία θανάτου όπως καταγράφεται στη βάση δεδομένων του νοσοκομείου. Το DOD_SSN είναι η ημερομηνία θανάτου από τη βάση δεδομένων κοινωνικής ασφάλισης. Σημειώνεται ότι το DOD συγχωνεύτηκε μαζί τα DOD_HOSP και DOD_SSN, δίνοντας προτεραιότητα στο DOD_HOSP εάν και οι δύο ήταν κατεστραμμένα.

EXPIRE_FLAG: Μια δυαδική σημαία που υποδεικνύει εάν ο ασθενής απεβίωσε, δηλαδή εάν το DOD είναι άκυρο ή όχι. Αυτοί οι θάνατοι περιλαμβάνουν τόσο τους θανάτους στο νοσοκομείο (DOD_HOSP) όσο και τους θανάτους που εντοπίστηκαν συνδυάζοντας τον ασθενή με τον κύριο δείκτη θανάτου της κοινωνικής ασφάλισης (DOD_SSN).

- Οι στήλες του πίνακα CAREGIVERS είναι:

Name	Postgres data type
ROW_ID	INT
CGID	INT
LABEL	VARCHAR(15)
DESCRIPTION	VARCHAR(30)

Όπου,

CGID: Είναι ένα μοναδικό αναγνωριστικό για κάθε ξεχωριστό φροντιστή που υπάρχει στη βάση δεδομένων.

LABEL: Ορίζει τον τύπο του φροντιστή: π.χ. RN, MD, PharmD κ.λπ. Σημειώνεται ότι το LABEL είναι ένα πεδίο ελεύθερου κειμένου και ως τέτοιο περιέχει πολλά τυπογραφικά λάθη και ορθολογικές παραλλαγές της ίδιας έννοιας (π.χ. MD, MDs, M.D.).

DESCRIPTION: Παρουσιάζεται λιγότερο συχνά από το LABEL και παρέχει επιπλέον πληροφορίες σχετικά με τον φροντιστή. Αυτή η στήλη είναι πολύ πιο δομημένη και περιέχει μόνο 17 μοναδικές τιμές από το MIMIC-III v1.0.

- Οι στήλες του πίνακα LABEVENTS είναι:

Name	**Postgres data type**
ROW_ID	INT
SUBJECT_ID	INT
HADM_ID	INT
ITEMID	INT
CHARTTIME	TIMESTAMP(0)
VALUE	VARCHAR(200)
VALUENUM	DOUBLE PRECISION
VALUEUOM	VARCHAR(20)
FLAG	VARCHAR(20)

Όπου

SUBJECT_ID, HADM_ID: Το SUBJECT_ID είναι μοναδικό για τον ασθενή, το HADM_ID είναι μοναδικό για κάθε διαμονή ασθενούς στο νοσοκομείο.

ITEMID: Αναγνωριστικό για έναν τύπο μέτρησης στη βάση δεδομένων. Κάθε σειρά που σχετίζεται με ένα στοιχείο ITEMID (π.χ. 212) αντιστοιχεί σε μια τιμή σε κάποια χρονική στιγμή της ίδιας μέτρησης (π.χ. καρδιακό ρυθμό).

CHARTTIME: Καταγράφει τον χρόνο κατά τον οποίο έγινε μια παρατήρηση και συνήθως είναι το πλησιέστερη αντιστοίχιση με το χρόνο που πραγματικά μετρήθηκαν τα δεδομένα.

VALUE, VALUENUM: Η τιμή VALUE περιέχει την τιμή που μετρήθηκε για την έννοια που προσδιορίζεται από το ITEMID. Αν αυτή η τιμή είναι αριθμητική, τότε το VALUENUM περιέχει τα ίδια δεδομένα σε αριθμητική μορφή. Εάν τα δεδομένα αυτά δεν είναι αριθμητικά, το VALUENUM είναι μηδενικό.

VALUEUOM: Η μονάδα μέτρησης για το VALUE, εάν χρειάζεται.

το **FLAG** υποδηλώνει εάν η εργαστηριακή τιμή θεωρείται μέσα στο φυσιολογικό εύρος ή όχι, με βάση τα προκαθορισμένα όρια. Οι υπόλοιπες στήλες έχουν αναλυθεί και στους παραπάνω πίνακες.

- Οι στήλες του πίνακα D_LABEVENTS είναι:

Name	Postgres data type
ROW_ID	INT
ITEMID	INT
LABEL	VARCHAR(100)
FLUID	VARCHAR(100)
CATEGORY	VARCHAR(100)
LOINC_CODE	VARCHAR(100)

Όπου,
ITEMID: Υποψήφιο κλειδί στον πίνακα, το ITEMID είναι μοναδικό σε κάθε σειρά.

LABEL: Περιγράφει την έννοια που αντιπροσωπεύει το ITEMID.

FLUID: Η ουσία στην οποία έγινε η μέτρηση. Για παράδειγμα, συχνά πραγματοποιούνται μετρήσεις χημείας στο αίμα, το οποίο αναφέρεται στη στήλη αυτή ως “BLOOD”. Πολλές από αυτές τις μετρήσεις προέρχονται από άλλα υγρά, όπως τα ούρα, και αυτή η στήλη διαφοροποιεί αυτές τις διαφορετικές έννοιες.

CATEGORY: Παρέχει πληροφορίες υψηλότερου επιπέδου σχετικά με τον τύπο μέτρησης. Για παράδειγμα, μια κατηγορία "ABG" υποδεικνύει ότι η μέτρηση είναι αρτηριακό αέριο αίματος.

LOINC_CODE: Περιέχει τον κωδικό LOINC που σχετίζεται με το δεδομένο στοιχείο ITEMID. Το LOINC είναι μια οντολογία η οποία αρχικά καθόριζε εργαστηριακές μετρήσεις αλλά έκτοτε επεκτάθηκε για να καλύψει ένα ευρύ φάσμα κλινικά σχετικών εννοιών.

- Οι στήλες του πίνακα DIAGNOSES_ICD είναι:

Name	PostgreSQL data type	Modifiers
ROW_ID	INT	not null
SUBJECT_ID	INT	not null
HADM_ID	INT	not null
SEQ_NUM	INT	
ICD9_CODE	VARCHAR(10)	

Όπου, εκτός από τα SUBJECT_ID, HADM_ID που έχουν ήδη περιγραφεί:

SEQ_NUM: Η σειρά με την οποία οι διαγνώσεις ICD σχετίζονται με τον ασθενή. Οι διαγνώσεις ICD ταξινομούνται κατά προτεραιότητα - και η σειρά έχει αντίκτυπο στην επιστροφή για θεραπεία.

ICD9_CODE: Ο πραγματικός κωδικός που αντιστοιχεί στη διάγνωση που αντιστοιχεί στον ασθενή για τη συγκεκριμένη σειρά. Σημειώνεται ότι όλοι οι κωδικοί, που συμπεριλαμβάνονται στη MIMIC-III v1.0, είναι κωδικοί ICD-9.

- Οι στήλες του πίνακα D_ICD_DIAGNOSES είναι:

Name	Postgres data type
ROW_ID	INT
ICD9_CODE	VARCHAR(10)
SHORT_TITLE	VARCHAR(50)
LONG_TITLE	VARCHAR(300)

Όπου τα **SHORT_TITLE** και **LONG_TITLE** περιγράφουν σύντομα και πιο αναλυτικά, τη διάγνωση που αντιστοιχεί στον κωδικό ICD9.

Στη συνέχεια, από τους συγκεκριμένους πίνακες θα κρατήσουμε τα στοιχεία που αφορούν ασθενείς με πρόβλημα στο αναπνευστικό σύστημα. Η διαδικασία του φιλτραρίσματος ξεκινάει από τον πίνακα DIAGNOSES_ICD, όπου εκεί συνδέονται οι διαγνώσεις με τον κωδικό ICD9_CODE με το κάθε περιστατικό, αλλά και τους ασθενείς. Επειδή οι κωδικοί ICD9_CODE είναι σε μορφή VARCHAR(10) και δε μπορούν να ταξινομηθούν σωστά, δημιουργήθηκε ένας νέος πίνακας μόνο με τις διαγνώσεις του αναπνευστικού ο D_ICD_DIAGNOSES_RESP, στηριζόμενος στον αριθμό της γραμμής του πίνακα D_ICD_DIAGNOSES που έχει τη μορφή INT.

Οι κωδικοί ICD9_CODE που αντιστοιχούν σε διαγνώσεις σχετικές με το αναπνευστικό [7] στον πίνακα D_ICD_DIAGNOSES είναι οι εξής:

Icd9 range	Diagnoses Categories	Mimic ICD9_CODE	Mimic ROW_ID
460-466	Acute Respiratory Infections	460-46619	5412-5437
470-478	Other Diseases of Upper Respiratory Tract	470-4789	5438-5493
480-487	Pneumonia And Influenza	4800-4878	5494-5531
490-496	Chronic Obstructive Pulmonary Disease and Allied Conditions	490-4920 & 4928-496	5541-5549 & 5096-5123
500-508	Pneumoconioses And Other Lung Diseases Due to External Agents	500-5089	5124-5143
510-519	Other Diseases of Respiratory System	5100-5199	5144-5294

Επομένως, οι γραμμές του πίνακα D_ICD_DIAGNOSES, που χρειαζόμαστε, είναι οι: 5096-5294, 5412-5531 και 5541-5549.

Εργαζόμενοι στο SQL shell έχουμε τις εξής εντολές:

```
CREATE TABLE d_icd_diagnoses_resp AS TABLE d_icd_diagnoses;
        SELECT 12167
DELETE FROM d_icd_diagnoses_resp WHERE row_id < 5096;
        DELETE 5095
DELETE FROM d_icd_diagnoses_resp WHERE row_id > 5294 AND row_id < 5412;
        DELETE 117
DELETE FROM d_icd_diagnoses_resp WHERE row_id > 5531 AND row_id < 5541;
        DELETE 9
DELETE FROM d_icd_diagnoses_resp WHERE row_id > 5549;
        DELETE 6618
SELECT COUNT (*) FROM d_icd_diagnoses_resp;
```

```
count
-------
  328
(1 row)
```

Στη συνέχεια με βάση την επιλογή που έγινε, δημιουργήθηκε ένας νέος πίνακας, ο DIAGNOSES_ICD_RESP, μόνο με τα στοιχεία που μας αφορούν, δηλαδή 141 περιστατικά.

```
CREATE TABLE diagnoses_icd_resp AS TABLE diagnoses_icd;
        SELECT 1761
DELETE FROM diagnoses_icd_resp WHERE NOT EXISTS(SELECT NULL FROM d_icd_diagnoses_resp
WHERE diagnoses_icd_resp.icd9_code=d_icd_diagnoses_resp.icd9_code);
        DELETE 1620
SELECT COUNT(*) FROM diagnoses_icd_resp;
         count
-------
  141
(1 row)
```

Με τον ίδιο τρόπο θα κρατήσουμε μόνο τα στοιχεία που χρειάζονται από τον πίνακα LABEVENTS αντιστοιχίζοντας το hadm_id για να βρούμε τις εργαστηριακές μετρήσεις για τα συγκεκριμένα περιστατικά και ομοίως στον πίνακα ADMISSIONS. Ομοίως κάνοντας χρήση του κλειδιού subject_id θα διατηρήσουμε μόνο τους ασθενείς του πίνακα PATIENTS τους οποίους αφορούσαν οι διαγνώσεις.

```
DELETE FROM labevents_resp WHERE NOT EXISTS(SELECT NULL FROM diagnoses_icd_resp
WHERE labevents_resp.hadm_id = diagnoses_icd_resp.hadm_id);
    DELETE 29537
SELECT COUNT(*) FROM labevents_resp;
     count
-------
 46537
(1 row)
DELETE FROM admissions_resp WHERE NOT EXISTS(SELECT NULL FROM diagnoses_icd_resp
WHERE admissions_resp.hadm_id = diagnoses_icd_resp.hadm_id);
SELECT COUNT(*) FROM admissions_resp;
     count
-------
   24583
(1 row)
DELETE FROM patients_resp WHERE NOT EXISTS(SELECT NULL FROM diagnoses_icd_resp
WHERE patients_resp.subject_id = diagnoses_icd_resp.subject_id);
SELECT COUNT(*) FROM patients_resp;
     count
-------
   19425

(1   row)
```

4. Μηχανική Μάθηση και εξόρυξη δεδομένων

Η μηχανική μάθηση (Machine Learning ML) είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης (Artificial Intelligence AI) στον τομέα της επιστήμης των υπολογιστών, το οποίο συχνά χρησιμοποιεί στατιστικές τεχνικές για να δώσει στους υπολογιστές τη δυνατότητα να «μαθαίνουν» (δηλαδή να βελτιώνουν σταδιακά την απόδοση σε μια συγκεκριμένη εργασία) με δεδομένα, χωρίς να είναι προγραμματισμένοι με ακρίβεια. Η μηχανική μάθηση συχνά συνδέεται στενά –αν δεν χρησιμοποιείται ως εναλλακτικός όρος– σε τομείς όπως η εξόρυξη δεδομένων (η διαδικασία της ανεύρεσης μοτίβων σε μεγάλα σύνολα δεδομένων που περιλαμβάνει μεθόδους που ανήκουν ταυτόχρονα στους τομείς της μηχανικής μάθησης, της στατιστικής και των συστημάτων βάσεων δεδομένων), η αναγνώριση μοτίβων, η στατιστική εξαγωγή ή στατιστική εκμάθηση. Όλοι αυτοί οι τομείς χρησιμοποιούν συχνά τις ίδιες μεθόδους και ίσως το όνομα αλλάζει με βάση την εμπειρία του επαγγελματία ή τον τομέα της εφαρμογής.

Τα κύρια καθήκοντα της μηχανικής μάθησης ταξινομούνται συνήθως σε δύο μεγάλες κατηγορίες, ανάλογα με το αν υπάρχει ή όχι το γνώρισμα βάση του οποίου θα εκπαιδευτεί ο υπολογιστής.

- **Εποπτευόμενη Μάθηση** (Supervised Learning): Το σύστημα παρουσιάζεται με παραδείγματα εισόδων και τις επιθυμητές εξόδους που παρέχει ο «δάσκαλος» και ο στόχος του αλγορίθμου μηχανικής μάθησης είναι να δημιουργήσει μια χαρτογράφηση από τις εισόδους στις εξόδους. Η χαρτογράφηση μπορεί να θεωρηθεί ως μια συνάρτηση ότι εάν δοθεί ως είσοδος ένα από τα δείγματα εκπαίδευσης θα πρέπει να εξάγει την επιθυμητή τιμή.
- **Μη εποπτευόμενη μάθηση** (Unsupervised Learning): Στην περίπτωση αυτή, στον αλγόριθμο μηχανικής μάθησης δεν δίδονται παραδείγματα επιθυμητών αποτελεσμάτων και αφήνεται μόνος του να βρει δομή στην είσοδό του.

Τρεις βασικά διαδικασίες μηχανικής μάθησης, διαχωρισμένες με βάση το τι προσπαθεί να επιτύχει τελικά το σύστημα στο τέλος, είναι οι εξής:

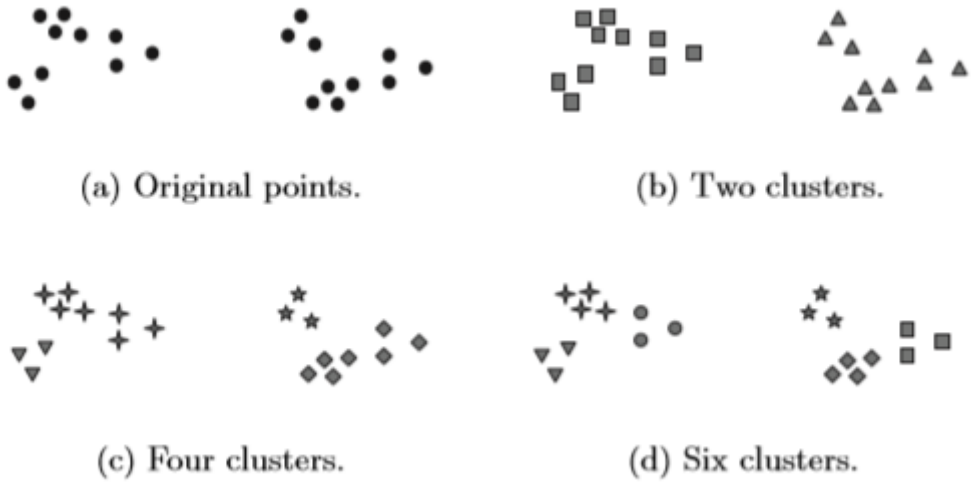
- **Ταξινόμηση** (Classification): οι είσοδοι χωρίζονται σε δύο ή περισσότερες κατηγορίες και ο υπολογιστής πρέπει να παράγει ένα μοντέλο που εκχωρεί αόρατες εισόδους σε μία ή περισσότερες ταξινομήσεις. Αυτό συνήθως αντιμετωπίζεται υπό εποπτεία.
- **Παλινδρόμηση** (Regression): επίσης ένα εποπτευόμενο πρόβλημα, οι εξόδους είναι συνεχείς παρά διακριτές.
- **Ομαδοποίηση** (Clustering): ένα σύνολο εισόδων πρέπει να χωριστεί σε ομάδες. Σε αντίθεση με την ταξινόμηση, οι ομάδες δεν είναι γνωστές εκ των προτέρων, καθιστώντας αυτό συνήθως μια εργασία χωρίς επίβλεψη. [8]

4.1. Συσταδοποίηση-Clustering

Η ανάλυση συστάδων (cluster analysis) ομαδοποιεί τα στοιχεία δεδομένων με βάση πληροφορίες που βρίσκονται μόνο στο σύνολο δεδομένων και είτε περιγράφουν τα ίδια, είτε τις σχέσεις τους. Ο σκοπός είναι τα στοιχεία μιας ομάδας να είναι παρόμοια (ή να συσχετίζονται) μεταξύ τους και να διαφέρουν από (ή να μην σχετίζονται με) τα στοιχεία σε άλλες ομάδες. Όσο μεγαλύτερη είναι η ομοιότητα (ή ομοιογένεια) μέσα σε μια ομάδα και όσο μεγαλύτερη είναι η διαφορά μεταξύ αυτών, τόσο καλύτερη ή πιο διακριτή είναι η συσταδοποίηση.

Σε πολλές εφαρμογές, η έννοια μιας συστάδας δεν είναι καλά ορισμένη. Για παράδειγμα έχουμε 20 σημεία και τρεις διαφορετικούς τρόπους διαίρεσης τους σε ομάδες. Τα σχήματα των δεικτών υποδηλώνουν την

ένταξη σε κάθε συστάδα. Τα σχήματα 4.1(b) και 4.1(d) χωρίζουν τα δεδομένα σε δύο και έξι μέρη, αντίστοιχα. Ωστόσο, η φαινομενική κατανομή του καθενός από τα δύο μεγαλύτερα σμήνη σε τρεις υποκλάσεις μπορεί απλά να είναι ένα τεχνούργημα του ανθρώπινου οπτικού συστήματος. Επίσης, μπορεί να μην είναι παράλογο να πούμε ότι τα σημεία σχηματίζουν τέσσερα σμήνη, όπως φαίνεται στο Σχήμα 4.1(c).



Σχήμα 4.1: Τρεις διαφορετικοί τρόποι συσταδοποίησης ενός συνόλου σημείων [9]

Αυτό δείχνει ότι ο καθορισμός μιας συστάδας είναι ασαφής και ότι ο καλύτερος διαχωρισμός των δεδομένων εξαρτάται από τη φύση τους και τα επιθυμητά αποτελέσματα.

Η ανάλυση συστάδας σχετίζεται με άλλες τεχνικές που χρησιμοποιούνται για τη διαίρεση των δεδομένων σε ομάδες. Για παράδειγμα, η συσταδοποίηση μπορεί να θεωρηθεί ως μια μορφή ταξινόμησης (classification), καθώς δημιουργεί μια επισήμανση των αντικειμένων με ετικέτες (συστάδες). Ωστόσο, αυτές οι ετικέτες προέρχονται μόνο από τα δεδομένα. Αντίθετα, η ταξινόμηση –όπως θα αναλύσουμε στη συνέχεια– αποτελεί εποπτευόμενη ταξινόμηση, καθώς στα νέα μη επισημασμένα αντικείμενα έχει εκχωρηθεί μια ετικέτα κλάσης χρησιμοποιώντας ένα μοντέλο που αναπτύχθηκε από αντικείμενα με γνωστές ετικέτες κλάσης. Για το λόγο αυτό, η συσταδοποίηση αναφέρεται μερικές φορές ως μη εποπτευόμενη ταξινόμηση. Όταν χρησιμοποιείται ο όρος ταξινόμηση χωρίς καμία εξειδίκευση στο πλαίσιο της εξόρυξης δεδομένων, συνήθως αναφέρεται στην εποπτευόμενη ταξινόμηση.

4.1.1. Διάφορα είδη συσταδοποίησης

Η συσταδοποίηση στοχεύει να εντοπίσει χρήσιμες ομάδες μεταξύ των δεδομένων (συστάδες), όπου η χρησιμότητα καθορίζεται από τους στόχους της ανάλυσης δεδομένων. Δεν αποτελεί έκπληξη το γεγονός ότι αρκετές διαφορετικές έννοιες μιας συστάδας αποδεικνύονται χρήσιμες στην πράξη. Προκειμένου να απεικονιστούν οπτικά οι διαφορές μεταξύ αυτών των τύπων συστάδων, χρησιμοποιούμε δισδιάστατα σημεία, όπως φαίνεται στο σχήμα 4.2, ως τα στοιχεί των δεδομένων μας.

- **Καλά διαχωρισμένη (Well-Separated):** Μια συστάδα είναι ένα σύνολο αντικειμένων στην οποία κάθε αντικείμενο είναι πιο κοντά (ή έχει περισσότερες ομοιότητες) με κάθε άλλο αντικείμενο στη συστάδα παρά με οποιοδήποτε αντικείμενο εκτός αυτής. Μερικές φορές ένα όριο χρησιμοποιείται για να καθορίσει ότι όλα τα αντικείμενα σε μια συστάδα πρέπει να είναι επαρκώς κοντά (ή παρόμοια) μεταξύ τους. Αυτός ο ιδεατός ορισμός ικανοποιείται, μόνο όταν τα δεδομένα διαχωρίζονται σε

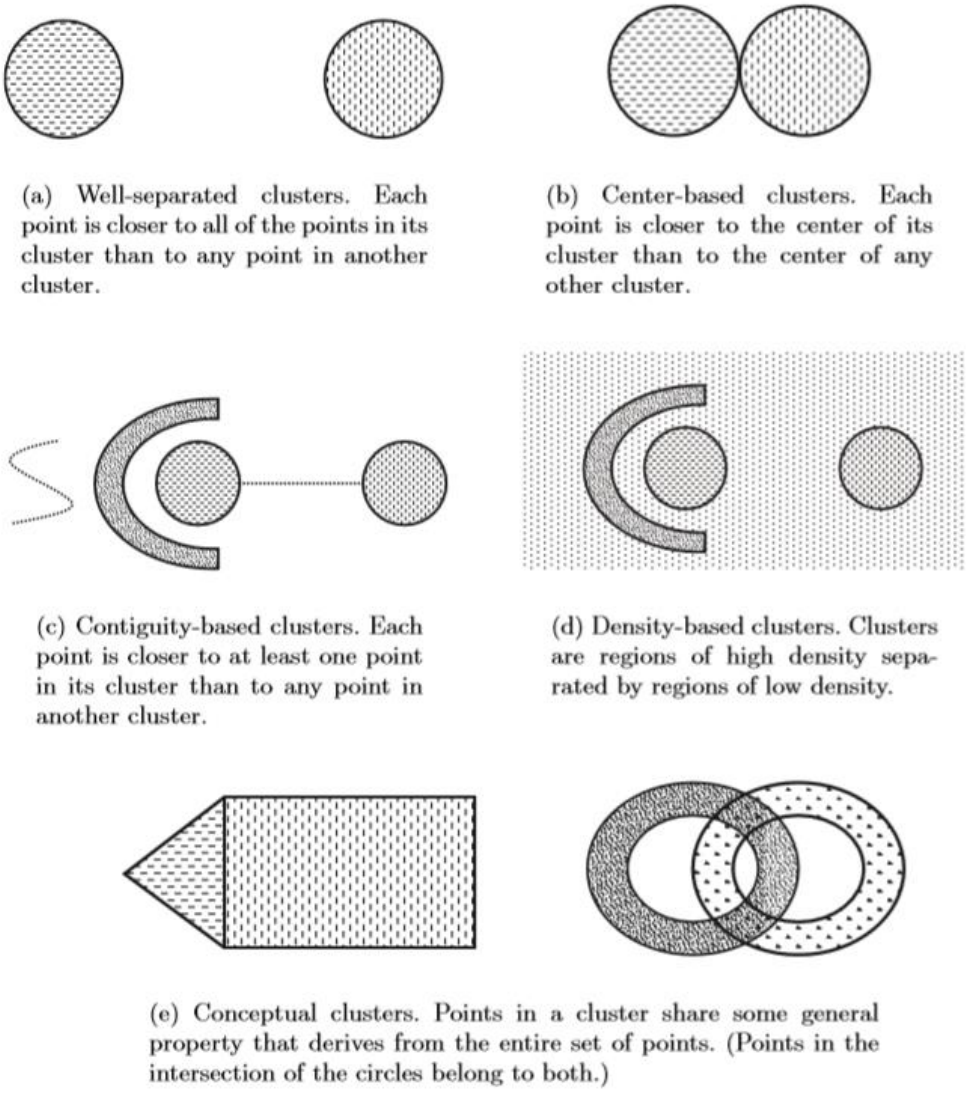
φυσικές συστάδες που απέχουν πολύ μεταξύ τους. Το σχήμα 4.2(a) δίνει ένα παράδειγμα καλά διαχωρισμένων συστάδων που αποτελείται από δύο ομάδες σημείων σε έναν δισδιάστατο χώρο. Η απόσταση μεταξύ οποιωνδήποτε δύο σημείων σε διαφορετικές ομάδες είναι μεγαλύτερη από την απόσταση μεταξύ οποιωνδήποτε δύο σημείων μέσα σε μια ομάδα. Οι καλά διαχωρισμένες συστάδες δεν χρειάζεται να είναι σφαιροειδείς, αλλά μπορούν να έχουν οποιοδήποτε σχήμα.

- **Βασισμένη σε Πρωτότυπα (Prototype-Based):** Μια συστάδα είναι ένα σύνολο αντικειμένων στην οποία κάθε αντικείμενο είναι πιο κοντά (ή έχει περισσότερες ομοιότητες) με το πρωτότυπο που ορίζει τη συστάδα, παρά από το πρωτότυπο οποιασδήποτε άλλης συστάδας. Για δεδομένα με συνεχή χαρακτηριστικά, το πρωτότυπο ενός συμπλέγματος είναι συχνά ένα κεντροειδές (centroid), δηλαδή το μέσο όλων των σημείων του συμπλέγματος. Όταν ένα κεντροειδές δεν έχει νόημα, όπως όταν τα δεδομένα έχουν κατηγορηματικά χαρακτηριστικά, το πρωτότυπο είναι συχνά το πιο αντιπροσωπευτικό σημείο ενός συμπλέγματος (medoid). Για πολλούς τύπους δεδομένων, το πρωτότυπο μπορεί να θεωρηθεί ως το πιο κεντρικό σημείο, και σε τέτοιες περιπτώσεις, συνήθως αναφερόμαστε σε συστάδες βάσει πρωτότυπων (prototype-based clusters) ως center-based clusters. Δεν αποτελεί έκπληξη το γεγονός ότι τέτοιες συστάδες τείνουν να είναι σφαιρικές, όπως φαίνεται στο σχήμα 4.2(b).

- **Βασισμένη σε γράφους (Graph-Based):** Αν τα δεδομένα αντιπροσωπεύονται ως γράφος, όπου οι κορυφές είναι αντικείμενα και οι ακμές αντιπροσωπεύουν τις συνδέσεις τους, τότε μια συστάδα μπορεί να οριστεί ως ένα συνδεδεμένο στοιχείο, δηλαδή μια ομάδα αντικειμένων που συνδέονται μεταξύ τους, αλλά δεν έχουν καμία σύνδεση με αντικείμενα εκτός της ομάδας. Το σχήμα 4.2(c) δείχνει ένα παράδειγμα τέτοιων συστάδων για δισδιάστατα σημεία. Όπως τα συμπλέγματα που βασίζονται σε πρωτότυπα, τέτοιες συστάδες τείνουν να είναι σφαιρικές.

- **Βάση πυκνότητας (Density-Based):** Μια συστάδα είναι μια πυκνή περιοχή αντικειμένων που περιβάλλεται από μια περιοχή χαμηλής πυκνότητας. Το σχήμα 4.2(d) δείχνει ορισμένες συστάδες με βάση την πυκνότητα για δεδομένα που δημιουργήθηκαν με την προσθήκη θορύβου στα δεδομένα του Σχήματος 4.2(c). Οι δύο κυκλικές συστάδες δεν συγχωνεύονται, όπως στο Σχήμα 4.2(c), επειδή η γέφυρα μεταξύ τους εξασθενεί στον θόρυβο. Ομοίως, η καμπύλη που υπάρχει στο Σχήμα 4.2(c) επίσης εξασθενεί τον θόρυβο και δεν σχηματίζει συστάδα στο Σχήμα 4.2(d). Τέτοιου είδους συστάδας χρησιμοποιείται συχνά όταν οι συστάδες είναι ακανόνιστες ή και όταν υπάρχουν θόρυβοι και ακραίες τιμές. Αντίθετα, η επιλογή συσταδοποίησης βάση πυκνότητας δεν θα λειτουργούσε καλά για τα δεδομένα του Σχήματος 4.2(d), επειδή ο θόρυβος θα τείνει να σχηματίζει γέφυρες μεταξύ των συστάδων.

- **Κοινής ιδιότητας (Conceptual Clusters):** Γενικότερα, μπορούμε να ορίσουμε μια συστάδα ως σύνολο αντικειμένων που μοιράζονται κάποια ιδιότητα. Ο ορισμός αυτός περιλαμβάνει όλους τους προηγούμενους ορισμούς μιας συστάδας. Ωστόσο, η προσέγγιση κοινής ιδιότητας περιλαμβάνει επίσης νέους τύπους συστάδων. Για παράδειγμα στο σχήμα 4.2(e), μια τριγωνική περιοχή (συστάδα) είναι δίπλα σε μια ορθογώνια και υπάρχουν δύο αλληλένδετοι κύκλοι (συστάδες). Και στις δύο περιπτώσεις, ένας αλγόριθμος ομαδοποίησης θα χρειαζόταν μια πολύ συγκεκριμένη έννοια για την επιτυχή ανίχνευση αυτών των συστάδων. Η διαδικασία ανίχνευσης τέτοιων ομάδων ονομάζεται εννοιολογική συσταδοποίηση.



Σχήμα 4.2: Διαφορετικοί τύποι συσταδοποίησης ενός συνόλου σημείων σε δύο διαστάσεις [9]

4.1.2. K-means

Υπάρχουν πολλές τέτοιες τεχνικές βασισμένες σε πρωτότυπα, αλλά δύο από τις πιο σημαντικές είναι οι K-means and K-medoid. Ο K-means –ο οποίος είναι ένας από τους παλαιότερους και πιο ευρέως χρησιμοποιούμενους αλγόριθμους ομαδοποίησης– ορίζει ένα πρωτότυπο από την άποψη ενός κεντροειδούς, το οποίο είναι συνήθως ο μέσος όρος μιας ομάδας σημείων και εφαρμόζεται τυπικά σε αντικείμενα σε συνεχή n-διάστατο χώρο. Ο K-medoid ορίζει ένα πρωτότυπο από την άποψη ενός medoid, το οποίο είναι το πιο αντιπροσωπευτικό σημείο μιας ομάδας σημείων και μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα τύπων δεδομένων, διότι απαιτεί μόνο ένα μέτρο εγγύτητας για ένα ζευγάρι αντικειμένων. Ενώ ένα centroid σχεδόν ποτέ δεν αντιστοιχεί σε ένα πραγματικό σημείο δεδομένων, ένα medoid, από τον ορισμό του, πρέπει να είναι ένα πραγματικό σημείο δεδομένων.

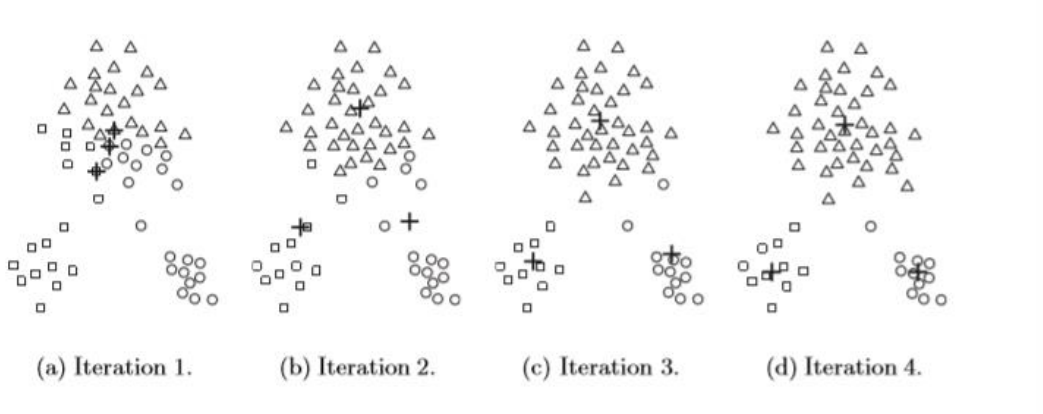
Ο Αλγόριθμος K-means

Η τεχνική συσταδοποίησης K-means είναι αρκετά απλή. Αρχικά επιλέγονται τα K αρχικά κεντροειδή, όπου το K είναι μια παράμετρος που έχει καθοριστεί από τον χρήστη, και ορίζει τον αριθμό των επιθυμητών συστάδων. Κάθε σημείο, στη συνέχεια, αντιστοιχίζεται στο πλησιέστερο κεντροειδές και κάθε ομάδα σημείων που αντιστοιχίστηκαν σε ένα κεντροειδές είναι μια συστάδα. Το κεντροειδές της κάθε συστάδας στη συνέχεια ενημερώνεται βάσει των σημείων που αντιστοιχίστηκαν σε αυτήν. Τα βήματα εκχώρησης και ενημέρωσης επαναλαμβάνονται, έως ότου κανένα σημείο δεν αλλάξει συστάδες, ή ισοδύναμα, έως ότου τα κεντροειδή παραμείνουν τα ίδια. Ο K-means περιγράφεται επισήμως από τον αλγόριθμο:

Basic K-means algorithm

1: Select K points as initial centroids.
2: repeat
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: until Centroids do not change.

Η λειτουργία του K-μέσου απεικονίζεται στο Σχήμα 4.3, το οποίο δείχνει πώς, ξεκινώντας από τρία τυχαία κεντροειδή με το σύμβολο «+», οι τελικές συστάδες κατασκευάζονται σε τέσσερα βήματα επικαιροποίησης-εκχώρησης.



Σχήμα 4.3. Χρήση του αλγορίθμου K-means για την εύρεση τριών συστάδων σε δείγμα δεδομένων [9]

Αντιστοίχιση σημείων στο πιο κοντινό κεντοειδές

Για να ορίσουμε ένα σημείο στο πλησιέστερο κεντροειδές, χρειαζόμαστε ένα μέτρο εγγύτητας που να ποσοτικοποιεί την έννοια του "πλησιέστερου" για τα συγκεκριμένα δεδομένα που εξετάζονται. Η ευκλείδεια (L_2) απόσταση χρησιμοποιείται συχνά για σημεία δεδομένων στον Ευκλείδειο χώρο, ενώ η ομοιότητα συνημίτονου είναι πιο κατάλληλη για τα κείμενα. Ωστόσο, διάφοροι τύποι μέτρων εγγύτητας μπορεί να είναι κατάλληλοι για δεδομένα τύπου κειμένου. Για παράδειγμα, η απόσταση Μανχάταν (L_1) μπορεί να χρησιμοποιηθεί για Ευκλείδεια δεδομένα, ενώ το μέτρο Jaccard χρησιμοποιείται συχνά για τα κείμενα.

Συνήθως, τα μέτρα ομοιότητας που χρησιμοποιούνται για τον K-means είναι σχετικά απλά, αφού ο αλγόριθμος υπολογίζει επανειλημμένα την ομοιότητα κάθε σημείου με κάθε κέντρο. Σε ορισμένες περιπτώσεις, ωστόσο, όπως όταν τα δεδομένα ανήκουν ευκλείδειο χώρο λίγων διαστάσεων, είναι δυνατόν

να αποφευχθεί η καταμέτρηση πολλών από τις ομοιότητες, επιταχύνοντας έτσι σημαντικά τον αλγόριθμο K-means.

Κεντροειδή και αντικειμενικές συναρτήσεις

Το βήμα 4 του αλγορίθμου K-means αναφέρθηκε γενικά ως «επανασχηματισμός του κεντροειδούς της κάθε συστάδας», αφού το κεντροειδές μπορεί να ποικίλει, ανάλογα με το μέτρο εγγύτητας για τα δεδομένα και τον στόχο της συσταδοποίησης. Ο στόχος της συσταδοποίησης τυπικά εκφράζεται από μια αντικειμενική συνάρτηση που εξαρτάται από τις γειτνιάσεις των σημείων μεταξύ τους ή με τα κεντροειδή του συμπλέγματος, π.χ. ελαχιστοποιώντας την τετραγωνική απόσταση κάθε σημείου μέχρι το πλησιέστερο κέντρο. Ωστόσο, το βασικό σημείο είναι ότι, αφού έχουμε καθορίσει ένα μέτρο εγγύτητας και μια αντικειμενική συνάρτηση, το κέντρο που πρέπει να επιλέξουμε μπορεί συχνά να προσδιοριστεί μαθηματικά.

Δεδομένα στον Ευκλείδειο Χώρο

Αν εξετάσουμε τα δεδομένα των οποίων το μέτρο εγγύτητας είναι η ευκλείδεια απόσταση, η αντικειμενική μας συνάρτηση, που υπολογίζει την ποιότητα μιας συσταδοποίησης, είναι το άθροισμα του τετραγωνικού σφάλματος (SSE), το οποίο είναι επίσης γνωστό ως scatter. Με άλλα λόγια, υπολογίζουμε το σφάλμα κάθε σημείου δεδομένων, δηλ. την ευκλείδεια απόσταση του από το πλησιέστερο κέντρο και στη συνέχεια υπολογίζουμε το συνολικό άθροισμα των τετραγωνικών σφαλμάτων. Λαμβάνοντας υπόψη δύο διαφορετικά σύνολα συστάδων που παράγονται από δύο διαφορετικές εκτελέσεις του K-means, προτιμάμε αυτό με το μικρότερο τετραγωνικό σφάλμα, αφού αυτό σημαίνει ότι τα πρωτότυπα (centroids) αυτής της συσταδοποίησης είναι μια καλύτερη αναπαράσταση των σημείων στο συστάδα τους. Έτσι η SSE ορίζεται τυπικά ως εξής:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

όπου *dist* είναι η τυπική ευκλείδεια απόσταση (L_2) μεταξύ δύο αντικειμένων στον Ευκλείδειο χώρο.

Με βάση αυτές τις υποθέσεις, μπορεί να φανεί ότι το κεντροειδές που ελαχιστοποιεί το SSE της συστάδας είναι το μέσο. Έτσι το κεντροειδές (μέσο) της συστάδας i καθορίζεται από την εξίσωση:

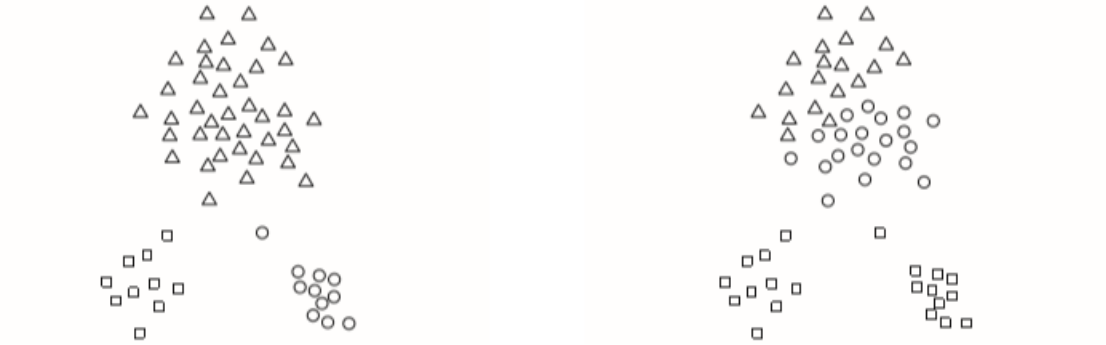
$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

Τα βήματα 3 και 4 του αλγορίθμου K-means προσπαθούν άμεσα να ελαχιστοποιήσουν το SSE. Το βήμα 3 σχηματίζει ομάδες, αναθέτοντας σημεία στο πλησιέστερο κέντρο τους, το οποίο ελαχιστοποιεί το SSE για το δεδομένο σύνολο κεντροειδών. Το βήμα 4 ανασυντάσσει τα κεντροειδή ώστε να ελαχιστοποιηθεί περαιτέρω το SSE. Εντούτοις, οι ενέργειες του K-means στα Βήματα 3 και 4 εγγυώνται ότι θα βρουν μόνο το τοπικό ελάχιστο σε σχέση με το SSE, επειδή βασίζονται στη βελτιστοποίηση του SSE για συγκεκριμένες επιλογές των κεντροειδών και των συστάδων, αντί για όλες τις πιθανές επιλογές.

Επιλέγοντας τα αρχικά κεντροειδή

Όταν χρησιμοποιείται τυχαία αρχικοποίηση των κεντροειδών, οι διαφορετικές διεργασίες K-means τυπικά παράγουν διαφορετικά συνολικά SSE. Αυτό αποτυπώνεται με το σύνολο σημείων δύο διαστάσεων που φαίνονται στο Σχήμα 4.3, το οποίο έχει τρεις φυσικές συστάδες. Το σχήμα 4.4(α) δείχνει μια λύση

συσταδοποίησης στην οποία το συνολικό SSE είναι το ελάχιστο για τις τρεις συστάδες, ενώ το Σχήμα 4.4(b) δείχνει μια υποσυστάδα που είναι μόνο ένα τοπικό ελάχιστο. Η επιλογή των κατάλληλων αρχικών κεντροειδών είναι το βασικό βήμα της βασικής διαδικασίας του K- means. Μια κοινή προσέγγιση είναι να επιλέγονται τυχαία τα αρχικά κεντροειδή.



(a) Optimal clustering. (b) Suboptimal clustering.

Σχήμα 4.4. Τρεις βέλτιστες και μη βέλτιστες συστάδες [9]

Μείωση του SSE με μεταγενέστερη επεξεργασία

Ένας προφανής τρόπος για να μειώσουμε το SSE είναι να βρούμε περισσότερες ομάδες, δηλ. Να χρησιμοποιήσουμε ένα μεγαλύτερο K. Ωστόσο, σε πολλές περιπτώσεις, θα θέλαμε να βελτιώσουμε το SSE, αλλά δεν θέλουμε να αυξήσουμε τον αριθμό των συστάδων. Αυτό είναι συχνά εφικτό, διότι ο K-means συνήθως συγκλίνει σε ένα τοπικό ελάχιστο. Διάφορες τεχνικές χρησιμοποιούνται για να «φτιάξουν» τις προκύπτουσες συστάδες προκειμένου να παράγουν μια ομαδοποίηση που έχει χαμηλότερη SSE. Η στρατηγική είναι να επικεντρωθούμε σε μεμονωμένες συστάδες, δεδομένου ότι το συνολικό SSE είναι απλώς το άθροισμα του SSE που συνεισφέρει κάθε συστάδα. Μπορούμε να αλλάξουμε το συνολικό SSE εκτελώντας διάφορες λειτουργίες στις συστάδες, όπως τον διαχωρισμό ή τη συγχώνευση. Μία συνήθως χρησιμοποιούμενη προσέγγιση είναι η χρήση εναλλακτικών φάσεων διάσπασης και συγχώνευσης συστάδων. Με αυτό τον τρόπο, είναι συχνά πιθανό να ξεφύγουμε από τα τοπικά ελάχιστα SSE και να παράγουμε ακόμα μια λύση συσταδοποίησης με τον επιθυμητό αριθμό ομάδων. Τα παρακάτω είναι μερικές τεχνικές που χρησιμοποιούνται στις φάσεις διαίρεσης και συγχώνευσης.

Δύο στρατηγικές που μειώνουν το συνολικό SSE αυξάνοντας τον αριθμό των συστάδων είναι οι ακόλουθες:

- Διαχωρισμός μιας συστάδας: Επιλέγεται συνήθως τη συστάδα με το μεγαλύτερο SSE, αλλά μπορούμε επίσης να χωρίσουμε τη συστάδα με τη μεγαλύτερη τυπική απόκλιση για ένα συγκεκριμένο χαρακτηριστικό.
- Εισαγωγή ενός νέου κεντροειδούς: Συχνά επιλέγεται το πιο απομακρυσμένο σημείο από οποιοδήποτε κέντρο συστάδας. Μπορούμε να το προσδιορίσουμε εύκολα εάν παρακολουθούμε το SSE που συνεισφέρει κάθε σημείο. Μια άλλη προσέγγιση είναι να επιλέξουμε τυχαία από όλα τα σημεία ή από τα σημεία με το υψηλότερο SSE σε σχέση με τα πλησιέστερα κεντροειδή τους.

Δύο στρατηγικές που μειώνουν τον αριθμό των συστάδων, ενώ προσπαθούν να ελαχιστοποιήσουν την αύξηση του συνολικού SSE, είναι οι εξής:

- Διασπορά μιας συστάδας: Αυτό επιτυγχάνεται αφαιρώντας το κεντροειδές που της αντιστοιχεί και επανατοποθετώντας τα σημεία σε άλλες συστάδες. Στην ιδανική περίπτωση, η συστάδα που είναι διασκορπισμένη πρέπει να είναι αυτή που αυξάνει το ελάχιστο συνολικό SSE.
- Συγχώνευση δύο συστάδων: Συνήθως επιλέγονται οι συστάδες με τα κοντινότερα κεντροειδή, αν και μια άλλη, ίσως καλύτερη, προσέγγιση είναι η συγχώνευση των δύο συστάδων που οδηγούν στη μικρότερη αύξηση του συνολικού SSE. Αυτές οι δύο στρατηγικές συγχώνευσης είναι οι ίδιες που χρησιμοποιούνται στις τεχνικές ιεραρχικής ομαδοποίησης γνωστές ως μέθοδος centroid και μέθοδος Ward, αντίστοιχα.

Πλεονεκτήματα και αδυναμίες

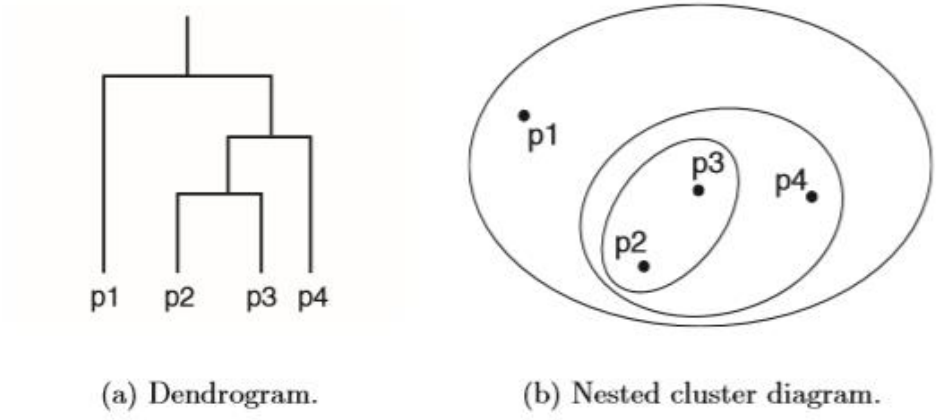
Ο K-means είναι απλός και μπορεί να χρησιμοποιηθεί για μια μεγάλη ποικιλία τύπων δεδομένων. Είναι επίσης αρκετά αποτελεσματικός, παρόλο που συχνά εκτελούνται πολλές δοκιμές. Ορισμένες παραλλαγές είναι ακόμα πιο αποτελεσματικές και είναι λιγότερο ευαίσθητες σε προβλήματα αρχικοποίησης. Εντούτοις, ο K-means δεν είναι κατάλληλος για όλους τους τύπους δεδομένων. Δεν μπορεί να χειριστεί συστάδες διαφορετικών μεγεθών και πυκνοτήτων ή μη σφαιρικές, αν και συνήθως μπορεί να εντοπίσει καθαρές υποσυστάδες εάν καθοριστεί ένας αρκετά μεγάλος αριθμός K. Ο K-means παρουσιάζει επίσης προβλήματα συσταδοποίησης δεδομένων που περιέχουν ακραίες τιμές. Η ανίχνευση και απομάκρυνση των ακραίων τιμών μπορεί να συμβάλει σημαντικά στις καταστάσεις αυτές. Τέλος, ο K-means περιορίζεται σε δεδομένα για τα οποία υπάρχει μια έννοια ενός κέντρου (κεντροειδούς). Η ομαδοποίηση K-medoid, δεν έχει αυτόν τον περιορισμό, αλλά είναι πιο ακριβή υπολογιστικά.

4.1.3. Ιεραρχική συσταδοποίηση

Οι τεχνικές ιεραρχικής συσταδοποίησης είναι μια δεύτερη σημαντική κατηγορία. Όπως και με τον K-means, αυτές οι προσεγγίσεις είναι σχετικά παλιές σε σύγκριση με πολλούς αλγορίθμους συσταδοποίησης, αλλά εξακολουθούν να έχουν ευρεία χρήση. Υπάρχουν δύο βασικές προσεγγίσεις για τη δημιουργία μιας ιεραρχικής ομαδοποίησης:

- Συσσωρευτικά: Ξεκινώντας με τα σημεία ως μεμονωμένα σύνολα και σε κάθε βήμα, συγχωνεύεται το πλησιέστερο ζευγάρι συστάδων. Αυτό απαιτεί να προσδιοριστεί η έννοια της εγγύτητας συστάδας.
- Διαιρετικά: Ξεκινώντας με μία ολοκληρωμένη συστάδα και σε κάθε βήμα, διαιρείται μια, έως ότου παραμείνουν μόνο διαχωρισμένα σύνολα μεμονωμένων σημείων. Σε αυτήν την περίπτωση, πρέπει να αποφασιστεί ποια συστάδα θα χωριστεί σε κάθε βήμα και πώς θα γίνει ο διαχωρισμός.

Οι συσσωρευτικές τεχνικές ιεραρχικής συσταδοποίησης είναι μακράν οι πιο συνηθισμένες και η αναφορά θα περιοριστεί σε αυτές τις μεθόδους. Μια ιεραρχική συσταδοποίηση εμφανίζεται συχνά γραφικά χρησιμοποιώντας ένα δέντρο, που μοιάζει με ένα δενδρόγραμμα, το οποίο εμφανίζει τόσο τις σχέσεις συστάδας-υποσυστάδας, όσο και τη σειρά με την οποία οι συστάδες συγχωνεύθηκαν (συσσωρευτική προβολή) ή χωρίστηκαν (διαιρετική προβολή). Για σύνολα δισδιάστατων σημείων, όπως εκείνα που θα χρησιμοποιήσουμε ως παραδείγματα, μια ιεραρχική ομαδοποίηση μπορεί επίσης να αναπαρασταθεί γραφικά χρησιμοποιώντας ένα ένθετο διάγραμμα συστάδας. Το σχήμα 4.5 δείχνει ένα παράδειγμα αυτών των δύο τύπων σχημάτων για ένα σύνολο τεσσάρων δισδιάστατων σημείων.



(a) Dendrogram. (b) Nested cluster diagram.

Σχήμα 4.5. Η Ιεραρχική συσταδοποίηση τεσσάρων σημείων που παρουσιάζονται σε δενδρόγραμμα και σε ένθετες συστάδες [9]

7.3.1 Αλγόριθμος ιεραρχικής συσταδοποίησης

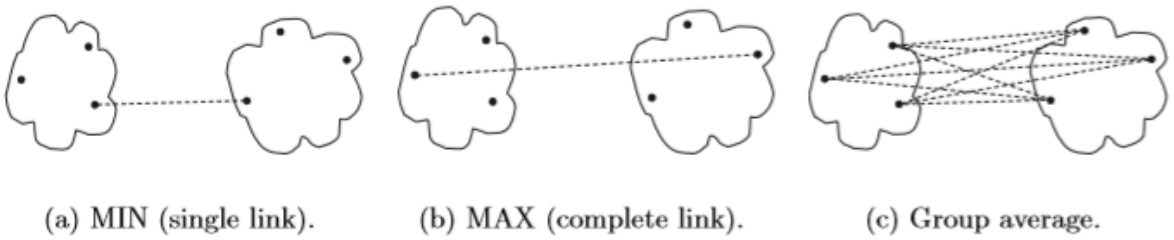
Πολλές τεχνικές ιεραρχικής συσταδοποίησης είναι παραλλαγές σε μια ενιαία προσέγγιση: ξεκινώντας με μεμονωμένα σημεία ως συστάδες, συγχωνεύουν διαδοχικά τις δύο κοντινότερες συστάδες μέχρι να παραμείνει μόνο μία. Αυτή η προσέγγιση εκφράζεται πιο τυπικά στον παρακάτω Αλγόριθμο.

Hierarchical clustering algorithm.

1: Compute the proximity matrix, if necessary.
2: repeat
3: Merge the closest two clusters.
4: Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5: until Only one cluster remains.

Καθορισμός της εγγύτητας μεταξύ συστάδων

Η βασική λειτουργία του Αλγόριθμου είναι ο υπολογισμός της εγγύτητας μεταξύ δύο συστάδων και ο ορισμός της εγγύτητας που διαφοροποιεί τις διάφορες ιεραρχικές τεχνικές. Η εγγύτητα της συστάδας τυπικά ορίζεται με έναν συγκεκριμένο τύπο συστάδας. Για παράδειγμα, πολλές τεχνικές ιεραρχικής συσταδοποίησης, όπως οι MIN, MAX, και Μέσος Όρος Ομάδας (Group Average), προέρχονται από μια οπτική διαχωρισμού των συστάδων βασισμένη σε γράφους. Η MIN ορίζει την εγγύτητα των συστάδων ως την εγγύτητα μεταξύ των δύο πιο κοντινών σημείων που βρίσκονται σε διαφορετικές συστάδες ή χρησιμοποιώντας τους όρους γράφων, τη συντομότερο ακμή μεταξύ δύο κορυφών σε διαφορετικά υποσύνολα κορυφών. Αυτό δίνει συστάδες που βασίζονται σε συγγένεια όπως φαίνεται στο σχήμα 4.2(c). Εναλλακτικά, το MAX λαμβάνει την εγγύτητα μεταξύ των πλέον απομακρυσμένων δύο σημείων σε διαφορετικές συστάδες ως εγγύτητα συστάδας ή χρησιμοποιώντας όρους γράφων, τη μεγαλύτερη ακμή μεταξύ δύο κορυφών σε διαφορετικά υποσύνολα κορυφών. Μια άλλη προσέγγιση που βασίζεται σε γράφους, η τεχνική του μέσου όρου, ορίζει ότι η εγγύτητα της συστάδας είναι η μέση απόσταση όλων των ζευγών σε διαφορετικές συστάδες (μέσο μήκος των ακμών). Το σχήμα 4.6 απεικονίζει αυτές τις τρεις προσεγγίσεις.



(a) MIN (single link). (b) MAX (complete link). (c) Group average.

Σχήμα 4.6. Ορισμοί της εγγύτητας βασιζόμενοι στους γράφους [9]

Αν, αντιθέτως, πάρουμε μια βασισμένη σε πρωτότυπα οπτική, στην οποία κάθε συστάδα αντιπροσωπεύεται από ένα κεντροειδές, οι διαφορετικοί ορισμοί της εγγύτητας συστάδας είναι πιο φυσικοί. Κατά τη χρήση των κεντροειδών, η εγγύτητα συστάδας ορίζεται συνήθως ως η εγγύτητα μεταξύ των κεντροειδών. Μια εναλλακτική τεχνική, η μέθοδος του Ward, υποθέτει επίσης ότι μια συστάδα αντιπροσωπεύεται από το κεντροειδές της, αλλά μετρά την εγγύτητα μεταξύ δύο συστάδων, σε όρους αύξησης του SSE που προκύπτει από τη συγχώνευση των δύο συστάδων. Όπως ο K-means, η μέθοδος του Ward επιχειρεί να ελαχιστοποιήσει το άθροισμα των τετραγωνικών αποστάσεων των σημείων από τα κεντροειδή των συστάδων τους.

Έλλειψη καθολικής αντικειμενικής συνάρτησης

Η ιεραρχική συσταδοποίηση δεν μπορεί να θεωρηθεί ότι βελτιστοποιεί συνολικά μια αντικειμενική συνάρτηση. Αντίθετα, οι τεχνικές ιεραρχικής συσταδοποίησης χρησιμοποιούν διάφορα κριτήρια για να αποφασίσουν τοπικά, σε κάθε βήμα, ποιες συστάδες πρέπει να συγχωνευθούν (ή να χωριστούν για διαχωριστικές προσεγγίσεις). Αυτή η προσέγγιση παράγει αλγόριθμους ομαδοποίησης που αποφεύγουν να προσπαθήσουν να επιλύσουν ένα δύσκολο συνδυαστικό πρόβλημα βελτιστοποίησης. (Μπορεί να φανεί ότι το γενικό πρόβλημα συσταδοποίησης για μια αντικειμενική συνάρτηση όπως η «ελαχιστοποίηση του SSE» δεν είναι υπολογιστικά εφικτό.) Επιπλέον, τέτοιες προσεγγίσεις δεν έχουν δυσκολία στην επιλογή των αρχικών σημείων. Παρόλα αυτά, η πολυπλοκότητα του χρόνου $O(m^2 logm)$ και η πολυπλοκότητα του χώρου $O(m^2)$ είναι απαγορευτικές σε πολλές περιπτώσεις.

Οι αποφάσεις συγχώνευσης είναι τελικές

Οι αλγόριθμοι ιεραρχικής συσταδοποίησης τείνουν να παίρνουν καλές τοπικές αποφάσεις για το συνδυασμό δύο ομάδων επειδή μπορούν να χρησιμοποιήσουν πληροφορίες σχετικά με την ομοιόμορφη ομοιότητα όλων των σημείων. Ωστόσο, όταν αποφασιστεί η συγχώνευση δύο ομάδων, δεν μπορεί να ανακληθεί αργότερα. Αυτή η προσέγγιση εμποδίζει ένα τοπικό κριτήριο βελτιστοποίησης, να γίνει ένα κριτήριο συνολικής βελτιστοποίησης. Πράγματι, οι συστάδες δεν είναι καν σταθερές, με την έννοια ότι ένα σημείο σε μια συστάδα μπορεί να είναι πιο κοντά στο κέντρο μιας άλλης απ' ό,τι στο κέντρο της δικής του.

Ακραία Σημεία

Καθώς δημιουργείται η ιεραρχική συσταδοποίηση τείνουν να σχηματίζονται ακραίες τιμές ή μικρές ομάδες τιμών που δεν συγχωνεύονται με άλλες συστάδες, έως πολύ αργότερα στη διαδικασία της συγχώνευσης. Με την απόρριψη μονάδων ή μικρών συστάδων που δεν συγχωνεύονται με άλλες, τα ακραία σημεία μπορούν να αφαιρεθούν.

Πλεονεκτήματα και αδυναμίες

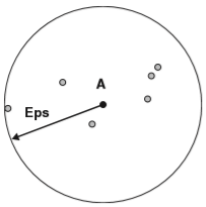
Οι δυνάμεις και οι αδυναμίες συγκεκριμένων αλγορίθμων ιεραρχικής συσσωμάτωσης συζητήθηκαν παραπάνω. Γενικότερα, αυτοί οι αλγόριθμοι χρησιμοποιούνται συνήθως επειδή η υποκείμενη εφαρμογή, π.χ. η δημιουργία μιας ταξινόμησης, απαιτεί ιεραρχία. Επίσης, μερικές μελέτες έχουν δείξει ότι αυτοί οι αλγόριθμοι μπορούν να παράγουν συστάδες καλύτερης ποιότητας. Ωστόσο, οι αλγόριθμοι ιεραρχικής συσταδοποίησης είναι δαπανηροί όσον αφορά τις απαιτήσεις υπολογισμού και αποθήκευσης. Το γεγονός ότι όλες οι συγχωνεύσεις είναι τελικές, μπορεί επίσης να προκαλέσει προβλήματα για θορυβώδη, μεγάλης κλίμακας δεδομένα, όπως τα δεδομένα κειμένων. Με τη σειρά τους, τα δύο αυτά προβλήματα μπορούν να αντιμετωπιστούν σε κάποιο βαθμό, αρχικά συγκεντρώνοντας μερικώς τα δεδομένα χρησιμοποιώντας μια άλλη τεχνική, όπως ο K-means.

4.1.4. DBSCAN

Η ομαδοποίηση με βάση την πυκνότητα εντοπίζει περιοχές υψηλής πυκνότητας που χωρίζονται μεταξύ τους από περιοχές χαμηλής πυκνότητας. Ο DBSCAN είναι ένας απλός και αποτελεσματικός αλγόριθμος συσταδοποίησης με βασίζεται την πυκνότητα, ο οποίος απεικονίζει μια σειρά από έννοιες που είναι σημαντικές για οποιαδήποτε προσέγγιση συσταδοποίησης βάσει πυκνότητας.

Παραδοσιακή πυκνότητα: προσέγγιση βασισμένη σε κέντρο

Στην προσέγγιση με βάση το κέντρο, η πυκνότητα υπολογίζεται για ένα συγκεκριμένο σημείο στο σύνολο δεδομένων μετρώντας τον αριθμό των σημείων μέσα σε μια συγκεκριμένη ακτίνα, Eps, του συγκεκριμένου σημείου. Αυτό περιλαμβάνει το ίδιο το σημείο. Η τεχνική αυτή απεικονίζεται γραφικά από το Σχήμα 4.7. Ο αριθμός των σημείων σε ακτίνα Eps του σημείου A είναι 7, συμπεριλαμβανομένου του ίδιου του A.

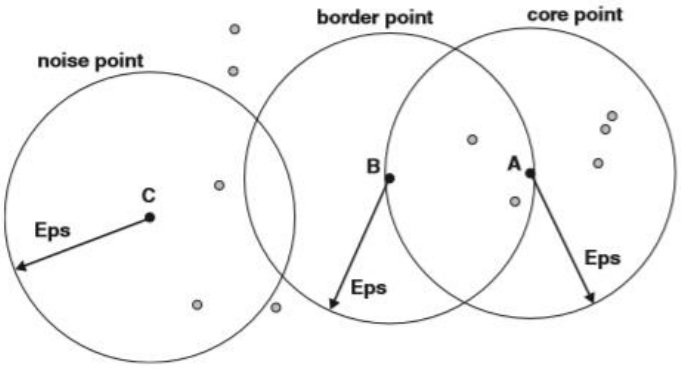


Σχήμα 4.7. [9]

Αυτή η μέθοδος είναι απλή στην εφαρμογή, αλλά η πυκνότητα οποιουδήποτε σημείου εξαρτάται από την συγκεκριμένη ακτίνα. Για παράδειγμα, εάν η ακτίνα είναι αρκετά μεγάλη, τότε όλα τα σημεία θα έχουν πυκνότητα m, τον αριθμό των σημείων στο σύνολο δεδομένων. Ομοίως, εάν η ακτίνα είναι πολύ μικρή, τότε όλα τα σημεία θα έχουν πυκνότητα 1. Μια προσέγγιση για να αποφασιστεί η κατάλληλη ακτίνα για δεδομένα χαμηλών διαστάσεων δίνεται παρακάτω.

Ταξινόμηση σημείων σύμφωνα με την πυκνότητα που βασίζεται στο κέντρο

Η προσέγγιση της πυκνότητας με βάση το κέντρο μας επιτρέπει να ταξινομήσουμε ένα σημείο ως: (1) στο εσωτερικό μιας πυκνής περιοχής (σημείο πυρήνα), (2) στην άκρη μιας πυκνής περιοχής (συνοριακό σημείο) ή (3) σε μια αραιή περιοχή (σημείο θορύβου ή υποβάθρου). Το Σχήμα 4.8 απεικονίζει γραφικά τις έννοιες των πυρήνων, των συνόρων και των σημείων θορύβου χρησιμοποιώντας μία συλλογή δισδιάστατων σημείων.



Σχήμα 4.8. [9]

Πιο αναλυτικά:

- **Σημεία πυρήνα:** Αυτά τα σημεία βρίσκονται στο εσωτερικό μιας συστάδας με βάση την πυκνότητα. Ένα σημείο είναι ένα σημείο πυρήνα, αν υπάρχουν κάποια τουλάχιστον MinPts μέσα σε μια απόσταση Eps, όπου τα MinPts και Eps είναι παράμετροι που καθορίζονται από το χρήστη. Στο Σχήμα 4.8, το σημείο A είναι ένα σημείο πυρήνα για την ακτίνα (Eps) αν *MinPts≥7*.
- **Συνοριακά σημεία**: Ένα σημείο συνόρων δεν είναι ένα σημείο πυρήνα, αλλά εμπίπτει στην περιοχή ενός πυρήνα. Στο Σχήμα 4.8, το σημείο B είναι ένα συνοριακό σημείο. Ένα συνοριακό σημείο μπορεί να εμπίπτει στις γειτονιές πολλών σημείων πυρήνα.
- **Σημεία θορύβου:** Ένα σημείο θορύβου είναι οποιοδήποτε σημείο που δεν είναι ούτε σημείο πυρήνα ούτε συνοριακό σημείο. Στο Σχήμα 4.8, το σημείο C είναι ένα σημείο θορύβου.

Ο Αλγόριθμος DBSCAN

Λαμβάνοντας υπόψη τους προηγούμενους ορισμούς των πυρήνων, των συνοριακών σημείων και των σημείων θορύβου, ο αλγόριθμος DBSCAN μπορεί να περιγραφεί ανεπίσημα ως εξής. Οποιαδήποτε δύο σημεία πυρήνα που είναι αρκετά κοντά –σε απόσταση Eps μεταξύ τους– τοποθετούνται στην ίδια ομάδα. Παρομοίως, κάθε συνοριακό σημείο που είναι αρκετά κοντά σε ένα σημείο πυρήνα τοποθετείται στην ίδια ομάδα με τον πυρήνα. (Οι δεσμοί πρέπει να επιλυθούν εάν ένα σημείο συνόρων είναι κοντά στα σημεία πυρήνα από διαφορετικές ομάδες.) Τα σημεία θορύβου απορρίπτονται. Οι επίσημες λεπτομέρειες δίνονται στον παρακάτω Αλγόριθμο.

DBSCAN algorithm.
1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points within a distance Eps of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.

Πολυπλοκότητα του χρόνου και του χώρου

Η βασική χρονική πολυπλοκότητα του αλγορίθμου DBSCAN είναι *O(m × χρόνος για να εντοπιστούν τα σημεία στην γειτονιά Eps)*, όπου *m* είναι ο αριθμός των σημείων. Στη χειρότερη περίπτωση, αυτή η πολυπλοκότητα είναι $O(m^2)$. Εντούτοις, σε χώρους λίγων διαστάσεων (ειδικά 2D χώρο), δομές δεδομένων όπως kd-δέντρα επιτρέπουν την αποτελεσματική ανάκτηση όλων των σημείων μέσα σε μια δεδομένη απόσταση ενός συγκεκριμένου σημείου και η πολυπλοκότητα του χρόνου μπορεί να είναι τόσο χαμηλή όσο *O(m logm)* μέση

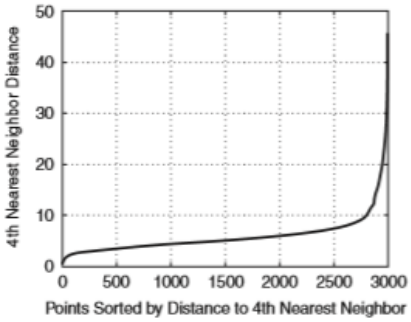
περίπτωση. Η απαίτηση χώρου του DBSCAN, ακόμη και για τα δεδομένα πολλών διαστάσεων, είναι *O(m)*, επειδή είναι απαραίτητο να διατηρείται μόνο ένα μικρό ποσό δεδομένων για κάθε σημείο, δηλαδή την ετικέτα συστάδας και τον προσδιορισμό κάθε σημείου ως πυρήνα, συνοριακό σημείο, ή σημείο θορύβου.

Επιλογή των παραμέτρων DBSCAN

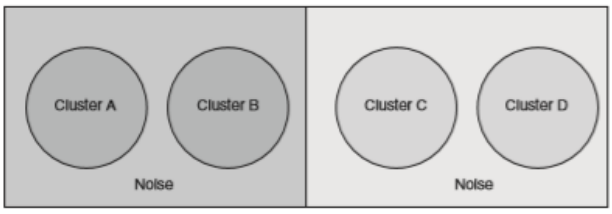
Υπάρχει, φυσικά, το ζήτημα του τρόπου προσδιορισμού των παραμέτρων Eps και MinPts. Η βασική προσέγγιση είναι να εξετάσουμε τη συμπεριφορά της απόστασης από ένα σημείο στον k πιο κοντινό γείτονα, το οποίο θα ονομάσουμε το k-dist. Για τα σημεία που ανήκουν σε κάποια συστάδα, η τιμή του k-dist θα είναι μικρή εάν το k δεν είναι μεγαλύτερο από το μέγεθος της συστάδας. Σημειώνεται ότι θα υπάρξουν κάποιες παραλλαγές, ανάλογα με την πυκνότητα της συστάδας και την τυχαία κατανομή των σημείων, αλλά κατά μέσο όρο, το εύρος της διακύμανσης δεν θα είναι τεράστιο εάν οι πυκνότητες των συστάδων δεν είναι ριζικά διαφορετικές. Ωστόσο, για σημεία που δεν βρίσκονται σε συστάδα, όπως σημεία θορύβου, το k-dist θα είναι σχετικά μεγάλο. Επομένως, αν υπολογίσουμε το k-dist για όλα τα σημεία δεδομένων για κάποια k, τα ταξινομήσουμε με αυξανόμενη σειρά και στη συνέχεια σχεδιάσουμε τις ταξινομημένες τιμές, περιμένουμε να δούμε μια απότομη μεταβολή στην τιμή του k-dist που αντιστοιχεί σε ένα κατάλληλο τιμή του Eps. Αν επιλέξουμε αυτήν την απόσταση ως την παράμετρο Eps και πάρουμε την τιμή k ως την παράμετρο MinPts, τότε τα σημεία για τα οποία το k-dist είναι μικρότερο από το Eps θα επισημαίνονται ως πυρήνα, ενώ άλλα σημεία θα επισημαίνονται ως θόρυβος ή συνοριακά σημεία.

Το Σχήμα 4.9. δείχνει ένα δείγμα του συνόλου δεδομένων, ενώ το γράφημα k-dist για τα δεδομένα δίνεται στο Σχήμα 4.10. Η τιμή του Eps που καθορίζεται με αυτόν τον τρόπο εξαρτάται από το k, αλλά δεν μεταβάλλεται δραματικά καθώς το k αλλάζει. Εάν η τιμή του k είναι πολύ μικρή, τότε ακόμη και ένας μικρός αριθμός κοντινών σημείων, που είναι θόρυβος ή ακραίες τιμές, θα χαρακτηριστούν εσφαλμένα ως συστάδες. Εάν η τιμή του k είναι πολύ μεγάλη, τότε μικρές συστάδες (μεγέθους μικρότερο του k) είναι πιθανό να χαρακτηριστούν ως θόρυβος. Ο αρχικός αλγόριθμος DBSCAN χρησιμοποίησε μια τιμή k = 4, η οποία φαίνεται να είναι μια λογική τιμή για τα περισσότερα δισδιάστατα σύνολα δεδομένων.

Σχήμα 4.9. Δείγμα συνόλου δεδομένων [9]



Σχήμα 4.10. K-οστό γράφημα για το δείγμα δεδομένων [9]



Σχήμα 4.11. Τέσσερεις συστάδες ενσωματωμένες στον θόρυβο [9]

Συστάδες ποικίλης πυκνότητας

Το DBSCAN μπορεί να έχει προβλήματα, εάν η πυκνότητα των συστάδων ποικίλλει ευρέως. Το Σχήμα 4.11 δείχνει τέσσερεις συστάδες ενσωματωμένες στο θόρυβο. Η πυκνότητα των συστάδων και των περιοχών θορύβου υποδεικνύεται από την τονικότητά τους. Ο θόρυβος γύρω από το ζεύγος των πυκνότερων συστάδων Α και Β έχει την ίδια πυκνότητα με τις συστάδες C και D. Για σταθερό *MinPts*, αν το κατώφλι του Eps επιλέγεται έτσι ώστε το DBSCAN να κατατάσσει τα C και D ως ξεχωριστές συστάδες, με τα σημεία που τις περιβάλλουν ως θόρυβο , τότε τα Α και Β και τα σημεία που τα περιβάλλουν θα γίνουν ένα ενιαίο συγκρότημα. Εάν το όριο του Eps έχει οριστεί έτσι ώστε ο DBSCAN να κατατάσσει τα Α και Β ως ξεχωριστές συστάδες και τα σημεία που τις περιβάλλουν χαρακτηρίζονται ως θόρυβος, τότε τα C, D και τα σημεία που τις περιβάλλουν θα σημειωθούν επίσης ως θόρυβος.

Πλεονεκτήματα και αδυναμίες

Επειδή ο DBSCAN χρησιμοποιεί έναν ορισμό πυκνότητας μιας συστάδας, είναι σχετικά ανθεκτικός στον θόρυβο και μπορεί να χειριστεί συστάδες αυθαίρετων σχημάτων και μεγεθών. Έτσι, ο DBSCAN μπορεί να βρει πολλές συστάδες που δεν μπορούσαν να βρεθούν χρησιμοποιώντας K-means, όπως αυτές του Σχήματος 4.9. Όπως αναφέρθηκε προηγουμένως, ωστόσο, ο DBSCAN έχει προβλήματα όταν οι συστάδες έχουν πολύ διαφορετικές πυκνότητες. Έχει επίσης πρόβλημα με τα δεδομένα μεγάλης διάστασης, επειδή η πυκνότητα είναι δυσκολότερη για τον προσδιορισμό αυτών των δεδομένων. Τέλος, ο DBSCAN μπορεί να είναι ακριβός όταν ο υπολογισμός των πλησιέστερων γειτόνων απαιτεί υπολογισμό όλων των ζευγών γειτνιάσεων, όπως συμβαίνει συνήθως για τα δεδομένα μεγάλης διάστασης.

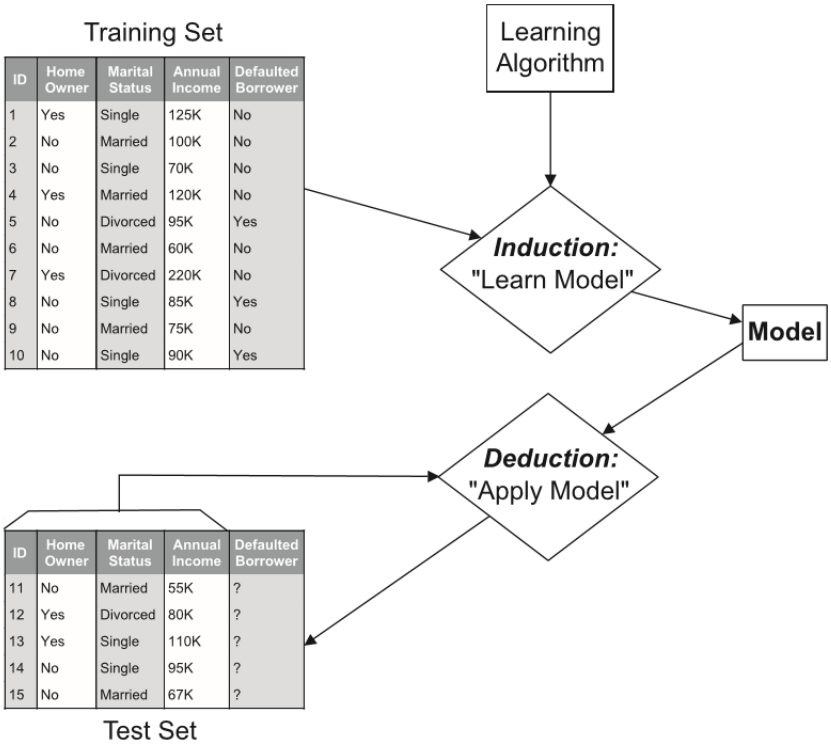
4.2. Ταξινόμηση-Classification

Μια τεχνική ταξινόμησης (ή ταξινομητής) είναι μια συστηματική προσέγγιση για την κατασκευή μοντέλων ταξινόμησης από ένα σύνολο δεδομένων εισόδου. Κάθε τεχνική χρησιμοποιεί έναν αλγόριθμο μάθησης για τον εντοπισμό ενός μοντέλου που προσαρμόζεται καλύτερα στη σχέση μεταξύ του συνόλου χαρακτηριστικών και της ετικέτας κλάσης των δεδομένων εισόδου. Το μοντέλο που παράγεται από έναν αλγόριθμο μάθησης θα πρέπει να ταιριάζει καλά στα δεδομένα εισόδου, αλλά και να προβλέπει σωστά τις ετικέτες κλάσης των στοιχείων που δεν έχει ξαναδεί ποτέ. Ως εκ τούτου, ένας βασικός στόχος του αλγορίθμου μάθησης είναι η δημιουργία μοντέλων με καλή ικανότητα γενίκευσης.

Το σχήμα 4.12 δείχνει μια γενική προσέγγιση για την επίλυση προβλημάτων ταξινόμησης. Πρώτον, πρέπει να παρέχεται ένα σύνολο εκπαίδευσης (training set) που αποτελείται από αρχεία των οποίων οι ετικέτες κλάσης είναι γνωστές. Το σύνολο εκπαίδευσης χρησιμοποιείται για την κατασκευή ενός μοντέλου ταξινόμησης, το οποίο στη συνέχεια εφαρμόζεται στο σύνολο επαλήθευσης (test set), το οποίο αποτελείται από εγγραφές με άγνωστες ετικέτες κλάσης.

Η αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης βασίζεται στις μετρήσεις του αριθμού των στοιχείων του συνόλου επαλήθευσης που προβλέφθηκαν σωστά ή λανθασμένα από το μοντέλο. Αυτές οι μετρήσεις καταγράφονται σε πίνακα που είναι γνωστός ως μήτρα σύγχυσης (confusion matrix). Ο Πίνακας 4.13 απεικονίζει τη μήτρα σύγχυσης για ένα δυαδικό πρόβλημα κατηγοριοποίησης. Κάθε εγγραφή f_{ij} σε αυτόν τον πίνακα υποδηλώνει τον αριθμό των εγγραφών από την κλάση i που προβλέπεται να είναι της κλάσης j. Για παράδειγμα, το f_{01} είναι ο αριθμός των εγγραφών από την κλάση 0 που έχουν προβλεφθεί λανθασμένα ως

κλάση 1. Με βάση τις καταχωρίσεις στο μήτρα σύγχυσης, ο συνολικός αριθμός σωστών προβλέψεων του μοντέλου είναι ($f_{11} + f_{00}$) και ο συνολικός αριθμός λανθασμένων προβλέψεων είναι ($f_{10} + f_{01}$).



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	55K	?
12	Yes	Divorced	80K	?
13	Yes	Single	110K	?
14	No	Single	95K	?
15	No	Married	67K	?

Σχήμα 4.12. Γενικό πλαίσιο για τη δημιουργία μοντέλου ταξινόμησης. [9]

		Predicted Class	
		Class = 1	Class = 0
Actual Class	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

Πίνακας 4.13. Μήτρα σύγχυσης για ένα δυαδικό πρόβλημα κατηγοριοποίησης [9]

Παρόλο που μια μήτρα σύγχυσης παρέχει τις πληροφορίες που απαιτούνται για να προσδιοριστεί πόσο καλά εκτελείται ένα μοντέλο ταξινόμησης, η περίληψη αυτών των πληροφοριών σε έναν μόνο αριθμό καθιστά πιο βολικό να συγκρίνουμε τη σχετική απόδοση των διαφόρων μοντέλων. Αυτό μπορεί να γίνει χρησιμοποιώντας μια μετρική αξιολόγησης όπως η ακρίβεια (accuracy), η οποία υπολογίζεται με τον ακόλουθο τρόπο:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Για δυαδικά προβλήματα ταξινομήσεως, η ακρίβεια ενός μοντέλου δίνεται από

$$Accuracy = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Το ποσοστό σφάλματος (error rate) είναι μια άλλη σχετική μετρική, η οποία ορίζεται ως εξής για δυαδικά προβλήματα ταξινόμησης:

$$Error\ rate = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Οι αλγόριθμοι εκμάθησης των περισσότερων τεχνικών ταξινόμησης έχουν σχεδιαστεί για να εκπαιδεύουν μοντέλα που επιτυγχάνουν την υψηλότερη ακρίβεια ή ισοδύναμα το χαμηλότερο ποσοστό σφάλματος όταν εφαρμόζονται στο σετ δοκιμών.

4.2.1. Ταξινομητής δέντρου απόφασης-Decision Tree Classifier

Για να περιγραφεί ο τρόπος λειτουργίας ενός δέντρου απόφασης, θα εξεταστεί το πρόβλημα της ταξινόμησης του διαχωρισμού των θηλαστικών από μη θηλαστικά χρησιμοποιώντας το σύνολο δεδομένων σπονδυλωτών που παρουσιάζεται στον Πίνακα 4.14. Ας υποθέσουμε ότι ένα νέο είδος ανακαλύπτεται από επιστήμονες. Πώς μπορούμε να πούμε εάν είναι θηλαστικό ή μη θηλαστικό; Μια προσέγγιση είναι να δημιουργηθεί μια σειρά ερωτήσεων σχετικά με τα χαρακτηριστικά του είδους. Το πρώτο ερώτημα που μπορούμε να θέσουμε είναι εάν το είδος είναι ψυχρόαιμο ή θερμόαιμο. Εάν είναι ψυχρόαιμο, τότε δεν είναι σίγουρα θηλαστικό. Διαφορετικά, είναι είτε πτηνό είτε θηλαστικό. Στην τελευταία περίπτωση, πρέπει να θέσουμε μια επόμενη ερώτηση: Τα θηλυκά του είδους γεννούν τα παιδιά τους; Εκείνα που γεννιούνται είναι απολύτως θηλαστικά, ενώ εκείνα που δεν είναι, πιθανό να είναι μη θηλαστικοί (με εξαίρεση τα θηλαστικά ωοτοκίας όπως ο πλατύποδας και ο κελυφωτός μυρηγκοφάγος).

Vertebrate Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Πίνακας 4.14. [9]

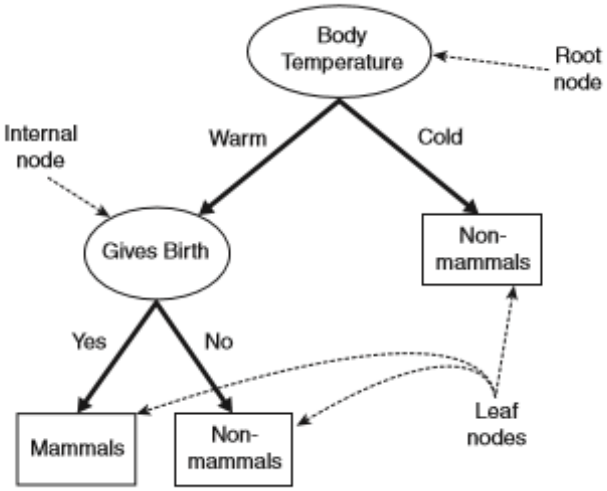
Το προηγούμενο παράδειγμα δείχνει πώς μπορούμε να λύσουμε ένα πρόβλημα ταξινόμησης ζητώντας μια σειρά από προσεκτικά δημιουργημένες ερωτήσεις σχετικά με τα χαρακτηριστικά στοιχείων δοκιμής. Κάθε φορά που λαμβάνουμε μια απάντηση, θα μπορούσαμε να ζητήσουμε μια επόμενη ερώτηση μέχρι να μπορέσουμε να αποφασίσουμε οριστικά για την ετικέτα της κατηγορίας της. Η σειρά ερωτήσεων και οι πιθανές απαντήσεις τους μπορούν να οργανωθούν σε μια ιεραρχική δομή που ονομάζεται δέντρο απόφασης. Το σχήμα 4.15 δείχνει ένα παράδειγμα του δέντρου απόφασης για το πρόβλημα ταξινομήσεως θηλαστικών. Το δέντρο έχει τρεις τύπους κόμβων:

- **Κόμβος ρίζας (root node)**, χωρίς εισερχόμενες συνδέσεις-κλαδιά και μηδέν ή περισσότερες εξερχόμενες συνδέσεις.

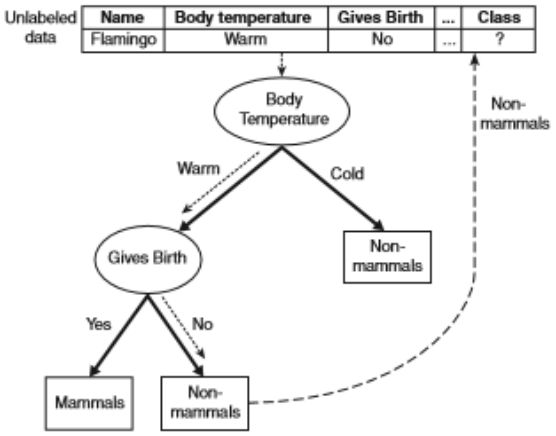
- **Εσωτερικοί κόμβοι (internal nodes)**, κάθε ένας από τους οποίους έχει ακριβώς μια εισερχόμενη σύνδεση και δύο ή περισσότερες εξερχόμενες συνδέσεις.
- **Φύλλα ή τερματικοί κόμβοι (Leaf or terminal nodes)**, ο καθένας από τους οποίους έχει ακριβώς μια εισερχόμενη σύνδεση και δεν έχει εξερχόμενες συνδέσεις.

Κάθε φύλλο στο δέντρο απόφασης συσχετίζεται με μια ετικέτα κλάσης. Οι μη τερματικοί κόμβοι, στους οποίους περιλαμβάνονται οι ρίζες και οι εσωτερικοί κόμβοι, περιέχουν συνθήκες δοκιμής χαρακτηριστικών που τυπικά ορίζονται χρησιμοποιώντας ένα μοναδικό χαρακτηριστικό. Κάθε πιθανή έκβαση της κατάστασης δοκιμής χαρακτηριστικών σχετίζεται με ακριβώς ένα παιδί αυτού του κόμβου. Για παράδειγμα, η ρίζα του δέντρου που φαίνεται στο Σχήμα 4.15 χρησιμοποιεί το χαρακτηριστικό «Θερμοκρασία σώματος» για να προσδιορίσει μία συνθήκη που έχει δύο αποτελέσματα, «θερμόαιμα» και «ψυχρόαιμα», με αποτέλεσμα δύο θυγατρικούς κόμβους.

Δεδομένου ενός δέντρου απόφασης, η ταξινόμηση ενός δοκιμαστικού στοιχείου είναι απλή. Ξεκινώντας από τον κόμβο ρίζας, εφαρμόζουμε τη δοκιμαστική συνθήκη του χαρακτηριστικού και ακολουθούμε το κατάλληλο κλαδί βάσει του αποτελέσματος της δοκιμής. Αυτό θα μας οδηγήσει είτε σε έναν άλλο εσωτερικό κόμβο, για τον οποίο εφαρμόζεται μια νέα δοκιμαστική συνθήκη χαρακτηριστικών, είτε σε ένα φύλλο. Μόλις φτάσουμε σε ένα φύλλο, αναθέτουμε την ετικέτα κλάσης που συσχετίζεται με τον κόμβο στην δοκιμαστικό στοιχείο. Για παράδειγμα, το Σχήμα 4.16 ανιχνεύει τη διαδρομή που χρησιμοποιείται για την πρόβλεψη της ετικέτας κλάσης ενός φλαμίνγκο. Η διαδρομή τερματίζεται σε ένα φύλλο που έχει επισημανθεί ως «μη θηλαστικά».



Σχήμα 4.15. Ένα δέντρο απόφασης για το πρόβλημα ταξινόμησης των θηλαστικών [9]



Σχήμα 4.16. Ταξινόμηση ενός μη χαρακτηρισμένου στοιχείου. Η διακεκομμένη γραμμή απεικονίζει το αποτέλεσμα της εφαρμογής διάφορων συνθηκών χαρακτηριστικών για το μη χαρακτηρισμένο στοιχείο, το οποίο σταδιακά χαρακτηρίζεται ως «μη θηλαστικό». [9]

Ο αλγόριθμος του Hunt

Μια από τις πρώτες μεθόδους για τη δημιουργία δέντρου απόφασης είναι ο αλγόριθμος του Hunt, που είναι η βάση για πολλές τρέχουσες υλοποιήσεις, συμπεριλαμβανομένων των **ID3**, **C4.5** και **CART**.

Στον αλγόριθμο του Hunt, ένα δέντρο αποφάσεων αναπτύσσεται με αναδρομικό τρόπο, χωρίζοντας τα στοιχεία του συνόλου εκπαίδευσης σε διαδοχικά καθαρότερα υποσύνολα. Ας θεωρήσουμε ότι το D_t είναι το σύνολο των στοιχείων εκπαίδευσης που σχετίζονται με τον κόμβο t και $y = \{y_1, y_2, ..., y_c\}$ είναι οι ετικέτες κλάσης. Το παρακάτω είναι ένας επαναληπτικός ορισμός του αλγόριθμου του Hunt.

> **Βήμα 1:** Εάν όλα τα αρχεία σε D_t ανήκουν στην ίδια κλάση y_t, τότε t είναι ένας κόμβος φύλλων που έχει επισημανθεί ως y_t.
>
> **Βήμα 2:** Εάν το D_t περιέχει εγγραφές που ανήκουν σε περισσότερες από μία κλάσεις, επιλέγεται μια συνθήκη δοκιμής χαρακτηριστικών για να χωρίσει τις εγγραφές σε μικρότερα υποσύνολα. Για κάθε αποτέλεσμα της δοκιμαστικής συνθέσεως δημιουργείται ένας θυγατρικός κόμβος και τα στοιχεία στο Dt διανέμονται βάσει των αποτελεσμάτων. Ο αλγόριθμος εφαρμόζεται στη συνέχεια σε κάθε θυγατρικό κόμβο.

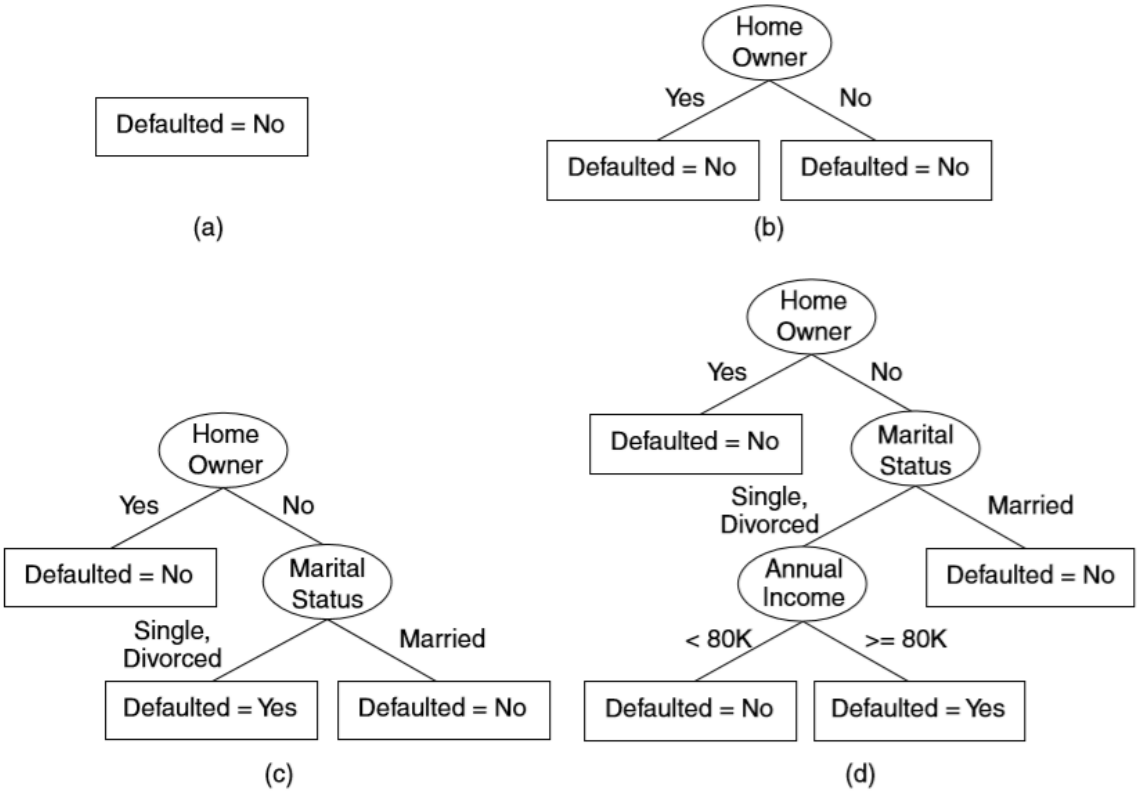
Για να αποτυπωθεί πώς λειτουργεί ο αλγόριθμος, εξετάζουμε το σύνολο εκπαίδευσης που παρουσιάζεται στον Πίνακα 4.17 για το πρόβλημα κατηγοριοποίησης δανειοληπτών, όπου εξετάζεται αν θα αποπληρώσουν το δάνειο ή όχι και συνεπώς, αν θα τους χορηγηθεί. Ας υποθέσουμε ότι εφαρμόζουμε τον αλγόριθμο του Hunt για να καταχωρήσουμε τα δεδομένα εκπαίδευσης.

Το αρχικό δέντρο για το πρόβλημα κατηγοριοποίησης περιέχει έναν μοναδικό κόμβο με την ετικέτα κατηγορίας Defaulted = No (βλ. Σχήμα 4.18(a)), πράγμα που σημαίνει ότι οι περισσότεροι δανειολήπτες αποπλήρωσαν με επιτυχία τα δάνειά τους. Το δέντρο, ωστόσο, πρέπει να επαναληφθεί αφού ο κόμβος ρίζας περιέχει εγγραφές και από τις δύο κατηγορίες. Τα αρχεία διακρίνονται στη συνέχεια σε μικρότερα υποσύνολα με βάση την τιμή του χαρακτηριστικού Home Owner, όπως φαίνεται στο σχήμα 4.18(b). Η αιτιολόγηση αυτού του χαρακτηριστικού θα συζητηθεί αργότερα. Προς το παρόν, θα υποθέσουμε ότι αυτό είναι το καλύτερο κριτήριο για τον διαχωρισμό των δεδομένων σε αυτό το σημείο. Ο αλγόριθμος του Hunt

εφαρμόζεται στη συνέχεια αναδρομικά σε θυγατρικό κόμβο του κόμβου ρίζας. Από το σύνολο εκπαίδευσης που δίνεται στον Πίνακα 4.17, παρατηρούμε ότι όλοι οι δανειολήπτες που είναι ιδιοκτήτες σπιτιού αποπληρώνουν με επιτυχία τα δάνειά τους. Ο αριστερός θυγατρικός κόμβος της ρίζας είναι επομένως ένας κόμβος φύλλων με την ένδειξη Defaulted = No (βλ. Σχήμα 4.18(b)). Για τον σωστό θυγατρικό κόμβο, πρέπει να συνεχίσουμε να εφαρμόζουμε το αναδρομικό βήμα του αλγορίθμου του Hunt μέχρις ότου όλα τα αρχεία να ανήκουν στην ίδια τάξη. Τα δέντρα που προκύπτουν από κάθε επαναληπτικό βήμα παρουσιάζονται στα σχήματα 4.18(c) και (d).

ID	Home Owner	Marital Status	Annual Income	Defaulted?
1	Yes	Single	125000	No
2	No	Married	100000	No
3	No	Single	70000	No
4	Yes	Married	120000	No
5	No	Divorced	95000	Yes
6	No	Single	60000	No
7	Yes	Divorced	220000	No
8	No	Single	85000	Yes
9	No	Married	75000	No
10	No	Single	90000	Yes

Πίνακας 4.17. Δείμα δεδομένων για πρόβλημα ταξινόμησης δανειοληπτών [9]



Σχήμα 4.18. Ο Αλγόριθμος του Hunt για τη δημιουργία δέντρων απόφασης [9]

Ο αλγόριθμος του Hunt θα λειτουργήσει εάν κάθε συνδυασμός των τιμών των χαρακτηριστικών υπάρχει στα δεδομένα εκπαίδευσης και κάθε συνδυασμός έχει μια μοναδική ετικέτα κλάσης. Αυτές οι υποθέσεις είναι

πολύ αυστηρές για χρήση στις περισσότερες πρακτικές καταστάσεις. Επιπλέον, απαιτούνται πρόσθετοι όροι για την αντιμετώπιση των ακόλουθων περιπτώσεων:

1. Είναι δυνατόν ορισμένοι από τους θυγατρικούς κόμβους που δημιουργήθηκαν στο Βήμα 2 να είναι κενοί. δηλαδή, δεν υπάρχουν αρχεία που να συνδέονται με αυτούς τους κόμβους. Αυτό μπορεί να συμβεί αν κανένα από τα αρχεία εκπαίδευσης δεν έχει το συνδυασμό των χαρακτηριστικών τιμών που σχετίζονται με τέτοιους κόμβους. Σε αυτή την περίπτωση ο κόμβος δηλώνεται ένας κόμβος φύλλου με την ίδια ετικέτα κατηγορίας με την τάξη πλειοψηφίας των αρχείων κατάρτισης που σχετίζονται με τον γονικό κόμβο του.

2. Στο Βήμα 2, εάν όλα τα αρχεία που σχετίζονται με το Dt έχουν τις ίδιες τιμές χαρακτηριστικών (εκτός από την ετικέτα κλάσης), τότε δεν είναι δυνατό να χωρίσετε αυτές τις εγγραφές περαιτέρω. Σε αυτή την περίπτωση, ο κόμβος δηλώνεται ως κόμβος φύλλου με την ίδια ετικέτα κλάσης με την τάξη πλειοψηφίας των αρχείων κατάρτισης που σχετίζονται με αυτόν τον κόμβο.

Μέτρα για την επιλογή μιας συνθήκης δοκιμής ιδιοτήτων

Υπάρχουν πολλά μέτρα που μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της καλής συνθήκης για την δοκιμή χαρακτηριστικών. Αυτά τα μέτρα προσπαθούν να δώσουν προτεραιότητα στη συνθήκη δοκιμής χαρακτηριστικών που χωρίζει τα στοιχεία, που θα οδηγηθούν στους θυγατρικούς κόμβους εκπαίδευσης, σε καθαρότερα υποσύνολα, τα οποία συνήθως έχουν τις ίδιες ετικέτες κατηγορίας. Είναι χρήσιμη η δημιουργία καθαρότερων κόμβων, δεδομένου του ότι ένας κόμβος που έχει όλα τα στοιχεία εκπαίδευσης από την ίδια κλάση, δεν χρειάζεται να επεκταθεί περαιτέρω. Αντίθετα, ένας μη κάθαρτος κόμβος που περιέχει στοιχεία εκπαίδευσης από πολλαπλές κλάσεις είναι πιθανό να απαιτήσει αρκετά επίπεδα εσωτερικών κόμβων, αυξάνοντας έτσι σημαντικά το βάθος του δέντρου. Τα μεγαλύτερα δέντρα είναι λιγότερο επιθυμητά, καθώς είναι πιο ευαίσθητα στην υπερπροσαρμογή μοντέλου με τα δεδομένα εκπαίδευσης (overfitting), γεγονός που μπορεί να υποβαθμίσει την απόδοση ταξινόμησης στα δεδομένα επικύρωσης. Είναι επίσης πιο δύσκολο εξηγηθούν και συνεπάγονται περισσότερο χρόνο εκπαίδευσης και επικύρωσης, σε σύγκριση με μικρότερα δέντρα.

Μέτρα προσμείξεων για έναν κόμβο

Η μη καθαρότητα (impurity) ενός κόμβου μετρά πόσο διαφορετικές είναι οι ετικέτες των κλάσεων για τα στοιχεία δεδομένων που ανήκουν σε έναν κοινό κόμβο. Ακολουθούν παραδείγματα μέτρων που μπορούν να χρησιμοποιηθούν για την αξιολόγηση της μη καθαρότητας ενός κόμβου t:

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) \log_2 p_i(t)$$

$$Gini\ index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

$$Classification\ error = 1 - \max_i [p_i(t)]$$

όπου *pi (t)* είναι η σχετική συχνότητα των στοιχείων εκπαίδευσης που ανήκουν στην κλάση *i* στον κόμβο *t*, *c* είναι ο συνολικός αριθμός των κλάσεων και *0 log₂0 = 0* σε υπολογισμούς εντροπίας. Και τα τρία μέτρα δίνουν

μια τιμή μηδενικής πρόσμειξης εάν ένας κόμβος περιέχει στοιχεία από μία κλάση και μέγιστης μη καθαρότητας αν ο κόμβος έχει ίσο ποσοστό συμμετοχών από πολλαπλές κλάσεις.

Το Σχήμα 4.19 συγκρίνει το σχετικό μέγεθος των μέτρων μη καθαρότητας όταν εφαρμόζεται σε δυαδικά προβλήματα ταξινόμησης. Δεδομένου ότι υπάρχουν μόνο δύο κατηγορίες,
$p_0(t) + p_1(t) = 1$. Ο οριζόντιος άξονας p αναφέρεται στο κλάσμα των στοιχείων των δεδομένων που ανήκουν σε μία από τις δύο κατηγορίες. Παρατηρήστε ότι και τα τρία μέτρα επιτυγχάνουν τη μέγιστη τιμή τους όταν η κατανομή τάξης είναι ομοιόμορφη
(δηλαδή $p_0(t) = p_1(t) = 0.5$) και ελάχιστη τιμή όταν όλα τα στοιχεία ανήκουν σε μία κλάση (δηλαδή το $p_0(t)$ ή το $p_1(t)$ ισούται με 1). Τα παρακάτω παραδείγματα επεξηγούν τον τρόπο με τον οποίο οι τιμές των μέτρων ακαθαρσίας ποικίλλουν καθώς αλλάζουμε την κατανομή της κλάσης.

Με βάση αυτούς τους υπολογισμούς, ο κόμβος N_1 έχει τη χαμηλότερη τιμή μη καθαρότητας, ακολουθούμενος από τους N_2 και N_3. Αυτό το παράδειγμα, μαζί με το Σχήμα 4.17, δείχνει τη συνοχή μεταξύ των μέτρων μη καθαρότητας, δηλαδή εάν ένας κόμβος N_1 έχει χαμηλότερη εντροπία από τον κόμβο N_2, τότε ο δείκτης Gini και το ποσοστό σφάλματος του N_1 θα είναι επίσης χαμηλότερα από αυτά του N_2. Παρά τη συμφωνία τους, το χαρακτηριστικό που επιλέγεται ως κριτήριο διαίρεσης από τα μέτρα προσμείξεων μπορεί να εξακολουθεί να είναι διαφορετικό.

Node N_1	Count
Class=0	0
Class=1	6

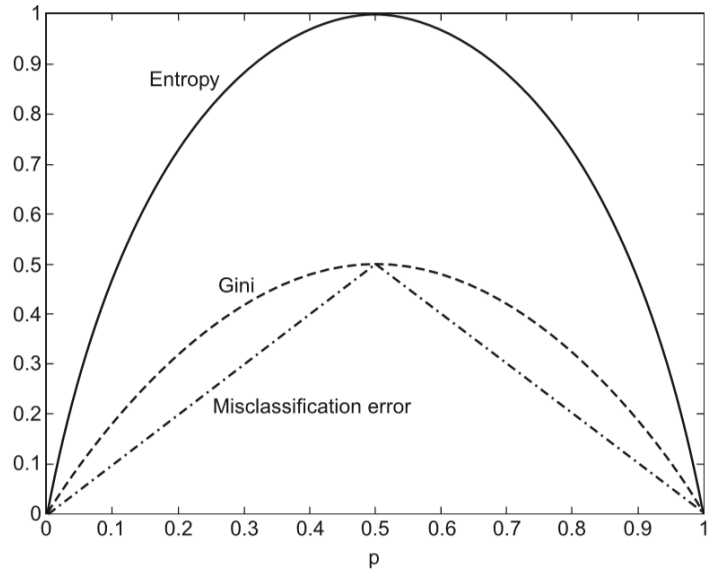
$\text{Gini} = 1 - (0/6)^2 - (6/6)^2 = 0$
$\text{Entropy} = -(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$
$\text{Error} = 1 - \max[0/6, 6/6] = 0$

Node N_2	Count
Class=0	1
Class=1	5

$\text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$
$\text{Entropy} = -(1/6)\log_2(1/6) - (5/6)\log_2(5/6) = 0.650$
$\text{Error} = 1 - \max[1/6, 5/6] = 0.167$

Node N_3	Count
Class=0	3
Class=1	3

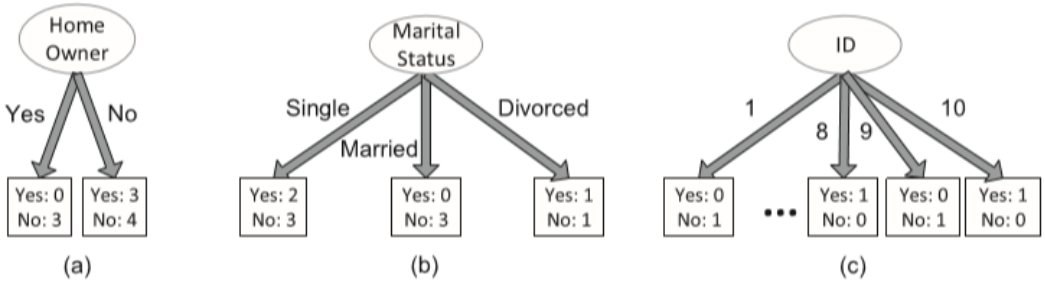
$\text{Gini} = 1 - (3/6)^2 - (3/6)^2 = 0.5$
$\text{Entropy} = -(3/6)\log_2(3/6) - (3/6)\log_2(3/6) = 1$
$\text{Error} = 1 - \max[3/6, 3/6] = 0.5$



Σχήμα 4.19. Σύγκριση του σχετικού μεγέθους των μέτρων μη καθαρότητας όταν εφαρμόζεται σε δυαδικά προβλήματα ταξινόμησης [9]

Λόγος κέρδους-Gain Ratio

Ένας δυνητικός περιορισμός των μέτρων πρόσμειξης όπως η εντροπία και ο δείκτης Gini είναι ότι τείνουν να ευνοούν ποιοτικά χαρακτηριστικά με μεγάλο αριθμό ξεχωριστών τιμών. Το Σχήμα 4.20 παρουσιάζει τρία χαρακτηριστικά υποψήφιου για τη διαίρεση του συνόλου δεδομένων που δίνεται στον Πίνακα 4.17. Το χαρακτηριστικό «Οικογενειακή κατάσταση» είναι μια καλύτερη επιλογή από την ιδιότητα «Ιδιοκτήτης ακινήτου», επειδή παρέχει μεγαλύτερο κέρδος πληροφορίας. Ωστόσο, αν τα συγκρίνουμε με το «ID», το τελευταίο παράγει τον καθαρότερο διαχωρισμό με το μέγιστο κέρδος πληροφορίας, αφού η σταθμισμένη εντροπία και ο δείκτης Gini είναι ίσα με μηδέν για τα παιδιά του. Ωστόσο, το «ID» δεν είναι ένα καλό χαρακτηριστικό για το διαχωρισμό επειδή έχει μια μοναδική τιμή για κάθε περίπτωση.



Σχήμα 4.20. Παραδείγματα χαρακτηριστικών υποψήφιου δανειολήπτη για τη διαίρεση του συνόλου δεδομένων [9]

Παρόλο που μια δοκιμαστική συνθήκη που περιλαμβάνει το «ID» θα ταξινομεί με ακρίβεια κάθε περίπτωση στα δεδομένα εκπαίδευσης, δεν μπορούμε να χρησιμοποιήσουμε μια τέτοια δοκιμαστική συνθήκη σε νέες περιπτώσεις δοκιμών με τιμές «ID» που δεν έχουν εμφανιστεί κατά τη διάρκεια της εκπαίδευσης. Αυτό το παράδειγμα υποδεικνύει ότι μόνο μια χαμηλή τιμή μη καθαρότητας είναι ανεπαρκής για να διαπιστώσει μια καλή κατάσταση δοκιμής χαρακτηριστικών για έναν κόμβο. Ο μεγαλύτερος αριθμός θυγατρικών κόμβων μπορεί να καταστήσει ένα δέντρο αποφάσεων πιο περίπλοκο και συνεπώς πιο επιρρεπές σε υπερπροσαρμογή. Ως εκ τούτου, ο αριθμός των παιδιών που παράγονται από το χαρακτηριστικό διάσπασης θα πρέπει επίσης να ληφθεί υπόψη κατά την επιλογή της καλύτερης κατάστασης δοκιμής χαρακτηριστικών. Υπάρχουν δύο τρόποι για να ξεπεραστεί αυτό το πρόβλημα. Ένας τρόπος είναι να δημιουργηθούν μόνο δυαδικά δέντρα αποφάσεων, αποφεύγοντας έτσι τη δυσκολία χειρισμού χαρακτηριστικών με διαφορετικό αριθμό κλαδιών. Αυτή η στρατηγική χρησιμοποιείται από τους ταξινομείς των δέντρων αποφάσεων όπως το CART. Ένας άλλος τρόπος είναι να τροποποιηθεί το κριτήριο διαίρεσης για να ληφθεί υπόψη ο αριθμός των κλαδιών που παράγονται από το χαρακτηριστικό. Για παράδειγμα, στον αλγόριθμο δέντρων αποφάσεων C4.5, χρησιμοποιείται ένα μέτρο γνωστό ως λόγος κέρδους για την αντιστάθμιση χαρακτηριστικών που παράγουν ένα μεγάλο αριθμό θυγατρικών κόμβων. Το μέτρο αυτό υπολογίζεται ως εξής:

$$Gain\ ratio = \frac{\Delta_{info}}{Split\ Info} = \frac{Entropy(Parent) - \sum_{i=1}^{k} \frac{N(v_i)}{N} Entropy(v_i)}{-\sum_{i=1}^{k} \frac{N(v_i)}{N} \log_2 \frac{N(v_i)}{N}}$$

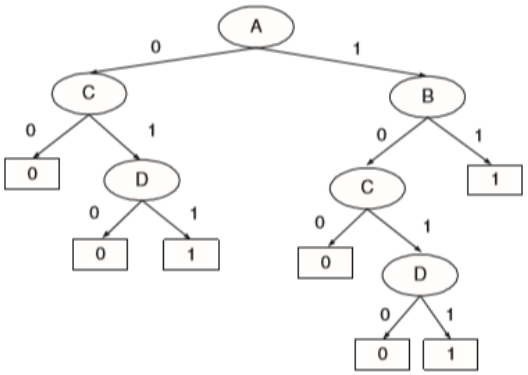
όπου *N(vi)* είναι ο αριθμός των περιπτώσεων που αντιστοιχούν στον κόμβο *vi* και *k* είναι ο συνολικός αριθμός των διαχωρισμών. Το *Split info* μετρά την εντροπία του διαχωρισμού ενός κόμβου στους θυγατρικούς του κόμβους και αξιολογεί εάν ο διαχωρισμός έχει ως αποτέλεσμα μεγαλύτερο αριθμό θυγατρικών κόμβων ίσου μεγέθους ή όχι. Για παράδειγμα, εάν κάθε κλαδί έχει τον ίδιο αριθμό περιπτώσεων, τότε $\forall i : N(v_i) / N = 1 / k$

και η πληροφορία διάσπασης θα είναι ίση με το $\log_2 k$. Έτσι, αν ένα χαρακτηριστικό παράγει ένα μεγάλο αριθμό κλαδιών, το Split Info του είναι επίσης μεγάλο, γεγονός που με τη σειρά του μειώνει τον λόγο κέρδους (Gain Ratio).

Χαρακτηριστικά των ταξινομητών δέντρων αποφάσεων

1. **Εφαρμογή:** Τα δέντρα αποφάσεων αποτελούν μη παραμετρική προσέγγιση για την κατασκευή μοντέλων ταξινόμησης. Αυτή η προσέγγιση δεν απαιτεί προηγούμενη παραδοχή σχετικά με την κατανομή πιθανότητας που διέπει την κλάση και τις ιδιότητες των δεδομένων και επομένως ισχύει για μια μεγάλη ποικιλία συνόλων δεδομένων. Είναι επίσης εφαρμόσιμο τόσο στα κατηγορικά όσο και στα συνεχή δεδομένα, χωρίς να απαιτείται η μετατροπή των χαρακτηριστικών σε μια κοινή αναπαράσταση μέσω της διμερισμού, της κανονικοποίησης ή της τυποποίησης. Μπορούν επίσης να αντιμετωπίσουν προβλήματα πολλαπλών κλάσεων. Ένα άλλο ελκυστικό χαρακτηριστικό των ταξινομητών δένδρων αποφάσεων είναι ότι τα προκαλούμενα δέντρα, ειδικά τα μικρότερα, είναι σχετικά εύκολα ερμηνεύσιμα. Η ακρίβεια των δένδρων είναι επίσης αρκετά συγκρίσιμη με άλλες τεχνικές ταξινόμησης για πολλά απλά σύνολα δεδομένων.
2. **Εκφραστικότητα:** Ένα δέντρο απόφασης παρέχει μια καθολική αναπαράσταση για συναρτήσεις με διακριτές τιμές. Αυτό συμβαίνει επειδή κάθε συνάρτηση διακριτών τιμών μπορεί να εκπροσωπείται ως πίνακας αντιστοίχισης, όπου κάθε μοναδικός συνδυασμός διακριτών τιμών έχει εκχωρηθεί μια ετικέτα κλάσης. Δεδομένου ότι κάθε συνδυασμός χαρακτηριστικών μπορεί να αναπαρασταθεί ως φύλλο στο δέντρο αποφάσεων, μπορούμε πάντα να εντοπίσουμε ένα δέντρο αποφάσεων του οποίου οι αντιστοιχίες των ετικετών στους κόμβους των φύλλων να ταιριάζουν με τον πίνακα αντιστοίχισης της αρχικής συνάρτησης. Τα δέντρα αποφάσεων μπορούν επίσης να βοηθήσουν στην παροχή σύντομης αναπαράστασης των συναρτήσεων, όταν ορισμένοι από τους μοναδικούς συνδυασμούς χαρακτηριστικών μπορούν να αναπαρασταθούν από το ίδιο φύλλο, όπως απεικονίζεται στο Σχήμα 4.21.

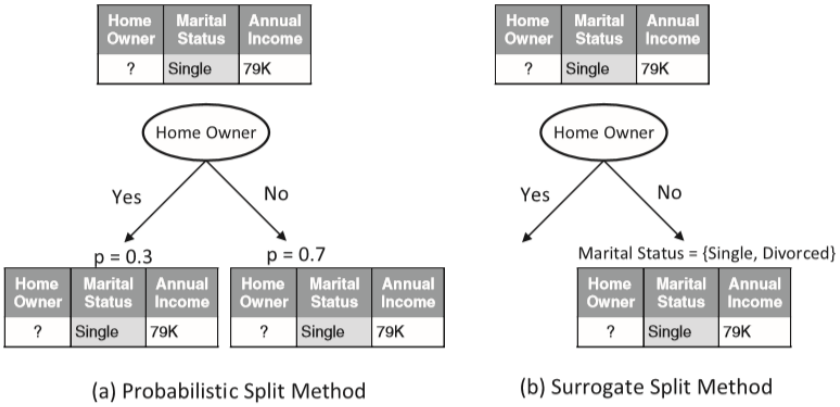
A	B	C	D	class
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1



Σχήμα 4.21. Δέντρο απόφασης της Boolean συνάρτησης (A ∧ B) ∨ (C ≤ D). [9]

3. **Υπολογιστική αποτελεσματικότητα:** Δεδομένου ότι ο αριθμός των πιθανών δέντρων αποφάσεων μπορεί να είναι πολύ μεγάλος, πολλοί αλγόριθμοι δέντρων αποφάσεων χρησιμοποιούν μια προσέγγιση βάσει ευρετικής προσέγγισης για να καθοδηγήσουν την έρευνά τους στον τεράστιο χώρο υποθέσεων. Για πολλά σύνολα δεδομένων, τέτοιες τεχνικές κατασκευάζουν γρήγορα ένα λογικά καλό δέντρο αποφάσεων ακόμα και όταν το μέγεθος του εκπαιδευτικού σετ είναι πολύ μεγάλο. Επιπλέον, μόλις δημιουργηθεί ένα δέντρο αποφάσεων, η ταξινόμηση ενός στοιχείου δοκιμής είναι εξαιρετικά γρήγορη, με τη χειρότερη περίπτωση πολυπλοκότητας του *O(w)*, όπου w είναι το μέγιστο βάθος του δέντρου.

4. **Διαχείριση των ελλειπουσών τιμών:** Ένας ταξινομητής δένδρου αποφάσεης μπορεί να χειριστεί τις τιμές χαρακτηριστικών που λείπουν με διάφορους τρόπους, τόσο στην εκπαίδευση όσο και στα σύνολα δοκιμών. Όταν υπάρχουν ελλείπουσες τιμές στο σετ εκπαίδευσης, ο ταξινομητής πρέπει να αποφασίσει ποιο κλαδί θα ακολουθήσει, εαν λείπει η τιμή ενός ενδιάμεσου κόμβου. Μια προσέγγιση, γνωστή ως probabilistic split method, η οποία χρησιμοποιείται από τον δέντρο ταξινομήσεως αποφάσεων C4.5, διανέμει το συγκεριμένο στοιχείο, σύμφωνα με την πιθανότητα το χαρακτηριστικό που λείπει να έχει μια συγκεκριμένη τιμή. Αντίθετα, ο αλγόριθμος CART χρησιμοποιεί τη μέθοδο surrogate split, όπου το στοιχείο που λείπει η τιμή του χαρακτηριστικού διαχωρισμού, εκχωρείται σε έναν από τους θυγατρικούς κόμβους με βάση την αξία ενός άλλου υφιστάμενου χαρακτηριστικού που δε λείπει, του οποίου οι διαχωρισμοί μοιάζουν περισσότερο με τους διαχωρισμούς που έγιναν για το χαρακτηριστικό που δεν υπάρχει. Το Σχήμα 4.22. δείχνει ένα παράδειγμα των δύο διαφορετικών τρόπων χειρισμού των ελλειπουσών τιμών.



Σχήμα 4.22. Μέθοδοι διαχείρισης ελλειπουσών τιμών χαρακτηριστικών σε ένα δέντρο απόφασης [9]

Άλλες στρατηγικές αντιμετώπισης των ελλειπουσών τιμών βασίζονται στην προεπεξεργασία δεδομένων, όπου η περίπτωση με ελλείπουσα τιμή είτε συμπληρώνεται π.χ. με τη μέση τιμή ή απορρίπτεται πριν εκπαιδευτεί ο ταξινομητής.

5. **Χειρισμός αλληλεπιδράσεων μεταξύ των ιδιοτήτων:** Τα χαρακτηριστικά θεωρούνται ότι αλληλεπιδρούν εάν μπορούν να διακρίθουν μεταξύ των κλάσεων όταν χρησιμοποιούνται μαζί, αλλά μεμονωμένα παρέχουν ελάχιστες ή καθόλου πληροφορίες. Λόγω της κοστοβόλας φύσης των κριτηρίων διάσπασης στα δέντρα απόφασης, τέτοια χαρακτηριστικά θα μπορούσαν να ξεπεραστούν χάρη σε άλλα χαρακτηριστικά που δεν είναι τόσο χρήσιμα. Αυτό θα μπορούσε να οδηγήσει σε πιο περίπλοκα δέντρα αποφάσεων από ό, τι είναι απαραίτητο. Ως εκ τούτου, τα δέντρα αποφάσεων μπορούν να έχουν χαμηλή απόδοση όταν υπάρχουν αλληλεπιδράσεις μεταξύ χαρακτηριστικών.
6. **Χειρισμός μη σχετικών χαρακτηριστικών:** Ένα χαρακτηριστικό δεν είναι σχετικό εάν δεν είναι χρήσιμο για την ταξινόμηση. Δεδομένου ότι τα μη σχετικά χαρακτηριστικά συνδέονται ελάχιστα με τις ετικέτες της στοχευόμενης κλάσης, θα παρέχουν ελάχιστο ή καθόλου κέρδο καθαρότητας και έτσι θα μεταφερθούν από άλλα πιο συναφή χαρακτηριστικά. Ως εκ τούτου, η παρουσία ενός μικρού αριθμού άσχετων χαρακτηριστικών δεν θα επηρεάσει τη διαδικασία κατασκευής του δέντρου απόφασης. Ωστόσο, δεν είναι άσχετα όλα τα χαρακτηριστικά που παρέχουν ελάχιστα κέρδη. Επομένως, εάν το πρόβλημα ταξινόμησης είναι σύνθετο (π.χ., περιλαμβάνει αλληλεπιδράσεις μεταξύ χαρακτηριστικών) και υπάρχει ένας μεγάλος αριθμός μη σχετικών χαρακτηριστικών, τότε κάποια από αυτά τα χαρακτηριστικά μπορούν να επιλεχτούν κατά λάθος κατά τη διάρκεια της διαδικασίας του δένδρου, καθώς μπορεί να αποφέρουν καλύτερο κέρδος από ένα σχετικό χαρακτηριστηκό τυχαία. Οι τεχνικές επιλογής χαρακτηριστικών μπορούν να βοηθήσουν στη βελτίωση της ακρίβειας των δέντρων αποφάσεων εξαλείφοντας τα μη σχετικά χαρακτηριστικά κατά την προεπεξεργασία.

7. **Χειρισμός περιττών χαρακτηριστικών:** Ένα χαρακτηριστικό είναι περιττό αν συσχετίζεται έντονα με ένα άλλο χαρακτηριστικό στα δεδομένα. Δεδομένου ότι τα πλεονάζοντα χαρακτηριστικά εμφανίζουν παρόμοια κέρδη καθαρότητας εάν επιλεχθούν για διαχωρισμό, μόνο ένα από αυτά θα επιλεγεί ως συνθήκη δοκιμής χαρακτηριστικών στον αλγόριθμο. Επομένως, τα δέντρα αποφάσεων μπορούν να χειριστούν την ύπαρξη πλεονασματικών χαρακτηριστικών.
8. **Επιλογή μέτρου μη καθαρότητας:** Πρέπει να σημειωθεί ότι η επιλογή του μέτρου μη καθαρότητας συχνά έχει ελάχιστη επίδραση στην απόδοση των ταξινομητών δέντρων αποφάσεων, καθώς πολλά από τα μέτρα μη καθαρότητας είναι αρκετά συνεκτικά μεταξύ τους. Αντ 'αυτού, η στρατηγική που χρησιμοποιείται για το κλάδεμα του δέντρου έχει μεγαλύτερη επίπτωση στο τελικό δέντρο από την επιλογή του μέτρου μη καθαρότητας.

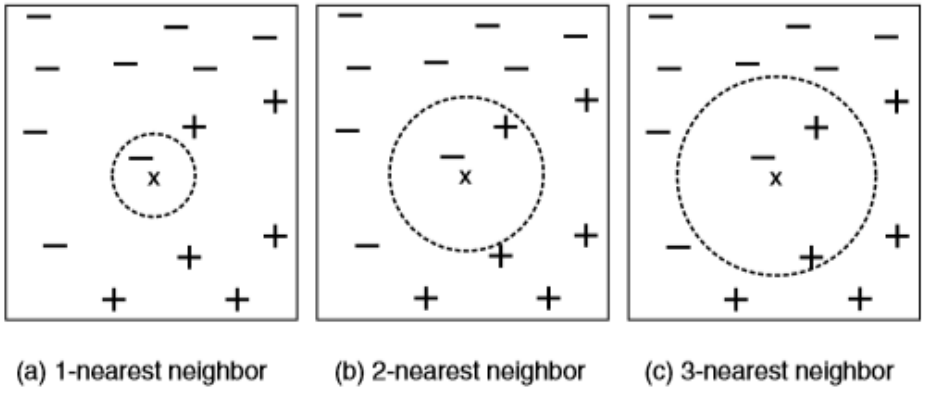
Υπερπροσαρμογή μοντέλου-Model Overfitting

Οι μέθοδοι που παρουσιάζονται μέχρι στιγμής προσπαθούν να μάθουν μοντέλα ταξινόμησης που παρουσιάζουν το χαμηλότερο σφάλμα στο σετ εκπαίδευσης. Ωστόσο, ακόμα και αν ένα μοντέλο βρίσκεται καλά πάνω στα δεδομένα της εκπαίδευσης, μπορεί να παρουσιάσει ακόμα κακή απόδοση γενίκευσης, ένα φαινόμενο που είναι γνωστό ως υπερπροσαρμογή μοντέλου. Αιτίες για το φαινόμενο αυτό είναι οι εξής:

- **Περιορισμένο μέγεθος εκπαίδευσης**
  Ένα σετ εκπαίδευσης που αποτελείται από ένα ορισμένο αριθμό στοιχείων μπορεί να παρέχει μόνο μια περιορισμένη αναπαράσταση των συνολικών δεδομένων. Ως εκ τούτου, είναι πιθανό τα σχέδια που αντλήθηκαν από ένα σετ εκπαίδευσης να μην αντιπροσωπεύουν πλήρως τα πραγματικά μοτίβα στα συνολικά δεδομένα, οδηγώντας σε υπερπροσαρμογή μοντέλου. Σε γενικές γραμμές, καθώς αυξάνουμε το μέγεθος ενός εκπαιδευτικού σετ (αριθμός στοιχείων εκπαίδευσης), τα πρότυπα που αντλήθηκαν από το εκπαιδευτικό σετ αρχίζουν να μοιάζουν με τα πραγματικά μοτίβα των συνολικών δεδομένων. Ως εκ τούτου, η επίδραση της υπερπροσαρμογής μπορεί να μειωθεί αυξάνοντας το μέγεθος της προπόνησης.
- **Υψηλή πολυπλοκότητα μοντέλου**
  Γενικά, ένα πιο πολύπλοκο μοντέλο έχει καλύτερη ικανότητα να αντιπροσωπεύει σύνθετα μοτίβα στα δεδομένα. Για παράδειγμα, τα δέντρα απόφασης με μεγαλύτερο αριθμό φύλλων μπορούν να αντιπροσωπεύουν πιο περίπλοκα όρια απόφασης από τα δέντρα απόφασης με λιγότερα φύλλα. Ωστόσο, ένα υπερβολικά πολύπλοκο μοντέλο έχει επίσης την τάση να μαθαίνει συγκεκριμένα μοτίβα στο σετ εκπαίδευσης που δεν γενικεύουν καλά στα δεδομένα επικύρωσης. Επομένως, τα μοντέλα με μεγάλη πολυπλοκότητα πρέπει να χρησιμοποιηθούν με σύνεση για να αποφευχθεί η υπερπροσαρμογή. Ένα μέτρο της πολυπλοκότητας του μοντέλου είναι ο αριθμός των "παραμέτρων" που πρέπει να συναχθούν από το σετ εκπαίδευσης.

4.2.2. Ταξινομητής πλησιέστερων γειτόνων-Nearest Neighbor Classifier

Το πλαίσιο ταξινόμησης που παρουσιάζεται στο Σχήμα 4.23 περιλαμβάνει μια διαδικασία δύο σταδίων: (1) ένα επαγωγικό βήμα για την κατασκευή ενός μοντέλου ταξινόμησης από τα δεδομένα, και (2) ένα αφαιρετικό βήμα για την εφαρμογή του μοντέλου σε παραδείγματα δοκιμών. Οι ταξινομητές δέντρων απόφασης είναι παραδείγματα πρόθυμων εκπαιδευομένων, επειδή έχουν σχεδιαστεί για να μάθουν ένα μοντέλο που χαρτογραφεί τα χαρακτηριστικά εισόδου στην ετικέτα της κλάσης, μόλις τα δεδομένα εκπαίδευσης είναι διαθέσιμα. Μια αντίθετη στρατηγική θα ήταν η καθυστέρηση της διαδικασίας μοντελοποίησης των δεδομένων εκπαίδευσης, μέχρι να χρειαστεί να ταξινομηθούν τα παραδείγματα των δοκιμών. Τεχνικές που χρησιμοποιούν αυτή τη στρατηγική είναι γνωστές ως τεμπέληδες εκπαιδευόμενοι. Ένα προφανές μειονέκτημα αυτής της προσέγγισης είναι ότι μερικές δοκιμαστικές εγγραφές μπορεί να μην ταξινομηθούν επειδή δεν ταιριάζουν με κανένα παράδειγμα εκπαίδευσης.



Σχήμα 4.23. 1, 2 και 3 κοντινότεροι γείτονες ενός στοιχείου [9]

Ένας τρόπος για να καταστήσουμε αυτήν την προσέγγιση πιο ευέλικτη είναι να εντοπίσουμε όλα τα παραδείγματα εκπαίδευσης που είναι σχετικά παρόμοια με τα χαρακτηριστικά του παραδείγματος δοκιμής. Αυτά τα παραδείγματα, τα οποία είναι γνωστά ως πλησιέστεροι γείτονες (nearest neighbors), μπορούν να χρησιμοποιηθούν για να προσδιορίσουν την ταξινόμηση των δειγμάτων. Η αιτιολόγηση για τους πλησιέστερους γείτονες περιγράφεται καλύτερα με το ακόλουθο ρητό: *"Αν περπατάει σαν πάπια, κράζει σαν πάπια και μοιάζει με πάπια, τότε είναι πιθανώς πάπια".* Ένας ταξινομητής πλησιέστερων γειτόνων αντιπροσωπεύει κάθε παράδειγμα ως σημείο δεδομένων σε d- διάστατο χώρο, όπου d είναι ο αριθμός των χαρακτηριστικών. Λαμβάνοντας υπόψη ένα παράδειγμα δοκιμής, υπολογίζουμε την εγγύτητά του με τα υπόλοιπα σημεία δεδομένων στο σύνολο εκπαίδευσης, χρησιμοποιώντας ένα από τα παρακάτω μέτρα εγγύτητας:

$$Ευκλείδια\ απόσταση: d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

όπου n είναι ο αριθμός διαστάσεων και x_k και y_k είναι, αντίστοιχα, τα k-οστά χαρακτηριστικά (συνιστώσες) των x και y, και γενικεύεται στην απόσταση Minkowsky

$$Απόσταση\ Minkowski: d(x,y) = \left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{1/r}$$

όπου r είναι μια παράμετρος.ο αριθμός διαστάσεων και x_k και y_k είναι, αντίστοιχα, τα k-οστά χαρακτηριστικά (συνιστώσες) των x και y. Τα πιο κλασικά παραδείγματα είναι:

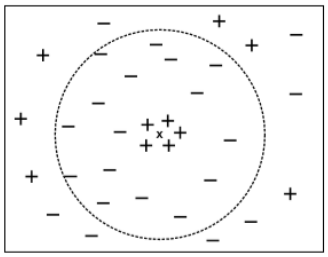
- r=1 απόσταση Hamilton
- r=2 Ευκλείδια απόσταση

Οι k-πλησιέστεροι γείτονες ενός δεδομένου παραδείγματος z αναφέρονται στα k σημεία που είναι πιο κοντά στο z.

Το Σχήμα 4.24 απεικονίζει τους 1, 2 και 3 πλησιέστερους γείτονες ενός σημείου δεδομένων που βρίσκεται στο κέντρο κάθε κύκλου. Το σημείο δεδομένων ταξινομείται βάσει των ετικετών κλάσης των γειτόνων του. Στην περίπτωση που οι γείτονες έχουν περισσότερες από μία ετικέτες, το σημείο αποδίδεται στην κλάση των

πλησιέστερων γειτόνων του που πλειοψηφεί. Στο Σχήμα 4.23(a), ο 1-πλησιέστερος γείτονας του σημείου είναι ένα αρνητικό παράδειγμα. Συνεπώς, το σημείο έχει εκχωρηθεί στην αρνητική κλάση. Εάν ο αριθμός των πλησιέστερων γειτόνων είναι τρεις, όπως φαίνεται στο Σχήμα 4.23(c), τότε η γειτονιά περιέχει δύο θετικά παραδείγματα και ένα αρνητικό. Χρησιμοποιώντας το πλειοψηφικό σχήμα ψηφοφορίας, το σημείο αντιστοιχίζεται στη θετική κλάση. Στην περίπτωση που υπάρχει ισοψηφία μεταξύ των τάξεων (βλ. Σχήμα 4.23(b)), μπορούμε να επιλέξουμε τυχαία μια από αυτές για να ταξινομήσουμε το σημείο.

Φαίνεται, λοιπον, πόσο καθοριστική είναι η σημασία της επιλογής της σωστής τιμής για το k. Εάν το k είναι πολύ μικρό, τότε ο πλησιέστερος ταξινομητής μπορεί να είναι επιρρεπής σε υπερπροσαρμογή λόγω θορύβου στα δεδομένα εκπαίδευσης. Από την άλλη πλευρά, αν το k είναι υπερβολικά μεγάλο, ο πλησιέστερος ταξινομητής μπορεί να λειτουργήσει λανθασμένα το δείγμα δοκιμής επειδή ο κατάλογος των πλησιέστερων γειτόνων μπορεί να περιλαμβάνει σημεία δεδομένων που βρίσκονται μακριά από τη γειτονιά του (βλ. Σχήμα 4.24).



Σχήμα 4.24. Απεικόνιση ταξινομητή K-NN που λειτουργεί λανθασμένα λόγω μεγάλου K [9]

The k-nearest neighbor classification algorithm.

1: Let k be the number of nearest neighbors and D be the set of training examples.
2: **for** each test example $z = (\mathbf{x}', y')$ **do**
3: Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every example, $(\mathbf{x}, y) \in D$.
4: Select $D_z \subseteq D$, the set of k closest training examples to z.
5: $y' = \underset{v}{\text{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
6: **end for**

Μόλις καθοριστεί η λίστα πλησιέστερων γειτόνων, το παράδειγμα δοκιμής ταξινομείται βάσει της πλειψηφούσας τάξης των πλησιέστερων γειτόνων του:

$$Majority\ Voting: y' = \underset{v}{argmax} \sum_{(x_i, y_i) \in D_z} I(v = y_i)$$

όπου *v* είναι μια ετικέτα κλάσης, y_i είναι η ετικέτα κλάσης για έναν από τους πλησιέστερους γείτονες και η *I(·)* είναι μια συνάρτηση δείκτη που επιστρέφει την τιμή 1 εάν το όρισμα της είναι αληθές και 0 διαφορετικά

Η στην προσέγγιση της πλειοψηφίας, κάθε γειτονιά έχει την ίδια επίδραση στην ταξινόμηση. Αυτό κάνει τον αλγόριθμο ευαίσθητο στην επιλογή του k, όπως φαίνεται στο Σχήμα 4.23. Ένας τρόπος για να μειωθεί η επιρροή του k είναι να σταθμίσει την επίδραση του πλησιέστερου γείτονα x_i ανάλογα με την απόσταση του: $w_i = 1/d(x', x_i)^2$. Ως αποτέλεσμα, παραδείγματα εκπαίδευσης που βρίσκονται μακριά από το z έχουν ασθενέστερη επίδραση στην ταξινόμηση σε σύγκριση με εκείνα που βρίσκονται κοντά του. Χρησιμοποιώντας το πρόγραμμα σταθμισμένης από απόσταση ψηφοφορίας, η ετικέτα κλάσης μπορεί να προσδιοριστεί ως εξής:

$$Distance - Weighted\ Voting: y' = \underset{v}{\text{argmax}} \sum_{(x_i, y_i) \in D_z} w_i \times I(v = y_i)$$

Χαρακτηριστικά των ταξινομητών πλησιέστερων γειτόνων

Τα χαρακτηριστικά του πλησιέστερου κατηγοριοποιητή συνοψίζονται παρακάτω:

- Η ταξινόμηση πλησιέστερων γειτόνων αποτελεί μέρος μιας γενικότερης τεχνικής γνωστής ως εκμάθησης βασισμένης σε παραδείγματα, η οποία χρησιμοποιεί συγκεκριμένες περιπτώσεις εκπαίδευσης για να κάνει προβλέψεις χωρίς να έχει να διατηρήσει μια αφαίρεση (ή ένα μοντέλο) που προέρχεται από δεδομένα. Οι αλγόριθμοι μάθησης βασισμένοι σε παραδείγματα, απαιτούν ένα μέτρο εγγύτητας για να προσδιοριστεί την ομοιότητα ή την απόσταση μεταξύ στοιχείων και μιας συνάρτησης ταξινόμησης που επιστρέφει την προβλεπόμενη κλάση ενός δοκιμαστικού παραδείγματος, με βάση την εγγύτητά του σε άλλα στοιχεία.
- Οι τεμπέληδες εκπαιδευόμενοι, όπως οι ταξινομητές πλησιέστερων γειτόνων, δεν απαιτούν οικοδόμηση μοντέλων. Ωστόσο, η ταξινόμηση ενός παραδείγματος δοκιμής μπορεί να είναι αρκετά δαπανηρή επειδή πρέπει να υπολογίσουμε τις τιμές εγγύτητας μεμονωμένα μεταξύ των παραδειγμάτων δοκιμής και εκπαίδευσης. Αντίθετα, οι πρόθυμοι εκπαιδευόμενοι ξοδεύουν συχνά το μεγαλύτερο μέρος των υπολογιστικών τους πόρων για τη δημιουργία μοντέλων. Μόλις κατασκευαστεί ένα μοντέλο, η ταξινόμηση ενός παραδείγματος δοκιμής είναι εξαιρετικά γρήγορη.
- Οι ταξινομητές πλησιέστερων γειτόνων κάνουν τις προβλέψεις τους βασισμένες σε τοπικές πληροφορίες, ενώ τα δέντρα απόφασης επιχειρούν να εντοπίσουν ένα γενικό μοντέλο που καλύπτει ολόκληρο το χώρο εισόδου. Επειδή οι αποφάσεις ταξινόμησης γίνονται τοπικά, οι ταξινομητές πλησιέστερων γειτόνων (με μικρές τιμές k) είναι αρκετά ευαίσθητοι στο θόρυβο.
- Οι ταξινομητές πλησιέστερων γειτόνων μπορούν να παράξουν αυθαίρετα διαμορφωμένα όρια απόφασης. Αυτά τα όρια παρέχουν μια πιο ευέλικτη αναπαράσταση του μοντέλου σε σύγκριση με τα δέντρα αποφάσεων, τα οποία συχνά περιορίζονται στα ευθύγραμμα (rectilineal) όρια απόφασης. Τα όρια απόφασης των πλησιέστερων κατηγοριοποιητών έχουν επίσης μεγάλη μεταβλητότητα επειδή εξαρτώνται από τη σύνθεση των παραδειγμάτων εκπαίδευσης. Η αύξηση του αριθμού των πλησιέστερων γειτόνων μπορεί να μειώσει αυτή τη μεταβλητότητα.
- Οι ταξινομητές πλησιέστερων γειτόνων μπορούν να δημιουργήσουν λανθασμένες προβλέψεις εκτός εάν ληφθούν τα κατάλληλα μέτρα εγγύτητας και γίνουν τα κατάλληλα βήματα προεπεξεργασίας των δεδομένων. Για παράδειγμα, ας υποθέσουμε ότι θέλουμε να ταξινομήσουμε μια ομάδα ανθρώπων με βάση χαρακτηριστικά όπως το ύψος (μετρημένο σε μέτρα) και το βάρος (μετρούμενο σε κιλά). Το χαρακτηριστικό ύψους έχει χαμηλή μεταβλητότητα που κυμαίνεται από 1,5 m έως 1,85 m, ενώ το χαρακτηριστικό βάρος μπορεί να κυμαίνεται από 45 kg έως 120 kg. Εάν η κλίμακα των χαρακτηριστικών δεν ληφθεί υπόψη, το μέτρο εγγύτητας μπορεί να κυριαρχείται από διαφορές στα βάρη ενός ατόμου.

4.2.3. Μπαγεσιανοί ταξινομητές-Bayesian Classifiers

Σε πολλές εφαρμογές η σχέση μεταξύ του συνόλου των χαρακτηριστικών και της μεταβλητής της κλάσης είναι μη-ντετερμινιστική. Με άλλα λόγια, η ετικέτα κλάσης ενός στοιχείου επικύρωσης δεν μπορεί να προβλεφθεί με βεβαιότητα ακόμα κι αν το σύνολο χαρακτηριστικών του είναι πανομοιότυπο με ορισμένα από τα παραδείγματα εκπαίδευσης. Η κατάσταση αυτή μπορεί να προκύψει εξαιτίας δεδομένων με θόρυβο ή της ύπαρξης ορισμένων συγχυτικών παραγόντων (confounding factors) που επηρεάζουν την ταξινόμηση και δεν έχουν συμπεριληφθεί στην ανάλυση. Για παράδειγμα, ας εξετάσουμε την πρόβλεψη αν ένα άτομο κινδυνεύει

από καρδιακές παθήσεις με βάση τη διατροφή του και τη συχνότητα άσκησης. Παρόλο που οι περισσότεροι άνθρωποι που τρώνε υγιεινά και ασκούνται τακτικά έχουν λιγότερες πιθανότητες εμφάνισης καρδιακών παθήσεων, μπορεί παρ΄όλα αυτά να εμφανίσουν τέτοιες παθήσεις λόγω άλλων παραγόντων όπως η κληρονομικότητα, το υπερβολικό κάπνισμα και η υπερβολική κατανάλωση αλκοόλ. Ο καθορισμός του εάν η διατροφή ενός ατόμου είναι υγιηνή ή η συχνότητα της άσκησης είναι ικανές και αναγκαίες συνθήκες, είναι ένα γεγονός που με τη σειρά του μπορεί να εισάγει αβεβαιότητες στο πρόβλημα εκπαίδευσης.

Εδώ θα αναφερθούμε σε μια προσέγγιση μοντελοποίησης πιθανών σχέσεων μεταξύ του συνόλου χαρακτηριστικών και της μεταβλητής της κλάσης, ξεκινώντας με μια εισαγωγή στο θεώρημα Bayes, μια στατιστική αρχή για τον συνδυασμό προηγούμενης γνώσης των τάξεων με νέα στοιχεία που συλλέγονται από τα δεδομένα.

Θεώρημα Bayes

Έστω X και Y ένα ζεύγος τυχαίων μεταβλητών. Η τομή των πιθανοτήτων τους, *P(X=x, Y=y)*, αναφέρεται στην πιθανότητα ότι η μεταβλητή X θα πάρει την τιμή x και η μεταβλητή Y θα πάρει την τιμή y. Μια δεσμευμένη πιθανότητα είναι η πιθανότητα μια τυχαία μεταβλητή να λάβει μια συγκεκριμένη τιμή, με δεδομένο ότι η τιμή μιας άλλης τυχαίας μεταβλητής είναι γνωστή. Για παράδειγμα, η δεσμευμένη πιθανότητα *P(Y=y|X=x)* αναφέρεται στην πιθανότητα ότι η μεταβλητή Y θα πάρει την τιμή y, δεδομένου ότι η μεταβλητή X παρατηρείται ότι έχει την τιμή x. Η τομή των πιθανοτήτων και η δεσμευμένη πιθανότητα για τα X και Y σχετίζονται με τον ακόλουθο τρόπο:

$$P(X,Y) = P(Y|X) \times P(X) = P(X|Y) \times P(Y)$$

Η αναδιάταξη των δύο τελευταίων εκφράσεων στην παραπάνω εξίσωση οδηγεί στον ακόλουθο τύπο γνωστό ως Θεώρημα Bayes:

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

Χρήση του Θεωρήματος του Bayes για Ταξινόμηση

Πριν περιγράψουμε πώς μπορεί να χρησιμοποιηθεί το Θεώρημα του Bayes για ταξινόμηση, ας ορίσουμε το πρόβλημα ταξινόμησης από στατιστική άποψη. Έστω ότι το X δηλώνει το σύνολο χαρακτηριστικών και το Y δηλώνει τη μεταβλητή της κλάσης. Εάν η μεταβλητή της τάξης έχει μια μη-ντετερμινιστική σχέση με τα χαρακτηριστικά, τότε μπορούμε να αντιμετωπίσουμε τα X και Y ως τυχαίες μεταβλητές και να καταγράψουμε την σχέση τους πιθανολογικώς χρησιμοποιώντας *P(Y|X).* Αυτή η δεσμευμένη πιθανότητα είναι επίσης γνωστή ως εκ των υστέρων πιθανότητα (posterior probability) για το Y, σε αντίθεση με την εκ των προτέρων πιθανότητα (prior probability), *P(Y)*.

Κατά τη διάρκεια της φάσης εκπαίδευσης, πρέπει να μάθουμε τις εκ των υστέρων πιθανότητες *P(Y|X)* για κάθε συνδυασμό X και Y με βάση τις πληροφορίες που συλλέγονται από τα δεδομένα εκπαίδευσης. Γνωρίζοντας αυτές τις πιθανότητες, ένα στοιχείο επικύρωσης X’ μπορεί να ταξινομηθεί βρίσκοντας την τάξη Y’ που μεγιστοποιεί την εκ των υστέρων πιθανότητα, *P(Y’|X’).*

Η εκτίμηση των εκ των υστέρων πιθανοτήτων με ακρίβεια για κάθε πιθανό συνδυασμό της ετικέτας της κλάσης και της τιμής των χαρακτηριστικών είναι ένα διφορούμενο πρόβλημα, επειδή απαιτεί ένα πολύ μεγάλο σύνολο εκπαίδευσης, ακόμη και για ένα μέτριο αριθμό χαρακτηριστικών. Το θεώρημα Bayes είναι

χρήσιμο επειδή μας επιτρέπει να εκφράσουμε την εκ των υστέρων πιθανότητα από την άποψη της εκ των προτέρων πιθανότητας P(Y), της δεσμευμένης πιθανότητας της κλάσης $P(\boldsymbol{X}|Y)$ και των στοιχείων $P(\boldsymbol{X})$:

$$P(Y|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|Y)\,P(Y)}{P(\boldsymbol{X})}$$

Κατά τη σύγκριση των εκ των υστέρων πιθανοτήτων για τις διαφορετικές τιμές του Y, ο όρος του παρονομαστή, $P(\boldsymbol{X})$, είναι πάντα σταθερός και επομένως μπορεί να αγνοηθεί. Η εκ των προτέρων πιθανότητα $P(Y)$ μπορεί εύκολα να εκτιμηθεί από το σύνολο εκπαίδευσης με τον υπολογισμό του κλάσματος των στοιχείων εκπαίδευσης που ανήκουν σε κάθε κλάση. Για την εκτίμηση των δεσμευμένων πιθανοτήτων της κλάσης $P(\boldsymbol{X}|Y)$, θα χρησιμοποιήσουμε: τον ταξινομητή Naïve Bayes.

Ταξινομητής Naïve Bayes- Naïve Bayes Classifier

Ένας ταξινομητής Naive Bayes υπολογίζει τη δεσμευμένη πιθανότητα της κλάσης, υποθέτωντας ότι τα χαρακτηριστικά είναι υπό συνθήκες ανεξάρτητα, δεδομένης της ετικέτας της κλάσης y. Η υπό συνθήκες ανεξάρτητη υπόθεση μπορεί να οριστεί επισήμως ως εξής:

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y)$$

όπου κάθε χαρακτηριστικό σύνολο **X** = {X_1, X_2, ..., X_d} αποτελείται από d χαρακτηριστικά.

Ανεξαρτησία υπό συνθήκες

Πριν εμβαθύνουμε στις λεπτομέρειες για το πώς λειτουργεί ένας Naive Bayes ταξονομητης, ας εξετάσουμε την έννοια της υπό συνθήκες ανεξαρτησίας. Έστω ότι τα **X, Y** και **Z** υποδηλώνουν τρία σύνολα τυχαίων μεταβλητών. Οι μεταβλητές στο **X** λέγεται ότι είναι υπό συνθήκες ανεξάρτητες από το **Y**, δεδομένου του **Z**, αν ισχύει η ακόλουθη συνθήκη:

$$P(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}) = P\,(\boldsymbol{X}|\boldsymbol{Z})$$

Ένα παράδειγμα της ανεξαρτησίας υπό συνθήκες είναι η σχέση μεταξύ του μήκους του βραχίονα ενός ατόμου και των δεξιοτήτων ανάγνωσής του. Θα μπορούσε κανείς να παρατηρήσει ότι τα άτομα με μακρύτερους βραχίονες τείνουν να έχουν καλύτερες δεξιότητες ανάγνωσης. Αυτή η σχέση μπορεί να εξηγηθεί από την ύπαρξη ενός συγχυτικού παράγοντα (confounding factor), ο οποίος είναι η ηλικία. Ένα μικρό παιδί τείνει να έχει μικρούς βραχίονες και δεν έχει τις ικανότητες ανάγνωσης ενός ενήλικα. Εάν η ηλικία ενός ατόμου είναι σταθερή, τότε η παρατηρούμενη σχέση μεταξύ του μήκους του βραχίονα και των δεξιοτήτων ανάγνωσης εξαφανίζεται. Έτσι, μπορούμε να καταλήξουμε στο συμπέρασμα ότι το μήκος του βραχίονα και οι δεξιότητες ανάγνωσης είναι υπό συνθήκες ανεξάρτητες όταν η μεταβλητή της ηλικίας είναι σταθερή.

Η υπό συνθήκες ανεξαρτησία μεταξύ των **X** και **Y** μπορεί επίσης να γραφτεί στη μορφή:

$$P(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}) = \frac{P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})}{P(\boldsymbol{Z})} = \frac{P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})}{P(\boldsymbol{Y},\boldsymbol{Z})} \times \frac{P(\boldsymbol{Y},\boldsymbol{Z})}{P(\boldsymbol{Z})} = P(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}) \times P(\boldsymbol{Y}|\boldsymbol{Z}) = P(\boldsymbol{X}|\boldsymbol{Z}) \times P(\boldsymbol{Y}|\boldsymbol{Z})$$

Πώς λειτουργεί ένας Ταξινομητής Naïve Bayes

Με την υπόθεση της υπο όρους ανεξαρτησίας, αντί να υπολογίσουμε την δεσμευμένη πιθανότητα της κλάσης για κάθε συνδυασμό του ***X***, πρέπει μόνο να εκτιμήσουμε την δεσμευμένη πιθανότητα κάθε X_i, δεδομένου του *Y*. Η τελευταία προσέγγιση είναι πιο πρακτική επειδή δεν απαιτεί πολύ μεγάλο σύνολο εκπαίδευσης για να αποκτήσει μια καλή εκτίμηση της πιθανότητας.

Για να ταξινομήσουμε ένα στοιχείο επικύρωσης, ο ταξινομητής Naïve Bayes υπολογίζει την εκ των υστέρων πιθανότητα για κάθε κλάση *Y*:

$$P(Y|\boldsymbol{X}) = \frac{P(Y)\prod_{I=1}^{d} P(X_i|\,Y)}{P(\boldsymbol{X})}$$

Δεδομένου ότι το $P(\boldsymbol{X})$ έχει οριστεί για κάθε Y, είναι αρκετό να επιλέξουμε την κλάση που μεγιστοποιεί τον αριθμητικό όρο, $P(Y)\prod_{I=1}^{d} P(X_i|\,Y)$.

Εκτίμηση των δεσμευμένων πιθανοτήτων για κατηγορηματικά χαρακτηριστικά

Για ένα κατηγορικό χαρακτηριστικό X_i, η δεσμευμένη πιθανότητα $P(X_i=x_i|Y=y)$ υπολογίζεται σύμφωνα με το κλάσμα των στοιχείων εκπαίδευσης που ανήκουν στην κλάση *y*, τα οποία λαμβάνουν μια συγκεκριμένη τιμή χαρακτηριστικού x_i.

Χαρακτηριστικά των ταξινομητών Naïve Bayes

Οι ταξινομητές Naïve Bayes έχουν γενικά τα ακόλουθα χαρακτηριστικά:

- Είναι ανθεκτικοί σε απομονωμένα σημεία θορύβου, επειδή τα εν λόγω σημεία υπολογίζονται κατά μέσο όρο όταν εκτιμώνται οι δεσμευμένες πιθανότητες των δεδομένων. Οι ταξινομητές Naïve Bayes μπορούν επίσης να χειριστούν τις ελλείπουσες τιμές αγνοώντας το παράδειγμα κατά τη διάρκεια της κατασκευής του μοντέλου και της ταξινόμησης.
- Είναι ανθεκτικοί σε μη σχετικά χαρακτηριστικά. Αν το X_i είναι ένα μη σχετικό χαρακτηριστικό, τότε το $P(X_i|Y)$ κατανέμεται σχεδόν ομοιόμορφα. Η δεσμευμένη πιθανότητα της κλάσης για το X_i δεν έχει καμία επίδραση στον συνολικό υπολογισμό της εκ των υστέρων πιθανότητας.
- Τα συσχετιζόμενα χαρακτηριστικά μπορούν να υποβαθμίσουν την απόδοση των ταξινομητών Naïve Bayes, διότι η υπόθεση της υπό συνθήκες ανεξαρτησίας δεν ισχύει πλέον για τέτοια χαρακτηριστικά. Για παράδειγμα, δεδομένων των ακολούθων πιθανοτήτων:

$$P(A=0|Y=0)=0.4, \qquad P(A=1|Y=0)=0.6,$$
$$P(A=0|Y=1)=0.6, \qquad P(A=1|Y=1)=0.4,$$

  όπου το Α είναι ένα δυαδικό χαρακτηριστικό και το Υ είναι μια δυαδική μεταβλητή κλάσης. Ας υποθέσουμε ότι υπάρχει ένα άλλο δυαδικό χαρακτηριστικό Β το οποίο είναι απολύτως συσχετισμένο με το Α όταν *Y=0*, αλλά είναι ανεξάρτητο από το Α όταν *Y=1*. Για λόγους απλότητας, υποθέτουμε ότι οι δεσμευμένες πιθανότητες της κλάσης για το Β είναι οι ίδιες με αυτές τγια ο Α. Δεδομένου ενός στοιχείου με χαρακτηριστικά *A=0*, *B=0*, μπορούμε να υπολογίσουμε τις εκ των υστέρων πιθανότητές του ως εξής:

$$P(Y=0|A=0,B=0)=\frac{P(A=0|Y=0)P(B=0|Y=0)P(Y=0)}{P(A=0,B=0)}=\frac{0.16\times P(Y=0)}{P(A=0,B=0)}$$

$$P(Y=1|A=0,B=0)=\frac{P(A=0|Y=1)P(B=0|Y=1)P(Y=1)}{P(A=0,B=0)}=\frac{0.36\times P(Y=1)}{P(A=0,B=0)}$$

Αν η *P(Y=0) = P(Y=1)*, τότε ο ταξινομητής Naïve Bayes θα αναθέσει το στοιχείο στην κλάση 1. Ωστόσο, η αλήθεια είναι,

$$P(A=0,B=0|Y=0)=P(A=0|Y=0)=0.4$$

διότι τα *A* και *B* συσχετίζονται απόλυτα όταν *Y=0*. Ως αποτέλεσμα, η εκ των υστέρων πιθανότητα *Y=0* είναι

$$P(Y=0|A=0,B=0)=\frac{P(A=0,B=0|Y=0)P(Y=0)}{P(A=0,B=0)}=\frac{0.4\times P(Y=0)}{P(A=0,B=0)}$$

η οποία είναι μεγαλύτερη από εκείνη του *Y=1*. Το στοιχείο, λοιπόν, έπρεπε να έχει ταξινομηθεί ως κλάση 0. [9]

5. Ανάλυση Δεδομένων

Για την ανάλυση των δεδομένων που αντλήθηκαν από την βάση δεδομένων MIMIC III, έγινε χρήση της γλώσσας R. Η γλώσσα R δημιουργήθηκε λίγο μετά την ανάπτυξη της S. Ένας βασικός περιορισμός της γλώσσας S ήταν ότι ήταν διαθέσιμη μόνο σε ένα εμπορικό πακέτο, το S-PLUS. Το 1991, η R δημιουργήθηκε από τον Ross Ihaka και τον Robert Gentleman στο Τμήμα Στατιστικής του Πανεπιστημίου του Auckland. Το 1993 η πρώτη δημοσίευση της R έγινε στο κοινό. Η εμπειρία των Ihaka και Gentleman κατά τη διάρκεια ανάπτυξης της R καταγράφηκε σε μια δημοσίευση του 1996 στο Journal of Computational and Graphical Statistics με τίτλο: «Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5 (3): 299-314, 1996»

Το 1995, ο Martin Mächler συνέβαλε σημαντικά πείθοντας τους Ihaka και Gentleman να χρησιμοποιήσουν την άδεια GNU General Public License[4] για να κάνουν την R δωρεάν λογισμικό. Αυτή η κίνηση ήταν κρίσιμη επειδή επέτρεψε στον αρχικό κώδικα για ολόκληρο το σύστημα της R να είναι προσβάσιμος σε όποιον ήθελε να πειραματιστεί μαζί της. Το 2000 R έκδοση 1.0.0 κυκλοφόρησε στο κοινό. [10]

5.1. Καθαρισμός δεδομένων και βασική ανάλυση

Τα δεδομένα των οποίων η επεξεργασία έγινε στην R προέκυψαν από τα queries που περιγράφηκαν στην ενότητα 3.1. Αρχικά φορτώθηκαν στη μνήμη τα αρχεία ADMISSIONS_RESP.csv, PATIENTS_RESP.csv, LABEVENTS_RESP.csv, DIAGNOSES_ICD_RESP.csv και D_ICD_DIAGNOSES.csv σε μορφή data frames, τα οποία αφορούν δεδομένα ασθενών που σχετίζονται με την πνευμονολογική κλινική.

Πιο συγκεκριμένα, όπως φαίνεται στο Σχήμα 5.1. ο πίνακας «admissions» αποτελείται από 24583 γραμμές (εγγραφές) και 19 στήλες (χαρακτηριστικά).

```
> str(admissions)
'data.frame':   24583 obs. of  19 variables:
 $ row_id              : int  28 30 31 32 36 37 38 39 42 456 ...
 $ subject_id          : int  28 31 32 33 36 36 36 37 41 357 ...
 $ hadm_id             : int  162569 128652 175413 176176 182104 122659 165660 188670 101757 174486 ...
 $ admittime           : Factor w/ 24542 levels "2100-06-07 19:59:00",..: 18570 1836 16812 3826 7346 7357 8181
 $ dischtime           : Factor w/ 24526 levels "2100-06-09 17:09:00",..: 18548 1828 16796 3812 7335 7347 8167
 $ deathtime           : Factor w/ 3972 levels "","2100-06-19 08:15:00",..: 1 299 1 1 1 1 1 1 1 ...
 $ admission_type      : Factor w/ 4 levels "ELECTIVE","EMERGENCY",..: 1 2 1 2 2 2 1 2 1 2 ...
 $ admission_location  : Factor w/ 8 levels "** INFO NOT AVAILABLE **",..: 4 5 4 3 2 3 4 3 4 3 ...
 $ discharge_location  : Factor w/ 17 levels "DEAD/EXPIRED",..: 6 1 5 5 6 14 12 6 3 14 ...
 $ insurance           : Factor w/ 5 levels "Government","Medicaid",..: 3 3 2 3 3 3 3 3 4 4 ...
 $ language            : Factor w/ 61 levels "","* FU","**SH",..: 1 1 1 1 40 40 40 1 1 40 ...
 $ religion            : Factor w/ 21 levels "","7TH DAY ADVENTIST",..: 5 5 21 18 16 16 16 12 18 16 ...
 $ marital_status      : Factor w/ 8 levels "","DIVORCED",..: 4 4 1 4 4 4 4 4 6 4 ...
 $ ethnicity           : Factor w/ 40 levels "AMERICAN INDIAN/ALASKA NATIVE",..: 36 36 36 35 36 36 36 36 36 36
 $ edregtime           : Factor w/ 15648 levels "","2100-06-07 13:14:00",..: 1 1 1 2437 1 4663 1 12739 1 14950
 $ edouttime           : Factor w/ 15644 levels "","2100-06-08 00:06:00",..: 1 1 1 2437 1 4660 1 12735 1 14946
 $ diagnosis           : Factor w/ 8419 levels "","   DUODENAL MASS/SDA",..: 2459 7504 7956 7207 2426 1679 8316
 $ hospital_expire_flag: int  0 1 0 0 0 0 0 0 0 0 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
```

Σχήμα 5.1. Η δομή του πίνακα admissions

Αντίστοιχα στο Σχήμα 5.2. απεικονίζονται τα στοιχεία του πίνακα «patients», που αποτελείται από 19425 γραμμές (εγγραφές) και 8 στήλες (χαρακτηριστικά). Όπως είναι λογικό, τα περιστατικά (admissions) είναι περισσότερα από τους ασθενείς, καθώς ένας ασθενής μπορεί να νοσηλευτεί στο νοσοκομείο περισσότερες από μια φορές.

```
> str(patients)
'data.frame':   19425 obs. of  8 variables:
 $ row_id     : int  234 235 237 238 240 245 246 249 251 252 ...
 $ subject_id : int  249 250 252 253 256 262 263 266 268 269 ...
 $ gender     : Factor w/ 2 levels "F","M": 1 1 2 1 2 2 2 1 1 2 ...
 $ dob        : Factor w/ 16249 levels "1800-07-02 00:00:00",..: 6956 16122 7368 9050 8570 9066
 $ dod        : Factor w/ 8702 levels "","2100-06-19 00:00:00",..: 1 7520 1 1 1 1 5781 1 8305 1
 $ dod_hosp   : Factor w/ 6019 levels "","2100-06-19 00:00:00",..: 1 5214 1 1 1 1 4043 1 5770 1
 $ dod_ssn    : Factor w/ 7525 levels "","2100-06-19 00:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ expire_flag: int  0 1 0 0 0 0 1 0 1 0 ...
```

Σχήμα 5.2. Η δομή του πίνακα patients

Από τους δύο αυτούς πίνακες, αφαιρέθηκαν οι στήλες που περιείχαν τον αύξοντα αριθμό της κάθε γραμμής, όπως και κάποιες στήλες που δεν περιέχουν χαρακτηριστικά χρήσιμα στην ανάλυση που ακολουθεί.

Από τις 19 στήλες του πίνακα admissions εκτός από το row_id αφαιρέθηκαν τα admission_type και admission_location, που σχετίζονται με τον τρόπο που έφτασε ο ασθενής στο νοσοκομείο και τα edregtime και edouttime, που είναι οι χρονικές στιγμές που περάστηκαν τα δεδομένα της εισόδου και της εξόδου του ασθενή στο σύστημα, διότι δεν έχουμε σκοπό να αξιολογήσουμε την αμεσότητα του νοσηλευτικού προσωπικού. Επίσης, αφαιρέθηκαν το deathtime γιατί στη συνέχεια θα κρατήσουμε το αντίστοιχο δεδομένο της ώρας θανάτου του ασθενή από τον αντίστοιχο πίνακα patients.

Από τις 8 στήλες του πίνακα patients εκτός από το row_id αφαιρέθηκαν dod_hosp kai dod_ssn διότι και αφορούν στοιχεία των διαφορετικών πληροφοριακών συστημάτων που χρησιμοποιεί το νοσοκομείο και έχουν συγχωνευτεί στο dod.

Μετά από την παραπάνω διαδικασία οι δύο πίνακες admissions και patients συγχωνεύτηκαν σε έναν νέο (data frame) με βάση το κλειδί subject_id, που αποτελεί τον κωδικό του κάθε ασθενή. Ο νέος πίνακας «adm_pat», αποτελείται επίσης από 24583 γραμμές και 19 στήλες, του οποίου τα χαρακτηριστικά φαίνονται στο Σχήμα 5.3. Επιπλέον, δημιουργήθηκε μια νέα στήλη η admitdate που μαζί με την dob μετατράπηκαν από Factor σε Date.

```
> str(adm_pat)
'data.frame':   24583 obs. of  19 variables:
 $ subject_id          : int  9 17 18 21 21 28 31 32 33 36 ...
 $ hadm_id             : int  150750 161087 188822 111970 109451 162569 128652 175413 176176 182104 ...
 $ admittime           : Factor w/ 24542 levels "2100-06-07 19:59:00",..: 11891 8413 16193 8346 8259 18570 1836
 $ dischtime           : Factor w/ 24526 levels "2100-06-09 17:09:00",..: 11870 8394 16167 8334 8249 18548 1828
 $ deathtime           : Factor w/ 3972 levels "","2100-06-19 08:15:00",..: 1962 1 1 1392 1 1 299 1 1 1 ...
 $ discharge_location  : Factor w/ 17 levels "DEAD/EXPIRED",..: 1 6 5 1 14 6 1 5 5 6 ...
 $ insurance           : Factor w/ 5 levels "Government","Medicaid",..: 2 4 4 3 3 3 3 2 3 3 ...
 $ language            : Factor w/ 61 levels "","* FU","**SH",..: 1 40 1 1 1 1 1 1 1 40 ...
 $ religion            : Factor w/ 21 levels "","7TH DAY ADVENTIST",..: 21 5 5 12 12 5 5 21 18 16 ...
 $ marital_status      : Factor w/ 8 levels "","DIVORCED",..: 1 4 4 4 4 4 4 4 1 4 4 ...
 $ ethnicity           : Factor w/ 40 levels "AMERICAN INDIAN/ALASKA NATIVE",..: 35 36 36 36 36 36 36 36 35 36 .
 $ diagnosis           : Factor w/ 8419 levels "","   DUODENAL MASS/SDA",..: 3565 5618 3781 7131 2080 2459 7504
 $ hospital_expire_flag: int  1 0 0 1 0 0 1 0 0 0 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
 $ admitdate           : Date, format: "2149-11-09" "2135-05-09" "2167-10-02" ...
 $ gender              : Factor w/ 2 levels "F","M": 2 1 2 2 2 2 2 2 2 2 ...
 $ dob                 : Date, format: "2108-01-26" "2087-07-14" "2116-11-29" ...
 $ dod                 : Factor w/ 8702 levels "","2100-06-19 00:00:00",..: 4177 1 1 2892 2892 1 588 1 1 1 ...
 $ expire_flag         : int  1 0 0 1 1 0 1 0 0 0 ...
```

Σχήμα 5.3. Η δομή του πίνακα adm_pat

Το επόμενο βήμα ήταν ο υπολογισμός της ηλικίας. Επειδή τα δεδομένα μας έχουν τροποποιηθεί για να είναι μη ταυτοποιήσιμα, έπρεπε να βρεθούν δύο σχετικές ημερομηνίες. Έτσι, αφαιρέθηκε από το admitdate το dob. Σε αυτό το σημείο η σύνοψη των δεδομένων μας φαίνεται στο Σχήμα 5.4.

```
> summary(adm_pat)
   subject_id       hadm_id                    admittime                  dischtime                  deathtime
 Min.   :    9   Min.   :100006   2102-11-14 18:50:00:    2   2100-11-08 14:30:00:    2                   :20611
 1st Qu.:13622   1st Qu.:125350   2104-03-25 07:15:00:    2   2101-09-22 13:00:00:    2   2108-02-22 15:15:00:    2
 Median :27089   Median :150341   2109-05-22 08:00:00:    2   2103-01-15 15:45:00:    2   2100-06-19 08:15:00:    1
 Mean   :37156   Mean   :150146   2109-08-05 07:15:00:    2   2103-10-05 19:00:00:    2   2100-07-21 10:34:00:    1
 3rd Qu.:60622   3rd Qu.:175035   2110-01-06 07:15:00:    2   2106-02-16 11:45:00:    2   2100-08-19 15:03:00:    1
 Max.   :99985   Max.   :199999   2110-01-17 07:15:00:    2   2106-02-18 13:00:00:    2   2100-09-03 12:35:00:    1
                                  (Other)            :24571   (Other)            :24571   (Other)            : 3966
              discharge_location      insurance        language                  religion      marital_status
 HOME HEALTH CARE        :5276     Government:  562   ENGL   :14265   CATHOLIC          :9193   MARRIED  :11060
 HOME                    :4286     Medicaid  : 2239          : 8131   NOT SPECIFIED     :4599   SINGLE   : 6240
 SNF                     :4207     Medicare  :14977   SPAN   :  495   PROTESTANT QUAKER:3374   WIDOWED  : 3851
 DEAD/EXPIRED            :3972     Private   : 6631   RUSS   :  395   UNOBTAINABLE      :2590   DIVORCED : 1715
 REHAB/DISTINCT PART HOSP:3745     Self Pay  :  174   PTUN   :  349   JEWISH            :2486            : 1224
 LONG TERM CARE HOSPITAL :1716                        CANT   :  208   OTHER             :1067   SEPARATED:  301
 (Other)                 :1381                        (Other):  740   (Other)           :1274   (Other)  :  192
                ethnicity                          diagnosis     hospital_expire_flag has_chartevents_data   admitdate
 WHITE                 :17638   PNEUMONIA               : 1499   Min.   :0.0000       Min.   :0.0000       Min.   :2100-06-07
 BLACK/AFRICAN AMERICAN: 2305   SEPSIS                  :  748   1st Qu.:0.0000       1st Qu.:1.0000       1st Qu.:2126-03-21
 UNKNOWN/NOT SPECIFIED : 1806   CONGESTIVE HEART FAILURE:  607   Median :0.0000       Median :1.0000       Median :2151-06-26
 HISPANIC OR LATINO    :  602   ALTERED MENTAL STATUS   :  420   Mean   :0.1616       Mean   :0.9815       Mean   :2151-07-30
 OTHER                 :  502   FEVER                   :  320   3rd Qu.:0.0000       3rd Qu.:1.0000       3rd Qu.:2177-01-27
 UNABLE TO OBTAIN      :  386   CHEST PAIN              :  309   Max.   :1.0000       Max.   :1.0000       Max.   :2210-08-17
 (Other)               : 1344   (Other)                 :20680
 gender         dob                          dod         expire_flag       dischdate              age
 F:11360   Min.   :1800-07-02                 :11405   Min.   :0.0000   Min.   :2100-06-09   Min.   :  0.00
 M:13223   1st Qu.:2055-11-22   2108-06-07 00:00:00:   24   1st Qu.:0.0000   1st Qu.:2126-04-08   1st Qu.: 54.99
           Median :2084-02-25   2127-12-13 00:00:00:   18   Median :1.0000   Median :2151-07-04   Median : 67.43
           Mean   :2073-08-18   2121-06-02 00:00:00:   14   Mean   :0.5361   Mean   :2151-08-11   Mean   : 77.84
           3rd Qu.:2110-10-06   2127-01-14 00:00:00:   13   3rd Qu.:1.0000   3rd Qu.:2177-02-06   3rd Qu.: 78.64
           Max.   :2201-04-18   2206-06-20 00:00:00:   13   Max.   :1.0000   Max.   :2210-08-24   Max.   :311.13
                                (Other)            :13096
```

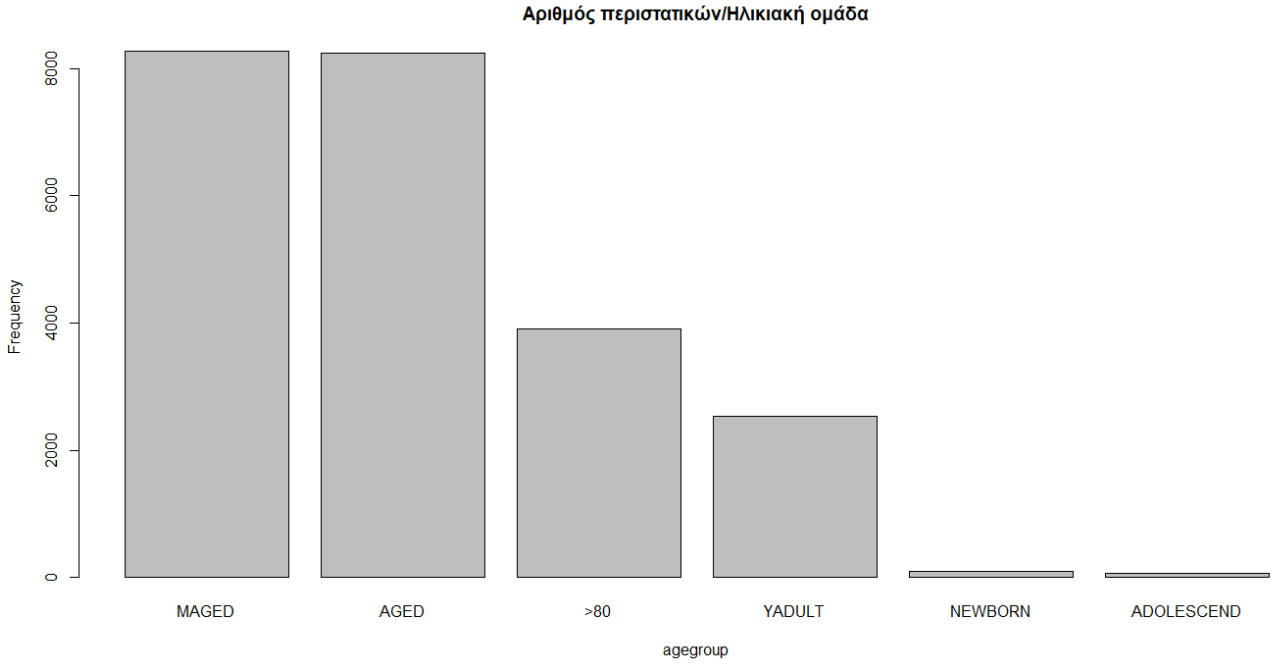
Σχήμα 5.4. Η σύνοψη του πίνακα adm_pat

Οι πρώτες παρατηρήσεις για τα χαρακτηριστικά των δεδομένων μας είναι ότι μετά τη νοσηλεία στο νοσοκομείο η πλειονότητα των ασθενών γυρίζει στο σπίτι τους, η επικρατέστερη ασφάλιση είναι το Medicare και ακολουθεί η ιδιωτική, γλώσσα που ομιλείται συχνότερα είναι τα αγγλικά, οι περισσότεροι ασθενείς είναι χριστιανοί καθολικοί, παντρεμένοι και λευκοί. Επίσης, τα περιστατικά αφορούν 11360 γυναίκες και 13223 άνδρες. Όσον αφορά στην αιτία για την οποία μπήκαν στην πνευμονολογική κλινική του νοσοκομείου, η επικρατέστερη είναι πνευμονία (PNEUMONIA) και ακολουθούν η σήψη (SEPSIS) με την συμφορητική καρδιακή ανεπάρκεια (CONGESTIVE HEART FAILURE).

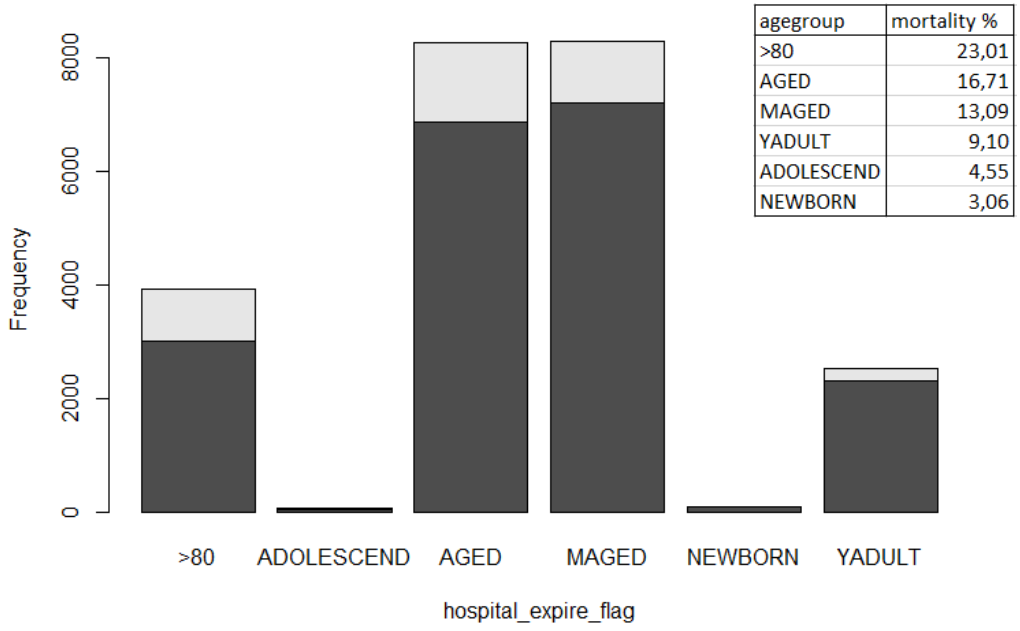
Παρατηρώντας τα συγκεντρωτικά χαρακτηριστικά βλέπουμε ότι η μέγιστη ηλικία είναι 311, 13 ετών, κάτι που δεν είναι λογικό. Επομένως συμπεραίνουμε ότι σε κάποιους ασθενείς η χρονική μεταφορά που έγινε είναι ασυνεπής. Αυτά τα περιστατικά, θα τα αφαιρέσουμε θέτοντας ως ανώτερο λογικό όριο για την ηλικία των ανθρώπων τα 110 χρόνια. Επιπλέον, δημιουργήσαμε μια νέα στήλη την «agegroup» όπου οι ηλικίες των ασθενών διακριτοποιούνται σε 7 κατηγορίες:

Newborn, για age ≤ 1
child, για age < 12
adolescent, για age < 19
Young Adult, για age < 12
Middle Aged, για age < 44
Aged, για age < 65
>80, για age > 80

Οπτικοποιώντας τα παραπάνω δεδομένα, βλέπουμε στο Σχήμα 5.5 τον αριθμό των περιστατικών ανά ηλικιακή ομάδα, στο Σχήμα 5.6 την αναλογία θνησιμότητας για κάθε ηλικιακή ομάδα και στο Σχήμα 5.7 για κάθε διαφορετική ασφάλιση μαζί με τα συνοδευτικά ποσοστά για κάθε κατηγορία ξεχωριστά.

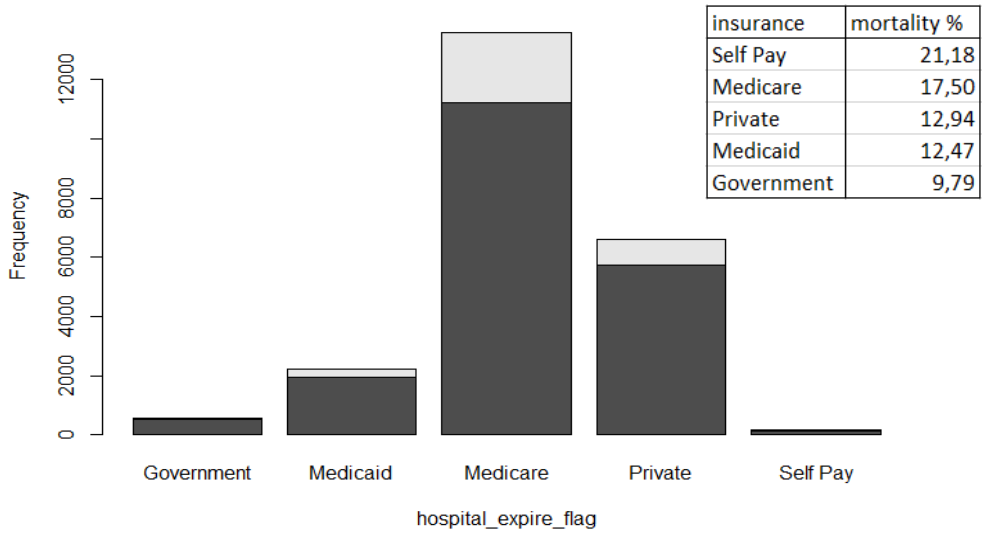


Σχήμα 5.5. Αριθμός περιστατικών ανά ηλικιακή ομάδα



agegroup	mortality %
>80	23,01
AGED	16,71
MAGED	13,09
YADULT	9,10
ADOLESCEND	4,55
NEWBORN	3,06

Σχήμα 5.6.Συσχετισμός ασθενών που απεβίωσαν για κάθε ηλικιακή ομάδα



insurance	mortality %
Self Pay	21,18
Medicare	17,50
Private	12,94
Medicaid	12,47
Government	9,79

Σχήμα 5.7. Συσχετισμός ασθενών που απεβίωσαν για κάθε τύπο ασφάλισης

Στη συνέχεια, εξερευνώντας τον πίνακα diagnoses_icd που φαίνεται στο Σχήμα 5.8 διαπιστώσαμε ότι αποτελείται από 42599 γραμμές, που οφείλεται στο ότι σε κάθε περιστατικό αντιστοιχούν, πολλές φορές, περισσότερες από μία διαγνώσεις. Οι 15 πιο συχνές τελικές διαγνώσεις παρουσιάζεται στον Πίνακα 5.9.

```
> str(diagnoses_icd)
'data.frame':   42599 obs. of  5 variables:
 $ row_id    : int  1298 1498 1509 1539 1542 1576 1577 1592 1601 1606 ...
 $ subject_id: int  109 114 115 117 117 124 124 124 124 124 ...
 $ hadm_id   : int  172335 178393 114585 164853 164853 112906 112906 134369 138376 138376 ...
 $ seq_num   : int  2 3 8 8 11 2 3 9 3 8 ...
 $ icd9_code : int  486 48283 5119 51882 486 486 496 496 49121 486 ...
```

Σχήμα 5.8. Η δομή του πίνακα diagnoses_icd

Το επόμενο βήμα είναι να ενώσουμε τον πίνακα «adm_pat» με τον πίνακα «diagnoses_icd». Γι' αυτό θα αφαιρέσουμε τις στήλες «row_id» και «subject_id». Ο νέος πίνακας που προκύπτει είναι ο «df», του οποίου η δομή φαίνεται στο Σχήμα 5.10. Ο df αποτελείται από 40143 γραμμές και 24 στήλες.

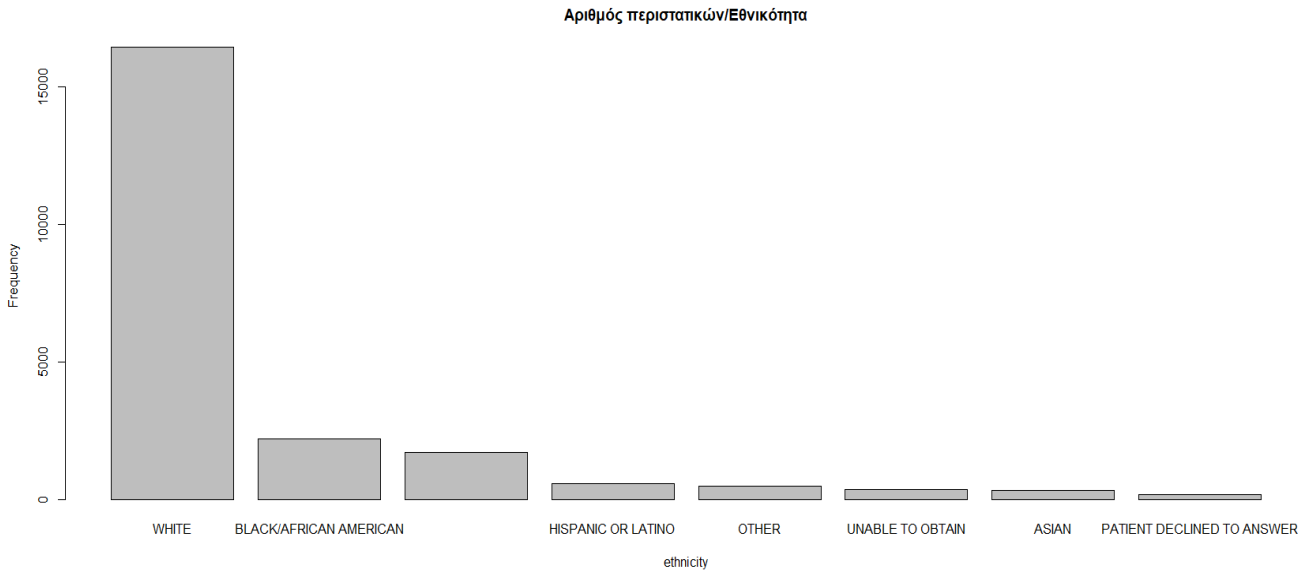
```
> str(df)
'data.frame':   40143 obs. of  24 variables:
 $ hadm_id             : int  100006 100006 100006 100007 100011 100016 100016 100016 100017 100018 ...
 $ subject_id          : int  9895 9895 9895 23018 87977 68591 68591 68591 16229 58128 ...
 $ admittime           : Factor w/ 24542 levels "2100-06-07 19:59:00",..: 1756 1756 1756 10791 18567 21185 1
5 564 18305 ...
 $ dischtime           : Factor w/ 24526 levels "2100-06-09 17:09:00",..: 1747 1747 1747 10772 18554 21161 1
1 555 18290 ...
 $ deathtime           : Factor w/ 3972 levels "","2100-06-19 08:15:00",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ discharge_location  : Factor w/ 17 levels "DEAD/EXPIRED",..: 5 5 5 5 14 16 16 16 2 12 ...
 $ insurance           : Factor w/ 5 levels "Government","Medicaid",..: 4 4 4 4 2 3 3 3 2 4 ...
 $ language            : Factor w/ 61 levels "","* FU","**SH",..: 1 1 1 1 40 40 40 40 1 40 ...
 $ religion            : Factor w/ 21 levels "","7TH DAY ADVENTIST",..: 16 16 16 12 16 18 18 18 5 18 ...
 $ marital_status      : Factor w/ 8 levels "","DIVORCED",..: 6 6 6 4 6 6 6 6 6 4 ...
 $ ethnicity           : Factor w/ 40 levels "AMERICAN INDIAN/ALASKA NATIVE",..: 13 13 13 36 17 36 36 36 35
 $ diagnosis           : Factor w/ 8419 levels "","   DUODENAL MASS/SDA",..: 2261 2261 2261 1278 4988 5764
5448 3626 ...
 $ hospital_expire_flag: int  0 0 0 0 0 0 0 0 0 0 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
 $ admitdate           : Date, format: "2108-04-06" "2108-04-06" "2108-04-06" "2145-03-31" ...
 $ gender              : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2 2 ...
 $ dob                 : Date, format: "2059-05-07" "2059-05-07" "2059-05-07" "2071-06-04" ...
 $ dod                 : Factor w/ 8702 levels "","2100-06-19 00:00:00",..: 690 690 690 1 1 7486 7486 7486 1
 $ expire_flag         : int  1 1 1 0 0 1 1 1 0 1 ...
 $ dischdate           : Date, format: "2108-04-18" "2108-04-18" "2108-04-18" "2145-04-07" ...
 $ age                 : num  49 49 49 74 21 55 55 55 27 55 ...
 $ agegroup            : Factor w/ 6 levels ">80","ADOLESCEND",..: 4 4 4 3 6 4 4 4 6 4 ...
 $ seq_num             : int  1 3 2 4 4 2 4 1 2 12 ...
 $ icd9_code           : int  49320 486 51881 486 48242 51881 47874 5070 51881 5119 ...
```

Σχήμα 5.9. Η δομή του πίνακα df

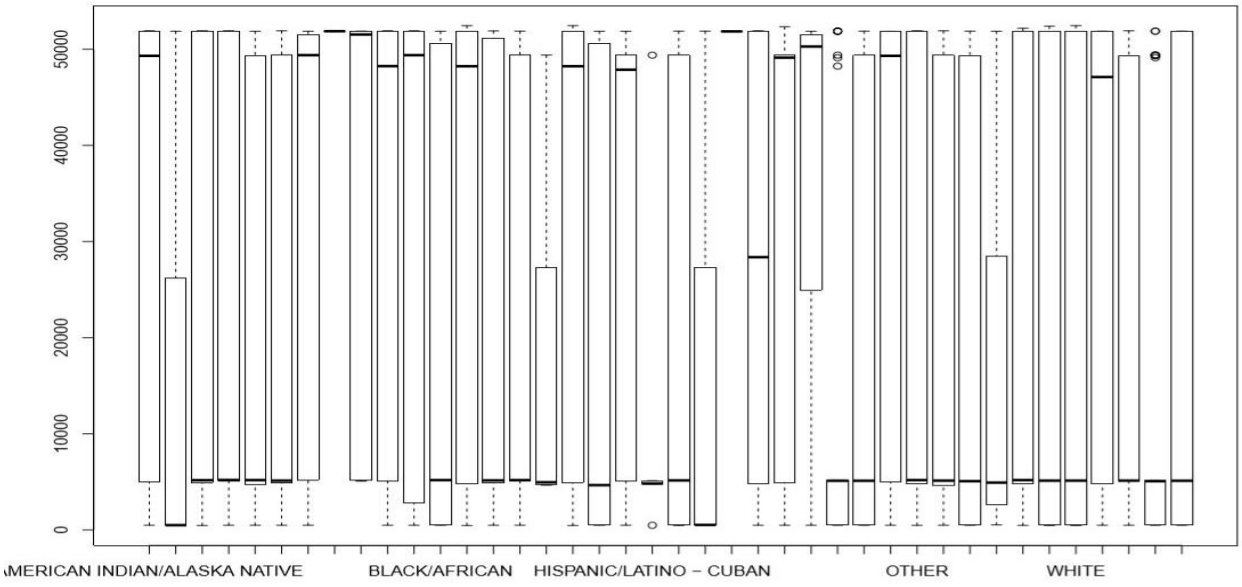
Αρχικά θα εξετάσουμε το χαρακτηριστικό της εθνικότητας, προκειμένου να εντοπίσουμε τυχόν συσχετίσεις με την τελική διάγνωση. Τα περιστατικά ανά εθνικότητα παρουσιάζονται στον πίνακα 5.10, το Σχήμα 5.11 και το θηκόγραμμα της εθνικότητας συναρτήσει του icd9_code στο Σχήμα 5.12.

	Ethnicity	**Admissions**	**Percentage**
1	WHITE	28831	71,82
2	BLACK/AFRICAN AMERICAN	3836	9,56
3	UNKNOWN/NOT SPECIFIED	2872	7,15
4	HISPANIC OR LATINO	945	2,35
5	OTHER	843	2,10
6	UNABLE TO OBTAIN	687	1,71
7	ASIAN	597	1,49
8	PATIENT DECLINED TO ANSWER	335	0,83
9	HISPANIC/LATINO - PUERTO RICAN	226	0,56
10	ASIAN - CHINESE	199	0,50
11	BLACK/CAPE VERDEAN	108	0,27
12	WHITE - RUSSIAN	86	0,21
13	PORTUGUESE	58	0,14
14	MULTI RACE ETHNICITY	57	0,14
15	BLACK/HAITIAN	52	0,13
16	HISPANIC/LATINO - DOMINICAN	46	0,11
17	WHITE - OTHER EUROPEAN	46	0,11
18	ASIAN - ASIAN INDIAN	45	0,11
19	MIDDLE EASTERN	33	0,08
20	BLACK/AFRICAN	32	0,08
21	WHITE - BRAZILIAN	28	0,07
22	ASIAN - VIETNAMESE	24	0,06
23	ASIAN - FILIPINO	23	0,06
24	HISPANIC/LATINO - GUATEMALAN	19	0,05
25	AMERICAN INDIAN/ALASKA NATIVE	16	0,04
26	WHITE - EASTERN EUROPEAN	16	0,04
27	ASIAN - CAMBODIAN	15	0,04
28	NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	12	0,03
29	ASIAN - OTHER	9	0,02
30	HISPANIC/LATINO - CENTRAL AMERICAN (OTHER)	7	0,02
31	ASIAN - KOREAN	6	0,01
32	HISPANIC/LATINO - COLOMBIAN	6	0,01
33	HISPANIC/LATINO - MEXICAN	6	0,01
34	HISPANIC/LATINO - CUBAN	5	0,01
35	CARIBBEAN ISLAND	4	0,01
36	HISPANIC/LATINO - SALVADORAN	4	0,01
37	SOUTH AMERICAN	4	0,01
38	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED	3	0,01
39	ASIAN - JAPANESE	1	0,00
40	HISPANIC/LATINO - HONDURAN	1	0,00

Πίνακας 5.10. Περιστατικά ανά εθνικότητα

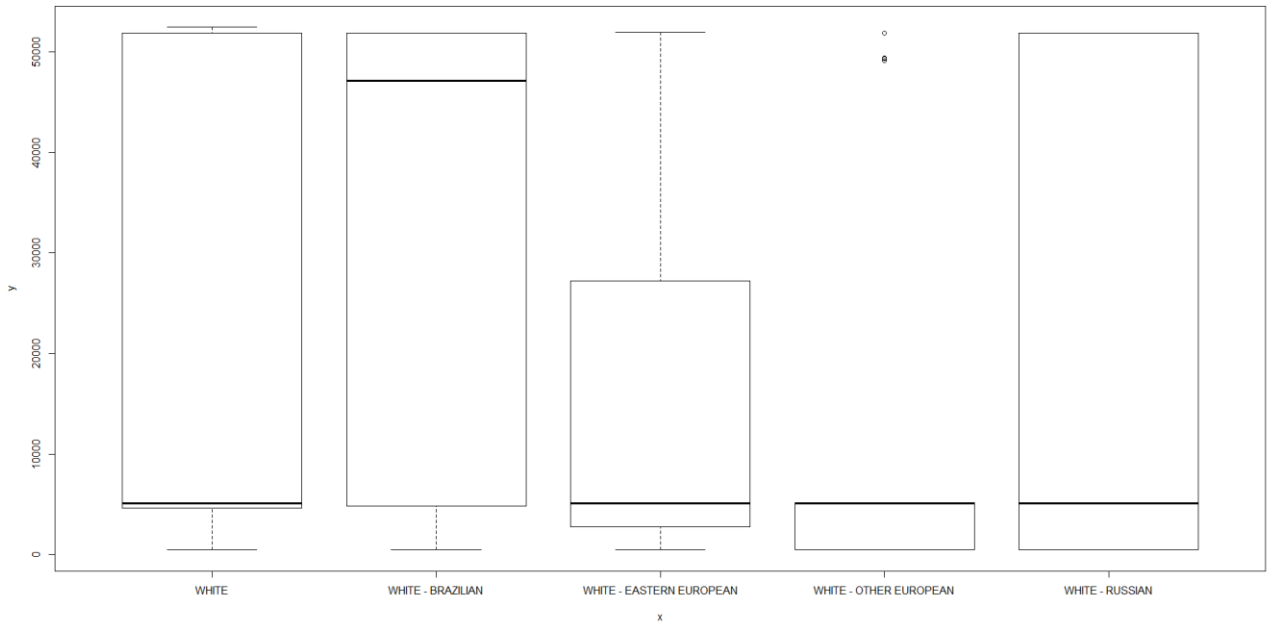


Σχήμα 5.11. Αριθμός περιστατικών ανά εθνικότητα

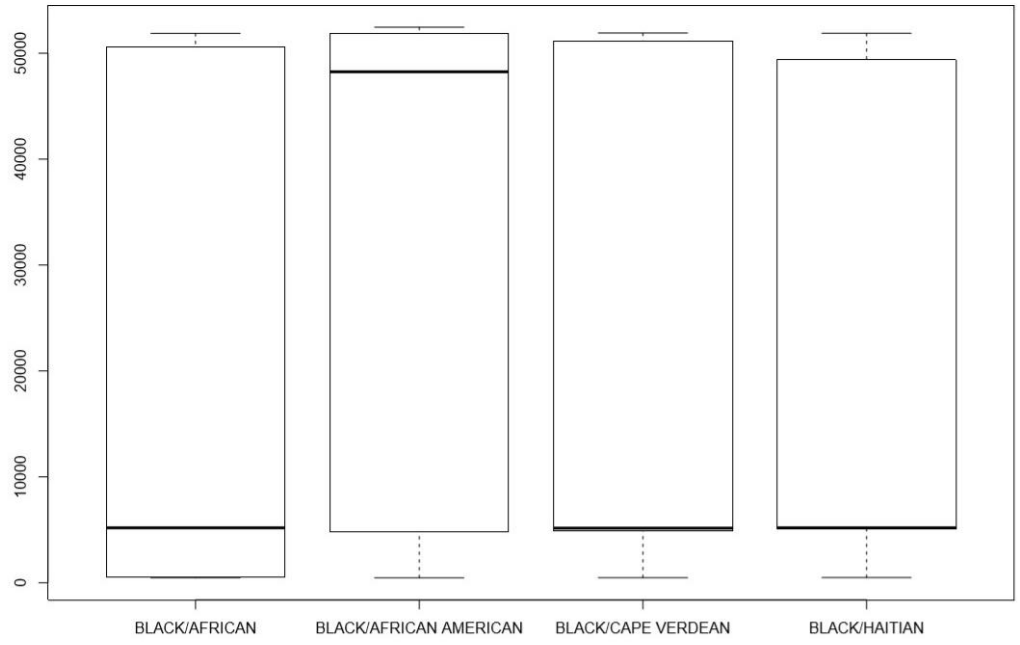


Σχήμα 5.12. Θηκόγραμμα ethnicity συναρτήσει του icd9_code

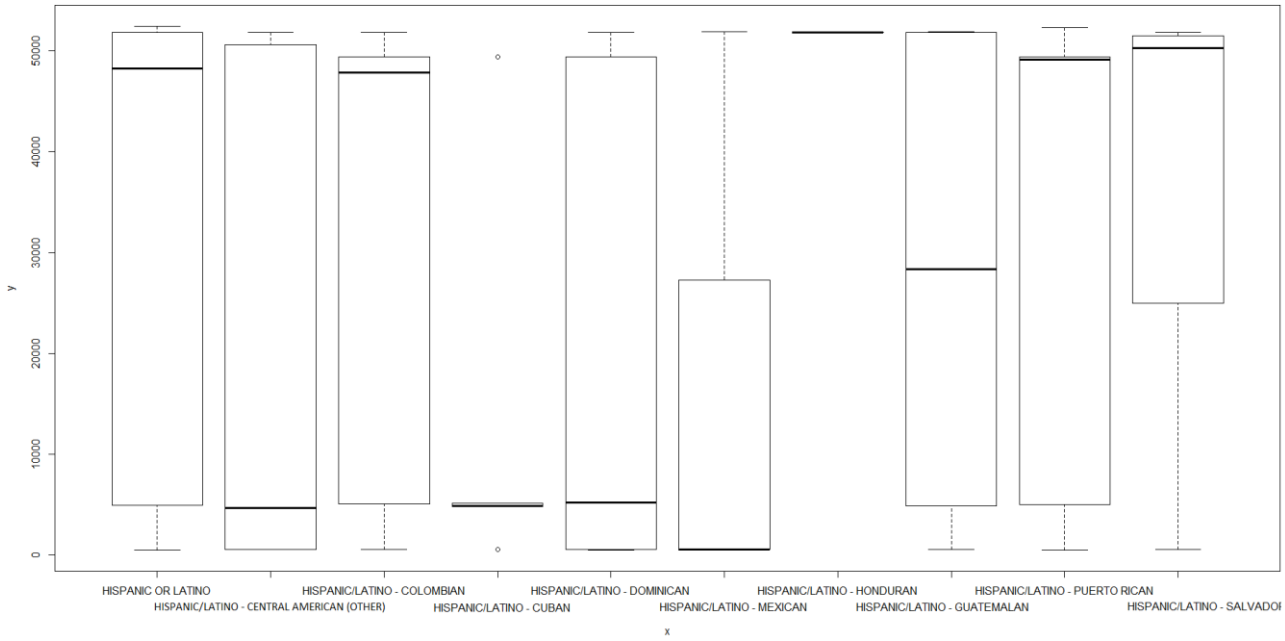
Λόγω των πολλών διαφορετικών τιμών για την μεταβλητή «ethnicity», δημιουργήθηκε μια νέα στήλη η «race», που αποτελείται από 4 τιμές: WHITE, BLACK, HISPANIC, ASIAN και OTHER. Έτσι ανά κατηγορία έχουμε:



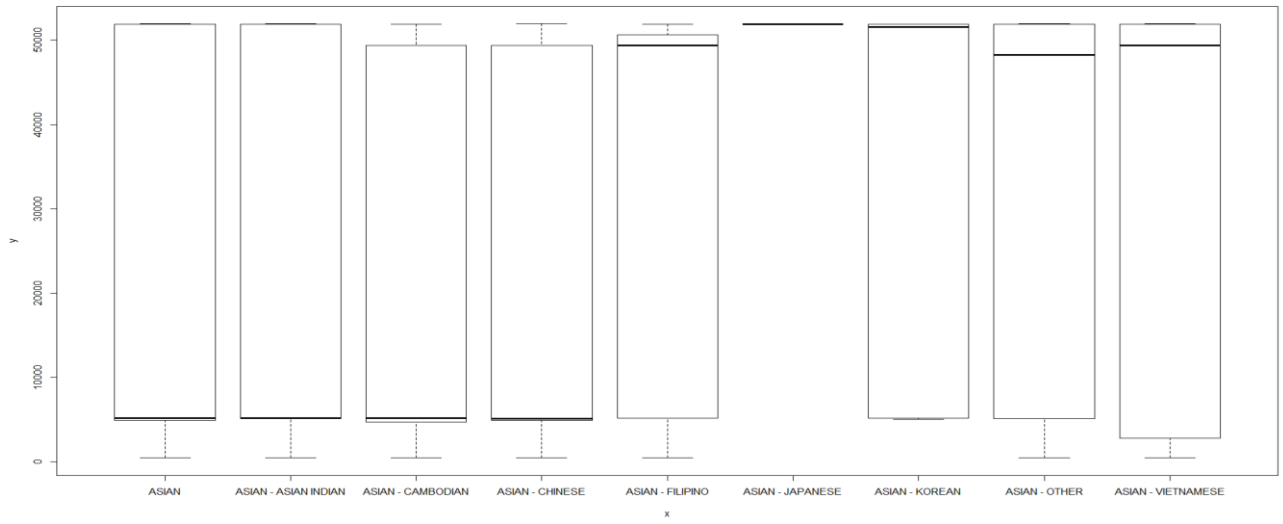
Σχήμα 5.13. Θηκόγραμμα ethnicity (white) συναρτήσει του icd9_code



Σχήμα 5.14. Θηκόγραμμα ethnicity (black) συναρτήσει του icd9_code



Σχήμα 5.15. Θηκόγραμμα ethnicity (hispanic) συναρτήσει του icd9_code



Σχήμα 5.16. Θηκόγραμμα ethnicity (asian) συναρτήσει του icd9_code

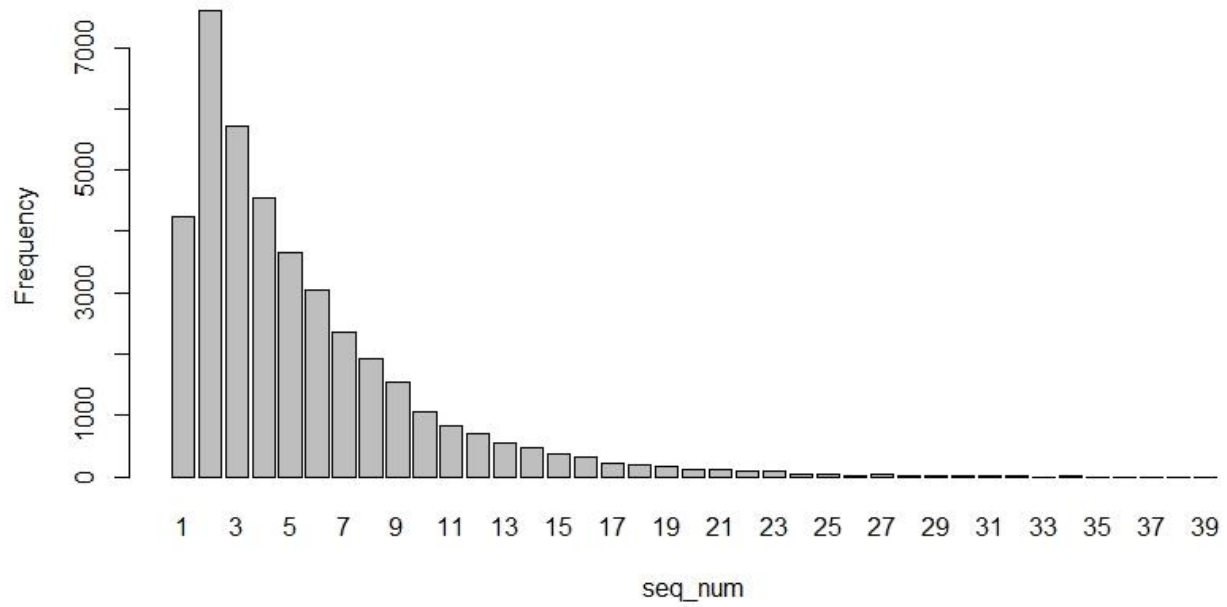
Στον Πίνακα 5.17 παρουσιάζονται και αριθμητικά οι πιο συχνοί κωδικοί icd9 ανά εθνικότητα.

Προκειμένου να συνεχίσουμε την ανάλυση, αφαιρέσαμε κάποια χαρακτηριστικά που πια δεν πρόκειται να χρειαστούμε. Αυτά είναι τα εξής: admittime, dod και admitdate διότι υπολογίσαμε την ηλικία του ασθενή, language και religion, που εκτός των άλλων έχουν και πολλές κενές τιμές και deathtime διότι είναι ίδια με το dishtime. Για να παραμείνει η πληροφορία του αν ο ασθενής είναι ζωντανός ή όχι διατηρήσουμε τη στήλη expire_flag.

ETHNICITY	ADMISSION	MOST COMMON	ICD9 FREQUENCY	ICD9
HISPANIC OR LATINO	945	51881	165	17,46
HISPANIC/LATINO CENTRAL AMERICAN (OTHER)	7		2	28,57
HISPANIC/LATINO - COLOMBIAN	6		1	16,67
HISPANIC/LATINO - CUBAN	5		1	20,00
HISPANIC/LATINO - DOMINICAN	46	486	6	13,04
HISPANIC/LATINO - GUATEMALAN	19	486	4	21,05
HISPANIC/LATINO - HONDURAN	1	51851	1	100,00
HISPANIC/LATINO - MEXICAN	6	51881	2	33,33
HISPANIC/LATINO - PUERTO RICAN	226	49390	34	15,04
HISPANIC/LATINO - SALVADORAN	4		1	25,00
BLACK/AFRICAN	32	51881	7	21,88
BLACK/AFRICAN AMERICAN	3836	51881	722	18,82
BLACK/CAPE VERDEAN	108	51881	20	18,52
BLACK/HAITIAN	52	486	8	15,38
WHITE	28831	51881	4908	17,02
WHITE - BRAZILIAN	28	51881	8	28,57
WHITE - EASTERN EUROPEAN	16	5121	3	18,75
WHITE - OTHER EUROPEAN	46	496	10	21,74
WHITE - RUSSIAN	86	51881	17	19,77
ASIAN	597	51881	122	20,44
ASIAN - INDIAN	45	5180	10	22,22
ASIAN - CAMBODIAN	15		2	13,33
ASIAN - CHINESE	199	486	27	13,57
ASIAN - FILIPINO	23	49390	7	30,43
ASIAN - JAPANESE	1	51881	1	100,00
ASIAN - KOREAN	6	51881	2	33,33
ASIAN - OTHER	9		1	11,11
ASIAN - VIETNAMESE	24	51881	8	33,33

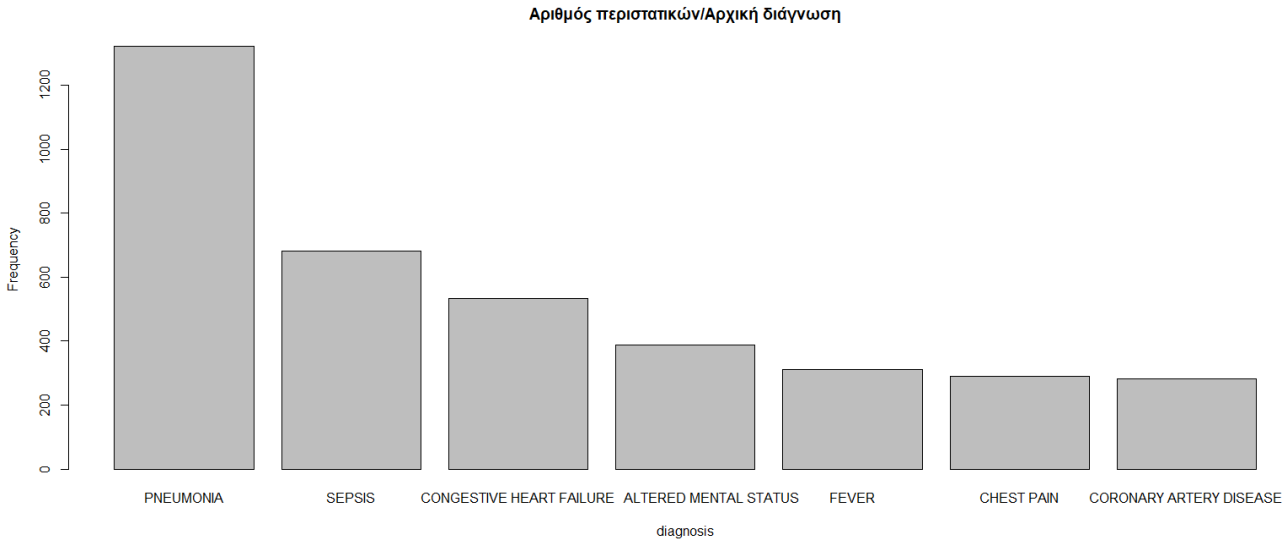
Πίνακας 5.17. Οι πιο συχνοί κωδικοί icd9 ανά εθνικότητα και τα ποσοστά επί των περιστατικών

Όπως αναφέρθηκε και στην ενότητα 3 το seq_num περιγράφει τη σειρά με την οποία οι διαγνώσεις ICD σχετίζονται με τον ασθενή. Οι διαγνώσεις ICD ταξινομούνται κατά προτεραιότητα και η σειρά έχει αντίκτυπο στην επιστροφή για θεραπεία. Οπτικοποιώντας, λοιπόν, τα δεδομένα μας, στο Σχήμα 5.18 βλέπουμε πόσες από τις διαγνώσεις έχουν ποια σειρά. Διαγνώσεις ICD με seq_num=1 αντιστοιχούν στο 10,57% των διαγνώσεων, με seq_num=2 στο 29,49%, με seq_num=3 στο 14,26% και με seq_num=4 στο 11,31%, που συνεπάγεται ότι συνολικά αντιστοιχούν στο 55,06%.



Σχήμα 5.18. Η συχνότητα της σειράς με την οποία οι διαγνώσεις ICD σχετίζονται με τον ασθενή

Οι συχνότητα των αρχικών διαγνώσεων για κάθε περιστατικό παρουσιάζεται στο Σχήμα 5.19. Σε αυτό το σημείο αξίζει να επισημανθεί ότι το κείμενο αυτό εισάγεται χειρόγραφα, που σημαίνει ότι ένας λίγο διαφορετικός τρόπος γραφής, για την ίδια αρχική διάγνωση αθροίζεται ξεχωριστά.



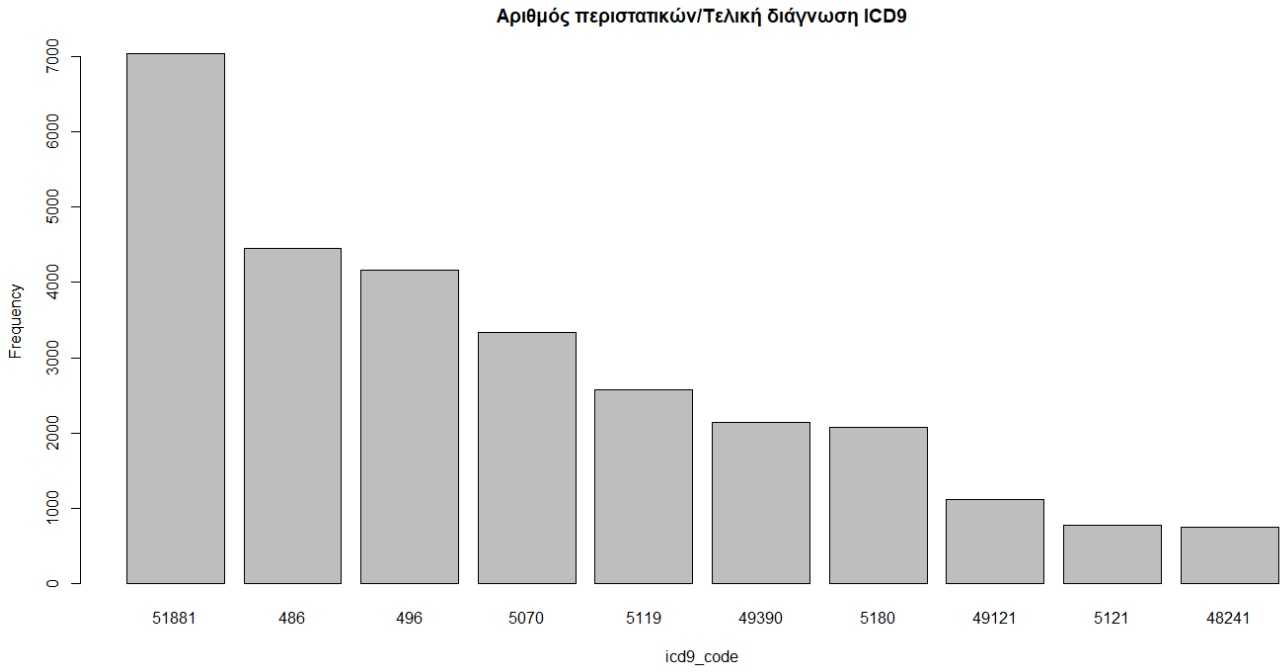
Σχήμα 5.19. Αριθμός περιστατικών ανά αρχική διάγνωση

Στον πίνακα 5.20 παρουσιάζονται οι 15 συχνότερες αρχικές διαγνώσεις. Στη θέση 13 βρίσκεται η εγγραφή ASTHMA;COPD, που υποδηλώνει μια αβεβαιότητα σχετικά με τον αν τα συμπτώματα προέρχονται από το Άσθμα ή από Χρόνια Αποφρακτική Πνευμονοπάθεια (COPD). Αυτή είναι και η αφορμή για να εφαρμόσουμε στην ενότητα 5.3 αλγορίθμους ταξινόμησης και να εξετάσουμε εάν μπορεί κάποιος από αυτούς να λειτουργήσει βοηθητικά.

	diagnosis	frequency
1	PNEUMONIA	1321
2	SEPSIS	682
3	CONGESTIVE HEART FAILURE	534
4	ALTERED MENTAL STATUS	389
5	FEVER	311
6	CHEST PAIN	290
7	CORONARY ARTERY DISEASE	281
8	ABDOMINAL PAIN	276
9	INTRACRANIAL	267
10	GASTROINTESTINAL BLEED	228
11	RESPIRATORY FAILURE	225
12	HYPOTENSION	213
13	ASTHMA;COPD	187
14	UPPER GI BLEED	184
15	DYSPNEA	181

Πίνακας 5.20. Οι 15 συχνότερες αρχικές διαγνώσεις

Στο Σχήμα 5.21 παρουσιάζεται ο κωδικός ICD9 για κάθε τελική διάγνωση. Σημειώνουμε ότι όπως προκύπτει και από το Σχήμα 5.19, σε κάθε περιστατικό μπορεί να αντιστοιχούν περισσότεροι από έναν κωδικοί icd9. Συγκρίνοντας τα δύο Σχήματα, παρατηρούμε ότι δεν υπάρχει απόλυτη ταύτιση.



Σχήμα 5.21. Συχνότητα εμφάνισης των διαγνώσεων ICD9 στα περιστατικά

	icd9_code	Freq	Final diagnosis
1	51881	7497	Acute respiratory failure
2	486	4839	Pneumonia, organism unspecified
3	496	4431	Chronic airway obstruction, not elsewhere classified
4	5070	3680	Pneumonitis due to inhalation of food or vomitus
5	5119	2734	Unspecified pleural effusion
6	49390	2195	Asthma, unspecified type, unspecified
7	5180	2165	Pulmonary collapse
8	49121	1198	Obstructive chronic bronchitis with (acute) exacerbation
9	48241	789	Methicillin susceptible pneumonia due to Staphylococcus aureus
10	5121	786	Iatrogenic pneumothorax
11	51882	761	Other pulmonary insufficiency, not elsewhere classified
12	49320	752	Chronic obstructive asthma, unspecified
13	51889	694	Other diseases of lung, not elsewhere classified
14	51884	684	Acute and chronic respiratory failure
15	4928	641	Other emphysema

Πίνακας 5.22. Οι 15 συχνότερες τελικές διαγνώσεις ICD9

5.2. Συσταδοποίηση των δεδομένων

Από το data frame «df» κρατήσαμε μόνο τις εγγραφές με diagnosis %in% c("ASTHMA","ASTHMA EXACERBATION","COPD EXACERBATION","ASTHMA;COPD EXACERBATION","COPD"). Επίσης προκειμένου να ετοιμάσουμε τα δεδομένα για συσταδοποίηση, αφαιρέσαμε τις στήλες: hadm_id, subject_id, dischtime, dishdate, marital_status, dob, age και race. Έτσι προέκυψε το data frame dfcluster με 675 γραμμές και 10 στήλες, η δομή του οποίου παρουσιάζεται στο Σχήμα 5.23.

```
> str(dfcluster)
'data.frame':   675 obs. of  10 variables:
 $ discharge_location  : Factor w/ 17 levels "DEAD/EXPIRED",..: 12 12 5 5 16 16 5 5 5 5 ...
 $ insurance           : Factor w/ 5 levels "Government","Medicaid",..: 3 3 3 3 3 3 3 4 4 ...
 $ ethnicity           : Factor w/ 40 levels "AMERICAN INDIAN/ALASKA NATIVE",..: 17 17 13 13 36 36 1
 ...
 $ diagnosis           : Factor w/ 8419 levels "","   DUODENAL MASS/SDA",..: 1014 1014 986 986 1014
1014 978 978 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
 $ gender              : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ expire_flag         : int  1 1 0 0 1 1 0 0 0 0 ...
 $ agegroup            : Factor w/ 6 levels ">80","ADOLESCEND",..: 3 3 4 4 1 1 4 4 4 4 ...
 $ seq_num             : int  1 2 1 2 1 3 1 2 1 2 ...
 $ icd9_code           : int  49322 51881 49392 51881 49121 51189 49392 486 49390 486 ...
```

Σχήμα 5.23. Η δομή του πίνακα dfcluster

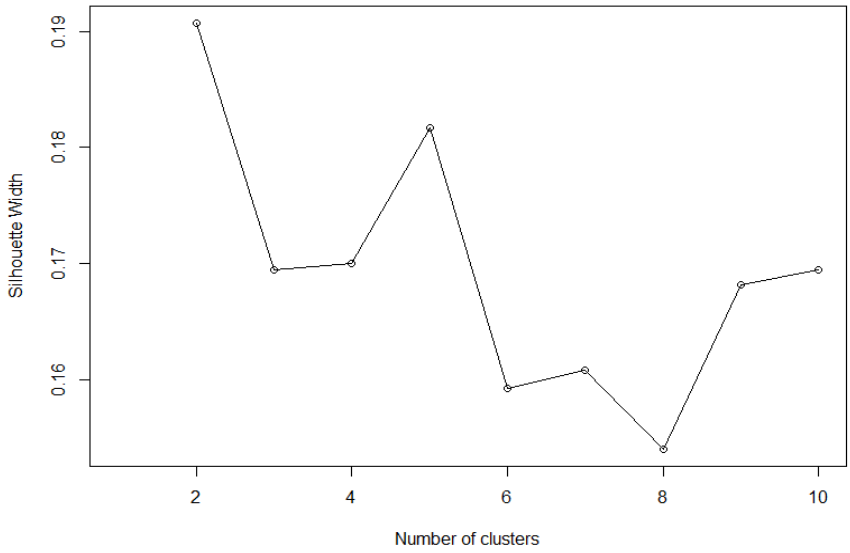
Για να μπορέσουμε να παρατηρήσουμε τα αποτελέσματα της συσταδοποίησης θα ξεκινήσουμε εφαρμόζοντας τον αλγόριθμο tsne. Ο t-Distributed Stochastic Neighbor Embedding είναι ένας μη γραμμικός αλγόριθμος μείωσης των διαστάσεων που χρησιμοποιείται για την εξερεύνηση δεδομένων πολλαπλών διαστάσεων. Χαρτογραφεί πολυδιάστατα δεδομένα σε δύο ή περισσότερες διαστάσεις κατάλληλες για ανθρώπινη παρατήρηση. [11]

K-means

Ο πρώτος αλγόριθμος που δοκιμάσαμε ήταν η παραλλαγή του K-means, ο K-medoid. Πρώτο βήμα ήταν ο υπολογισμός του συντελεστή σιλουέτας για να μπορέσουμε να επιλέξουμε το K, δηλαδή σε πόσες συστάδες

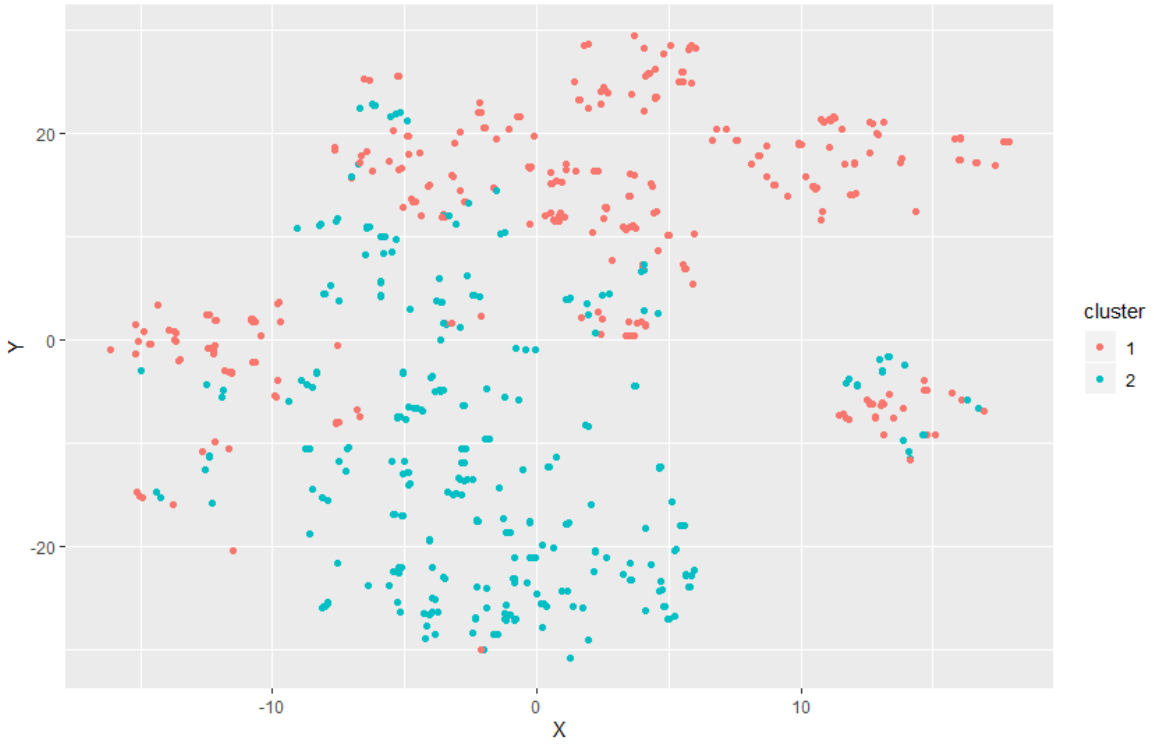
θα χωρίσουμε τα δεδομένα. Στο Σχήμα 5.24 φαίνεται το εύρος του συντελεστή σιλουέτας για διαφορετικά K. Βάσει αυτού δοκιμάζουμε για K=2 και K=5, αν και αρχικά το αποτέλεσμα δεν είναι ιδιαίτερα ικανοποιητικό.

Ο επόμενος στόχος μας είναι να καθορίσουμε τον βέλτιστο αριθμό συστάδων. Η επιλογή μας θα βασίζεται στις τιμές του συντελεστή σιλουέτας που χρησιμοποιείται ως μέτρο του πόσο παρόμοιο είναι ένα αντικείμενο με τη δική του συστάδα (συνοχή-cohesion) σε σύγκριση με τις υπόλοιπες (διαχωρισμός-separation).

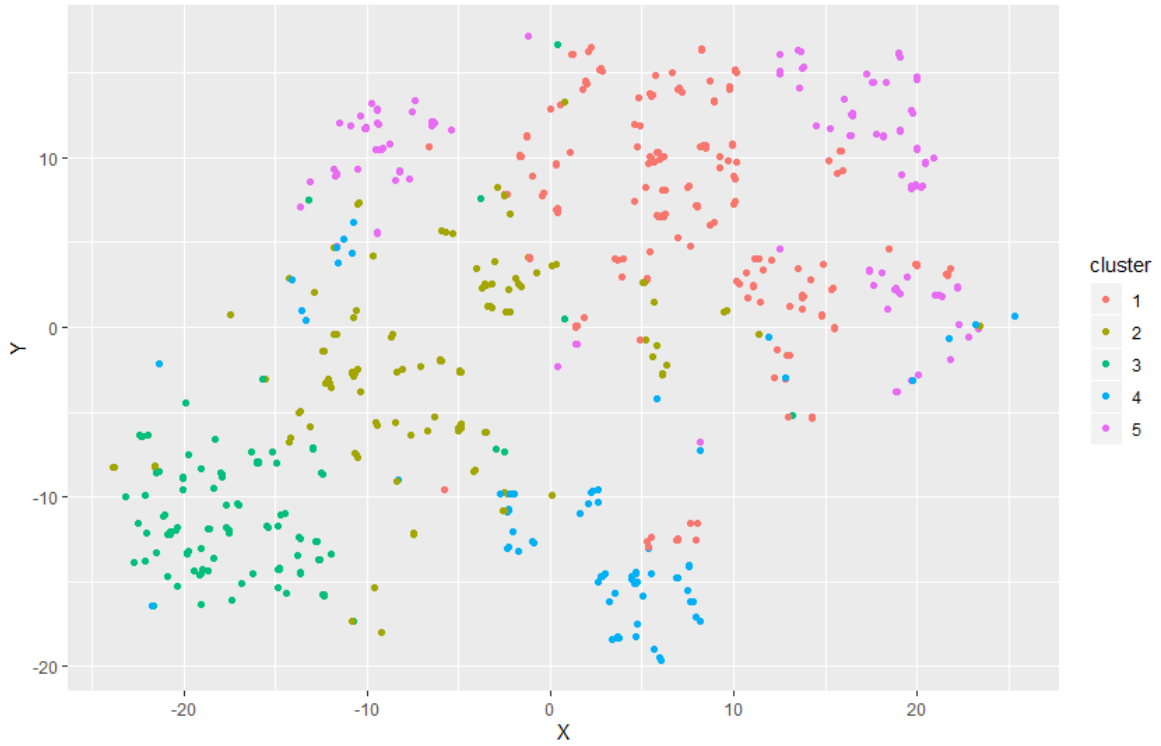


Σχήμα 5.24. Εύρος του συντελεστή σιλουέτας των δεδομένων dfcluster για διαφορετικό αριθμό συστάδων K

Έτσι, για K=2 μια απεικόνιση στο Σχήμα 5.25. Αντίστοιχα για K=5 προκύπτουν τα Σχήματα 5.26



Σχήμα 5.25. Απεικόνιση των συστάδων για K=2 για τα δεδομένα dfcluster



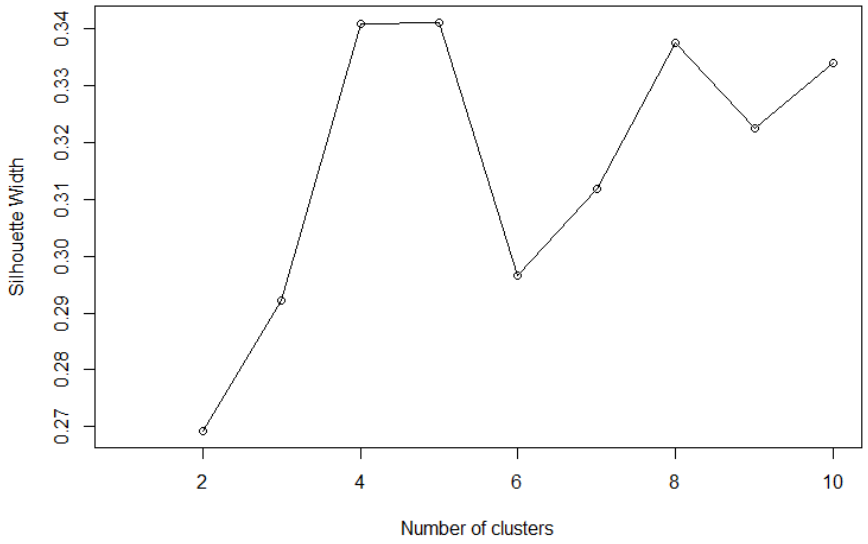
Σχήμα 5.26. Απεικόνιση των συστάδων για K=5 για τα δεδομένα dfcluster

Επειδή καμία από τις δύο προσεγγίσεις δεν είναι ικανοποιητικές θα μετατρέψουμε όλα τα χαρακτηριστικά του πίνακα σε αριθμητικά (Σχήμα 5.27). Στη συνέχεια θα τα ανάγουμε σε δύο διαστάσεις για να μπορέσουμε να τα οπτικοποιήσουμε (Σχήμα 5.28) και όπως προκύπτει από το Σχήμα 5.29, θα ακολουθήσουμε την παραπάνω διαδικασία με K=4 και για K=5 [12] Επιπλέον θα εφαρμόσουμε τα δεδομένα του Σχήματος 5.27 τον αλγόριθμο Ιεραρχικής συσταδοποίησης και τον αλγόριθμο DBSCAN

```
> str(data_tsne)
'data.frame':   675 obs. of  10 variables:
 $ discharge_location  : int  9 9 4 4 12 12 4 4 4 4 ...
 $ insurance           : int  3 3 3 3 3 3 3 3 4 4 ...
 $ ethnicity           : int  6 6 4 4 13 13 4 4 13 13 ...
 $ diagnosis           : int  3 3 2 2 3 3 3 3 1 1 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
 $ gender              : int  1 1 1 1 1 1 1 1 1 1 ...
 $ expire_flag         : int  1 1 0 0 1 1 0 0 0 0 ...
 $ agegroup            : int  3 3 4 4 1 1 4 4 4 4 ...
 $ seq_num             : int  1 2 1 2 1 3 1 2 1 2 ...
 $ icd9_code           : int  49322 51881 49392 51881 49121 51189 49392 486 49390 486 ...
```

Σχήμα 5.27. Δομή του πίνακα data_tsne

Σχήμα 5.28. Απεικόνιση των δεδομένων data_tsne σε δύο διαστάσεις

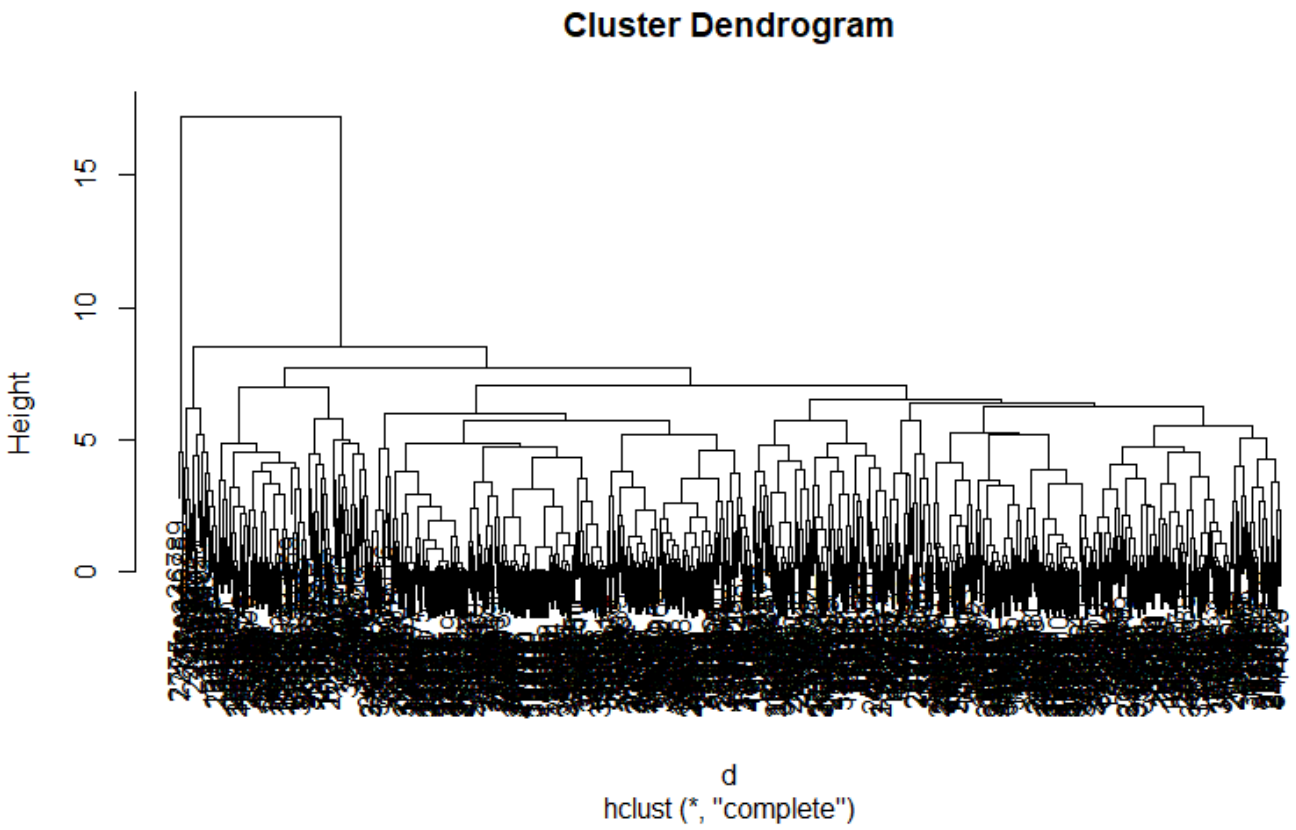


Σχήμα 5.29. Εύρος του συντελεστή σιλουέτας των δεδομένων data_tsne για διαφορετικό αριθμό K

Ιεραρχική συσταδοποίηση

Η ιεραρχική συσταδοποίηση δημιουργεί συστάδες μέσα σε συστάδες και δεν απαιτεί προκαθορισμένο αριθμό όπως ο K-means. Μια ιεραρχική συσταδοποίηση μπορεί να θεωρηθεί ως δέντρο και εμφανίζεται ως ένα δενδρόγραμμα. στην κορυφή υπάρχει μόνο μία συστάδα που αποτελείται από όλες τις παρατηρήσεις και στο κάτω μέρος κάθε παρατήρηση είναι μια ολόκληρη ομάδα. Στο μεταξύ υπάρχουν ποικίλα επίπεδα ομαδοποίησης.

Στα δεδομένα μας το δενδρόγραμμα δεν είναι ξεκάθαρο οπτικά (Σχήμα 5.30), λόγω του μεγάλου αριθμού των διαστάσεων. Το αποτέλεσμα της συσταδοποίησης φαίνεται καλύτερα στο Σχήμα 5.31.

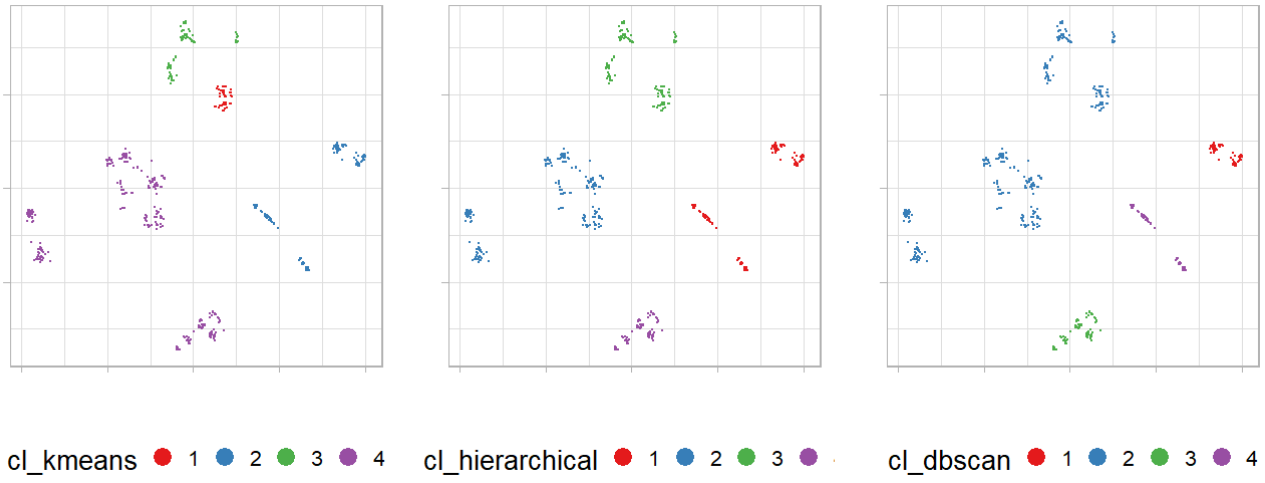


Σχήμα 5.30. Δενδρόγραμμα ιεραρχικής συσταδοποίησης

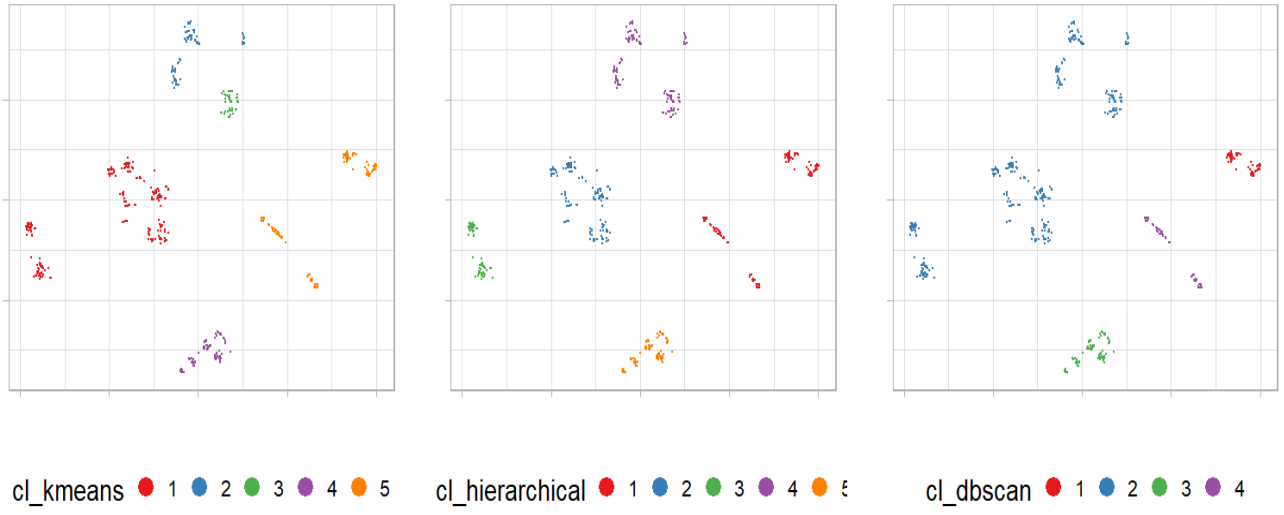
DBSCAN

Ο στόχος είναι να εντοπιστούν πυκνές περιοχές, οι οποίες μπορούν να μετρηθούν με τον αριθμό αντικειμένων που βρίσκονται κοντά σε ένα δεδομένο σημείο. Για τον DBSCAN απαιτούνται δύο σημαντικές παράμετροι: το epsilon ("eps") και τα ελάχιστα σημεία ("MinPts"). Η παράμετρος eps ορίζει την ακτίνα της γειτονιάς γύρω από ένα σημείο x. Η παράμετρος MinPts είναι ο ελάχιστος αριθμός γειτόνων εντός της ακτίνας "eps". Εμείς ορίσαμε eps = 1.0, MinPts = 4.

Τα αποτελέσματα και για τους τρεις αλγόριθμους οπτικοποιούνται στα Σχήματα 5.31 και 5.32, για τέσσερεις και πέντε συστάδες αντίστοιχα.



Σχήμα 5.31. Απεικόνιση των συστάδων για K=4 για τα δεδομένα data_tsne χρησιμοποιώντας τους αλγορίθμους K-means, hierarchical clustering και dbscan.



Σχήμα 5.32. Απεικόνιση των συστάδων για K=5 για τα δεδομένα data_tsne χρησιμοποιώντας τους αλγορίθμους K-means, hierarchical clustering και dbscan.

5.3. Ταξινόμηση-Χ.Α.Π. ή Άσθμα

Στον Πίνακα 5.20. παρατηρούμε ότι η δέκατη τρίτη πιο συχνή αρχική διάγνωση είναι η «ASTHMA/COPD EXACERBATION». Με βάση αυτό το στοιχείο δημιουργήθηκε ο προβληματισμός για το αν θα μπορούσε να γίνει μια ταξινόμηση βασισμένη στα δημογραφικά χαρακτηριστικά των ασθενών, αλλά και σε κάποιες τιμές από τον πίνακα των LABEVENTS. Έτσι, αρχικά, από τον πίνακα df κρατήσαμε τις εγγραφές όπου diagnosis %in% c("ASTHMA","ASTHMA EXACERBATION", "COPD EXACERBATION", "ASTHMA;COPD EXACERBATION", "COPD") και δημιουργήσαμε τον dfullAsthmaCOPD. Ακολούθως, αφαιρέθηκαν τα χαρακτηριστικά: dob, age, dishtime, dishdate, insurance,subject_id και diagnosis.

Επίσης, με βάση μόνο τους αριθμούς hadm_id των παραπάνω περιστατικών δημιουργήθηκε ένα data frame από τον πίνακα των LABEVENTS, από το οποίο αφαιρέθηκαν τα χαρακτηριστικά subject_id, valuenum,

valueuom και row_id. Επιπλέον, το charttime χαρακτηρίστηκε ως POSIXct (μια μορφή που συνδυάζει ημερομηνία με ώρα και αντιπροσωπεύει τον αριθμό δευτερολέπτων από τις αρχές του 1970 ως αριθμητικό διάνυσμα) αντί για factor και το flag που υποδηλώνει εάν η τιμή της κάθε εξέτασης είναι εντός των φυσιολογικών ορίων, πήρε τις τιμές 0 ή 1 με ένα να αντιστοιχεί στο «abnormal». Ακολούθησε η διατήρηση μόνο των εγγραφών item_id ∃ (50800,50802,50820,50804,50818,50821,50812,50825,50816) οι οποίες είναι εξετάσεις στα αέρια του αίματος, που σχετίζονται με το COPD [13]. Ο πίνακας «απλώθηκε» με κάθε την τιμή για κάθε διαφορετικό item_id να αποτελεί και μια ξεχωριστή στήλη και αφαιρέθηκαν οι εγγραφές με κενές τιμές, με αποτέλεσμα ο πίνακας να αποτελείται από 644 σειρές.

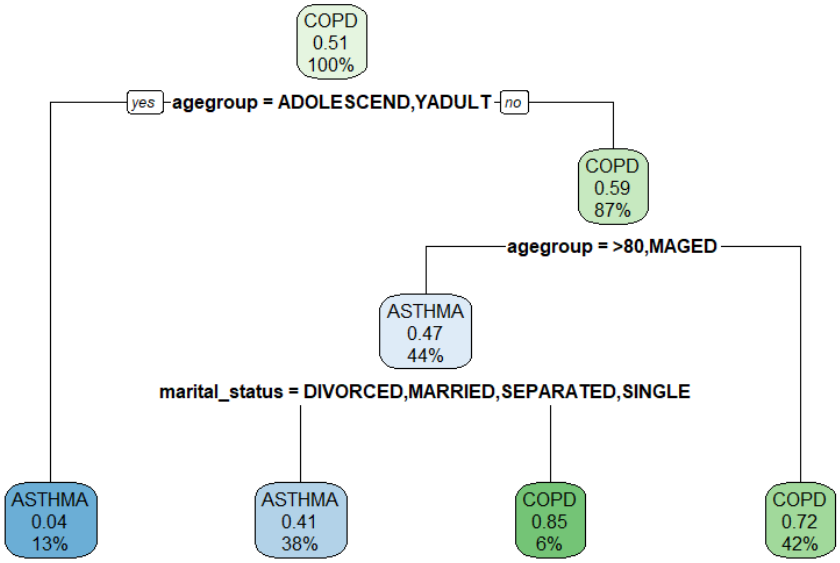
Το κομμάτι της προετοιμασίας για την ταξινόμηση ολοκληρώθηκε με την ένωση των data frame dfAsthmaCOPD και labAsthmaCOPDsp2, δημιουργώντας το dfclassif, από τον οποίο διατηρήθηκαν μόνο οι πρώτες χρονικά τιμές για κάθε item_id. Έτσι ο τελικός πίνακας πάνω στον οποίο θα εφαρμοστούν τα μοντέλα της ταξινόμησης αποτελείται από 261 στοιχεία με 15 χαρακτηριστικά, ένα εκ των οποίων είναι η κλάση Χ.Α.Π. ή Άσθμα. Η δομή τους φαίνεται στο Σχήμα 5.33.

```
> str(dfclassif.sub)
'data.frame':   261 obs. of  15 variables:
 $ marital_status      : Factor w/ 7 levels "","DIVORCED",..: 5 5 7 5 5 5 3 2 3 5 ...
 $ ethnicity           : Factor w/ 14 levels "AMERICAN INDIAN/ALASKA NATIVE",..: 6 4 13 13 4 4 13 13 4 13 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
 $ gender              : Factor w/ 2 levels "F","M": 1 1 1 1 2 1 2 1 2 2 ...
 $ agegroup            : Factor w/ 5 levels ">80","ADOLESCEND",..: 3 4 1 4 4 4 1 1 3 4 ...
 $ 50800               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50802               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50804               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50812               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50816               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50818               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50820               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50821               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ 50825               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ icd9_category       : chr  "ASTHMA" "ASTHMA" "COPD" "ASTHMA" ...
```

Σχήμα 5.33. Δομή του πίνακα dfclassif.sub

Δέντρο Απόφασης

Ο πρώτος αλγόριθμος ταξινόμησης που εφαρμόστηκε είναι το δέντρο απόφασης, που είναι ένας ισχυρός μη γραμμικός ταξινομητής εποπτευόμενης μηχανικής μάθησης. Αρχικά τα δεδομένα χωρίστηκαν σε training και testing σετ κατά αναλογία 80% προς 20% με τυχαίο τρόπο, έτσι ώστε να είναι όσο το δυνατόν πιο ομοιογενή. Έτσι προέκυψε το διάγραμμα του Σχήματος 5.34(α). Για την εκπαίδευση του μοντέλου χρησιμοποιήσαμε και τα 14 χαρακτηριστικά του πίνακα dfclassif.sub



Σχήμα 5.34(α). Δέντρο Απόφασης

Το βασικότερο κριτήριο που αποτελεί και τη ρίζα του δέντρου μας είναι η ηλικιακή ομάδα στην οποία ανήκει το περιστατικό. Αν κάποιος είναι κάτω από 45 ετών, το πιθανότερο είναι να πάσχει από Άσθμα. Αν είναι άνω των 80 ετών και μεταξύ 45 και 65 ετών, η οικογενειακή του κατάσταση φαίνεται να έχει και αυτή σημασία.

Αξιολογώντας το αλγόριθμο, παρατηρούμε ότι η ακρίβεια ταξινόμησης (Accuracy), είναι αρκετά κακή στο 67,92%, όπως φαίνεται στο Σχήμα 5.33(b). Στο επάνω κομμάτι του ίδιου σχήματος φαίνεται και η μήτρα σύγχυσης (Confusion Matrix), που μας δείχνει έναν πίνακα του τρόπου με τον οποίο συσχετίζονται οι προβλέψεις και οι πραγματικές κλάσεις. Έτσι τα διαγώνια κελιά, όπου η πρόβλεψη και η κλάση είναι οι ίδιες, αντιπροσωπεύουν τη σωστή πρόβλεψη. Έτσι (16+20)x100/53 μας δίνει την τιμή της ακρίβειας του αλγορίθμου.

Η ευαισθησία (Sensitivity), η μέτρηση δηλαδή που αντιστοιχεί στο ότι ασθενής πάσχει από άσθμα, όντως ταξινομήθηκε σωστά, είναι 84,21%. Η εξειδίκευση (Specificity) από την άλλη πλευρά, η μέτρηση δηλαδή που αντιστοιχεί στο ότι ένας ασθενής που πάσχει από Χ.Α.Π. όντως ταξινομήθηκε σωστά, είναι 58,82%.

```
Confusion Matrix and Statistics

          Reference
Prediction ASTHMA COPD
    ASTHMA     16   14
    COPD        3   20

               Accuracy : 0.6792
                 95% CI : (0.5368, 0.8008)
    No Information Rate : 0.6415
    P-Value [Acc > NIR] : 0.33802

                  Kappa : 0.3816

 Mcnemar's Test P-Value : 0.01529

            Sensitivity : 0.8421
            Specificity : 0.5882
         Pos Pred Value : 0.5333
         Neg Pred Value : 0.8696
             Prevalence : 0.3585
         Detection Rate : 0.3019
   Detection Prevalence : 0.5660
      Balanced Accuracy : 0.7152

       'Positive' Class : ASTHMA
```

Σχήμα 5.34(b). Μήτρα σύγχυσης και μετρικές για το δέντρο απόφασης

Naive Bayes

Στη συνέχεια θα δοκιμάσουμε τον αλγόριθμο Naive Bayes, που βασίζεται στο Θεώρημα Bayes που χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης ακολουθώντας μια πιθανοτική προσέγγιση. Βασίζεται στην ιδέα ότι οι μεταβλητές πρόβλεψης σε ένα μοντέλο μηχανικής μάθησης είναι ανεξάρτητες η μία από την άλλη. Αυτό σημαίνει ότι το αποτέλεσμα ενός μοντέλου εξαρτάται από ένα σύνολο ανεξάρτητων μεταβλητών που δεν έχουν καμία σχέση μεταξύ τους.

Όμως σε πραγματικά προβλήματα, οι μεταβλητές πρόβλεψης δεν είναι πάντα ανεξάρτητες μεταξύ τους, υπάρχουν πολλές φορές συσχετίσεις. Αυτός είναι και ο λόγος που το μοντέλο καλείται «Naive» Bayes

Αρχικά θα πρέπει να μετατρέψουμε τα χαρακτηριστικά του training set από «num» σε «factor» και «int», όπως στην μορφή σχήματος 5.35.

```
> str(train)
'data.frame':   208 obs. of  15 variables:
 $ marital_status      : Factor w/ 7 levels "","DIVORCED",..: 2 3 2 3 5 5 5 7 7 2 ...
 $ ethnicity           : Factor w/ 14 levels "AMERICAN INDIAN/ALASKA NATIVE",..: 9 4 13 4 8 4 13 13 13 13 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
 $ gender              : Factor w/ 2 levels "F","M": 2 1 1 1 2 1 2 1 1 2 ...
 $ agegroup            : Factor w/ 5 levels ">80","ADOLESCEND",..: 3 5 3 4 3 3 3 4 4 3 ...
 $ 50800               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50802               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50804               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50812               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50816               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50818               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50820               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50821               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50825               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ icd9_category       : Factor w/ 2 levels "ASTHMA","COPD": 2 1 2 1 2 2 2 2 1 2 ...
```

Σχήμα 5.35. μορφή του training set για την εφαρμογή του μοντέλου ταξινομητή Naive Bayes

Αφού τρέξαμε το κλασικό μοντέλο και το μοντέλο με εξομάλυνση Laplace υπολογίστηκαν οι πιθανότητες για κάθε χαρακτηριστικό, που παρουσιάζονται στα Σχήματα 5.36(α) και 5.36(b). Δυστυχώς και τα δύο μοντέλα του Naive Bayes έβγαλαν ακόμα χειρότερα αποτελέσματα από αυτά του δέντρου απόφασης, με ακρίβεια 60,38%, πράγμα που εν μέρει ήταν αναμενόμενο, για τους λόγους που αναφέρθηκαν παραπάνω. Αυτή η πολύ μέτρια ταξινόμηση οπτικοποιείται από την καμπύλη ROC του Σχήματος 5.38.

Η συνάρτηση naiveBayes υποθέτει κατανομές γκαουσιανές για τις αριθμητικές μεταβλητές. Επίσης, οι a priori πιθανότητες υπολογίζονται από την αναλογία των δεδομένων εκπαίδευσης. Οι τιμές Y είναι οι μέσοι και οι τυπικές αποκλίσεις των προγνωστικών σε κάθε κλάση.

Διαφορές στις μετρικές αξιολόγησης των δύο μοντέλων (accuracy, sensitivity, specificity) δεν υπάρχουν. Με την εξομάλυνση Laplace είναι υπολογισμένες λίγο διαφορετικά οι a priori πιθανότητες για των χαρακτηριστικών.

Η ευαισθησία (Sensitivity), του μοντέλου του Naive Bayes όπως φαίνεται στο Σχήμα 5.37 είναι 46,43% και η εξειδίκευση (Specificity) είναι 76,00%. Αυτό συνεπάγεται ότι παρ' όλο που η ακρίβεια του μοντέλου είναι χειρότερη από αυτή του δέντρου απόφασης, λειτουργεί καλύτερα στη σωστή πρόβλεψη ασθενών με Χ.Α.Π.

```
> nb_default

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
   ASTHMA      COPD
0.4423077 0.5576923

Conditional probabilities:
        marital_status
Y                    DIVORCED    MARRIED  SEPARATED     SINGLE UNKNOWN (DEFAULT)    WIDOWED
  ASTHMA 0.01086957 0.06521739 0.34782609 0.02173913 0.50000000        0.02173913 0.03260870
  COPD   0.02586207 0.14655172 0.31034483 0.01724138 0.30172414        0.00000000 0.19827586

        ethnicity
Y        AMERICAN INDIAN/ALASKA NATIVE      ASIAN ASIAN - CHINESE BLACK/AFRICAN AMERICAN BLACK/HAITIAN HISPANIC OR LATINO
  ASTHMA                    0.00000000 0.01086957      0.00000000             0.28260870    0.00000000          0.07608696
  COPD                      0.00862069 0.00000000      0.00862069             0.18103448    0.00862069          0.00000000
        ethnicity
Y        HISPANIC/LATINO - DOMINICAN HISPANIC/LATINO - PUERTO RICAN MULTI RACE ETHNICITY      OTHER PATIENT DECLINED TO ANSWER
  ASTHMA                   0.00000000                     0.02173913           0.00000000 0.02173913                   0.00000000
  COPD                     0.00862069                     0.05172414           0.00862069 0.00862069                   0.00000000
        ethnicity
Y        UNKNOWN/NOT SPECIFIED      WHITE WHITE - RUSSIAN
  ASTHMA            0.03260870 0.54347826      0.01086957
  COPD              0.00862069 0.69827586      0.00862069

        has_chartevents_data
Y        [,1] [,2]
  ASTHMA    1    0
  COPD      1    0

        gender
Y                F         M
  ASTHMA 0.6630435 0.3369565
  COPD   0.5431034 0.4568966

        agegroup
Y               >80 ADOLESCEND       AGED      MAGED     YADULT
  ASTHMA 0.05434783 0.01086957 0.18478261 0.45652174 0.29347826
  COPD   0.10344828 0.00000000 0.56896552 0.32758621 0.00000000

        50800
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

        50802
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

        50804
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

        50812
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

        50816
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

        50818
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

        50820
Y               [,1]      [,2]
  ASTHMA 0.03260870 0.1785834
  COPD   0.01724138 0.1307343

        50821
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

        50825
Y        [,1] [,2]
  ASTHMA    0    0
  COPD      0    0
```

Σχήμα 5.36(α). Naive Bayes: πιθανότητες για κάθε χαρακτηριστικό

```
> nb_laplace

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
   ASTHMA      COPD
0.4423077 0.5576923

Conditional probabilities:
       marital_status
Y                    DIVORCED     MARRIED   SEPARATED      SINGLE UNKNOWN (DEFAULT)     WIDOWED
  ASTHMA 0.020202020 0.070707071 0.333333333 0.030303030 0.474747475       0.030303030 0.040404040
  COPD   0.032520325 0.146341463 0.300813008 0.024390244 0.292682927       0.008130081 0.195121951

       ethnicity
Y       AMERICAN INDIAN/ALASKA NATIVE       ASIAN ASIAN - CHINESE BLACK/AFRICAN AMERICAN BLACK/HAITIAN HISPANIC OR LATINO
  ASTHMA                    0.009433962 0.018867925     0.009433962             0.254716981   0.009433962        0.075471698
  COPD                      0.015384615 0.007692308     0.015384615             0.169230769   0.015384615        0.007692308
       ethnicity
Y       HISPANIC/LATINO - DOMINICAN HISPANIC/LATINO - PUERTO RICAN MULTI RACE ETHNICITY       OTHER PATIENT DECLINED TO ANSWER
  ASTHMA                 0.009433962                    0.028301887          0.009433962 0.028301887                0.009433962
  COPD                   0.015384615                    0.053846154          0.015384615 0.015384615                0.007692308
       ethnicity
Y       UNKNOWN/NOT SPECIFIED       WHITE WHITE - RUSSIAN
  ASTHMA           0.037735849 0.481132075     0.018867925
  COPD             0.015384615 0.630769231     0.015384615

       has_chartevents_data
Y       [,1] [,2]
  ASTHMA    1    0
  COPD      1    0

       gender
Y              F         M
  ASTHMA 0.6595745 0.3404255
  COPD   0.5423729 0.4576271

       agegroup
Y               >80  ADOLESCEND        AGED       MAGED      YADULT
  ASTHMA 0.061855670 0.020618557 0.185567010 0.443298969 0.288659794
  COPD   0.107438017 0.008264463 0.553719008 0.322314050 0.008264463

       50800
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

       50802
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

       50804
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

       50812
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

       50816
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

       50818
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

       50820
Y              [,1]      [,2]
  ASTHMA 0.03260870 0.1785834
  COPD   0.01724138 0.1307343

       50821
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0

       50825
Y       [,1] [,2]
  ASTHMA    0    0
  COPD      0    0
```

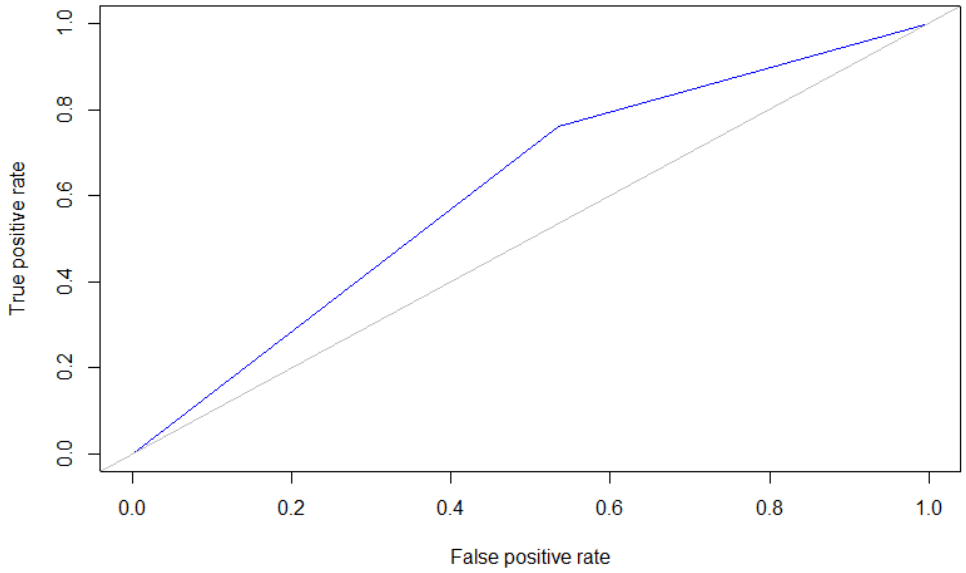
Σχήμα 5.36(b). Naive Bayes: Πιθανότητες για κάθε χαρακτηριστικό με εξομάλυνση Laplace

```
> confusionMatrix(data = default_pred, reference = test$icd9_category)
Confusion Matrix and Statistics

          Reference
Prediction ASTHMA COPD
    ASTHMA     13    6
    COPD       15   19

               Accuracy : 0.6038
                 95% CI : (0.46, 0.7355)
    No Information Rate : 0.5283
    P-Value [Acc > NIR] : 0.16784

                  Kappa : 0.22

 Mcnemar's Test P-Value : 0.08086

            Sensitivity : 0.4643
            Specificity : 0.7600
         Pos Pred Value : 0.6842
         Neg Pred Value : 0.5588
             Prevalence : 0.5283
         Detection Rate : 0.2453
   Detection Prevalence : 0.3585
      Balanced Accuracy : 0.6121

       'Positive' Class : ASTHMA
```

```
> confusionMatrix(data = laplace_pred, reference = test$icd9_category)
Confusion Matrix and Statistics

          Reference
Prediction ASTHMA COPD
    ASTHMA     13    6
    COPD       15   19

               Accuracy : 0.6038
                 95% CI : (0.46, 0.7355)
    No Information Rate : 0.5283
    P-Value [Acc > NIR] : 0.16784

                  Kappa : 0.22

 Mcnemar's Test P-Value : 0.08086

            Sensitivity : 0.4643
            Specificity : 0.7600
         Pos Pred Value : 0.6842
         Neg Pred Value : 0.5588
             Prevalence : 0.5283
         Detection Rate : 0.2453
   Detection Prevalence : 0.3585
      Balanced Accuracy : 0.6121

       'Positive' Class : ASTHMA
```

Σχήμα 5.37. Μήτρα σύγχυσης και μετρικές για τις δύο εκδοχές του Naive Bayes

Η καμπύλη ROC του σχήματος 5.38 μας δείχνει αυτό που διαπιστώσαμε και με τις αριθμητικές μετρικές παραπάνω, ότι έχουμε έναν πολύ μέτριο ταξινομητή που αποδίδει λίγο καλύτερα από το να μαντεύει τυχαία.

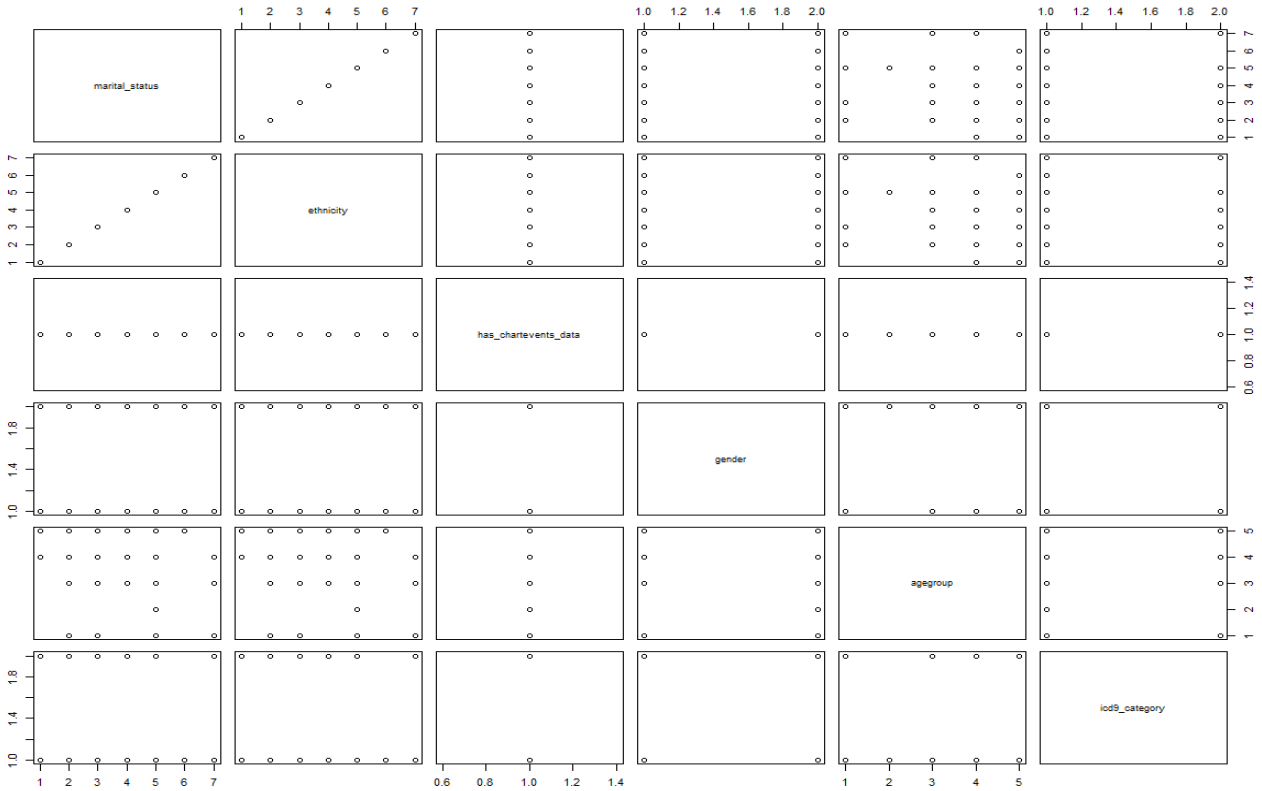


Σχήμα 5.38. Καμπύλη ROC

k-NN

Ο k-Nearest Neighbors (k-NN) είναι ένας αλγόριθμος που είναι χρήσιμος για την πραγματοποίηση ταξινομήσεων, όταν υπάρχουν δυνητικά μη γραμμικά όρια που διαχωρίζουν τις κλάσεις. Εννοιολογικά, το k-NN για να καθορίσει την κλάση ενός σημείου ενδιαφέροντος, εξετάζει τις κλάσεις των σημείων γύρω από αυτό (δηλ. Τους γείτονές του). Η πλειοψηφία ή η μέση τιμή θα αποδοθεί στο σημείο ενδιαφέροντος.

Προτού εφαρμόσουμε τον αλγόριθμο k-NN σχεδιάστηκαν τα σχετικά διαγράμματα των δεδομένων για τα δημογραφικά χαρακτηριστικά.



Σχήμα 5.39. Σχετικά διαγράμματα των δημογραφικών χαρακτηριστικών των δεδομένων

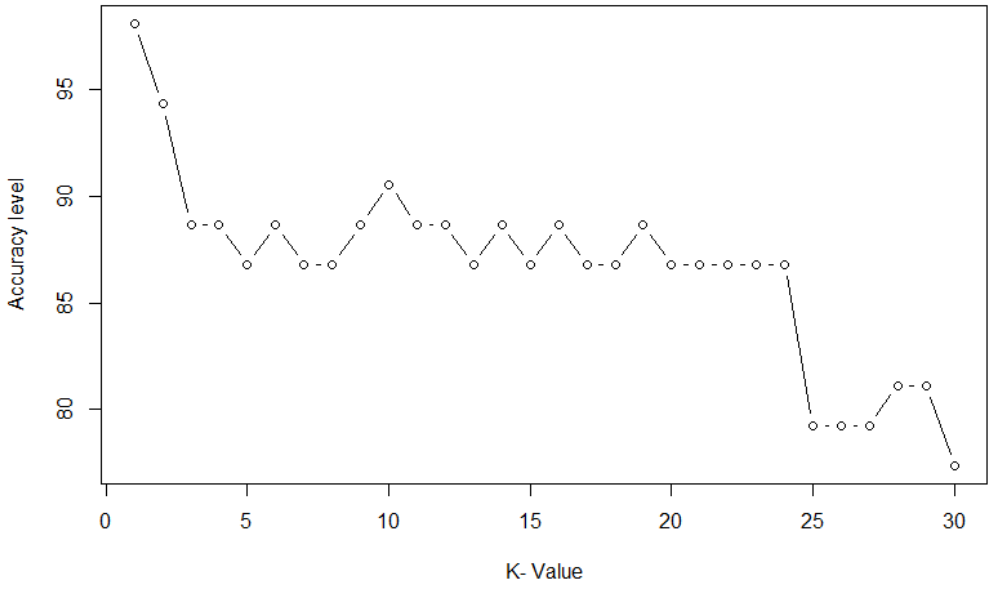
Επειδή οι διαφορετικές μεταβλητές έχουν διαφορετικές μονάδες, κανονικοποιήσαμε κάθε μία από τις μεταβλητές χρησιμοποιώντας τον τύπο (x-min(x))/(min(x) - max(x)). Η σύνοψη και η δομή των κανονικοποιημένων δεδομένων dfclass_norm φαίνονται στο Σχήμα 5.40. Αφού εφαρμόσαμε τον αλγόριθμο για διάφορες τιμές του k καταλήξαμε ότι η καλύτερη επιλογή είναι το k=10, με ακρίβειά του αλγορίθμου 90,57%, δηλαδή πολύ καλύτερη από αυτή των άλλων αλγορίθμων.

```
> summary(nn)
 marital_status    ethnicity     has_chartevents_data     gender          agegroup         50800        50802        50804
 Min.   :1.000   Min.   :1.000   Min.   :1            Min.   :1.000   Min.   :1.000   Min.   :0   Min.   :0   Min.   :0
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:1            1st Qu.:1.000   1st Qu.:3.000   1st Qu.:0   1st Qu.:0   1st Qu.:0
 Median :5.000   Median :5.000   Median :1            Median :1.000   Median :3.000   Median :0   Median :0   Median :0
 Mean   :4.268   Mean   :4.268   Mean   :1            Mean   :1.383   Mean   :3.418   Mean   :0   Mean   :0   Mean   :0
 3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:1            3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:0   3rd Qu.:0   3rd Qu.:0
 Max.   :7.000   Max.   :7.000   Max.   :1            Max.   :2.000   Max.   :5.000   Max.   :0   Max.   :0   Max.   :0
     50812        50816        50818        50820            50821        50825   icd9_category
 Min.   :0   Min.   :0   Min.   :0   Min.   :0.00000   Min.   :0   Min.   :0   Min.   :1.00
 1st Qu.:0   1st Qu.:0   1st Qu.:0   1st Qu.:0.00000   1st Qu.:0   1st Qu.:0   1st Qu.:1.00
 Median :0   Median :0   Median :0   Median :0.00000   Median :0   Median :0   Median :2.00
 Mean   :0   Mean   :0   Mean   :0   Mean   :0.01916   Mean   :0   Mean   :0   Mean   :1.54
 3rd Qu.:0   3rd Qu.:0   3rd Qu.:0   3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.:0   3rd Qu.:2.00
 Max.   :0   Max.   :0   Max.   :0   Max.   :1.00000   Max.   :0   Max.   :0   Max.   :2.00
> str(nn)
'data.frame':   261 obs. of  15 variables:
 $ marital_status      : num  5 5 7 5 5 5 3 2 3 5 ...
 $ ethnicity           : num  5 5 7 5 5 5 3 2 3 5 ...
 $ has_chartevents_data: int  1 1 1 1 1 1 1 1 1 1 ...
 $ gender              : num  1 1 1 1 2 1 2 1 2 2 ...
 $ agegroup            : num  3 4 1 4 4 4 1 1 3 4 ...
 $ 50800               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50802               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50804               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50812               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50816               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50818               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50820               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50821               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 50825               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ icd9_category       : num  1 1 2 1 1 1 2 2 2 2 ...
```

Σχήμα 5.40. Η σύνοψη και η δομή των κονονικοποιημένων δεδομένων dfclass_norm

```
> confusionMatrix(table(knn.10 ,df_test_cat))
Confusion Matrix and Statistics

      df_test_cat
knn.10  1  2
     1 18  0
     2  5 30

               Accuracy : 0.9057
                 95% CI : (0.7934, 0.9687)
    No Information Rate : 0.566
    P-Value [Acc > NIR] : 6.945e-08

                  Kappa : 0.803

 Mcnemar's Test P-Value : 0.07364

            Sensitivity : 0.7826
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8571
             Prevalence : 0.4340
         Detection Rate : 0.3396
   Detection Prevalence : 0.3396
      Balanced Accuracy : 0.8913

       'Positive' Class : 1
```

Σχήμα 5.41. Η μήτρα σύγχυσης και οι μετρικές για τον αλγόριθμο k-nn



Σχήμα 5.42. Ακρίβεια του αλγόριθμου k-NN συναρτήσει του k

Στο Σχήμα 5.42 βλέπουμε ότι για k=1 και για k=2 η ακρίβεια είναι ακόμα καλύτερη. Ο λόγος που δεν επιλέξαμε αυτές τις τιμές του k είναι γιατί κατά πάσα πιθανότητα ο αλγόριθμος έχει υπερπροσαρμοστεί στα δεδομένα μας.

Η ευαισθησία (Sensitivity), του μοντέλου k-nn όπως φαίνεται στο Σχήμα 5.41 είναι 78,23% και η εξειδίκευση (Specificity) είναι 100,00%. Αυτό συνεπάγεται ότι ταξινομεί 100% σωστά του ασθενείς με Χ.Α.Π. και κατά 78,23% τους ασθενείς με άσθμα.

6. Συμπεράσματα

Είναι προφανές ότι ο εμπλουτισμός των αναλυτικών τεχνικών των Πληροφοριακών Συστημάτων Υποστήριξης Διοίκησης με μεθοδολογίες Τεχνητής Νοημοσύνης αποφέρει νέες δυνατότητες για την εξαγωγή συμπερασμάτων και τη λήψη αποφάσεων, και αυξάνει την ακρίβεια, την αξιοπιστία και τη χρηστικότητα του συστήματος. Ειδικά στον τομέα της υγείας, η ουσία της υποστήριξης των αποφάσεων του νοσηλευτικού και ιατρικού προσωπικού, είναι αδιαμφισβήτητη.

Οι αλγόριθμοι Μηχανικής Μάθησης οι οποίοι εφαρμόστηκαν στα δεδομένα που αντλήθηκαν από την βάση δεδομένων MIMIC ανέδειξαν τον παράγοντα της ηλικίας, ως έναν από τους καθοριστικότερους για την τελική διάγνωση. Επίσης, ακόμα και από τη στατιστική ανάλυση των δεδομένων έγινε σαφής η αναγκαιότητα νοσηλείας σε αρκετά περιστατικά, καθώς η αρχική με την τελική διάγνωση διαφοροποιούνταν σε αρκετές περιπτώσεις. Επιπρόσθετα, δεν εντοπίστηκε κάποιος έντονος συσχετισμός μεταξύ της εθνικότητας και των πνευμονολογικών νοσημάτων.

Στην προσπάθεια υποστήριξης της απόφασης των ιατρών σχετικά με το αν η ο ασθενής πάσχει από Άσθμα η Χ.Α.Π. ο αλγόριθμος k-nn φαίνεται να είναι ένας ισχυρός σύμμαχος. Οι άλλοι δύο αλγόριθμοι δεν είχαν ιδιαίτερα θετικά αποτελέσματα, καθώς η ακρίβειά τους δεν ξεπέρασε το 70%. Παρ’ όλα αυτά οι αλγόριθμοι εφαρμόστηκαν στο στάδιο που υπάρχουν τιμές για εξετάσεις στα αέρια του αίματος. Αν θέλαμε να χρησιμοποιήσουμε αλγορίθμους μηχανικής μάθησης προτού γίνουν εξετάσεις, ίσως χρειαζόμασταν περισσότερα δεδομένα για την φυσική κατάσταση του ασθενή, όπως και πληροφορίες για το αν είναι καπνιστής ή όχι. Βέβαια, στα αποτελέσματά μας υπάρχει και το ποσοστό λανθασμένης διάγνωσης από τον θεράποντα ιατρό, το οποίο δεν μπορούμε να γνωρίζουμε.

Σε κάθε περίπτωση σε ένα πληροφοριακό σύστημα υγείας ο απόλυτος αυτοματισμός δεν θα ήταν δυνατός, διότι ένας αλγόριθμος μπορεί μόνο να λειτουργήσει συμβουλευτικά, καθώς την ευθύνη την έχει ο θεράπων ιατρός.

7. Βιβλιογραφία

[1] Health Metrics Network Framework and Standards for Country Health Information Systems, World Health Organization, January 2008

[2] Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων. Ευστάθιος Γ. Κύρκος ISBN: 978-960-603-109-0

[3] Πληροφοριακά Συστήματα-Προσεγγίσεις Ανάπτυξη-Πραγμάτωση. Γεώργιος Βασιλακόπουλος (2015) ISBN 978-960-93-6691-5

[4] Oracle BPM SUITE documentation

[5] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635

[6] Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov P, Mark RG, Mietus JE, Moody GB, Peng C, and Stanley HE. Circulation. 101(23), pe215–e220. 2000.

[7] http://www.icd9data.com/2007/Volume1/default.htm

[8] Practical Machine Learning in R. Kyriakos Chatzidimitriou, Themistoklis Diamantopoulos, Michail Papamichail and Andreas Symeonidis. http://leanpub.com/practical-machine-learning-r
© 2013 - 2018 Kyriakos Chatzidimitriou, Themistoklis Diamantopoulos, Michail Papamichail and Andreas Symeonidis

[9] Introduction to Data Mining, 2nd edition. PangNing Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. ISBN-13: 9780133128901

[10] R Programming for Data Science. Roger D. Peng http://leanpub.com/rprogramming
© 2014 - 2019 Roger D. Peng

[11] Comprehensive Guide on t-SNE algorithm with implementation in R & Python
https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/

[12] Data Science Live Book. Pablo Casas, January 2019 ISBN: 978-987-42-5911-0 (eBook version).
https://livebook.datascienceheroes.com/

[13] Arterial Blood Gas Analyses in Chronic Obstructive Pulmonary Disease: In the Clinical Laboratory or as Point-of-Care Testing? Paloma Oliver, Antonio Buno and Rodolfo Alvarez-Sala
https://austinpublishinggroup.com/pulmonary-respiratory-medicine/fulltext/ajprm-v2-id1024.php