ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

ΕΦΑΡΜΟΓΗ ΣΥΣΤΗΜΑΤΟΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρήστος Σ. Αλεξιάδης

Επιβλέπων : Γεώργιος Ματσόπουλος
Επίκουρος Καθηγητής Ε.Μ.Π

Αθήνα, Ιούνιος 2014

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

ΕΦΑΡΜΟΓΗ ΣΥΣΤΗΜΑΤΟΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρήστος Σ. Αλεξιάδης

Επιβλέπων : Γεώργιος Ματσόπουλος
Επίκουρος Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 5η Ιουνίου 2014

……..	……..	……..
Νικόλαος Ουζούνογλου Καθηγητής, ΣΗΜ&ΜΥ, ΕΜΠ	Δημήτριος Κουτσούρης Καθηγητής ΣΗΜ&Μ, ΕΜΠ	Γεώργιος Ματσόπουλος Επ. Καθηγητής, ΣΗΜ&ΜΥ, ΕΜΠ

Αθήνα, Ιούνιος 2014

.....
Χρήστος Σ. Αλεξιάδης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη ολοκληρωμένης μεθοδολογίας για την παροχή χρήσιμων στατιστικών πληροφοριών σε μαστογραφικά δεδομένα, καθώς και η δημιουργία εργαλείων πρόβλεψης καλοήθους και κακοήθους όγκου στον μαστό βάσει των διαθέσιμων ιατρικών δεδομένων. Αρχικά, συλλέχθηκαν από το διαδίκτυο και εξετάστηκαν δύο βάσεις δεδομένων μαστογραφιών από πραγματικά ιατρικά στοιχεία . Στη συνέχεια, με τη βοήθεια του σχεδιασμού και της ανάλυσης απαιτήσεων αναπτύχθηκαν και υλοποιήθηκαν οι τελικές βάσεις δεδομένων στις οποίες αποθηκεύτηκαν τα δεδομένα προς περαιτέρω επεξεργασία. Παράλληλα με το σχεδιασμό κάθε βάσης, πραγματοποιούταν η προεργασία για το σχεδιασμό και την ανάλυση απαιτήσεων των τεχνικών εξόρυξης που θα χρησιμοποιηθούν σε επόμενο .

Στη συνέχεια πραγματοποιήθηκε βιβλιογραφική ανασκόπηση των όρων «εξόρυξη δεδομένων» και «καρκίνος του μαστού» καθώς και οι υπάρχουσες τεχνικές διάγνωσης του καρκίνου του μαστού μέχρι σήμερα. Στη συνέχεια επελέγησαν μετά από έρευνα οι δύο καταλληλότεροι αλγόριθμοι για την μελέτη και αξιοποίηση των δεδομένων των δύο βάσεων. Η μεθοδολογία που χρησιμοποιήθηκε είναι η εξόρυξη δεδομένων με την πλέον αποτελεσματική τεχνική: «δένδρα αποφάσεων» (decision trees), καθώς επίσης και με την τεχνική ομαδοποίησης (clustering). Η τροποποίηση των αλγορίθμων και ο πειραματισμός με τη μεθοδολογία των εν λόγω τεχνικών οδήγησε στη σύσταση και τον καθορισμό των παραμέτρων των δύο αλγορίθμων για την αποτελεσματικότερη και σωστότερη πρόβλεψη του είδους του όγκου. Τέλος, έγινε σύγκριση μεταξύ των δύο τεχνικών με τρεις διαφορετικές τεχνικές μέτρησης των αποτελεσμάτων με κύριο κριτήριο την πρόβλεψη του όγκου (καλοήθης ή κακοήθης).

Λέξεις Κλειδιά

Δένδρα αποφάσεων, Ομαδοποίηση, Εξόρυξη δεδομένων, καρκίνος του μαστού, βάση δεδομένων

Abstract

The purpose of this thesis is to develop a methodology of useful statistical information on mammographic data as well as a technique for early diagnosis based on the records of benign and malignant breast tumor. Initially there were created two mammographic databases of true medical data available online. An analysis of the design and the requirements of the databases for data processing storage were initially developed. Furthermore, with the design for each data base, the groundwork for the design and analysis requirements of data mining techniques that would be required later on for the project were achieved.

Further study review was performed in order to define the terms “data mining” and “breast cancer” as well as the existing diagnostic techniques up to date. Then two algorithms were selected which were suitable for the study and use for the two databases. The methodology selected is the “data mining” with the most effective technic decision trees as well as the technic “clustering”. The modification of the algorithms and experiment on these techniques led to the recommendation but also to definition of the parameters of two algorithms for efficient and more accurate diagnosis of the type of tumor (benign or malignant). The two algorithmic techniques were compared with three different techniques for measuring their results, where the main criterion was their success in the diagnosis of the kind of tumor (benign or malignant).

Keywords

Decision trees, clustering, data mining, breast cancer, database

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, στο Εργαστήριο Μικροκυμάτων και Οπτικών Ινών. Η εργασία ήταν η αφορμή να ασχοληθώ εντατικότερα με εφαρμογές της πληροφορικής στον τομέα της υγείας και πιο συγκεκριμένα στο ιδιαίτερα ενδιαφέρον πεδίο της εξόρυξης στοιχείων από ιατρικά δεδομένα (Data Mining), με χρήση πολλών και διαφορετικών μεταξύ τους τεχνολογιών όπως προγραμματισμός, βάσεις δεδομένων, εξόρυξη γνώσης, αλγοριθμικές μέθοδοι, στατιστική, καθώς και ιατρικά θέματα.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα Επίκουρο Καθηγητή Ε.Μ.Π. κ. Γεώργιο Ματσόπουλο ο οποίος ήταν η αιτία και το έναυσμα να ασχοληθώ με την εκπόνηση αυτής της διδακτορικής διατριβής, καθώς επίσης και βοηθός, συμπαραστάτης και οδηγός σε όλη αυτή την προσπάθεια και σε κάθε δυσκολία της. Τον ευχαριστώ πολύ για τον πολύτιμο χρόνο τον οποίο αφιέρωσε αγόγγυστα για το σκοπό αυτό.

ΠΕΡΙΕΧΟΜΕΝΑ

Λίστα εικόνων και πινάκων

Κεφάλαιο 1

1. Εισαγωγή

Ο 21ος αιώνας έχει χαρακτηρισθεί ως ο αιώνας της τεχνολογίας και του υπολογιστή. Αναμφισβήτητα η χρήση του ηλεκτρονικού υπολογιστή έχει καθιερωθεί σε όλες τις επιστήμες, σύγχρονες και μη. Εξαιτίας της ραγδαίας αύξησης των πληροφοριών, ο όγκος των δεδομένων που πρέπει να αποθηκευτούν έχει πολλαπλασιαστεί. Από το 1980 μέχρι σήμερα καθημερινά δημιουργούνται περίπου 2.5 quintillion (2.5×10^{18}) bytes δεδομένων.[11] Αντιλαμβάνεται κανείς, ότι εξαιτίας αυτής της αύξησης της πληροφορίας, συμβατικοί τρόποι αποθήκευσης των σημαντικών δεδομένων έχουν καταργηθεί με αποτέλεσμα ο υπολογιστής να εμφανίζεται ως μοναδική λύση αξιόπιστης αποθήκευσης. Το ερώτημα, όμως, που αντιμετωπίζουν οι σύγχρονες επιστήμες είναι τι μπορεί να προσφέρει ο υπολογιστής, όσον αφορά την διαχείριση των δεδομένων. Τα σύνολα Δεδομένων που αποθηκεύονται είναι της τάξης των Terabytes. Επειδή οι πληροφορίες αυτές καταγράφονται απευθείας στον υπολογιστή χωρίς κάποια ιδιαίτερη προ-επεξεργασία είναι δύσκολο να διακριθεί η χρήσιμη και ουσιαστική γνώση-πληροφορία. Το πρόβλημα που πρέπει να αντιμετωπίσουμε είναι ο εντοπισμός και η εξαγωγή πληροφοριών που είναι χρήσιμες. Διότι σήμερα έχουμε πολύ μεγάλο όγκο πληροφοριών χωρίς όμως ποιότητα στο σύνολό τους.

Τη λύση στο παραπάνω πρόβλημα ήρθε να δώσει η «Εξόρυξη δεδομένων» (στα αγγλικά «data mining|) ένα μεγάλο πλέον διεπιστημονικό πεδίο της επιστήμης των υπολογιστών. Οι γνωστές μέθοδοι υλοποίησης της τεχνικής Εξόρυξης Δεδομένων είναι τουλάχιστον 10 με αποτέλεσμα οι συνδυασμοί μίξης και τροποποίησης των μεθόδων για την εύρεση της πιο γρήγορης και αποτελεσματικής μεθόδου να είναι χιλιάδες από αλγοριθμική άποψη.

Πρόβλημα εξαγωγής πληροφοριών αντιμετωπίζει και ο τομέας της υγείας με αποτέλεσμα να είναι ο κατεξοχήν τομέας ενδιαφέροντος εφαρμογής της εξόρυξης δεδομένων. Η βοήθεια την οποία μπορεί να προσφέρει η εν λόγω τεχνική στηρίζεται στη συμβολή της στη λήψη αποφάσεων που σχετίζονται με τη διάγνωση των ασθενειών ή ακόμα και την πρόβλεψή τους. Πολλά τέτοια συστήματα έχουν αναπτυχθεί και εφαρμοστεί για την υποβοήθηση της διάγνωσης σε παθήσεις καρκίνου ύπατος, σε προβλήματα ρύθμισης γλυκόζης, ταξινόμησης χρωμοσωμάτων κτλ. Στα πλαίσια αυτής της διπλωματικής θα γίνει προσπάθεια χρήσης και τροποποίησης της μεθόδου «δένδρα αποφάσεων» της εξόρυξης ιατρικών μαστογραφικών δεδομένων με σκοπό την καταγραφή χρήσιμων χαρακτηριστικών των εξεταζομένων περιπτώσεων των ασθενών, όπως η ηλικία, η προέλευση και η πρόβλεψης του κακοήθη όγκου.

1.1 Σκοπός της διπλωματικής

Σκοπός της παρούσας διπλωματικής εργασίας είναι η έκδοση πληροφοριών και αποτελεσμάτων με χρήση μεθόδων εξόρυξης δεδομένων για την κατηγοριοποίηση μαστογραφιών κατά καλοήθη ή κακοήθη όγκο.

Με την μέθοδο «δένδρα αποφάσεων» και την μέθοδο «ομαδοποίησης» εξήχθησαν αποτελέσματα και πρότυπα που βασίζονται σε ένα σύνολο χαρακτηριστικών. Με τον πειραματισμό και τα χαρακτηριστικά που ήταν διαθέσιμα από τις βάσεις δεδομένων έγιναν κατηγοριοποιήσεις και πρότυπα με σκοπό την παροχή πληροφοριών, οι οποίες μπορούν να υποστηρίξουν τους ραδιολόγους να επιβεβαιώσουν την διάγνωσή τους. Ένας από τους καλύτερους τρόπους αποφυγής του καρκίνου του μαστού είναι η γρήγορη ανίχνευσή του, που επιτυγχάνεται με την συλλογή και κατηγοριοποίηση των δεδομένων που θα μπορούν να συμβάλλουν στην έγκαιρη και αξιόπιστη διάγνωσή του.

Τα χαρακτηριστικά τα οποία χρησιμοποιήθηκαν στην κατηγοριοποίηση είναι η ηλικία του ασθενή, το σχήμα του όγκου, η πυκνότητα της μάζας δηλαδή στοιχεία BI-RADS (Breast Imaging-Reporting and Data System), ενός εργαλείου διασφάλισης της ποιότητας για την σωστή ανάγνωση της μαστογραφίας, το οποίο περιλαμβάνει την διάγνωση των ογκολόγων και επιτρέπει την σύγκριση της μεθοδολογίας που χρησιμοποιήθηκε από τον υπολογιστή για να καταλήξει σε διάγνωση με την μεθοδολογία που χρησιμοποίησε ο ιατρός και την κατάληξη της διάγνωσής του σε σχέση με το πραγματικό αποτέλεσμα που είναι η διαπίστωση του κατά πόσον ο όγκος είναι καλοήθης ή κακοήθης.

1.2 Οργάνωση Διπλωματικής Εργασίας

Η διπλωματική εργασία αποτελείται από τα παρακάτω κεφάλαια:

Στο **Κεφάλαιο 2**, περιγράφονται αρχικά τα στοιχεία, που σχετίζονται με την ιατρική επιστήμη δηλαδή η ανατομία, η φυσιολογία και διευκρινίζονται ιατρικοί όροι, απαραίτητοι για την κατανόηση της διπλωματικής. Στην συνέχεια γίνεται αναφορά στην ασθένεια του καρκίνου του μαστού, την συχνότητα και την σοβαρότητα εμφάνισής της νόσου, ενώ παρέχονται επίσης σημαντικές πληροφορίες από στατιστικές έρευνες, οι οποίες περιγράφουν, ποιές μέθοδοι εφαρμόζονται σήμερα για την διάγνωση και την αντιμετώπισή της. Παρέχονται πληροφορίες αξιολόγησης των μεθόδων καθώς και τα αίτια δημιουργίας βάσης μαστογραφικών δεδομένων, καθώς επίσης και ο λόγος στροφής της σύγχρονης διαγνωστικής ιατρικής στον ηλεκτρονικό υπολογιστή και στην Εξόρυξη δεδομένων για την διάγνωση του καρκίνου του μαστού.

Στο **Κεφάλαιο 3,** περιγράφεται η σύγχρονη αλγοριθμική μέθοδος εξόρυξης δεδομένων και γίνεται και εκτενής περιγραφή της. Στην συνέχεια περιγράφονται αναλυτικά οι τεχνικές που εφαρμόστηκαν στα πλαίσια της διπλωματικής εργασίας που είναι τα Δένδρα Απόφασης (Decision Trees) και η ομαδοποίηση (clustering). Στην συνέχεια γίνεται η σύνδεση μεταξύ ιατρικής επιστήμης και της μεθόδου εξόρυξης δεδομένων, που έχει καθιερωθεί ως μία από τις πιο αξιόλογες και χρήσιμες μεθόδους που συμβάλλουν και θα συμβάλλουν ακόμα περισσότερο στο μέλλον.

Στο **Κεφάλαιο 4**, περιγράφεται η έννοια της βάσης δεδομένων και η αναγκαιότητα της σύγχρονης εποχής που οδήγησε στην δημιουργία τους. Στην συνέχεια περιγράφονται οι υπολογιστικοί πόροι του συστήματος, τα προγράμματα που χρησιμοποιήθηκαν και γίνεται αναλυτική παρουσίαση του σχεδιασμού, της ανάλυσης απαιτήσεων και της ανάπτυξης - υλοποίησης των βάσεων δεδομένων για την αποθήκευση των δεδομένων. Στην συνέχεια παρουσιάζονται οι δυο τελικές βάσεις πάνω στις οποίες εφαρμόσθηκαν η μέθοδος εξόρυξης δεδομένων με τις προαναφερθείσες τεχνικές. Τέλος, γίνεται παρουσίαση των διαθέσιμων στοιχείων και των ιδιαιτεροτήτων τους.

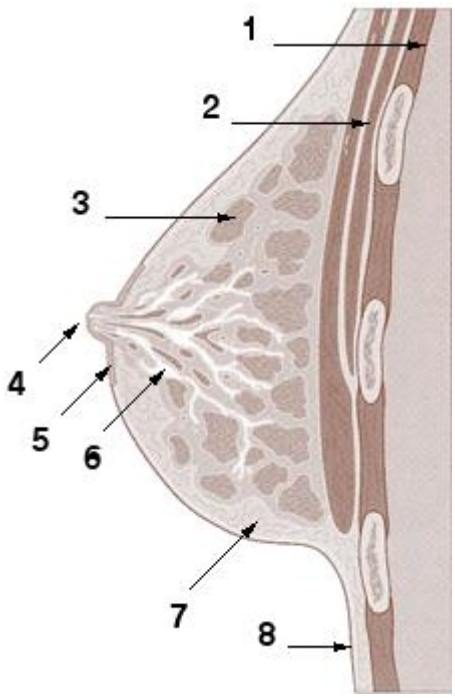
Στο **Κεφάλαιο 5**, παρουσιάζεται όλη η προετοιμασία για το χτίσιμο του μοντέλου εξόρυξης, η ρύθμιση των παραμέτρων των αλγορίθμων καθώς και η προετοιμασία των δύο βάσεων δεδομένων για εξόρυξη. Στην συνέχεια παρουσιάζονται αναλυτικά τα αποτελέσματα από κάθε βάση για κάθε αλγόριθμο που εφαρμόστηκε, και γίνεται σύγκριση των αποτελεσμάτων και της αποδοτικότητας και ευστοχίας των δύο αλγορίθμων μέσω τριών κριτηρίων: lift chart,classification matrix, και cross validation.

Τέλος, στο **Κεφάλαιο 6** παρουσιάζονται τα γενικά συμπεράσματα της διπλωματικής εργασίας και προτείνονται συγκεκριμένα πεδία για μελλοντικές επεκτάσεις της παρούσας διπλωματικής εργασίας, τόσο σε τεχνολογικό, όσο και σε ιατρικό επίπεδο.

Κεφάλαιο 2

2.1 Τι είναι ο μαστός

Ο **μαστός** αποτελεί ημισφαιρική λιπώδη πτυχή του δέρματος στο πρόσθιο θωρακικό τοίχωμα, ο οποίος περιέχει τον μαστικό ή μαζικό αδένα και είναι ιδιαίτερα ανεπτυγμένος στις γυναίκες μετά την εφηβεία (Η περιγραφή που ακολουθεί αφορά στον γυναικείο πλήρως ανεπτυγμένο μαστό)[1].



Η ανατομία του μαστού περιγράφεται στην παρακάτω εικόνα:

Τομή μαστού:

1.Μεσοπλεύριοι μύες

2. Θωρακικοί μύες

3. Λοβοί του μαστικού αδένα

4. Θηλή

5.Θηλαία άλως

6. Γαλακτοφόροι κόλποι

7. Περιμαστικό λίπος

Η ανατομία του μαστού

Η εξωτερική μορφολογία του μαστού περιλαμβάνει την θηλή, την θηλαία άλω και τα αλωαία οζίδια.

Η **θηλή** αποτελεί έπαρμα του δέρματος του μαστού που βρίσκεται λίγο πιο κάτω και έξω από το μέσο του μαστού. Στην κορυφή της υπάρχουν 15-20 στόμια όπου καταλήγουν οι γαλακτοφόροι πόροι. Το ύψος της θηλής είναι περίπου 1-1,5 cm και αυξάνει κατά την γαλουχία (θηλασμός). Το καστανέρυθρο χρώμα της θηλής οφείλεται στην άφθονη παρουσία μελανίνης ουσίας.[1]

Η **θηλαία άλως** αποτελεί υπο-στρόγγυλη και ελαφρά επηρμένη περιοχή γύρω από την θηλή με διάμετρο 1,5-6 cm. Στην επιφάνεια της θηλαίας άλω υπάρχουν μικρά

επάρματα, τα *θηλαία οζίδια* τα οποία έχουν ως υπόθεμα τους αλωαίους αδένες. Οι *αλωαίοι αδένες* είναι κυρίως οσμηγόνοι, αλλά και σμηγματογόνοι και υποτυπώδεις γαλακτικοί αδένες.[1]

Εσωτερικά ο μαστός αποτελείται από τον μαστικό ή μαζικό αδένα και το περιμαστικό λίπος. Το *περιμαστικό λίπος* είναι συνέχεια του υποδόριου λίπους το οποίο όμως είναι αφθονότερο στην πρόσθια περιοχή του μαστού, ανάμεσα στο δέρμα και τον μαστικό αδένα.[1]

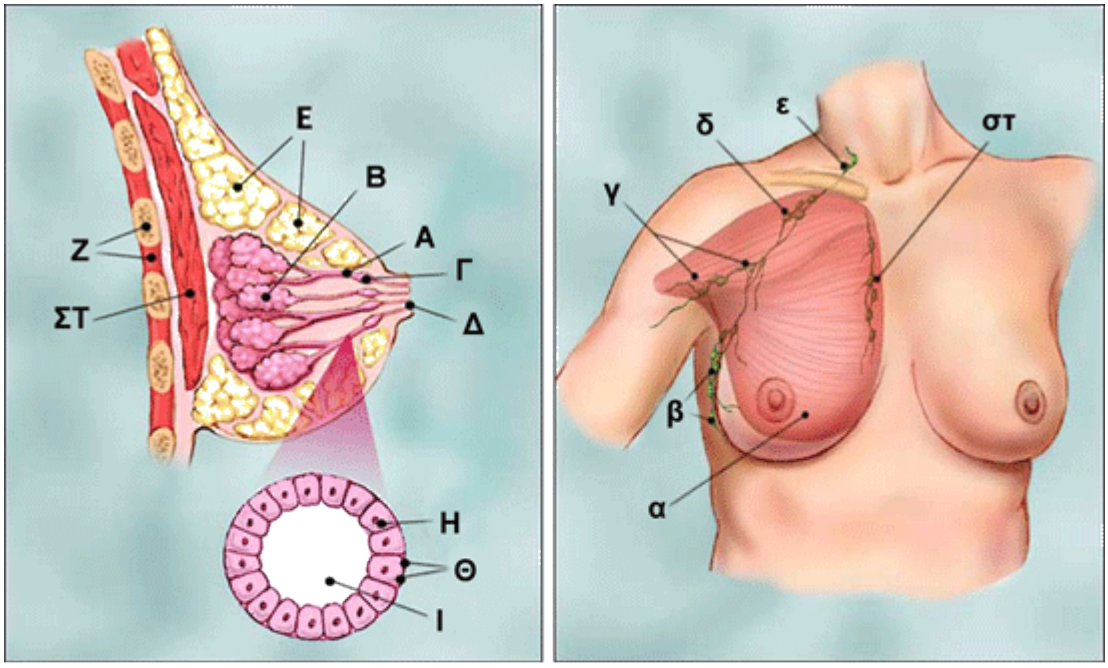
Ο **μαστικός αδένας** σε γυναίκα μη κυοφορούσα βρίσκεται πίσω από την θηλαία άλω και μόλις που υπερβαίνει τα όριά της. Ο μαστικός αδένας αποτελείται από τους λοβούς όπου παράγεται το γάλα και τους *γαλακτοφόρους πόρους* που μεταφέρουν το γάλα στους γαλακτοφόρους κόλπους. Οι *γαλακτοφόροι κόλποι* είναι ανευρύσματα των πόρων τα οποία λειτουργούν ως αποθήκη του γάλακτος, το οποίο και απελευθερώνουν μετά από πίεση της θηλής από το βρέφος.[1]

2.2 Ιστολογία του μαστού

Ο μαστός περιλαμβάνει 15-25 γαλακτοφόρους πόρους οι οποίοι ξεκινούν από την θηλή, διακλαδίζονται σε μικρότερους πόρους και καταλήγουν στους λοβούς. Ο κάθε λοβός αποτελείται από έναν τελικό πόρο και πολλούς μικρούς πόρους (λόβια). Οι μεγάλοι και οι μικροί πόροι καλύπτονται εσωτερικά από μία στιβάδα κυβοειδών και κυλινδρικών κυττάρων και εξωτερικά από μία στιβάδα μυοεπιθηλιακών κυττάρων. Ο συνδετικός ιστός μέσα στο λοβό συνίσταται από ινοβλάστες μέσα σε ένα υπόστρωμα από κολλαγόνο και όξινης βλέννης με ιστιοκύτταρα και περιστασιακά με λεμφοκύτταρα. Το στρώμα ανάμεσα στα λόβια είναι υποκυτταρικό και συντίθεται από ινολιπώδη ιστό.[2]

Το επιθήλιο και το στρώμα των λοβών είναι ορμονοευαίσθητα. Κατά την διάρκεια της εγκυμοσύνης υπάρχει αξιοσημείωτη υπερπλασία των μικρών πόρων, έχοντας ως αποτέλεσμα την δημιουργία μεγάλων λοβών, και τα επιθηλιακά κύτταρα έχουν άφθονο κυτταρόπλασμα γεμάτο με εκκριτικά κοκκία.[2]

Παρακάτω παρατίθεται εικόνα με όλα τα προαναφερθέντα:



Α. Γαλακτοφόροι πόροι
Β. Λοβία
Γ. Διευρύνσεις των πόρων που αποθηκεύεται το γάλα
Δ. Θηλή
Ε. Λίπος
ΣΤ. Μείζων θωρακικός μυς
Ζ. Θωρακικό τοίχωμα
Η. Φυσιολογικά κύτταρα των πόρων
Θ. Βασική μεμβράνη
Ι. Αυλός των γαλακτοφόρων πόρων

α. Μείζων θωρακικός μυς
β. Μασχαλιαίοι λεμφαδένες – επίπεδο Ι
γ. Μασχαλιαίοι λεμφαδένες – επίπεδο ΙΙ
δ. Μασχαλιαίοι λεμφαδένες – επίπεδο ΙΙΙ
ε. Υπερκλείδιοι λεμφαδένες
στ. Έσω μαστικοί λεμφαδένες

Ιστολογία του μαστού

2.3 Καρκίνος του μαστού-αίτια

Ο όρος «**καρκίνος του μαστού**» αναφέρεται στην ανάπτυξη κακοήθους όγκου στην περιοχή του μαστού. Αποτελεί μία από τις συχνότερα εμφανιζόμενες μορφές καρκίνου παγκοσμίως και είναι η πρώτη σε αριθμό κρουσμάτων στο γυναικείο πληθυσμό. Προκαλείται από ανεξέλεγκτο πολλαπλασιασμό παθολογικών κυττάρων, που ως αποτέλεσμα προκαλούν το σχηματισμό κακοήθους όγκου στην περιοχή του μαστού και ουσιαστικά αποτελεί κυτταρική νόσο. Τα παθολογικά αυτά κύτταρα έχουν τη δυνατότητα εξάπλωσης σε γειτονικούς ιστούς σε δυσάρεστες συνέπειες για ολόκληρο τον οργανισμό. Η πιθανότητα εμφάνισης της νόσου σε άρρενες είναι υπαρκτή αλλά πολύ μικρή. Όσον αφορά στις γυναίκες όλες αντιμετωπίζουν τον κίνδυνο εμφάνισης της νόσου – όχι, όμως στον ίδιο βαθμό.[18]

Στατιστικά στοιχεία

Σύμφωνα με τη Διεθνή Έκθεση για τις καρκινικές νόσους που παρουσιάστηκε στη Γενεύη της Ελβετίας τον Απρίλιο του 2003 από την ***IARC*** (***International Agency for Research on Cancer***), ο καρκίνος του μαστού αποτελεί την πιο κοινή μορφή καρκίνου μεταξύ των γυναικών, με περίπου 1.000.000 νέα κρούσματα παγκοσμίως. Στις Η.Π.Α. τα κρούσματα ξεπερνούν τα 200.000 ετησίως, ενώ στην Ελλάδα διαγιγνώσκονται περίπου 4.500 γυναίκες με καρκίνο του μαστού κάθε χρόνο. Υπολογίζεται (βάσει στοιχείων του έτους 2002) ότι 1 στις 8 γυναίκες στην Αμερική και 1 στις 9 στην Ευρώπη θα νοσήσει από κακοήθη μαστική νεοπλασία κάποια στιγμή στη ζωής της. Μόλις το 30% από τις γυναίκες αυτές έχει ιστορικό κληρονομικότητας στην οικογένειά των. Στην Ελλάδα το αναλογικό ποσοστό κρουσμάτων είναι 1 στις 12.

Είναι ενδιαφέρον ότι, στην Ευρώπη, το **60% των κρουσμάτων καρκίνου του μαστού διαγιγνώσκεται σε πρώιμο στάδιο** – ποσοστό που **στην Ελλάδα μόλις και μετά βίας εγγίζει το 5%.** Τα στοιχεία αυτά καταδεικνύουν πόσο ελλιπής είναι η σχετική ενημέρωση μεταξύ των Ελληνίδων, γεγονός εξαιρετικά λυπηρό, αν λάβουμε υπόψη τις δυνατότητες πλήρους ίασης που παρέχει μία έγκαιρη διάγνωση.[19]

Το πενταετές ποσοστό επιβίωσης σε περιπτώσεις διάγνωσης σε πρώιμο στάδιο φθάνει ως και το 95%, στοιχείο που υποδηλώνει πως ο καρκίνος του μαστού μπορεί να αντιμετωπιστεί επιτυχώς για την πλειονότητα των γυναικών που φροντίζουν να τον εντοπίσουν έγκαιρα, μέσω συχνών προληπτικών ελέγχων και ψηλάφησης του στήθους. Αν και μόλις το 20% των διογκώσεων που μπορεί να εντοπιστούν μέσω της ψηλάφησης ενδέχεται να είναι καρκινογόνο, όλες πρέπει να εξετάζονται προσεκτικά από γιατρό αμέσως μετά τον εντοπισμό τους.[19]

Τα ποσοστά θανάτου από καρκίνο του μαστού χαρακτηρίζονται από πτωτική τάση από τις αρχές του 1990, με τις μεγαλύτερες μειώσεις να εντοπίζονται στις γυναίκες κάτω των 50. Οι ερευνητές αποδίδουν την πτώση αυτή στην έγκαιρη διάγνωση μέσω μαστογραφιών καθώς και στις βελτιώσεις που έχουν επέλθει στις σχετικές θεραπευτικές αγωγές. Ο αριθμός των ατόμων που έχουν αντιμετωπίσει με επιτυχία τον καρκίνο του μαστού αυξάνεται συνεχώς - από τον Ιανουάριο του 2006, υπήρξαν περίπου 2,5 εκατομμύρια γυναίκες στις ΗΠΑ που, βάσει της έκθεσης, έχουν ξεπεράσει με επιτυχία την περιπέτεια του καρκίνου του μαστού.[20]

Η παγκόσμια επιδημία του καρκίνου του μαστού έχει πολλούς αιτιολογικούς παράγοντες από τους οποίους οι σημαντικότεροι είναι: [18]

1. **Ηλικία**: ο καρκίνος του μαστού μπορεί να προκύψει σε οποιαδήποτε ηλικία μετά την εφηβεία αλλά τα ποσοστά αυξάνονται όσο αυξάνονται και οι ηλικιακές κλίμακες. Οι περισσότερες περιπτώσεις παρουσιάζονται μετά από την ηλικία των 50 ετών, ενώ είναι σπάνιος σε γυναίκες ηλικίας κάτω των 35 ετών (5% των περιπτώσεων), με εξαίρεση τις γυναίκες που έχουν κληρονομική προδιάθεση.
2. **Κληρονομικότητα**: Υπολογίζεται ότι μόλις το 5 - 10% των κρουσμάτων καρκίνου του μαστού σχετίζεται όντως με παράγοντες κληρονομικότητας. Ωστόσο, δύο γονίδια, γνωστά ως BRCA 1 και BRCA 2, έχουν προσδιοριστεί ως παράγοντες που συμβάλλουν στην εμφάνιση καρκίνου του μαστού. Επίσης, γυναίκες με εξ αίματος συγγενείς που έχουν νοσήσει αντιμετωπίζουν αυξημένο κίνδυνο εμφάνισης καρκίνου στο μαστό.

3. **Διαταραχές της έμμηνου ρύσης**: Στοιχεία υποδηλώνουν πως γυναίκες με πρώιμη έναρξη της εμμήνου ρύσης (πριν από το 12ο έτος της ηλικίας τους) ή με καθυστερημένη εμμηνόπαυση (μετά τα 55) αντιμετωπίζουν αυξημένο κίνδυνο εμφάνισης καρκίνου στο μαστό. Επίσης, η λήψη οιστρογόνων μετά την εμμηνόπαυση έχει συσχετιστεί με αυξημένα ποσοστά εμφάνισης της νόσου, με τον κίνδυνο να είναι ανάλογος του διαστήματος λήψης των οιστρογόνων. Ανάλογος συσχετισμός έχει προκύψει και για γυναίκες που παρέμειναν άτεκνες, που δεν είχαν πλήρεις κυήσεις (διάρκεια εννέα μηνών) ή που γέννησαν μετά τα τριανταπέντε τους χρόνια.
4. **Αλκοόλ**: Τα οινοπνευματώδη ποτά αυξάνουν τη συγκέντρωση των οιστρογόνων στο αίμα. Σύμφωνα με μελέτη που παρουσιάστηκε το Δεκέμβριο του 2009 στο Διεθνές Συνέδριο ογκολογίας του Σαν Αντόνιο, άτομα που νόσησαν από καρκίνο του μαστού και καταναλώνουν με μετριοπάθεια αλκοόλ διατρέχουν μεγαλύτερο κίνδυνο επανεμφάνισής του από εκείνα που πίνουν λίγο ή καθόλου οινοπνευματώδη.
5. **Παχυσαρκία**: Η παχυσαρκία αυξάνει τον κίνδυνο καρκίνου του μαστού καθώς αυξάνει τα επίπεδα των οιστρογόνων. Η παραγωγή των οιστρογόνων στις γυναίκες μετά την εμμηνόπαυση γίνεται κυρίως μέσα σε λιπώδη ιστό (μετατροπή των επινεφριδικών ανδρογόνων σε οιστρογόνα από την αρωματάση, ένα ένζυμο που βρίσκεται κυρίως στο λίπος). Τον Ιούνιο του 2009 ανακοινώθηκε από το Αμερικανικό Ίδρυμα για την έρευνα του Καρκίνου (AICR) ότι η συσσώρευση σωματικού λίπους σε ποσοστά άνω του κανονικού ευθύνεται για το 17% των κρουσμάτων καρκίνου του μαστού στις Η.Π.Α.
6. **Κάπνισμα**: πρόσφατες μελέτες απέδειξαν ότι η κατανάλωση ενός πακέτου τσιγάρων ημερησίως, από γυναίκες προ της εμμηνόπαυσης για εννέα περίπου χρόνια, αυξάνει δραστικά τον κίνδυνο εμφάνισης καρκίνου του μαστού κατά σχεδόν 60%.
7. **Άτυπη υπερπλασία ή άλλη προ-κακοήθης κατάσταση:** Μολονότι δεν είναι όλες οι αλλοιώσεις ή όγκοι στο στήθος κακοήθεις και καρκινικοί, ορισμένες προετοιμάζουν το έδαφος για την εμφάνιση παθογόνων καρκινικών κυττάρων και καρκίνου στην ευρύτερη περιοχή του μαστού.
8. **Λήψη αντισυλληπτικών χαπιών**: με επιφύλαξη αναφερόμαστε σε αυτόν τον παράγοντα καθώς μελέτες δεν έχουν αποδείξει ακόμα τη συσχέτιση της λήψης αντισυλληπτικών με την εμφάνιση καρκίνου του μαστού.
9. **Ιστορικό Καρκίνου**: γυναίκες που έχουν εμφανίσει προηγουμένως καρκίνο της μήτρας, των ωοθηκών ή του μαστού έχουν αυξημένες πιθανότητες να εμφανίσουν έναν 2ο καρκίνο στο μαστό.
10. **Καθιστική Ζωή**: Η τακτική άσκηση πριν την έναρξη της έμμηνου ρύσης μπορεί να μειώσει τον κίνδυνο καρκίνου του μαστού μιας γυναίκας, κυρίως διότι μπορεί να καθυστερήσει την έναρξη της έμμηνου ρύσεως, να επιμηκύνει τον χρόνο μεταξύ των περιόδων ή να ελαττώσει τον αριθμό των εμμηνορρυσιακών κύκλων, μειώνοντας έτσι την έκθεση της γυναίκας στα οιστρογόνα.
11. **Έκθεση σε ακτινοβολία**
12. **Ατεκνία**
13. **Θεραπεία Ορμονικής Υποκατάστασης**: προσφέρει ανακούφιση από τα συμπτώματα της εμμηνόπαυσης, ωστόσο, η μακροχρόνια χρήση της μετά την εμμηνόπαυση αυξάνει τον κίνδυνο ανάπτυξης καρκίνου του μαστού.

2.4 Συμπτώματα-Εξετάσεις-Διάγνωση

Ο καρκίνος ως ομάδα ασθενειών χαρακτηρίζεται από ορισμένα γενικά συμπτώματα, ενώ κάθε διακριτή του μορφή συσχετίζεται με ορισμένα ειδικά συμπτώματα. Τα ειδικά συμπτώματα που χαρακτηρίζουν εν γένει τις εκάστοτε παραλλαγές του καρκίνου του μαστού - η φύση και η έκταση των οποίων ποικίλλει – είναι τα εξής:[18]

- Εξόγκωμα, ογκίδιο ή σκλήρυνση στην ευρύτερη περιοχή του μαστού και/ή της μασχάλης.
- Έκκριση υγρών ή αίματος από τη θηλή του μαστού.
- Διόγκωση λεμφαδένων της μασχάλης.
- Έλξη του δέρματος ή της θηλής προς το εσωτερικό του μαστού (εισολκή δέρματος).
- Αλλοιώσεις του δέρματος (όψη φλοιού πορτοκαλιού).
- Ερυθρότητα, φλόγωση, ευαισθησία ή πόνοι στο στήθος.

Όμως όλες οι αλλοιώσεις, διογκώσεις ή σκληρύνσεις δεν είναι καρκινικές και κακοήθεις. Υπολογίζεται πως το 90% των μαστικών όγκων σε γυναίκες ηλικίας 20 – 50 ετών προκαλούνται από κάποια πάθηση καλοήθους φύσης. [18]

Καλοήθεις παθήσεις του μαστού

Ινοαδένωμα

Το ινοαδένωμα είναι το συνηθέστερο καλοήθες εξόγκωμα στην περιοχή του μαστού. Θωρείται καλοήθης νεοπλασία, αν και πολλοί είναι αυτοί που τη συνδέουν με υπερπλασία των μαστικών λοβίων. Συνίσταται στο σχηματισμό ενός όγκου από ινώδη και αδενικό ιστό. [3]

Κύστη

Κύστη είναι μια παθολογική, καλοήθης ανάπτυξη ενός θύλακος σε κάποιο όργανο του σώματος, με ρευστό ή ημίρρευστο περιεχόμενο. Οι μαστικές κύστεις σχηματίζονται από τοπική συλλογή υγρού. [3]

Λίπωμα

Το λίπωμα είναι μία καλοήθης μάζα λιπώδους σύστασης που εμφανίζεται συνήθως υποδόρια (αμέσως κάτω από το δέρμα) αλλά ορισμένες φορές και σε βαθύτερα μυϊκά στρώματα. Σε περίπτωση συσχέτισής του με αδένες, τότε αναφερόμαστε σε αυτό ως αδενολίπωμα. [3]

Φυλλοειδής όγκος

Ο φυλλοειδής όγκος είναι μια σπάνια μορφή (μόλις 0,3 – 0,5% επί του συνόλου των μαστικών νεοπλασμάτων) μαστικού νεοπλάσματος, η οποία μπορεί να αναπτυχθεί είτε ως καλοήθης είτε ως κακοήθης (κυστεοσάρκωμα). Μορφολογικά μοιάζει πολύ με το ινοαδένωμα, από το οποίο είναι συχνά δύσκολο να διακριθεί. [3]

Θήλωμα

Το θήλωμα είναι ένα είδος καλοήθους όγκου που αναπτύσσεται στο μαστό ή και σε άλλα όργανα. Ουσιαστικά αποτελεί μια καλοήθη υπερπλασία του επιθηλίου (υποδόριων κυτταρικών υμένων) των γαλακτοφόρων πόρων του μαστού. [3]

Φλεγμονή-απόστημα

Το απόστημα είναι ουσιαστικά η συσσώρευση πυώδους υγρού σε μια νεοσχηματισμένη κοιλότητα και συχνά συνοδεύεται από φλεγμονή.[3]

Ινοκυστική μαστοπάθεια (κυστική νόσος –χρόνια κυστική μαστίτιδα – δυσπλασία μαστών)

Η ινοκυστική μαστοπάθεια είναι μια αμιγώς καλοήθης πάθηση που περιλαμβάνει αλλοιώσεις στο μαστό χωρίς δυσάρεστες επιπλοκές. Συνίσταται στην παρουσία μικρών όγκων – κύστεων, που προκαλούνται από διάταση (τέντωμα) των πόρων του μαστικού αδένος.[3]

Γαλακτοφορεκτασία

Γαλακτοφοροεκτασία ονομάζεται η διάταση (τέντωμα) ορισμένων πόρων στο μαστικό αδένα. [3]

Μασταλγία

Η μασταλγία συνίσταται στον πόνο που προκαλείται στο μαστό από διάφορες αιτίες. Διακρίνεται σε διάφορες μορφές, η πιο κοινή από τις οποίες είναι η επονομαζόμενη «κυκλική μασταλγία». Η κυκλική μασταλγία σχετίζεται με τις ορμονικές αλλαγές που σχετίζονται με την έμμηνο ρύση. Συχνά οι αλλαγές αυτές προκαλούν πόνο και στους δύο μαστούς κάποιες ημέρες πριν την έναρξη της περιόδου εμμήνου ρύσης, ο οποίος υποχωρεί σταδιακά με τη λήξη της περιόδου. Η κυκλική μασταλγία δεν σχετίζεται με κάποια μορφή κακοήθειας ή με κάποια άλλη σοβαρή μορφή πάθησης.[3]

Γαλακτοκήλη

Η γαλακτοκήλη συνίσταται σε διάταση των πόρων του μαστού, που υπερπληρώνονται με γάλα. Παρατηρείται σε γυναίκες που θηλάζουν, και ιδιαίτερα σε πολύτεκνες.[3]

Μαστίτιδα

Η μαστίτιδα είναι μια φλεγμονή του μαστού και διακρίνεται σε πολλές μορφές και σπάνια σχετίζεται με καρκίνο στο μαστό. Η πλέον συνήθης μορφή της πλήττει τους γαλακτοφόρους πόρους και αφορά γυναίκες που θηλάζουν.[3]

Αμάτρωμα - Αδενολίπωμα

Το αμάτρωμα είναι μια σπάνια κατάσταση σχηματισμού καλοηθών όγκων στο μαστό που συναποτελούνται από αδενικό, λιπώδη και συνδετικό ιστό. Δεν υπάρχουν στοιχεία που αποδεικνύουν την εξέλιξή τους σε καοήθη πάθηση.[3]

Ο καρκίνος του μαστού διακρίνεται σε διάφορες μορφές. Ακολουθούν επεξηγηματικές πληροφορίες για τους κακοήθεις όγκους.

Οι κακοήθεις όγκοι διακρίνονται σε:

Επί τόπου Πορογενές Καρκίνωμα (DCIS)

Είναι η πιο κοινή μορφή μη διηθητικού καρκίνου στο μαστό. Δεν θεωρείται ιδιαίτερα επικίνδυνη ή επιθετική και αντιμετωπίζεται σχετικά εύκολα. Ωστόσο, αυξάνει τον κίνδυνο εμφάνισης μιας πιο επιθετικής μορφής μαστικού καρκίνου. Μία γυναίκα που νόσησε με επί τόπου πορογενές καρκίνωμα αντιμετωπίζει περίπου 25% πιθανότητες επανεμφάνισης της νόσου σε διάστημα 5 έως 10 ετών.[3]

Πορογενές διηθητικό καρκίνωμα (Invasive Ductal Carcinoma [IDC])

Το Πορογενές Διηθητικό Καρκίνωμα (IDC) είναι μακράν η πιο συνήθης μορφή καρκίνου του μαστού, καταλαμβάνοντας σχεδόν το 80% των κρουσμάτων παγκοσμίως. Μπορεί να εμφανιστεί σε γυναίκες κάθε ηλικίας, αν και φαίνεται πως τα

ποσοστά των κρουσμάτων αυξάνονται όσο ανεβαίνουμε τις ηλικιακές βαθμίδες. Μπορεί να εμφανιστεί ακόμη και σε ανδρικούς μαστικούς αδένες.[3]

Σωληνώδες Καρκίνωμα του μαστού

Το Σωληνώδες καρκίνωμα του μαστού είναι μία σπάνια παραλλαγή του Διηθητικού Πορογενούς Καρκινώματος (IDC), καθώς καταλαμβάνει ποσοστό που δεν ξεπερνά το 2% των κρουσμάτων παγκοσμίως, και εμφανίζεται σε γυναίκες ηλικίας 40 έως 60 ετών. Ο όρος «καρκίνωμα» αναφέρεται σε έναν κακοήθη καρκινικό όγκο, ο οποίος αποτελείται από επιθηλιακό ιστό – τον ιστό που σαν υμένας καλύπτει ή περιβάλλει τα όργανα του σώματος. Περιγράφεται ως «σωληνώδες» λόγω του επιμήκους σχήματος των καρκινικών του κυττάρων.[3]

Μυελώδες Καρκίνωμα στο στήθος

Το Μυελώδες Καρκίνωμα (***Medullary Carcinoma***) στο στήθος είναι μια σπάνια (μόλις 3 – 5% των κρουσμάτων καρκίνου του μαστού) μορφή Διηθητικού Πορογενούς Καρκινώματος. Είναι μία από τις πλέον εύκολα αντιμετωπίσιμες μορφές καρκίνου του μαστού.[3]

Βλεννοπαράγωγο Μαστικό Καρκίνωμα

Το Βλεννοπαράγωγο Μαστικό Καρκίνωμα (Mucinous breast carcinoma) είναι επίσης γνωστό και Κολλοειδές Καρκίνωμα, και αποτελεί μία σπανιότατη μορφή καρκίνου στο στήθος. Ουσιαστικά αποτελεί παραλλαγή του Διηθητικού Πορογενούς Καρκινώματος, μιας μορφής καρκίνου του μαστού που εμφανίζεται στους γαλακτοφόρους πόρους του αδένα και επεκτείνεται σταδιακά και στον υπόλοιπο μαστό.[3]

Θηλώδες (Θηλοειδές) Καρκίνωμα στο μαστό (Papillary Carcinoma of the Brest)

Το διηθητικό Θηλώδες Καρκίνωμα (Papillary Carcinoma) του μαστού είναι μια σπάνια μορφή καρκίνου που καταλαμβάνει μόλις το 1 -2 % των κρουσμάτων διηθητικού μαστικού καρκίνου παγκοσμίως.[3]

Διάτρητο Καρκίνωμα του Μαστού [Cribriform Carcinoma of the Breast]

Το Διάτρητο Καρκίνωμα (Cribriform Carcinoma) του μαστού είναι μια σπάνια μορφή μαστικού καρκίνου. Κατά την ανάπτυξή του, τα καρκινικά κύτταρα εισβάλλουν στους συνδετικούς ιστούς του στήθους και σχηματίζουν ελλειπτικές συσσωματώσεις ανάμεσα στους πόρους και στα λοβία. Ιδιαίτερο χαρακτηριστικό αυτού του τύπου καρκινώματος είναι η ιδιότυπη δομή του εσωτερικού, που δεν είναι συμπαγής αλλά διάτρητη, γεμάτη κενά.[3]

Διηθητικό Λοβιακό Καρκίνωμα (Invasive lobular carcinoma [ILC])

Το Διηθητικό Λοβιακό Καρκίνωμα (ILC) αποτελεί τη δεύτερη συνηθέστερη μορφή καρκίνου του μαστού μετά το Διηθητικό Πορογενές Καρκίνωμα. Υπολογίζεται πως το 10% όλων των μορφών διηθητικού μαστικού καρκίνου αφορά Διηθητικά Λοβιακά Καρκινώματα.[3]

Φλεγμονώδης Καρκίνος του Μαστού (Inflammatory breast cancer [IBC])

Ο Φλεγμονώδης Καρκίνος του Μαστού (Inflammatory Brest Cancer) είναι μια σπάνια αλλά εξαιρετικά επιθετική μορφή καρκίνου. Σύμφωνα με Εθνικό Καρκινικό Κέντρο των Η.Π.Α. ο Φλεγμονώδης Καρκίνος του Μαστού καταλαμβάνει ποσοστό 1 – 5% επί όλων των κρουσμάτων μαστικού καρκίνου.[3]

Διηθητικό Λοβιακό Καρκίνωμα (Invasive lobular carcinoma [ILC])

Το Διηθητικό Λοβιακό Καρκίνωμα (ILC) αποτελεί τη δεύτερη συνηθέστερη μορφή καρκίνου του μαστού μετά το Διηθητικό Πορογενές Καρκίνωμα. Υπολογίζεται πως το 10% όλων των μορφών διηθητικού μαστικού καρκίνου αφορά Διηθητικά Λοβιακά Καρκινώματα.[3]

ΔΙΑΓΝΩΣΗ

Ο καρκίνος του μαστού δεν προλαμβάνεται ακόμα πρωτογενώς διότι δεν γνωρίζουμε ακόμα τον παράγοντα δημιουργίας του αλλά προλαμβάνεται δευτερογενώς με έγκαιρη διάγνωση, πρόληψη και θεραπεία. Μια πρώτη διάγνωση για την παρουσία του καρκίνου του μαστού μπορεί να γίνει από την ίδια την ασθενή, όταν παρατηρήσει μια αλλοιωμένη υφή στο μαστό, όπως για παράδειγμα ένα μαλακό όγκο. Πέραν του 80% των περιπτώσεων έχουν ανιχνευθεί με αυτό τον τρόπο στις Ηνωμένες Πολιτείες Αμερικής Η δευτερογενής πρόληψη περιλαμβάνει κλινική εξέταση από ιατρό και μαστογραφία.[21] Η αξία της μαστογραφίας στη μάχη ενάντια στον καρκίνο του μαστού έχει αποδειχτεί ανεκτίμητη. Ας σημειωθεί πως μια μαστογραφία μπορεί να εντοπίσει αναπτυσσόμενους καρκινικούς όγκους που θα γίνονταν αντιληπτοί και ψηλαφητοί από την ασθενή ή το γιατρό περίπου δύο χρόνια αργότερα. [5]**Η μαστογραφία** – δηλαδή η εξέταση των μαστών μέσω ακτινογραφίας - διακρίνεται, ως προς τη μέθοδο, σε τρεις κατηγορίες:

Αναλογική Μαστογραφία

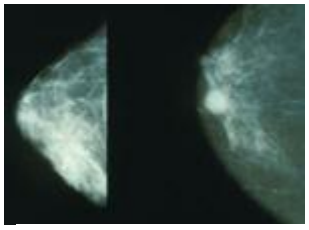
Η Αναλογική Μαστογραφία είναι η πλέον κοινή εξέταση μαστογραφίας παγκοσμίως. Η ανακάλυψή της συνέβαλε στη δραστική μείωση της ακτινοβολίας που δεχόταν το ανθρώπινο σώμα από παλαιότερες μορφές ακτινογραφικής εξέτασης των μαστών.[5]

Ψηφιακή Μαστογραφία

Η ψηφιακή μαστογραφία – γνωστή και ως FFDM (Full-Field Digital Mammography) - περιλαμβάνει ακριβώς την ίδια απεικονιστική μέθοδο, χρησιμοποιεί, όμως, ψηφιακή τεχνολογία και όχι αναλογικό φιλμ. Ως εξέταση, περιλαμβάνει ακριβώς την ίδια μέθοδο με την αναλογική· το μόνο που αλλάζει είναι η απεικονιστική τεχνολογία.[5]

Γαλακτογραφία

Η Γαλακτογραφία είναι η ακτινογραφική εξέταση του μαστού που αποτυπώνει με εξαιρετική διαύγεια τη δομή και το εσωτερικό των Γαλακτοφόρων αδένων του μαστού. .[5]



Γαλακτογραφία

Αποτελέσματα της μαστογραφίας: αριστερά διακρίνεται υγιής μαστός και δεξιά μαστός με καρκίνο.

Για να ισχύουν βέβαια όλα αυτά, θα πρέπει η μαστογραφία να ερμηνεύεται όσο ακριβέστερα γίνεται από τον ακτινολόγο, γεγονός που απαιτεί μεγάλη προσοχή κατά την εξέταση της και αποκτάται με την εμπειρία. Σε αυτό το σημείο εισέρχεται να βοηθήσει ο υπολογιστής για μια πιο έγκυρη διάγνωση.

Παραδοσιακά οι μαστογραφίες διαβάζονται και ερμηνεύονται από τους ακτινολόγους. Παρόλα’ αυτά με την πρόοδο της τεχνολογίας, οι μαστογραφίες έχουν πλέον ψηφιοποιηθεί. Συνάμα, έχουν αναπτυχθεί υπολογιστικά συστήματα υποβοηθούμενης διάγνωσης (Computer-Aided Diagnosis systems – CADs). Τα συστήματα αυτά χρησιμοποιούνται για να βελτιώσουν την αξιοπιστία της διάγνωσης των μαστογραφιών, προσφέροντας μια αξιόπιστη δεύτερη γνώμη στους ιατρούς, ιδιαίτερα σε περιπτώσεις όπου η διάγνωση των μαστογραφικών ευρημάτων είναι δυσδιάκριτη ένεκα διαφόρων παραγόντων.[22-24]
Η έγκαιρη διάγνωση είναι η καλύτερη δυνατή άμυνα ενάντια στην ασθένεια αυτή, αφού διαδραματίζει βαρυσήμαντο ρόλο στην αύξηση του ποσοστού επιβίωσης των ασθενών, ενώ παράλληλα ελαττώνει το συναισθηματικό τους φορτίο.
Οι πλέον συνήθεις μέθοδοι που ακολουθούνται είναι η μαστογραφία, η υπερηχοτομογραφία, η μαγνητική τομογραφία, οι ραδιοϊσοτοπικές τεχνικές και οι επεμβατικές τεχνικές.

Η διάγνωση ξεκινά όταν ανακαλυφθεί μια μάζα στο μαστό μετά από ψηλάφηση ή μαστογραφία ή υπέρηχο και στη συνέχεια ακολουθούν εξετάσεις ώστε να διευκρινιστεί η διαφορά ανάμεσα σε ένα συμπαγή όγκο και μια κύστη γεμάτη υγρό.

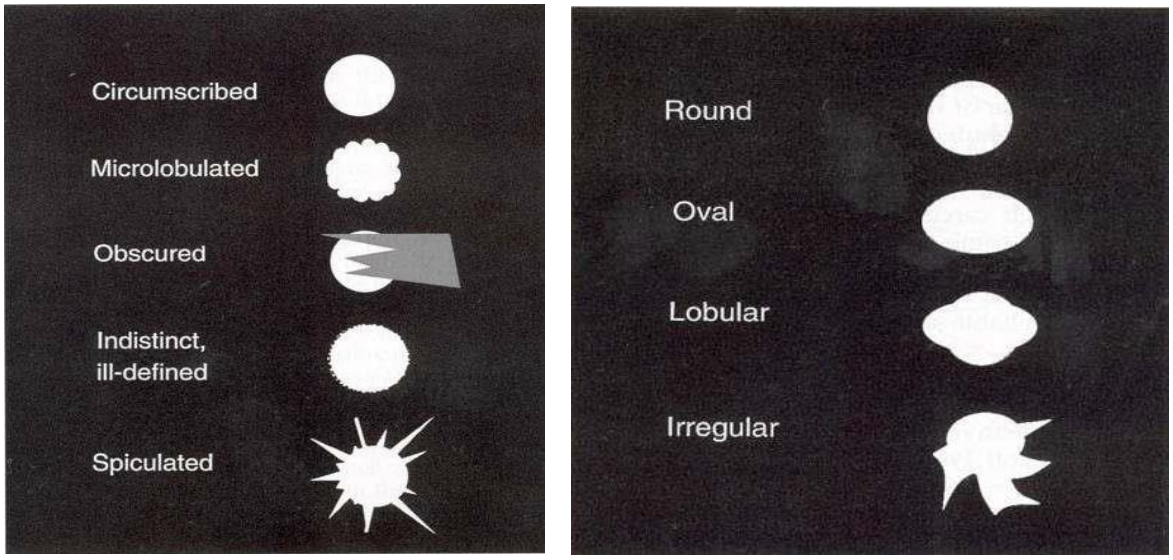
Υπάρχουν επίσης 4 χειρουργικές τεχνικές βιοψίας του μαστού:[12]

1. **Κυτταρολογική βιοψία**: Γίνεται με μια λεπτή βελόνα (FNA - Fine Needle Aspiration Biopsy) που προσαρμόζεται σε μια σύριγγα και επιτρέπει την αναρρόφηση κυττάρων από την ύποπτη περιοχή του μαστού.
2. **Ιστολογική βιοψία**: η βελόνα σε αυτή την περίπτωση είναι μεγαλύτερη, απαιτείται τοπική αναισθησία και τομή και το υλικό που μελετάται είναι ιστολογικό.
3. **Στερεοτακτική βιοψία**: η λήψη του ιστολογικού υλικού γίνεται με την καθοδήγηση ειδικών μηχανημάτων
4. **Ανοιχτή χειρουργική βιοψία**: η ταυτοποίηση της φύσης της ύποπτης περιοχής γίνεται με ταχεία βιοψία κατά τη διάρκεια του χειρουργείου πριν την αφαίρεση του όγκου.

Σημείωση: Η FNA, η παρακέντηση με λεπτή βελόνα και αναρρόφηση υλικού προς κυτταρολογική εξέταση τείνει να καταργηθεί γιατί δε προσφέρει επαρκείς και αληθείς πληροφορίες.

2.5 Μάζα

Ως μάζα αποκαλείται η χωροκατακτητική επεξεργασία (τρισδιάστατη δομή) που απεικονίζεται σε δύο τουλάχιστον μαστογραφικές προβολές. Εφ' όσον διαπιστώσουμε μάζα, θα πρέπει να τη χαρακτηρίσουμε περεταίρω ως προς το σχήμα, τα όρια και την πυκνότητά της. Το σχήμα μιας μάζας μπορεί να είναι στρογγυλό, ωοειδές, λοβωτό ή ανώμαλο, ενώ τα όριά της περιγράφονται ως περιγεγραμμένα, ομαλά, μικρολοβωτά, εν μέρει αποκρυπτόμενα από το παρακείμενο μαζικό παρέγχυμα, ασαφή καθώς και με προεκβολές προς τους γύρω ιστούς. Η πυκνότητα της μάζας συγκρίνεται με αυτήν του αδενικού ιστού και χαρακτηρίζεται ως ίση, μικρότερη ή μεγαλύτερη. Στο σχήμα 4, παρουσιάζονται τα σχήματα και τα όρια που χαρακτηρίζουν μια μάζα αντίστοιχα.



Σχήματα και όρια που χαρακτηρίζουν μια μάζα

Η μάζα μοιάζει συνήθως με την υγιή περιοχή υψηλής πυκνότητας, τόσο ως προς τη μορφολογία όσο και ως προς τη φωτεινότητά της. Συνεπώς, η ανίχνευσή της καθίσταται δύσκολη. Ωστόσο, κάποια χαρακτηριστικά της βοηθούν τους ακτινολόγους να εκτιμήσουν εάν πρόκειται για καλοήθεια ή κακοήθεια. Συνήθως η καλοήθεια έχει ευκρινές περίγραμμα, είναι συμπαγής και το σχήμα της είναι σχεδόν κυκλικό ή ελλειπτικό. Τουναντίον, η κακοήθης μάζα έχει ασαφές περίγραμμα, έχει ανομοιόμορφο σχήμα και η εξωτερική της επιφάνεια δύναται να παρουσιάζει ακτινικούς σχηματισμούς. Παρόλα αυτά όμως, πιθανό κάποια καλοήθεια να παρουσιάζει ασαφές περίγραμμα ή ακτινωτούς σχηματισμούς στην επιφάνειά της, με αποτέλεσμα να ταξινομηθεί σε λανθασμένη κατηγορία (False Positive) [7]. Δηλαδή, ενώ πρόκειται για καλοήθεια, ταξινομείται ως κακοήθεια.

2.6 Κατηγοριοποίηση μαστογραφικών ευρημάτων κατά BI-RADS καθώς και τα πλεονεκτήματα και περιορισμοί της κατηγοριοποίησης

Η αναφορά και ταξινόμηση των μαστογραφικών ευρημάτων μπορεί να γίνει με διάφορους τρόπους. Μια πρώτη κατηγοριοποίηση γίνεται σύμφωνα με την ακριβή σύνθεση του στήθους ανάλογα με την πυκνότητά του . Σύμφωνα με αυτόν τον τρόπο ταξινόμησης, υπάρχουν οι εξής τέσσερις κατηγορίες :[25]

- Τύπου 1: λιπώδης μαστός (πυκνότητα ιστού μικρότερη από 10%)
- Τύπου 2: ινοαδένωμα (πυκνότητα ιστού 10-49%)
- Τύπου 3: ανομοιογενής πυκνότητα (πυκνότητα ιστού 49-90%)
- Τύπου 4: πυκνότητα κα ανομοιογένεια (πυκνότητα μεγαλύτερη από 90%)

Η ακρίβεια της μαστογραφίας για να ανιχνεύσει τις ύποπτες ανωμαλίες, μειώνεται για τις κατηγορίες τύπου 3 και 4 .[26]

Το εργαλείο BI-RADS είναι ένα οδηγός διασφάλισης της ποιότητας της αναφοράς από την μαστογραφία με στόχο την σταθεροποίηση και διευκόλυνση της παρακολούθησης αποτελεσμάτων.

Το εργαλείο BI-RADS χρησιμεύει ως ένα ολοκληρωμένος οδηγός που θα παρέχει τυποποιημένη ορολογία της απεικόνισης του μαστού, την οργάνωση και τη δομή έκθεσης αξιολόγησης, καθώς και σύστημα ταξινόμησης για την μαστογραφία, για το υπερηχογράφημα και την μαγνητική τομογραφία (MRI) του μαστού. Πρόκειται για μια συστηματική μέθοδος για τους ακτινολόγους για να αναφέρουν τα ευρήματά από την μαστογραφία με χρήση 7 τυποποιημένων κατηγοριών ή επιπέδων. Κάθε BI-RADS κατηγορία έχει και μία σύσταση συσχετισμένη με αυτήν για να βοηθήσει τους ακτινολόγους και άλλους γιατρούς να διαχειρίζονται κατάλληλα τη φροντίδα του ασθενούς. Οι 7 κατηγορίες είναι:[9]

BI-RADS

Breast Imaging Reporting and Database System (BI-RADS)		
Κατηγορία	Αξιολόγηση	Οι επακόλουθες συστάσεις
Α) Η αξιολόγηση δεν έχει ολοκληρωθεί		
0	Χρειάζεστε πρόσθετη αξιολόγηση Εικόνας και / ή προηγούμενη μαστογραφία για σύγκριση	Πρόσθετες απεικονίσεις ή / και προηγούμενες εικόνες είναι απαραίτητες πριν να μπορεί να ανατεθεί τελική αξιολόγηση

Breast Imaging Reporting and Database System (BI-RADS)		
Κατηγορία	Αξιολόγηση	Οι επακόλουθες συστάσεις
Β) Η αξιολόγηση έχει ολοκληρωθεί		
1	Αρνητικό	Τακτική ετήσια προληπτική μαστογραφία (για γυναίκες άνω των 40 ετών)
2	Καλοήθη ευρήματα	Τακτική ετήσια προληπτική μαστογραφία (για γυναίκες άνω των 40 ετών)
3	Πιθανόν Καλοήθη ευρήματα	αρχική βραχυπρόθεσμη εξέταση (συνήθως 6-μηνών)
4	Ύποπτη ανωμαλία-θα πρέπει να ληφθεί υπόψη η βιοψία Προαιρετικά υποδιαιρέσεις: 4Α: Η εύρεση χρειάζεται παρέμβαση με μια χαμηλή υποψία για κακοήθεια 4Β: Αλλοιώσεις με μια μέτρια υπόνοια κακοήθειας 4C: Ευρήματα μέτριων ανησυχιών, αλλά δεν είναι κλασική περίπτωση για κακοήθεια	Συνήθως χρειάζεται βιοψία
5	Ισχυρές ενδείξεις κακοήθειας-Κατάλληλα μέτρα πρέπει να ληφθούν	Απαιτείται βιοψία ή χειρουργική θεραπεία
6	Γνωστή Βιοψία-Αποδεδειγμένη κακοήθεια - Κατάλληλα μέτρα πρέπει να ληφθούν	Κατηγορία προορίζεται για τις αλλοιώσεις που εντοπίσθηκαν σε απεικονιστική εξέταση, με την απόδειξη της βιοψίας κακοήθειας πριν από την οριστική

		θεραπεία

Ακολούθως, τα μαστογραφικά ευρήματα πρέπει να περιέχουν και την περιγραφή όλων των καλοηθειών, ύποπτων ανωμαλιών ή κακοηθειών που εμφανίζονται στις μαστογραφίες. Αυτές οι ανωμαλίες είναι που κατηγοριοποιούνται κατά BI-RADS (Breast Imaging- Reporting and Data System). Το λεξικό BI-RADS αποτελεί ένα εργαλείο που ορίστηκε για να μειώσει τη μεταβλητότητα μεταξύ των ακτινολόγων και χρησιμοποιείται σήμερα ευρέως στις πλείστες χώρες που χρησιμοποιούν τη μαστογραφία ως απεικονιστική μέθοδο για τον καρκίνο του μαστού. Αποτελεί ένα ενδιαφέρον εργαλείο για την εκπαίδευση των νέων ακτινολόγων. Το λεξικό αυτό προέκυψε από τη συνεργασία πολλών ομάδων που ασχολούνται με την υγεία, αλλά δημοσιεύτηκε και κατατέθηκε ως εμπορικό σήμα από το Αμερικάνικο Κολλέγιο Ακτινολογίας (American College of Radiology)[25].

2.6.1 Πλεονεκτήματα και περιορισμοί της κατηγοριοποίησης κατά BI-RADS

Η ταξινόμηση των μαστογραφικών ευρημάτων κατά BI-RADS έχει βασικά πλεονεκτήματα όπως:

- Η αύξηση της σαφήνειας της μαστογραφικής έκθεσης.
- Η βελτίωση της επικοινωνίας μεταξύ των διάφορων ιατρικών ειδικοτήτων που ασχολούνται με τις παθήσεις του μαστού.
- Η προώθηση της έρευνας.
- Η εξαγωγή χρήσιμων στατιστικών για την σωστή και αξιόπιστη διάγνωση σε συνδυασμό με άλλα χαρακτηριστικά των ασθενών όπως είναι η ηλικία το σχήμα και η πυκνότητα του όγκου στην μαστογραφία.

Επιπλέον η κατηγοριοποίηση των μαστρογραφικών ευρημάτων κατά BI-RADS υφίσταται. Ακόμα κι αν δεν αποτελεί την τέλεια μέθοδο, καθορίζει έναν οδηγό ερμηνείας για τις μαστογραφικές εικόνες, λιγότερο σχετικό με την υποκειμενικότητα του ακτινολόγου. Επιτρέπει επίσης μια ομογενοποίηση της γλώσσας μεταξύ των ακτινολόγων, αλλά και μεταξύ των ακτινολόγων και των ιατρών άλλων ειδικοτήτων. Υπάρχουν έτσι λιγότερες παρερμηνείες στις αναφορές που κάνουν για τα μαστογραφικά ευρήματα [27]. Συνεπώς, η ύπαρξη ενός καθορισμένου οδηγού διευκολύνει τη σύγκριση, η οποία είναι χρήσιμη στη μαστογραφία.

Αντιθέτως, η κατηγοριοποίηση κατά BI-RADS εμφανίζει και κάποιους περιορισμούς. Όπως κάθε άλλη μέθοδος ταξινόμησης, έτσι κι αυτή δεν είναι τέλεια. Μερικοί ακτινολόγοι συνηθίζουν να γράφουν τους δικούς τους όρους γιατί είναι επιφυλακτικοί

απέναντι στην ακριβή ορολογία που τους έχει επιβληθεί από το λεξικό BIRADS. Εντούτοις, αυτή η μέθοδος ταξινόμησης παραμένει μια ακτινολογική ταξινόμηση και έτσι δεν υπολογίζει κάποιους κλινικούς ή προγνωστικούς παράγοντες, οι οποίοι θα μπορούσαν να αλλάξουν κατηγορία σε μερικές μαστογραφίες [28]. Τέλος, αυτή η ταξινόμηση παρουσιάζει μεγάλη μεταβλητότητα παρατηρήσεων για εικόνες που είναι δυσκολότερο να ταξινομηθούν.

Κεφάλαιο 3

3.1Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων (το βήμα ανάλυση της διαδικασίας «Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων» ,ή KDD), [29] ένα διεπιστημονικό υποπεδίο της επιστήμης των υπολογιστών, [30] [31] [32] είναι η υπολογιστική διαδικασία που ανακαλύπτει μοτίβα σε μεγάλα σύνολα δεδομένων διασταυρώνοντας μεθόδους της τεχνητής νοημοσύνης, της μηχανική μάθηση, καθώς και στατιστικά στοιχεία, και συστήματα βάσεων δεδομένων. [30] Ο γενικός στόχος της διαδικασίας εξόρυξης δεδομένων είναι η εξαγωγή πληροφοριών από ένα σύνολο δεδομένων και ταυτόχρονα η μετατροπή σε μια κατανοητή δομή για περαιτέρω χρήση. [30] Εκτός από την πρώτη βαθμίδα ανάλυσης, περιλαμβάνει στοιχεία αξιοποίησης της παραχθείσας βάσης δεδομένων και πτυχές διαχείρισης αυτής, προ-επεξεργασία των δεδομένων, μοντελοποίηση και εκτίμηση συμπερασμάτων, χρήσιμες μετρήσεις , πολύπλοκες εκτιμήσεις που δεν είναι φανερές στον άνθρωπο, καθώς και μετέπειτα επεξεργασία των ανακαλυφθεισών δομών, απεικόνιση των αποτελεσμάτων για ευκολότερη κατανόηση από τον άνθρωπο, και άμεση ενημέρωση της βάσης και των εκτιμήσεων. [30]

Το πραγματικό έργο της εξόρυξης δεδομένων είναι η αυτόματη ή ημι-αυτόματη ανάλυση των μεγάλων ποσοτήτων δεδομένων για την εξαγωγή προηγουμένως άγνωστων ενδιαφερόντων μοτίβων όπως οι ομάδες αρχείων δεδομένων (cluster analysis), ασυνήθιστα αρχεία (anomaly detection) και εξαρτήσεις (association rule mining). Αυτό συνήθως περιλαμβάνει τη χρήση τεχνικών δεδομένων, όπως η χωρικοί δείκτες. Αυτά τα μοτίβα μπορεί στη συνέχεια να θεωρηθούν ως ένα είδος σύνοψης των δεδομένων εισόδου, και μπορεί να χρησιμοποιηθούν σε περαιτέρω ανάλυση ή, για παράδειγμα, στη μάθηση μηχανής και προγνωστική ανάλυση. Για παράδειγμα, το στάδιο εξόρυξης δεδομένων θα μπορούσε να εντοπίσει πολλαπλές ομάδες στα δεδομένα, τα οποία μπορούν στη συνέχεια να χρησιμοποιηθούν για την απόκτηση ακριβέστερων αποτελεσμάτων πρόβλεψης από ένα σύστημα υποστήριξης αποφάσεων. Ούτε η συλλογή δεδομένων, η προετοιμασία, ούτε και η ερμηνεία των αποτελεσμάτων και η υποβολή εκθέσεων είναι μέρος του σταδίου της εξόρυξης δεδομένων, αλλά ανήκουν στη γενική KDD (Knowledge Discovery in Databases) διαδικασία ως πρόσθετα μέτρα.[33]

Η εξόρυξη δεδομένων χρησιμοποιεί πληροφορίες από το παρελθόν για να αναλύσει το αποτέλεσμα της σε ένα συγκεκριμένο πρόβλημα ή μια κατάσταση που μπορεί να προκύψει. Η εξόρυξη δεδομένων χρησιμοποιείται για να αναλύσει τα δεδομένα που

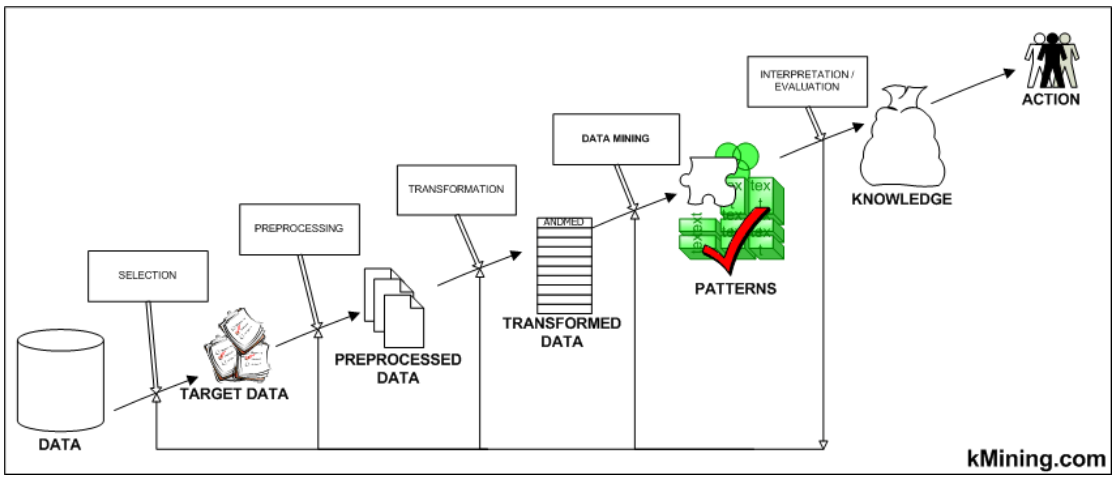
αποθηκεύονται στις αποθήκες δεδομένων, όπου ο όγκος τους είναι αδύνατον να επεξεργασθεί αποτελεσματικά και γρήγορα από τον άνθρωπο.

Το 2011, η δικαστική υπόθεση του Sorrell κατά IMS Health, Inc, που αποφασίστηκε από το Ανώτατο Δικαστήριο των Ηνωμένων Πολιτειών, έκρινε, ότι τα φαρμακεία μπορούν να ανταλλάσσουν πληροφορίες με εξωτερικές εταιρίες. Η πρακτική αυτή έχει εγκριθεί με την 1η τροπολογία του Συντάγματος, την προστασία της «ελευθερίας του λόγου». [34]

Ο κύριος σκοπός της εξόρυξης δεδομένων λοιπόν είναι, να εξαχθούν τα κρυμμένα πρότυπα από τα δεδομένα και να αυξηθεί η εγγενής αξία τους μετατρέποντας τα δεδομένα σε γνώση. Μπορεί να τεθεί το ερώτημα, γιατί δεν γίνεται η αναζήτηση αυτής της γνώσης με τη χρήση των κλασικών ερωτήσεων SQL, ή αλλιώς ποιές είναι οι θεμελιώδεις διαφορές μεταξύ εξόρυξης δεδομένων και τεχνολογιών σχεσιακών βάσεων δεδομένων. Το ακόλουθο παράδειγμα μπορεί να δώσει μια ικανοποιητική προσέγγιση. Ας υποθέσουμε ότι έχουμε ένα πίνακα με στοιχεία ασθενών, όπως γένος, ηλικία, μετρήσεις αίματος, DNA, ιστορικό ασθενειών και ιστορικό καρκίνων σε συγγενείς. Θέλουμε να δούμε τι είναι αυτό που οδηγεί κάποιον στο να εμφανίσει καρκίνο. Μπορεί να γραφεί μια SQL ερώτηση για να καταδείξει, πόσοι άνδρες σε σχέση με γυναίκες έχουν καρκίνο. Μπορεί επίσης να γραφεί μια SQL ερώτηση που θα δείξει ποιά είναι η επιρροή της παραμέτρου «ιστορικό καρκίνου σε συγγενικά πρόσωπα». Αλλά τι γίνεται με τους άνδρες πάνω από 30 που έχουν συγγενείς με ιστορικό ασθενειών ή με τις γυναίκες που δεν έχουν; Θα πρέπει να γραφούν εκατοντάδες αυτών των ερωτήσεων για να καλυφθούν όλοι οι πιθανοί συνδυασμοί. [35][36].

Ακόμα δυσκολότερη καθίσταται η επεξεργασία στην περίπτωση που τα δεδομένα δίδονται σε αριθμητικές τιμές. Τότε θα πρέπει να επιλεγούν αυθαίρετες περιοχές αριθμητικών τιμών, να υπάρχουν εκατοντάδες στήλες στον πίνακα και θα καταλήγαμε γρήγορα σε έναν αριθμό SQL ερωτήσεων δύσκολα διαχειρίσιμων για να απαντηθεί μια βασική ερώτηση. Αντίθετα, η προσέγγιση του προβλήματος με μεθόδους εξόρυξης δεδομένων είναι μάλλον απλή. Γίνεται επιλογή του σωστού αλγόριθμου εξόρυξης δεδομένων και επιλέγεται η χρήση κάθε στήλης δεδομένων, σαν στήλης εισόδου ή στήλης πρόβλεψης. Σύμφωνα με το προηγούμενο παράδειγμα, ένας αλγόριθμος δέντρων απόφασης θα μπορούσε να αναδείξει την επιρροή του ιστορικού καρκίνου σε συγγενικά πρόσωπα στην εμφάνιση καρκίνου στο προς εξέταση άτομο. Στην προκειμένη περίπτωση, θα επιλέγαμε όλες τις στήλες σαν στήλες εισόδου εκτός από την στήλη «εμφάνιση καρκίνου», που θα ήταν η στήλη πρόβλεψης. Ο αλγόριθμος θα επεξεργαστεί τα δεδομένα, θα αναλύσει την επίδραση κάθε ιδιότητας εισόδου και θα επιλέξει την πιο σημαντική ιδιότητα ώστε να διαχωρίσει τα δεδομένα σε δύο υποσύνολα, έτσι ώστε οι τιμές της παραμέτρου εξόδου (ύπαρξη καρκίνου) να είναι όσο το δυνατόν πιο διαφορετικές στα δύο υποσύνολα. Αυτή η διαδικασία επαναλαμβάνεται σε κάθε υποσύνολο έως ότου έχει ολοκληρωθεί το δέντρο. Μόλις η διαδικασία εκπαίδευσης του δέντρου ολοκληρωθεί, μπορεί κάποιος να ανακαλύψει τα πρότυπα διατρέχοντας το δέντρο. Κάθε πορεία από

τη ρίζα σε έναν κόμβο φύλλων διαμορφώνει έναν κανόνα με αποτέλεσμα να μπορούν να εξαχθούν συμπεράσματα, όπως για παράδειγμα, ότι ασθενείς συγκεκριμένου ηλικιακού εύρους, με συγκεκριμένο σχήμα όγκου παράλληλα με την συγκεκριμένη κατηγοριοποίηση BI-RADS, που έγινε από τον εξεταστή ογκολόγο, παρουσιάζει συγκεκριμένο ποσοστό πιθανότητας εμφάνισης κακοήθους όγκου. Με αυτόν τον τρόπο, εξάγεται πληροφορία από τα υπάρχοντα δεδομένα. Η προ αναφερθείσα διαδικασία απεικονίζεται πολύ καλά στην παρακάτω εικόνα:



Εξόρυξη Δεδομένων

Στην συνέχεια θα εξετάσουμε τις βασικές κατηγορίες μεθόδων εξόρυξης δεδομένων με σκοπό να κατηγοριοποιηθούν οι τεχνικές που χρησιμοποιήθηκαν στα πλαίσια αυτής της διπλωματικής.

3.2 Η διαδικασία και τα στάδια της εξόρυξης δεδομένων

Για τη διεξαγωγή συστηματικής ανάλυσης των δεδομένων εξόρυξης, συνήθως ακολουθείται μία γενική διαδικασία. Υπάρχουν μερικές τυποποιημένες διαδικασίες, η πιο συνηθισμένη από αυτές είναι η CRISP (Cross-Industry Standard Process for Data Mining). Η CRISP είναι μια βιομηχανοποιημένη διαδικασία-πρότυπο, που αποτελείται από μια ακολουθία βημάτων που συνήθως εμπλέκονται σε μια μελέτη εξόρυξης δεδομένων.[42]

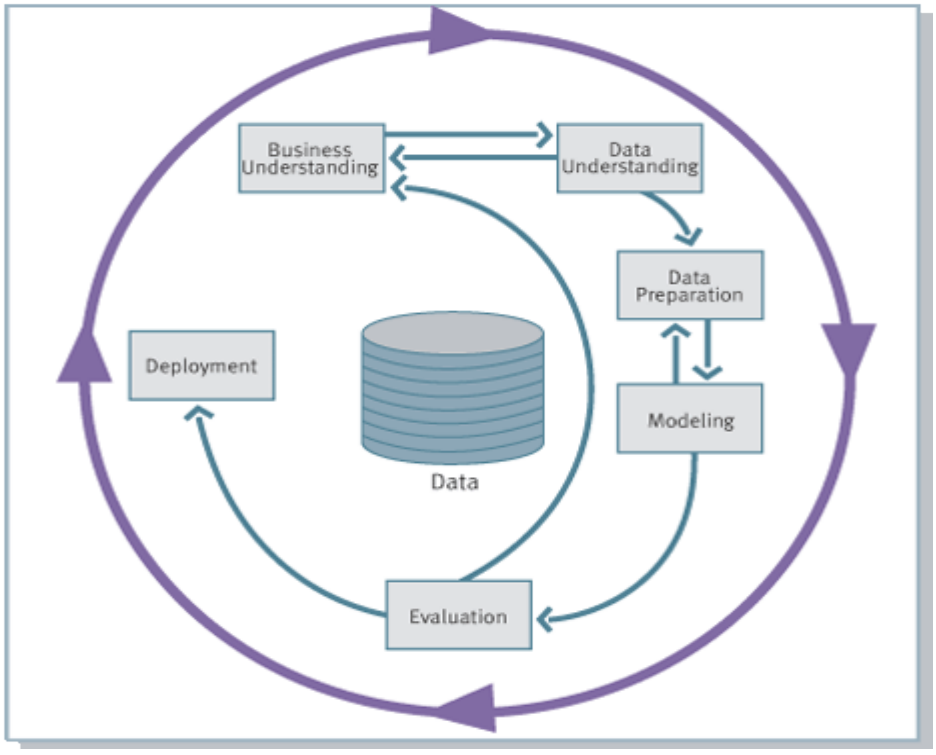
Το μοντέλο αυτό αποτελείται από έξι στάδια που προορίζονται ως μια κυκλική διαδικασία (βλέπε Σχήμα παρακάτω):

- **Κατανόηση του προβλήματος** (Business understanding): περιλαμβάνει τον καθορισμό των επιχειρηματικών στόχων, αξιολόγηση της τρέχουσας κατάστασης, ίδρυση στόχους για την μέθοδο εξόρυξης, καθώς και την ανάπτυξη ενός σχεδίου (project).[42]

- **Κατανόηση των δεδομένων**(data understanding): Μόλις καθοριστούν οι επιχειρηματικοί στόχοι και το σχέδιο του έργου, με την κατανόηση των δεδομένων υπολογίζονται οι απαιτήσεις δεδομένων. Αυτό το βήμα μπορεί να περιλαμβάνει την αρχική συλλογή δεδομένων, περιγραφή, διερεύνηση δεδομένων, και την επαλήθευση της ποιότητας των δεδομένων. Δεδομένα εξερεύνησης, όπως η προβολή συνοπτικά των στατιστικών στοιχείων (η οποία περιλαμβάνει την οπτική απεικόνιση των μεταβλητών κατηγοριοποίησης) μπορεί να συμβεί στο τέλος αυτής της φάσης. Μοντέλα, όπως το σύμπλεγμα ανάλυση μπορούν επίσης να εφαρμοστούν κατά τη διάρκεια αυτής της φάσης, με την πρόθεση να προσδιοριστούν τα μοτίβα στα δεδομένα.[42]
- **Προετοιμασία των δεδομένων** (data preparation) : Μόλις προσδιοριστούν τα δεδομένα και οι πόροι που είναι διαθέσιμοι, αυτά πρέπει να επιλέγουν, να καθαριστούν, να χτιστούν με βάση την επιθυμητή μορφή, και να μορφοποιηθούν. Ο καθαρισμός των δεδομένων και η μετατροπή των δεδομένων στο πλαίσιο της προετοιμασίας της μοντελοποίησης δεδομένων πρέπει να συμβεί σε αυτή την φάση. Δεδομένα εξερεύνησης σε μεγαλύτερο βάθος μπορεί να εφαρμόζονται κατά τη διάρκεια αυτής της φάσης, και μπορούν να χρησιμοποιηθούν επιπλέον μοντέλα, και πάλι παρέχοντας την ευκαιρία να δούμε τα πρότυπα που βασίζονται στην κατανόηση των επιχειρήσεων.[42]
- **Μοντελοποίηση δεδομένων** (Modeling) εργαλείων λογισμικού εξόρυξης, όπως οπτικοποίηση (εκτυπώνοντας δεδομένα και τη θέσπιση σχέσεων) και ανάλυση διασποράς (για τον εντοπισμό μεταβλητών που πηγαίνουν καλά μαζί) είναι χρήσιμη για την αρχική ανάλυση. Εργαλεία όπως η γενικευμένη επαγωγή κανόνα μπορεί να αναπτύξει αρχικούς κανόνες συσχέτισης. Μόλις επιτευχθεί μεγαλύτερη κατανόηση των δεδομένων (συχνά μέσω μοτίβου αναγνώρισης, που προκλήθηκε από την προβολή της παραγωγής μοντέλο), μπορούν να εφαρμοστούν πιο λεπτομερή μοντέλα κατάλληλα για το είδος των δεδομένων. Η διαίρεση των δεδομένων σε σύνολα εκπαίδευσης(training) και test sets είναι επίσης απαραίτητη για τη μοντελοποίηση.[42]
- Τα αποτελέσματα του **Μοντέλου Αξιολόγησης** (evaluation) θα πρέπει να αξιολογηθούν στο πλαίσιο των στόχων που έχουν διατυπωθεί στο πρώτο στάδιο (κατανόηση του προβλήματος). Αυτό θα οδηγήσει στην ταυτοποίηση άλλων αναγκών (όπως για παράδειγμα μέσω της αναγνώρισης προτύπων), που συχνά επανέρχεται στην προηγούμενη φάσεις CRISP-DM. Επιτυγχάνοντας την κατανόηση του προβλήματος επιτυγχάνεται ταυτόχρονα μια επαναληπτική διαδικασία εξόρυξης δεδομένων, όπου τα αποτελέσματα των διαφόρων εργαλείων όπως οπτικοποίηση, στατιστικές, και τεχνητή νοημοσύνη δείχνουν στον χρήστη νέες σχέσεις που παρέχουν μια βαθύτερη κατανόηση των οργανωτικών εργασιών.[42]
- **Ανάπτυξη** : η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί τόσο για την επαλήθευση προηγουμένων υποθέσεων, ή για την απόκτηση της γνώσης

(εντοπισμός των απρόβλεπτων και χρήσιμων σχέσεων). Μέσω της απόκτησης της γνώσης στις προηγούμενες φάσεις της διαδικασίας CRISP-DM , μοντέλα ήχου μπορούν να αποκτηθούν που μπορεί στη συνέχεια να εφαρμοστούν σε επιχειρηματικές δραστηριότητες για πολλούς σκοπούς, συμπεριλαμβανομένης της πρόβλεψης ή προσδιορισμού των βασικών καταστάσεων. Αυτά τα μοντέλα θα πρέπει να παρακολουθούνται για αλλαγές στις συνθήκες λειτουργίας, διότι αυτό, που θα μπορούσε να ισχύει σήμερα μπορεί να μην ισχύει σε ένα χρόνο από σήμερα. Εάν υπάρξουν σημαντικές αλλαγές, το μοντέλο θα πρέπει να επαναληφθεί. Είναι επίσης σκόπιμο να καταγράφονται τα αποτελέσματα των έργων (projects) δεδομένων εξόρυξης, ώστε τεκμηριωμένα αποδεικτικά στοιχεία να είναι διαθέσιμα για μελλοντικές μελέτες.[42]

Αυτή η έξι φάσεων διαδικασία δεν είναι μια άκαμπτη, αυστηρή διαδικασία. Υπάρχει συχνά επαναφορά σε προηγούμενη φάση και ξανά σε επόμενη μέχρι την επίτευξη του αποτελέσματος. Επιπλέον, έμπειροι αναλυτές μπορεί να μην χρειάζεται να εφαρμόζουν κάθε φάση για κάθε μελέτη. Αλλά η CRISP-DM παρέχει ένα χρήσιμο πλαίσιο για την εξόρυξη δεδομένων. Όλα τα στάδια που περιγράφηκαν αναπαρίστανται στην παρακάτω εικόνα:



Στάδια Εξόρυξης Δεδομένων

3.3 Βασικές κατηγορίες μεθόδων εξόρυξης δεδομένων

Οι δύο κύριοι στόχοι της εξόρυξης δεδομένων είναι η περιγραφή και η πρόβλεψη. Η περιγραφή αφορά στην αναπαράσταση των δεδομένων μιας πολύπλοκης βάσης δεδομένων με ένα κατανοητό και αξιοποιήσιμο τρόπο και η πρόβλεψη στην ανεύρεση κρυμμένων προτύπων, αποκάλυψη μη αναμενόμενων σχέσεων και πρόβλεψη μελλοντικών συνθηκών, συμπεριφορών και τάσεων. [35]

Ταξινόμηση (classification): Η ταξινόμηση είναι μία από τις δημοφιλέστερες κατηγορίες εξόρυξης δεδομένων. Επιχειρησιακά προβλήματα, όπως η διαχείριση κινδύνου επιλύονται συνήθως με την ταξινόμηση. Η ταξινόμηση αφορά στην κατάταξη των περιπτώσεων σε κατηγορίες με βάση μια προβλέψιμη ιδιότητα. Κάθε περίπτωση έχει ένα σύνολο ιδιοτήτων, μια από τις οποίες χαρακτηρίζεται ως ιδιότητα κατηγορίας (προβλέψιμη ιδιότητα). Σκοπός είναι να παραχθεί ένα μοντέλο το οποίο θα περιγράφει την ιδιότητα κατηγορίας σαν συνάρτηση των ιδιοτήτων εισόδου. Στην περίπτωση του συνόλου δεδομένων ασθενών που χρησιμοποιήθηκαν στην παρούσα διδακτορική διατριβή δύο κατηγορίες εξετάστηκαν : αυτές στις οποίες υπάρχει και αυτές στις οποίες δεν υπάρχει καρκίνος. Για να εκπαιδευτεί το μοντέλο (supervised), πρέπει για τα δεδομένα εκπαίδευσης να είναι γνωστή η κατάσταση της ιδιότητας «ύπαρξη καρκίνου», κάτι το οποίο προκύπτει από το ιστορικό των ασθενών. Αντίστοιχα και για την περίπτωση «μη ύπαρξη καρκίνου» Χαρακτηριστικοί αλγόριθμοι ταξινόμησης είναι τα δέντρα απόφασης, τα τεχνητά νευρωνικά δίκτυα, τα Naïve Bayes κ.α. [36][37][38][39][40].

Ομαδοποίηση (clustering): Η ομαδοποίηση ή αλλιώς κατάτμηση χρησιμοποιείται για να προσδιορίσει τους φυσικούς σχηματισμούς ομάδων από τα δεδομένα με βάση ένα σύνολο κοινών ιδιοτήτων. Οι περιπτώσεις μέσα στην ίδια ομάδα έχουν περισσότερες ή λιγότερες παρόμοιες τιμές ιδιοτήτων. Συγκεκριμένα, έστω ένα σύνολο δεδομένων πελατών που περιέχει δύο ιδιότητες: ηλικία και εισόδημα. Ένας αλγόριθμος ομαδοποίησης συγκεντρώνει το σύνολο δεδομένων με βάση αυτές τις ιδιότητες. Η ομάδα 1 περιέχει το νεώτερο πληθυσμό με χαμηλό εισόδημα. Η ομάδα 2 περιέχει τους μέσης ηλικίας πελάτες με υψηλό εισόδημα. Η ομάδα 3 τους μεγαλύτερης ηλικίας με χαμηλό εισόδημα και ούτω καθεξής. Οι αλγόριθμοι ομαδοποίησης χαρακτηρίζονται ως μη επιβλέψιμοι (non supervised). Καμία ιδιότητα δεν χρησιμοποιείται για να καθοδηγήσει τη διαδικασία κατάρτισης, όλες οι ιδιότητες εισόδου αντιμετωπίζονται εξίσου [36][37][38][39][40].

Ένωση (association): Η ένωση είναι επίσης μια από τις δημοφιλείς μεθόδους εξόρυξης δεδομένων. Μια χαρακτηριστική εφαρμογή της ένωσης είναι η ανάλυση του πίνακα πωλήσεων μιας επιχείρησης ώστε προσδιορισθούν εκείνα τα προϊόντα τα οποία πωλούνται συχνά μαζί στο ίδιο καλάθι αγορών. Η συνήθης χρήση της ένωσης

είναι να προσδιορίσει τα κοινά σύνολα στοιχείων (συχνά σετ στοιχείων) και τους κανόνες δημιουργίας αυτών με σκοπό την στοχευόμενη πώληση. Από την άποψη της ένωσης, κάθε προϊόν, ή γενικότερα κάθε ιδιότητα/αξία θεωρείται στοιχείο. Οι περισσότεροι αλγόριθμοι τύπου ένωσης βρίσκουν τα συχνά σετ στοιχείων με την ανίχνευση του συνόλου δεδομένων πολλές φορές. Το κατώτατο όριο συχνότητας (υποστήριξης) καθορίζεται από το χρήστη πριν την επεξεργασία. Παραδείγματα τέτοιων αλγορίθμων είναι οι αλγόριθμοι που βασίζονται στην υποστήριξη κανόνα (support rule) και στην εμπιστοσύνη κανόνα (confidence rule) [36][37][38][39][40][41].

Παλινδρόμηση (regression): Η μέθοδος της παλινδρόμησης είναι παρόμοια με την ταξινόμηση. Η κύρια διαφορά είναι ότι η προβλέψιμη ιδιότητα είναι συνεχής αριθμός. Οι τεχνικές παλινδρόμησης χρησιμοποιούνται ευρέως στον τομέα της στατιστικής. Η γραμμική συμμεταβολή και η λογιστική παλινδρόμηση αποτελούν δημοφιλέστερες μεθόδους παλινδρόμησης. Άλλες μέθοδοι παλινδρόμησης περιλαμβάνουν τα δέντρα παλινδρόμησης και τα τεχνητά νευρωνικά δίκτυα. Η παλινδρόμηση μπορεί να εφαρμοστεί σε διάφορους τομείς, όπως στη μετεωρολογία για να προβλεφθούν οι ταχύτητες ανέμου με βάση τη θερμοκρασία, την πίεση αέρα, και την υγρασία [36][37][38][39][40][41].

Πρόβλεψη (forecasting): Η πρόβλεψη είναι μια ακόμα σημαντική μέθοδος εξόρυξης δεδομένων. Μπορεί να βοηθήσει στην απάντηση ερωτημάτων όπως ποια θα είναι η αξία ενός αποθεματικού αύριο; Ποιό θα είναι το ποσοστό πωλήσεων αναψυκτικών για τον επόμενο μήνα; Τα δεδομένα εισόδου είναι τύπου χρονικής σειράς. Οι μέθοδοι πρόβλεψης εξετάζουν γενικές τάσεις και περιοδικότητα. Ως δημοφιλέστερη μέθοδος πρόβλεψης θεωρείται η ARIMA, η οποία υλοποιεί τη μεθοδολογία Auto Regressive Integrated Moving Average Model [36][37][38][39][40][41].

Ανάλυση Ακολουθίας (sequence analysis): Η ανάλυση ακολουθίας χρησιμοποιείται για τη δημιουργία μοντέλων σε μια διακριτή σειρά. Μια ακολουθία αποτελείται από μια σειρά διακριτών τιμών (ή καταστάσεων). Μια ακολουθία DNA είναι μια μακρά ακολουθία από τέσσερα διαφορετικά μέρη: Αδενίνη (Adenine), Θυμίνη (Thymine), Κυτοσίνη (Cytosine) και Γουανίνη (Guanine). Μια επιλογή στον Παγκόσμιο Ιστό είναι μια ακολουθία από ιστοσελίδες. Οι αγορές πελατών μπορούν επίσης να διαμορφωθούν ως στοιχεία ακολουθίας. Χαρακτηριστικό παράδειγμα αποτελεί ένας πελάτης που αγοράζει αρχικά έναν υπολογιστή, έπειτα ένα μικρόφωνο και τελικά μια Web κάμερα. Το κοινό των μεθόδων ανάλυσης ακολουθίας και ένωσης (association) είναι ότι κάθε μεμονωμένη περίπτωση περιέχει ένα σύνολο στοιχείων ή καταστάσεων. Η διαφορά τους είναι ότι οι μέθοδοι ανάλυσης ακολουθίας αναλύουν τις μεταβάσεις καταστάσεων ενώ η μέθοδος ένωσης θεωρεί κάθε στοιχείο ίσο και ανεξάρτητο. Σύμφωνα με τη μέθοδο ανάλυσης ακολουθίας, το να αγοράσει κάποιος έναν υπολογιστή προτού αγοράσει μικρόφωνο είναι μια διαφορετική ακολουθία από το να αγοράσει μικρόφωνο πριν από έναν υπολογιστή. Για έναν αλγόριθμο ένωσης, αυτά θεωρούνται όμοια. Η ανάλυση ακολουθίας είναι ένας σχετικά νέος τρόπος εξόρυξης δεδομένων. Είναι αρκετά σημαντική σε δύο κυρίως τύπους εφαρμογών:

Ανάλυση Παγκόσμιου Ιστού και ανάλυση DNA. Υπάρχουν διάφορες τεχνικές ανάλυσης ακολουθίας διαθέσιμες όπως οι αλυσίδες Markov και άλλες [36][37][38][39][40][41].

3.4 Αλγόριθμοι και τεχνικές εξόρυξης δεδομένων

Επιλογή του αλγόριθμου

Η επιλογή του καλύτερου αλγόριθμου που θα χρησιμοποιηθεί για ένα συγκεκριμένο αναλυτικό έργο μπορεί να είναι μια πρόκληση για έναν αναλυτή. Ενώ μπορεί να χρησιμοποιήσει διαφορετικούς αλγορίθμους για να εκτελέσει την ίδια εργασία, κάθε αλγόριθμος παράγει ένα διαφορετικό αποτέλεσμα, και μερικοί αλγόριθμοι μπορούν να παράγουν περισσότερα από ένα είδος αποτελέσματος. Για παράδειγμα, μπορεί να χρησιμοποιήσετε τα δέντρα απόφασης όχι μόνο για την πρόβλεψη, αλλά και ως έναν τρόπο για να μειωθεί ο αριθμός των στηλών σε ένα σύνολο δεδομένων, επειδή το δέντρο απόφασης μπορεί να εντοπίσει στήλες που δεν επηρεάζουν το τελικό μοντέλο εξόρυξης .[43]

1.Επιλέγοντας έναν αλγόριθμο με βάση τον τύπο

Οι Υπηρεσίες ανάλυσης περιλαμβάνουν τους ακόλουθους τύπους αλγόριθμων

- Αλγόριθμοι ταξινόμησης προβλέπουν μία ή περισσότερες διακριτές μεταβλητές , με βάση τις άλλες ιδιότητες στο σύνολο δεδομένων.

- Αλγόριθμοι παλινδρόμησης προβλέπουν μία ή περισσότερες συνεχείς μεταβλητές, όπως είναι το κέρδος ή ζημία, με βάση άλλα χαρακτηριστικά στο σύνολο δεδομένων.

- Αλγόριθμοι τμηματοποίησης χωρίζουν τα δεδομένα σε ομάδες, ή ομάδες , από τα στοιχεία που έχουν παρόμοιες ιδιότητες.

- Σύνδεσμος αλγορίθμων βρίσκουν συσχετισμούς μεταξύ διαφορετικών χαρακτηριστικών σε ένα σύνολο δεδομένων. Η πιο κοινή εφαρμογή αυτού του είδους είναι ο αλγόριθμος για τη δημιουργία κανόνων συσχέτισης, ο οποίος μπορεί να χρησιμοποιηθεί σε μια ανάλυση καλαθιού αγοράς.

- Αλγόριθμοι ανάλυσης αλληλουχίας συνοψίζουν συχνές αλληλουχίες ή κομμάτια σε δεδομένα, όπως μια διαδρομή ροής Web.

2.Επιλέγοντας έναν αλγόριθμο με βάση την εργασία

Ο παρακάτω πίνακας παρέχει προτάσεις για τους τύπους των εργασιών για τα οποία κάθε αλγόριθμος χρησιμοποιείται παραδοσιακά.[43]

Εργασίες και αλγόριθμοι

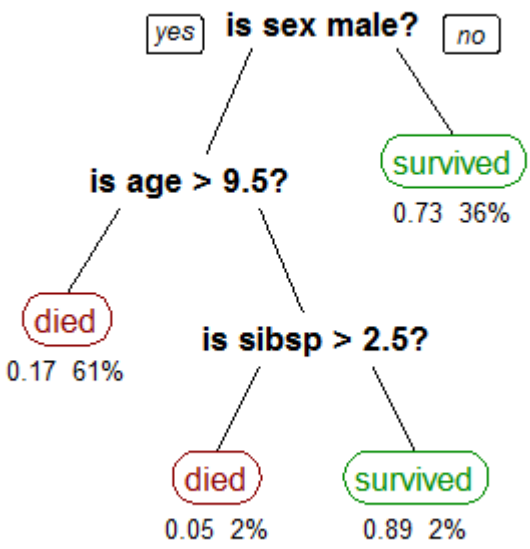
Παραδείγματα εργασιών	Αλγόριθμος που προτείνεται για χρήση
Προβλέποντας ένα διακριτό χαρακτηριστικό • Επισήμανε τους πελάτες σε μια προοπτική λίστα αγοραστών ως καλούς ή κακούς με βάση τις προοπτικές. • Υπολογίστε την πιθανότητα ότι ένας διακομιστής θα αποτύχει μέσα στους επόμενους 6 μήνες. • Κατηγοριοποίηση των αποτελεσμάτων των ασθενών και να εξερευνήσετε σχετικών παραγόντων.	Αλγόριθμος Δένδρα αποφάσεων Αλγόριθμος Naive Bayes Αλγόριθμος νευρωνικού δικτύου Αλγόριθμος ομαδοποίησης
Πρόβλεψη ενός συνεχούς χαρακτηριστικού • Πρόγνωση πωλήσεων του επόμενου έτους. • Προβλέψτε επισκέπτες της ιστοσελίδας λόγω των παλαιότερων ιστορικών και εποχιακές τάσεις. • Δημιουργήστε ένα σκορ κινδύνου, λαμβάνοντας υπόψη τα δημογραφικά στοιχεία.	Αλγόριθμος Δένδρα αποφάσεων Αλγόριθμος Χρονικής Σειράς Αλγόριθμος γραμμικής παλινδρόμησης
Προβλέποντας μια ακολουθία • Εκτελέστε clickstream ανάλυση του Web site της εταιρείας. • Αναλύστε τους παράγοντες που οδηγούν στην αποτυχία του διακομιστή. • Σύλληψη και ανάλυση ακολουθιών των δραστηριοτήτων κατά τη διάρκεια επισκέψεων στα εξωτερικά ιατρεία, για τη διαμόρφωση βέλτιστων πρακτικών γύρω από κοινές δραστηριότητες.	Αλγόριθμος ακολουθίας ομαδοποίησης
Η εύρεση ομάδων με κοινά στοιχεία στις συναλλαγές	Αλγόριθμος Συσχέτισης

• Χρησιμοποιήστε ανάλυση καλαθιού αγορών για να καθοριστεί η τοποθέτηση προϊόντων. • Προτείνετε επιπλέον προϊόντα σε έναν πελάτη για την αγορά. • Αναλύστε τα στοιχεία της έρευνας από τους επισκέπτες σε ένα γεγονός, με σκοπό την εύρεση δραστηριοτήτων που συσχετίστηκαν, με σκοπό την σχεδίαση μελλοντικών δραστηριοτήτων.	Αλγόριθμος Δένδρα αποφάσεων
Εύρεση ομάδες των παρόμοιων ειδών • Δημιουργία προφίλ κινδύνου για ομάδες ασθενών που βασίζονται σε χαρακτηριστικά όπως η δημογραφία και οι συμπεριφορές τους. • Αναλύστε τους χρήστες από την περιήγηση και τα αγοραστικά πρότυπα. • Προσδιορίστε servers που έχουν παρόμοια χρήση χαρακτηριστικών.	Αλγόριθμος ακολουθίας ομαδοποίησης Αλγόριθμος ομαδοποίησης

Με βάση τον παραπάνω πίνακα έγινε η επιλογή του αλγορίθμου δένδρα αποφάσεων (decision trees) και του αλγορίθμου ομαδοποίησης (clustering).

3.5 Τεχνική : δένδρα αποφάσεων της εξόρυξης δεδομένων

Η τεχνική εκμάθησης «δένδρα αποφάσεων» είναι μια μέθοδος, που χρησιμοποιείται ευρέως στην εξόρυξη δεδομένων. [44] Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που προβλέπει την τιμή μιας μεταβλητής στόχου που βασίζεται σε διάφορες μεταβλητές εισόδου. Ένα παράδειγμα φαίνεται στην παρακάτω εικόνα. Κάθε εσωτερικός κόμβος αντιστοιχεί σε μία από τις μεταβλητές εισόδου. Υπάρχουν ακμές στους κόμβους παιδιά για κάθε μία από τις πιθανές τιμές αυτής της μεταβλητής εισόδου. Κάθε φύλλο αντιπροσωπεύει μια τιμή της μεταβλητής στόχου έχοντας τις τιμές των μεταβλητών εισόδου που αντιπροσωπεύονται από το μονοπάτι της ρίζας στο φύλλο.



Ένα δέντρο που δείχνει το ποσοστό επιβίωσης των επιβατών του Τιτανικού («sibsp" είναι ο αριθμός των συζύγων ή παιδιών στο πλοίο). Οι αριθμοί κάτω από τα φύλλα δείχνουν την πιθανότητα επιβίωσης και το ποσοστό των παρατηρήσεων στο φύλλο.

Παράδειγμα Decision Tree

Το δέντρο απόφασης αποτελείται από κόμβους που διαμορφώνουν ένα δέντρο με ρίζα, το οποίο σημαίνει ότι είναι ένα κατευθυνόμενο δέντρο με έναν κόμβο αποκαλούμενο "ρίζα" ο οποίος δεν έχει καμία εισερχόμενη άκρη. Όλοι οι άλλοι κόμβοι έχουν ακριβώς μια εισερχόμενη άκρη. Ένας κόμβος με εξερχόμενες άκρες ονομάζεται εσωτερικός ή κόμβος δοκιμής. Όλοι οι άλλοι κόμβοι ονομάζονται φύλλα (επίσης γνωστοί ως κόμβοι τερματικοί ή απόφασης). Σε ένα δέντρο απόφασης, κάθε εσωτερικός κόμβος χωρίζει το σύνολο των περιπτώσεων σε δύο ή περισσότερα υποσύνολα σύμφωνα με μια συγκεκριμένη συνάρτηση διαχωρισμού των τιμών εισόδου των χαρακτηριστικών παραμέτρων. Στην απλούστερη και συχνότερη περίπτωση, με κάθε δοκιμή εξετάζεται μια χαρακτηριστική παράμετρος, έτσι ώστε το σύνολο των δεδομένων να διαχωριστεί σύμφωνα με την τιμή αυτής της χαρακτηριστικής παραμέτρου. Η αναδρομή ολοκληρώνεται, όταν το υποσύνολο σε ένα κόμβο έχει την ίδια αξία (τις ίδιες τιμές) της μεταβλητής -στόχου, ή όταν ο διαχωρισμός δεν προσθέτει τιμές στις προβλέψεις. Αυτή η διαδικασία της top-down επαγωγής δέντρων απόφασης (TDIDT) [45] είναι ένα παράδειγμα ενός άπληστου αλγόριθμου, και είναι μακράν η πιο κοινή στρατηγική για την εκμάθηση δέντρων απόφασης από τα δεδομένα, αλλά δεν είναι η μόνη στρατηγική. Στην πραγματικότητα, κάποιες προσεγγίσεις έχουν αναπτυχθεί πρόσφατα, επιτρέποντας επαγωγή δένδρου από bottom-up διαδικασία. [46]
Στην εξόρυξη δεδομένων, η τεχνική «δέντρα απόφασης» μπορεί να περιγραφεί επίσης ως ο συνδυασμός των μαθηματικών και υπολογιστικών μεθόδων για να βοηθήσουν την περιγραφή, την κατηγοριοποίηση και τη γενίκευση ενός συγκεκριμένου συνόλου δεδομένων.

Η μορφή των δεδομένων συνήθως έχει την παρακάτω μορφή:

$$(x, Y) = (x_1, x_2, x_3, x_4, \dots\dots, x_k, Y)$$

Η εξαρτημένη μεταβλητή, Y, είναι η μεταβλητή-στόχος που προσπαθούμε να κατανοήσουμε, να ταξινομήσουμε ή να γενικεύσουμε. Ο φορέας x αποτελείται από τις μεταβλητές εισόδου, x1, x2, x3, κλπ., που χρησιμοποιούνται για την διεργασία και εκροή των αποτελεσμάτων.

Υπάρχουν δύο τύποι για την τεχνική «δένδρα αποφάσεων» που χρησιμοποιούνται για εξόρυξη δεδομένων:

- Ταξινομητική ανάλυση δέντρου είναι όταν η προβλεπόμενη έκβαση είναι η τάξη στην οποία ανήκουν τα δεδομένα.
- Ανάλυση παλινδρόμησης δέντρου υπάρχει, όταν ως προβλεπόμενη έκβαση μπορούμε να θεωρήσουμε έναν πραγματικό αριθμό (π.χ. η τιμή ενός σπιτιού, ή η διάρκεια παραμονής του ασθενούς σε νοσοκομείο).

Ο όρος ταξινόμησης και παλινδρόμησης Δένδρου ανάλυσης (CART: Classification And Regression Tree) είναι ένας γενικός όρος που χρησιμοποιείται για να αναφερθεί σε δύο από τις παραπάνω διαδικασίες, για πρώτη φορά από τον Breiman et al [47]. Τα δένδρα, που χρησιμοποιούνται για την παλινδρόμηση και τα δέντρα, που χρησιμοποιούνται για την ταξινόμηση παρουσιάζουν κάποιες ομοιότητες αλλά επίσης κάποιες διαφορές, όπως η διαδικασία, που χρησιμοποιείται για να προσδιοριστεί πού πρέπει να γίνει ο διαχωρισμός .[47]

Μερικές τεχνικές, που συχνά αποκαλούνται μέθοδοι συνόλου, δημιουργούν περισσότερα από ένα δέντρο απόφασης :

- **Ενσάκιση** (Bagging) των δέντρων αποφάσεων, μια πρώιμη μέθοδος συνόλου, δημιουργεί πολλαπλά δέντρα απόφασης επαναλαμβάνοντας το στάδιο της δειγματοληψίας δεδομένων εκπαίδευσης με την αντικατάσταση, και στο τέλος η μέθοδος ψηφίζει τα παραχθέντα δέντρα για μια πρόβλεψη συναίνεσης . [48]
- Ένας **τυχαίος ταξινομητής Forest** χρησιμοποιεί μια σειρά από δέντρα απόφασης , προκειμένου να βελτιώσει το ποσοστό κατάταξης .
- **Ενισχυμένα δέντρα** μπορούν να χρησιμοποιηθούν για την παλινδρόμηση και ταξινόμηση τύπου προβλήματα. [49] [50]
- **Εναλλαγή Forest** - στην οποία κάθε δέντρο απόφασης έχει εκπαιδευτεί από την πρώτη εφαρμογή Ανάλυση Κύριων Συνιστωσών (PCA: principal component analysis) σε ένα τυχαίο υποσύνολο των χαρακτηριστικών εισόδου [51] .

Η εκμάθηση δέντρων απόφασης είναι η κατασκευή ενός δέντρου απόφασης από την επισημανθείσα κατηγορία των πλειάδων εκπαίδευσης. Ένα δέντρο απόφασης είναι ένα διάγραμμα ροής - δομής, όπου κάθε εσωτερικός (μη φύλλο) κόμβος υποδηλώνει μια δοκιμή σε ένα χαρακτηριστικό, κάθε κλάδος αντιπροσωπεύει το αποτέλεσμα μιας δοκιμής, και κάθε φύλλο (ή τερματικό) κόμβος κατέχει μια κατηγορία ετικέτα. Ο κορυφαίος κόμβος σε ένα δέντρο είναι ο κόμβος ρίζα.

Υπάρχουν πολλοί αλγόριθμοι δένδρα αποφάσεων. Οι πιο αξιοσημείωτοι περιλαμβάνουν :

- ID3 (Iterative Dichotomiser 3)
- C4.5 (successor of ID3)
- CART (Classification And Regression Tree)
- CHAID (CHi-squared Automatic Interaction Detector). Εκτελεί πολυεπίπεδη χωρίζει κατά τον υπολογισμό δέντρα ταξινόμησης . [52]
- MARS : επεκτείνει τα δέντρα απόφασης για να χειριστεί καλύτερα τα αριθμητικά δεδομένα .

Οι ID3 και CART εφευρέθηκαν ανεξάρτητα, περίπου, όμως, την ίδια χρονική περίοδο (1970-1980), αλλά παρόλα αυτά ακολουθούν μια παρόμοια προσέγγιση για την εκμάθηση δέντρων απόφασης από πλειάδες εκπαίδευσης .

Η φόρμουλα της τεχνικής δένδρα αποφάσεων του αλγορίθμου που χρησιμοποιήθηκε στα πλαίσια αυτής της διπλωματικής

3.6 Ο αλγόριθμος Decision trees της Microsoft

Ο αλγόριθμος της Microsoft «Δένδρα αποφάσεων» είναι ένας υβριδικός αλγόριθμος που ενσωματώνει διαφορετικές μεθόδους για τη δημιουργία ενός δέντρου, και υποστηρίζει πολλαπλά αναλυτικά καθήκοντα, συμπεριλαμβανομένης της παλινδρόμησης, ταξινόμησης, και του συνεταιρίζεσθαι. Ο αλγόριθμος της Microsoft «Δένδρα αποφάσεων» υποστηρίζει την μοντελοποίηση τόσο των διακριτών όσο και των συνεχών χαρακτηριστικών.

Στην συνέχεια θα εξηγηθεί η ενσωμάτωση του αλγορίθμου, καθώς και η περιγραφή της προσαρμογής της συμπεριφοράς του αλγορίθμου σε διάφορες εργασίες.

Ενσωμάτωση του αλγορίθμου δένδρα απόφασης

Ο αλγόριθμος της Microsoft Δένδρα αποφάσεων εφαρμόζει την Bayesian προσέγγιση στη μάθηση μοντέλων αλληλεπίδρασης με την προσέγγιση προηγούμενων κατανομών για τα μοντέλα .[54]

Η μεθοδολογία για την εκτίμηση της πληροφοριακής αξίας των προηγουμένων που απαιτούνται για την κατανόηση βασίζεται στην παραδοχή της πιθανότητας ισοδυναμίας. Αυτή η παραδοχή υποδηλώνει, ότι τα δεδομένα δεν θα πρέπει να βοηθήσουν στο να διακριθούν δομές δικτύου, που αντιπροσωπεύουν διαφορετικά τους ίδιους ισχυρισμούς της υπό όρους ανεξαρτησίας. Κάθε περίπτωση θεωρείται ότι έχει εκ των προτέρων ένα δίκτυο Bayesian και ένα ενιαίο μέτρο εμπιστοσύνης για το εν λόγω δίκτυο.

Χρησιμοποιώντας αυτά τα προηγούμενα δίκτυα, ο αλγόριθμος υπολογίζει στη συνέχεια τις σχετικές, *posterior*, πιθανότητες των δομών του δικτύου για τα εκάστοτε δεδομένα εκπαίδευσης, και προσδιορίζει τις δομές του δικτύου που έχουν την υψηλότερη *posterior* πιθανότητα.[55]

Ο αλγόριθμος της Microsoft Δένδρα αποφάσεων χρησιμοποιεί διαφορετικές μεθόδους για να υπολογίσει το καλύτερο δέντρο. Η μέθοδος, που χρησιμοποιείται εξαρτάται από την εργασία, η οποία μπορεί να είναι γραμμική παλινδρόμηση, ταξινόμηση, ή ένωση ανάλυσης. Ένα απλό μοντέλο μπορεί να περιέχει πολλαπλά δένδρα για διάφορα αναμενόμενα χαρακτηριστικά. Επιπλέον, κάθε δέντρο μπορεί να περιέχει πολλαπλούς κλάδους, ανάλογα με το πόσα πολλά χαρακτηριστικά και τιμές υπάρχουν στα δεδομένα. Το σχήμα και το βάθος του δέντρου χτισμένο σε ένα συγκεκριμένο μοντέλο εξαρτάται από τη μέθοδο βαθμολόγησης και άλλες παραμέτρους που χρησιμοποιούνται. Αλλαγές στις παραμέτρους μπορούν επίσης να επηρεάσουν τον διαχωρισμό των κόμβων.[55]

Χτίζοντας το Δέντρο

Όταν ο αλγόριθμος της Microsoft «Δένδρα αποφάσεων» δημιουργεί το σύνολο των πιθανών δυνατών τιμών εισόδου, εκτελεί την επιλογή χαρακτηριστικών, για τον προσδιορισμό των χαρακτηριστικών και τιμών που παρέχουν τις περισσότερες πληροφορίες, και αφαιρεί από την εξέταση τις τιμές που είναι πολύ σπάνιες. Ο αλγόριθμος ομαδοποιεί τις αξίες σε κάδους, για να δημιουργήσει ομάδες τιμών που μπορούν να υποβληθούν σε επεξεργασία ως μονάδα για τη βελτιστοποίηση της απόδοσης .[55]

Ένα δέντρο είναι χτισμένο με τον καθορισμό των συσχετισμών μεταξύ μιας εισόδου και του επιδιωκομένου αποτελέσματος. Αφού όλα τα χαρακτηριστικά έχουν συσχετισθεί, ο αλγόριθμος προσδιορίζει το μοναδικό χαρακτηριστικό, που διαχωρίζει καθαρότερα τα αποτελέσματα. Αυτό το σημείο του καλύτερου διαχωρισμού μετράται χρησιμοποιώντας μια εξίσωση που υπολογίζει την αξία των πληροφοριών. Το χαρακτηριστικό που έχει την καλύτερη βαθμολογία για την αξία των πληροφοριών χρησιμοποιείται για να διαιρέσει τις περιπτώσεις σε υποσύνολα, τα οποία στην συνέχεια αναλύονται αναδρομικά με την ίδια διαδικασία, μέχρις ότου το δέντρο δεν μπορεί να χωριστεί πια.[55]

Η ακριβής εξίσωση που χρησιμοποιείται για την αξιολόγηση οφέλους (αξίας) των πληροφοριών εξαρτάται από τις παραμέτρους που ορίζει ο χρήστης-αναλυτής, όταν δημιουργεί τον αλγόριθμο, από τον τύπο δεδομένων της στήλης που θα προβλεφθεί, και από τον τύπο δεδομένων της εισόδου .[55]

Διακριτές και Συνεχείς είσοδοι

Όταν το προβλέψιμο χαρακτηριστικό είναι διακριτό και οι είσοδοι είναι διακριτές, για να μετρηθεί το αποτέλεσμα ανά είσοδο, δημιουργείται μια μήτρα και βαθμολογίες για κάθε κελί της μήτρας.

Ωστόσο, όταν το προβλέψιμο χαρακτηριστικό είναι διακριτικό και οι είσοδοι είναι συνεχείς, η είσοδος των συνεχών στηλών αυτόματα διαχωρίζονται. Οι υπηρεσίες ανάλυσης μπορούν στην συνέχεια να βρουν τον βέλτιστο αριθμό των κάδων ή μπορεί ο αναλυτής που χρησιμοποιεί τον αλγόριθμο να ελέγξει τον τρόπο με τον οποίο οι συνεχείς εισροές διαχωρίζονται με τον καθορισμό του Discretization Method και Discretization Bucket Count .[56][57]

Για τα συνεχή χαρακτηριστικά, ο αλγόριθμος χρησιμοποιεί γραμμική παλινδρόμηση για να προσδιορίσει που θα χωριστεί το δέντρο απόφασης.

Όταν το προβλέψιμο χαρακτηριστικό είναι μια συνεχής, αριθμητική μεταβλητή δεδομένων, η επιλογή χαρακτηριστικών εφαρμόζεται στις εξόδους, με σκοπό να μειώσει τον πιθανό αριθμό των αποτελεσμάτων και να βοηθήσει στην κατασκευή του μοντέλου ταχύτερα. Είναι στην ευχέρεια του χρήστη-αναλυτή να αλλάξει το όριο για την επιλογή χαρακτηριστικών και έτσι να αυξήσει ή να μειώσει τον αριθμό των πιθανών τιμών ορίζοντας την παράμετρο MAXIMUM_OUTPUT_ATTRIBUTES. [56][57]

3.7 Θετικά και περιορισμοί της τεχνικής δένδρα αποφάσεων

Πολλά πλεονεκτήματα της τεχνικής «δένδρα απόφασης», ως εργαλείο ταξινόμησης έχουν τονιστεί στη βιβλιογραφία:[53]

1. Τα δέντρα απόφασης είναι αυτό-εξηγούμενα και όταν συμπιεστούν είναι επίσης εύκολο να ακολουθήσει κάποιος την λογική τους. Με άλλα λόγια, εάν το δέντρο απόφασης έχει έναν λογικό αριθμό φύλλων, μπορεί να γίνει αντιληπτό από μη επαγγελματίες χρήστες. Επί πλέον τα δέντρα αποφάσεων μπορεί να μετατραπούν σε ένα σύνολο κανόνων. Έτσι, αυτή η αναπαράσταση μπορεί να θεωρηθεί, ως κατανοητή.
2. Τα δέντρα απόφασης, είναι δυνατόν, να χειρίζονται τόσο ονομαστικά όσο και αριθμητικά χαρακτηριστικά εισόδου.

3. Η αναπαράσταση των δένδρων απόφασης είναι αρκετά πλούσια για να αντιπροσωπεύει οποιαδήποτε διακριτού ταξινομητή αξία.

4. Τα δέντρα απόφασης είναι σε θέση να χειρίζονται σύνολα δεδομένων που μπορεί να έχουν λάθη.

5. Τα δέντρα απόφασης είναι σε θέση να χειρίζονται σύνολα δεδομένων που μπορεί να τους λείπουν τιμές.

6. Τα δέντρα απόφασης θεωρείται ότι είναι μια μη παραμετρική μέθοδος. Αυτό σημαίνει ότι τα δέντρα απόφασης δεν έχουν υποθέσεις αναφορικά με την κατανομή στο χώρο και τη δομή του ταξινομητή.

7. Απαιτείται για την τεχνική ελάχιστη προετοιμασία των δεδομένων. Άλλες τεχνικές απαιτούν συχνά κανονικοποίηση, δημιουργία ψευδό-μεταβλητών και αφαίρεση των κενών τιμών.

8. Επίσης είναι πιθανή η επικύρωση του μοντέλου με την χρήση στατιστικών δοκιμών. Αυτό καθιστά δυνατό να λογοδοτήσει κάποιος για την αξιοπιστία του μοντέλου.

9. Στιβαρό. Εκτελείται καλά ακόμα και αν οι υποθέσεις του κάπως παραβιάζονται από το πραγματικό μοντέλο από το οποίο δημιουργούνται τα δεδομένα του.

10.Εκτελείται καλά σε μεγάλα σύνολα δεδομένων. Μεγάλες ποσότητες δεδομένων μπορούν να αναλυθούν με τη χρήση τυποποιημένων υπολογιστικών πόρων σε εύλογο χρονικό διάστημα.

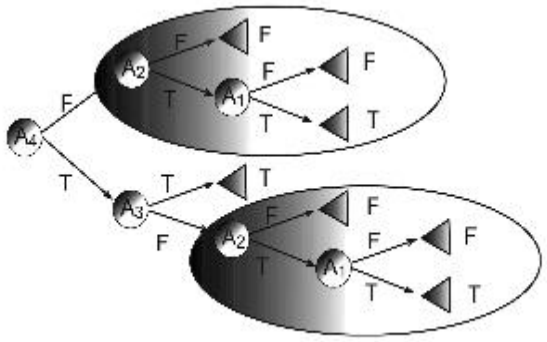
Από την άλλη πλευρά, η τεχνική «δένδρα αποφάσεων» έχει και κάποιους περιορισμούς- μειονεκτήματα που πρέπει να αναφερθούν:[53]

1.Οι περισσότεροι από τους αλγορίθμους (όπως ID3 και C4.5) απαιτούν ο στόχος στοιχείο να έχει μόνο διακριτές τιμές.

2. Καθώς τα δέντρα απόφασης χρησιμοποιούν την "διαίρει και βασίλευε" μέθοδο, τείνουν να αποδίδουν καλά, αν υπάρχουν μερικά πολύ σημαντικά χαρακτηριστικά, ενώ αποδίδουν σε μικρότερο βαθμό όταν διαφορετικές και πολύπλοκες αλληλεπιδράσεις είναι παρούσες. Ένας από τους λόγους, που συμβαίνει αυτό είναι ότι με άλλες τεχνικές μπορεί να περιγραφεί πλέον αποτελεσματικά ένας ταξινομητής, ενώ αυτό θα ήταν δυσκολότερο μέσω της τεχνικής «δένδρα αποφάσεων». Ένα απλό παράδειγμα αυτού είναι το πρόβλημα αναπαραγωγής των δέντρων απόφασης (Pagallo και Huassler, 1990). Δεδομένου ότι τα περισσότερα δέντρα απόφασης χωρίζουν το χώρο κατά περίπτωση και αλληλοαναιρούνται περιφέρειες που αντιπροσωπεύουν μια έννοια, σε ορισμένες περιπτώσεις, το δέντρο πρέπει να περιέχει αρκετές επαναλήψεις του ίδιου υπο-δέντρου, προκειμένου να αντιπροσωπεύεται αποτελεσματικά ο ταξινομητής. Για παράδειγμα, αν η έννοια ακολουθεί την ακόλουθη δυαδική συνάρτηση:

$$y = (A_1 \cap A_2) \cup (A_3 \cap A_4)$$

τότε το ελάχιστο μονομεταβλητό δέντρο απόφασης που αντιπροσωπεύει αυτή τη συνάρτηση απεικονίζεται στο παρακάτω σχήμα. Σημειώστε ότι το δέντρο περιέχει δύο αντίγραφα του ίδιου υπο-δέντρου.



απεικόνιση του δέντρου απόφασης με αναδιπλασιασμό.

3. Το χαρακτηριστικό greedy των δέντρων απόφασης οδηγεί σε ένα άλλο μειονέκτημα που θα πρέπει να επισημανθεί. Πρόκειται για την υπερευαισθησία του στο σύνολο εκπαίδευσης τόσο σε άσχετα χαρακτηριστικά όσο και στο θόρυβο (Quinlan, 1993).

3.8 Τεχνική ομαδοποίησης της εξόρυξης δεδομένων

Η ανάλυση συμπλέγματος ή ομαδοποίηση είναι η διαδικασία της ομαδοποίησης ενός συνόλου αντικειμένων κατά τέτοιο τρόπο ώστε τα αντικείμενα στην ίδια ομάδα (που ονομάζεται σύμπλεγμα/συστάδα) να είναι παρόμοια (κατά τον ένα ή τον άλλο τρόπο) μεταξύ τους και όχι με αυτά άλλων ομάδων (συμπλεγμάτων/ συστάδων). Είναι μια κύρια διαδικασία των διερευνητικών πεδίων της εξόρυξης δεδομένων και μια κοινή τεχνική για την στατιστική ανάλυση των δεδομένων, που χρησιμοποιείται σε πολλούς τομείς, συμπεριλαμβανομένης της μηχανικής μάθησης, αναγνώρισης προτύπων, ανάλυσης εικόνας, ανάκτησης πληροφοριών, και της βιο-πληροφορικής.

Η ανάλυση ομαδοποίησης ανιχνεύει συστάδες των αντικειμένων δεδομένων, που προσομοιάζουν μεταξύ τους. Τα μέλη της συστάδας παρουσιάζουν μεγαλύτερη ομοιότητα προς τα μέλη της συγκεκριμένης συστάδας από ό, τι προς αυτά άλλης. Ο στόχος της ανάλυσης ομαδοποίησης είναι να εντοπίσει υψηλής ποιότητας συστάδες, έτσι ώστε η ομοιότητες με άλλες συστάδες να είναι χαμηλή ενώ η ομοιότητα ένδον της συστάδας να είναι υψηλή. [68]

Η ομαδοποίηση, όπως η ταξινόμηση, χρησιμοποιείται για την κατάτμηση των δεδομένων. Σε αντίθεση με την ταξινόμηση, η ομαδοποίηση μοντέλων δεδομένων διαχωρίζει τα δεδομένα σε ομάδες που δεν είχαν οριστεί προηγουμένως. Μοντέλα ταξινόμησης δεδομένων διαχωρίζουν τα δεδομένα, κατατάσσοντας τα σε

προκαθορισμένες κατηγορίες, οι οποίες καθορίζονται από έναν στόχο. Μοντέλα ομαδοποίησης δεν χρησιμοποιούν στόχο.[68]

Η ομαδοποίηση είναι χρήσιμη για την διερεύνηση των δεδομένων. Εάν υπάρχουν πολλές περιπτώσεις και δεν δημιουργούνται προφανείς ομάδες, οι αλγόριθμοι ομαδοποίησης μπορεί να χρησιμοποιηθούν για να βρεθούν οι φυσικές ομαδοποιήσεις. Η ομαδοποίηση μπορεί επίσης να χρησιμεύσει ως ένα χρήσιμο βήμα προεπεξεργασίας δεδομένων για τον προσδιορισμό ομοιογενών ομάδων για την κατασκευή μοντέλων εποπτείας.[68]

Η ομαδοποίηση μπορεί επίσης να χρησιμοποιηθεί για την ανίχνευση ανωμαλίας. Μόλις τα δεδομένα έχουν υποδιαιρεθεί σε ομάδες, μπορείτε να διαπιστωθεί σε ορισμένες περιπτώσεις ότι δεν εντάσσονται επιτυχώς σε κάποια ομάδα (συστάδα) . Οι περιπτώσεις αυτές είναι οι ανωμαλίες ή ακραίες τιμές .[68]

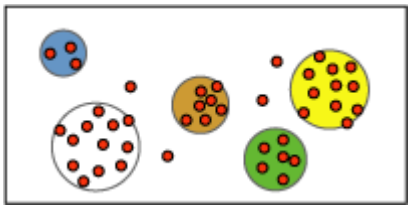
3.9 Ο αλγόριθμος Clustering της Microsoft

Ο αλγόριθμος Clustering της Microsoft είναι ένας αλγόριθμος κατάτμησης που παρέχεται από τις υπηρεσίες ανάλυσης. Ο αλγόριθμος χρησιμοποιεί επαναληπτικές τεχνικές για να ομαδοποιήσει τις περιπτώσεις σε ένα σύνολο δεδομένων, σε συστάδες, που εμφανίζουν παρόμοια χαρακτηριστικά. Αυτές οι ομαδοποιήσεις είναι χρήσιμες για την κατανόηση των δεδομένων, τον εντοπισμό ανωμαλιών στα δεδομένα , και, τη δημιουργία προβλέψεων.

Μοντέλα ομαδοποίησης αναγνωρίζουν τις σχέσεις σε ένα σύνολο δεδομένων που μπορεί να μην προκύπτουν λογικά μέσω μίας περιστασιακής παρατήρησης.

Πώς λειτουργεί ο αλγόριθμος

Ο αλγόριθμος Clustering Microsoft εντοπίζει πρώτα τις σχέσεις σε ένα σύνολο δεδομένων και παράγει μια σειρά από ομάδες που βασίζονται σε αυτές τις σχέσεις . Ένα διάγραμμα διασποράς είναι μία χρήσιμη μέθοδος για την οπτική αντιπροσώπευση του τρόπου με τον οποίο ο αλγόριθμος ομαδοποιεί τα δεδομένα, όπως φαίνεται στο ακόλουθο διάγραμμα . Το διάγραμμα διασποράς εμφανίζει όλες τις περιπτώσεις του σύνολο δεδομένων, και κάθε περίπτωση είναι ένα σημείο στο γράφημα. Οι συστάδες ομαδοποιούν τα σημεία στο γράφημα και απεικονίζουν τις σχέσεις που ο αλγόριθμος εντοπίζει .



Διάγραμμα διασποράς των περιπτώσεων σε ένα σύνολο δεδομένων

Μετά τον πρώτο καθορισμό των συστάδων, ο αλγόριθμος υπολογίζει πόσο καλά οι ομάδες αντιπροσωπεύουν ομάδες από τα σημεία, και, στη συνέχεια επαναπροσδιορίζει τις ομάδες για τη δημιουργία συνεργατικών σχηματισμών που αντιπροσωπεύουν καλύτερα τα δεδομένα. Ο αλγόριθμος επαναλαμβάνεται μέσα από αυτή τη διαδικασία έως ότου δεν μπορεί να βελτιώσει τα αποτελέσματα πάνω από τον επαναπροσδιορισμό των συστάδων.

Μπορεί φυσικά ο αναλυτής να προσαρμόσει τον τρόπο που ο αλγόριθμος λειτουργεί με την επιλογή συγκεκριμένης τεχνικής ομαδοποίησης, περιορίζοντας το μέγιστο αριθμό των συστάδων, ή την αλλαγή του ποσού της ενίσχυσης που απαιτείται για τη δημιουργία ενός συμπλέγματος.

Ενσωμάτωση του αλγορίθμου ομαδοποίησης[69]

Ο αλγόριθμος Clustering της Microsoft παρέχει δύο μεθόδους για τη δημιουργία συνεργατικών σχηματισμών και την ανάθεση σημεία δεδομένων για τις συστάδες. Η πρώτη , ο K - means αλγόριθμος , είναι μια σκληρή μέθοδο ομαδοποίησης . Αυτό σημαίνει ότι ένα σημείο δεδομένων μπορεί να ανήκει σε μία μόνο συστάδα, και ότι μια μόνο πιθανότητα υπολογίζεται για τη σύνθεση του κάθε σημείου δεδομένων στο εν λόγω σύμπλεγμα. Η δεύτερη μέθοδος , η μεγιστοποίηση Προσδοκία (EM) μέθοδος , είναι μια μαλακή μέθοδος ομαδοποίησης . Αυτό σημαίνει ότι ένα σημείο δεδομένων ανήκει πάντα σε πολλαπλές συστάδες, και ότι μια πιθανότητα υπολογίζεται για κάθε συνδυασμό των δεδομένων σημείου και συμπλέγματος .

Μπορεί ο αναλυτής να επιλέξει ποιόν αλγόριθμο θα χρησιμοποιήσει ορίζοντας την CLUSTERING_METHOD παράμετρο. Η προεπιλεγμένη μέθοδος για την ομαδοποίηση είναι επεκτάσιμη EM .

EM Clustering

Στην EM ομαδοποίηση, ο αλγόριθμος βελτιώνει επαναληπτικά ένα αρχικό μοντέλο διασποράς για να ταιριάξει τα δεδομένα και καθορίζει την πιθανότητα ότι ένα σημείο υπάρχει σε ένα σύμπλεγμα. Ο αλγόριθμος τερματίζει τη διαδικασία, όταν το μοντέλο πιθανοτήτων ταιριάξει τα δεδομένα. Η συνάρτηση που χρησιμοποιείται για τον προσδιορισμό του ταιριάσματος είναι η λογαριθμική πιθανότητα των δεδομένων που δίνονται στο μοντέλο.

Εάν κενά clusters δημιουργούνται κατά τη διάρκεια της διαδικασίας ή αν η συμμετοχή ενός ή περισσοτέρων από τα clusters πέφτει κάτω από ένα συγκεκριμένο όριο, οι συστάδες με μικρούς πληθυσμούς ανατροφοδοτούνται σε νέα σημεία και ο αλγόριθμος EM επαναλαμβάνεται.

Τα αποτελέσματα της μεθόδου ομαδοποίησης EM είναι πιθανολογικά. Αυτό σημαίνει ότι κάθε σημείο δεδομένων ανήκει μεν σε όλες τις ομάδες, αλλά τα στοιχεία (χαρακτηριστικά γνωρίσματα) κάθε σημείου εμφανίζει διαφορετική πιθανότητα

(ανάλογα με την ομάδα στην οποία εντάσσεται κάθε φορά). Επειδή η μέθοδος επιτρέπει διαφορετικές ομάδες/συστάδες να αλληλεπικαλύπτονται, το άθροισμα των στοιχείων σε όλες τις συστάδες μπορεί να υπερβαίνει το σύνολο των στοιχείων στο σύνολο εκπαίδευσης. Στα αποτελέσματα του μοντέλου εξόρυξης, τα αποτελέσματα που δείχνουν υποστήριξη προσαρμόζονται ώστε να καλύπτουν αυτήν την περίπτωση.

Ο αλγόριθμος EM είναι ο προεπιλεγμένος αλγόριθμος, που χρησιμοποιείται στα Microsoft μοντέλα ομαδοποίησης. Αυτός ο αλγόριθμος χρησιμοποιείται ως προεπιλογή επειδή προσφέρει πολλαπλά πλεονεκτήματα σε σύγκριση με την ομαδοποίηση k- μέσα:

- Απαιτεί μία σάρωση της βάσης δεδομένων κατά μέγιστο.
- Θα λειτουργήσει παρά την περιορισμένη μνήμη (RAM) .
- Έχει τη δυνατότητα να χρησιμοποιήσει ένα προς τα εμπρός μόνο δρομέα.
- Υπερτερεί στις προσεγγίσεις δειγματοληψίας .

Η εφαρμογή Microsoft παρέχει δύο επιλογές : επεκτάσιμη και μη επεκτάσιμη EM . Από προεπιλογή, στην κλιμακούμενη EM, τα πρώτα 50.000 αρχεία χρησιμοποιούνται για τη διασπορά της αρχικής σάρωσης. Αν αυτή είναι επιτυχής, το μοντέλο χρησιμοποιεί αυτά τα δεδομένα μόνο. Εάν το μοντέλο δεν μπορεί να ταιριάξει χρησιμοποιώντας 50.000 εγγραφές, διαβάζονται επιπλέον 50.000 εγγραφές. Σε μη - επεκτάσιμη EM, ολόκληρο το σύνολο δεδομένων διαβάζεται ανεξάρτητα από το μέγεθός του. Η μέθοδος αυτή θα μπορούσε να δημιουργήσει πιο ακριβή συμπλέγματα/συστάδες, αλλά οι απαιτήσεις μνήμης μπορεί να είναι μεγάλες. Επειδή η επεκτάσιμη EM λειτουργεί σε ένα τοπικό buffer, η επανάληψη μέσω των δεδομένων είναι πολύ πιο γρήγορη, και ο αλγόριθμος επιτυγχάνει πολύ καλύτερη χρήση της μνήμης cache της CPU από την μη επεκτάσιμη EM. Επιπλέον, η επεκτάσιμη EM είναι τρεις φορές πιο γρήγορη από ό, τι η μη - επεκτάσιμη EM, ακόμη και αν όλα τα στοιχεία μπορούν να χωρέσουν στην κύρια μνήμη. Στην πλειονότητα των περιπτώσεων, η βελτίωση των επιδόσεων δεν οδηγεί σε χαμηλότερη ποιότητα του πλήρους μοντέλου .

K -means Clustering

Η K -means ομαδοποίηση είναι μία ευρέως γνωστή μέθοδος της ένταξης των μελών ενός συμπλέγματος στο σύμπλεγμα, που επιτυγχάνεται με την ελαχιστοποίηση των διαφορών μεταξύ των μελών/αντικειμένων στο σύμπλεγμα αυτό και παράλληλα μεγιστοποιεί την απόσταση μεταξύ των ομάδων. Το "means" στο k -means αναφέρεται στο κεντροειδές της ομάδος, που είναι ένα σημείο δεδομένων, που έχει επιλεγεί αυθαίρετα και στη συνέχεια εξευγενίζεται επαναληπτικά μέχρι να αντιπροσωπεύει την πραγματική μέση τιμή όλων των μελών/αντικειμένων (σημείων δεδομένων) του συμπλέγματος. Το "K" αναφέρεται σε έναν αυθαίρετο αριθμό των σημείων που χρησιμοποιούνται για τη διασπορά της διαδικασίας ομαδοποίησης. Ο k - means αλγόριθμος υπολογίζει το τετράγωνο της Ευκλείδειας απόστασης μεταξύ των εγγραφών των δεδομένων σε ένα σύμπλεγμα και το φορέα που αντιπροσωπεύει τη

μέση τιμή συμπλέγματος, και συγκλίνει σε ένα τελικό σύνολο των k clusters, όταν το ποσό αυτό φθάνει στην ελάχιστη τιμή του.

Ο k - means αλγόριθμος αναθέτει (εντάσσει) κάθε σημείο δεδομένων ακριβώς σε μία συστάδα, και δεν επιτρέπει την αβεβαιότητα στην ένταξη. Η σύνθεση σε ένα σύμπλεγμα εκφράζεται ως μία απόσταση από το κεντροειδές .

Ο αλγόριθμος k-means παρουσιάζει το πλεονέκτημα ότι είναι απλός και αρκετά αποτελεσματικός αλλά εμφανίζει και αρκετά μειονεκτήματα:[70]

- Τα τελικά κέντρα των clusters δεν αντιπροσωπεύουν ένα ολικό ελάχιστο, αλλά μόνο ένα τοπικό.
- Τα αποτελέσματα μπορεί να διαφέρουν σημαντικά με βάση την αρχική επιλογή της διασποράς.
- Εντελώς διαφορετικές τελικές ομάδες μπορεί να προκύψουν από διαφορές που μπορεί να υπάρχουν από τα αρχικά τυχαίως επιλεγμένα cluster- κέντρα.
- Ο αλγόριθμος μπορεί εύκολα να αποτύχει στο να βρει μια λογική ομαδοποίηση.

3.10 Νέες τάσεις στην εξόρυξη δεδομένων

Ότι η εξόρυξη δεδομένων έχει αυξηθεί σε ωριμότητα, ως επιστήμη, μπορεί επίσης να επιβεβαιωθεί από το γεγονός, ότι οι περισσότεροι προμηθευτές βάσεων δεδομένων προσφέρουν ολοκληρωμένες λύσεις εξόρυξης δεδομένων. Τα εργαλεία αυτά διευκολύνουν τη διαδικασία δημιουργίας της γνώσης και συμβάλλουν στην περαιτέρω εξάπλωση της εξόρυξης δεδομένων. Σε ορισμένους τομείς, όπως η πρόληψη του εγκλήματος ή βιο-πληροφορική, η εξόρυξη δεδομένων βρίσκεται ακόμη σε πρώιμο στάδιο. Μία αόρατη αισιοδοξία χαρακτηρίζει τις περισσότερες από αυτές τις νέες εφαρμογές με τους γνωστούς περιορισμούς της εξόρυξης δεδομένων. Εκτός από τον πολλαπλασιασμό των πιθανών τομέων εφαρμογής, μπορεί επίσης να παρατηρηθεί ότι συνεχώς επιτυγχάνονται βελτιώσεις στις υπάρχουσες τεχνικές. Οι βελτιώσεις αυτές λαμβάνουν χώρα σε πολλούς διαφορετικούς τομείς: κατανόηση, προβλέψεις, ενσωμάτωση με υπάρχουσες βάσεις δεδομένων, σε πραγματικό χρόνο, αναλύσεις, κ.τ.λ. [58]

Α. Η καταπολέμηση της τρομοκρατίας: η βελόνα στα άχυρα[58]:

Με τα επακόλουθα των επιθέσεων της 11ης Σεπτεμβρίου, πολλές χώρες έχουν εγκρίνει νέους νόμους για την καταπολέμηση της τρομοκρατίας. Αυτοί οι νόμοι επιτρέπουν σε υπηρεσίες πληροφοριών να συγκεντρώσουν όλες τις πληροφορίες που κρίνονται απαραίτητες για την πρόληψη νέων επιθέσεων και να εντοπίσουν γρήγορα πιθανούς τρομοκράτες. Στον τομέα αυτό, οι Ηνωμένες Πολιτείες της Αμερικής

έπαιξαν πρωτοποριακό ρόλο με το «Total Information Awareness» πρόγραμμα τους. Ο στόχος αυτού του προγράμματος ήταν η δημιουργία μιας τεράστιας κεντρικής βάσης δεδομένων που συγκεντρώνει όλες τις διαθέσιμες πληροφορίες για τον πληθυσμό. Παρόμοια έργα είχαν ανακοινωθεί στην Ευρώπη και τον υπόλοιπο κόσμο. Αν και μερικά από αυτά τα προγράμματα ακυρώθηκαν λόγω της μαζικής αντίστασης των οργανώσεων της ιδιωτικής ζωής, τα περισσότερα από τα σχέδια αυτά, ωστόσο, φαίνεται ότι «αναστήθηκαν» αργότερα με ένα ελαφρώς διαφορετικό όνομα. Για παράδειγμα, το "Total Information Awareness " πρόγραμμα μετονομάστηκε σε «Terrorist Information Awareness».

Συνδυάζοντας τις πληροφορίες που προέρχονται από τον ιδιωτικό τομέα, όπως ο τραπεζικός τομέας και πληροφορίες αγοράς, σε συνδυασμό με τις πληροφορίες της κυβέρνησης παράγεται ένας πιθανός θησαυρός πληροφοριών, αλλά εκτός των ήδη αναφερθέντων προβλημάτων παραβίασης της ιδιωτικής ζωής, δημιουργούνται και πολλά άλλα εμπόδια. Ένα πρώτο πρόβλημα αφορά την ποικιλία και την ετερογένεια των δεδομένων. Εκτός από δομημένες πληροφορίες, η κεντρική βάση δεδομένων πρέπει να είναι σε θέση να ασχοληθεί με το κείμενο και τα αντικείμενα πολυμέσων. Αυτή η ποικιλομορφία των δεδομένων δημιουργεί ειδικά προβλήματα για τους περισσότερους αλγορίθμους εξόρυξης δεδομένων που συνήθως αναπτύσσονται για να αναγνωρίσουν πρότυπα σε δομημένα δεδομένα. Επιπλέον, η ποσότητα των δεδομένων περιορίζει σοβαρά την επεκτασιμότητα των αλγορίθμων. Ο χρόνος εκτέλεσης μπορεί να αυξηθεί μόνο σε ορισμένο βαθμό, όταν ο όγκος των διαθέσιμων δεδομένων αυξάνεται. Εφαρμογές πραγματικού χρόνου αποτελούν επίσης αυστηρά όρια σχετικά με τον επιτρεπόμενο χρόνο εκτέλεσης των αλγορίθμων. Για παράδειγμα, 230 κάμερες είχαν τοποθετηθεί στην πόλη του Λονδίνου για τον έλεγχο της κυκλοφορίας προς το κέντρο, (ανάγνωση αυτόματη των πινακίδων κυκλοφορίας των διερχόμενων οχημάτων). Με περίπου 40.000 οχήματα που περνούν από την κάμερα κάθε ώρα, το σύστημα πρέπει να αναγνωρίσει 10 οχήματα ανά δευτερόλεπτο. Ένα τεράστιο έργο, που θέτει υψηλές απαιτήσεις τόσο σε υλικό και λογισμικό (Transport for London (2004) .

Τέλος, πρέπει κανείς να λάβει υπόψη τα έξοδα που σχετίζονται με κάθε απόφαση. Συστήματα που παρέχουν τις πιο ακριβείς προβλέψεις θα είναι συχνά κατώτερα από λιγότερο ακριβή συστήματα, όταν λαμβάνεται υπόψη το κόστος ταξινόμησης. Για παράδειγμα, ένα σύστημα που είναι σε θέση να εντοπίσει όλους τους πιθανούς αεροπειρατές σε ένα αεροπλάνο, αλλά ταξινομεί εσφαλμένα κάποιους κανονικούς επιβάτες, ως τρομοκράτες, θα πρέπει να προτιμηθεί από ένα σύστημα που ταξινομεί περισσότερους από τους επιβάτες σωστά, αλλά ταξινομεί εσφαλμένα μερικούς από τους τρομοκράτες. Η αναστάτωση, που θα έχει να αντιμετωπίσει από κανονικούς επιβάτες, που εσφαλμένα ταξινόμησε ως τρομοκράτες, είναι μικρότερης σημασίας σε σύγκριση με τη ζημία από έναν τρομοκράτη που δεν εντοπίστηκε. Όμως, η επιλογή μεταξύ των διαφόρων συστημάτων δεν είναι εύκολη : για την ανάπτυξη βέλτιστων συστημάτων, πρέπει να είναι σε θέση να υπολογίσει όλα τα έξοδα και τα οφέλη και αυτό μπορεί να είναι ένα δύσκολο έργο. Για παράδειγμα, πώς μετράμε την

ταλαιπωρία για τους επιβάτες ; Εκτός από τις μετρήσιμες καθυστερήσεις κατά το check -in, πολλοί άλλοι παράγοντες παίζουν ρόλο : τα συναισθήματα του (αν) ασφάλειας, την αποτροπή πιθανών τρομοκρατών , κ.τ.λ.

Είναι σαφές ότι πολλά είναι τα δεδομένα και τα προβλήματα που πρέπει να ξεπεραστούν πριν η δυναμική της εξόρυξης δεδομένων μπορέσει να αξιοποιηθεί πλήρως σε αυτόν τον τομέα. Τα προβλήματα που είναι πιο εύκολο να αντιμετωπίσει κανείς είναι τα προβλήματα τεχνικής φύσεως. Η περαιτέρω έρευνα σε αυτόν τον ταχέως εξελισσόμενο τομέα θα παράσχει λύσεις στην πλειονότητα των σημερινών περιορισμών. Τα προβλήματα που είναι δύσκολο να ξεπεραστούν, όμως, είναι αυτά της κοινωνικής φύσης, διότι η προσωπική ελευθερία αγκαλιάζεται από τους περισσότερους ανθρώπους ως μια από τις πλέον θεμελιώδεις αρχές προστασίας της προσωπικότητας.

Βιο-πληροφορική: η αναζήτηση για θεραπείες

Ένας δεύτερος τομέας στον οποίο η εξόρυξη δεδομένων έχει έντονα αγκαλιαστεί είναι η βιο-πληροφορική. Βιο-πληροφορική είναι η επιστήμη που αφορά την διαχείριση, την εξόρυξη και την ερμηνεία των βιολογικών ακολουθιών και δομών. Έργα (projects) με γονιδιακές αλληλουχίας, που συνέβαλαν σε μια εκθετική αύξηση σε πλήρη ή μερική βάση δεδομένων ακολουθίας. Το Structural Genome Initiative έχει στόχο την καταγραφή της δομής-λειτουργίας πληροφοριών των πρωτεϊνών. Η πρόοδος στον τομέα των τεχνολογιών, όπως οι μικροσυστοιχίες, είχε ως αποτέλεσμα την δημιουργία των υποτομέων της γονιδιωματικής και πρωτεϊνωματικής. Αυτά τα πεδία αφορούν στην μελέτη των γονιδίων, των πρωτεϊνών και του κυκλώματος στο εσωτερικό του κυττάρου που ρυθμίζει την γονιδιακή έκφραση. Πολλά δεδομένα δημιουργούνται, δεδομένα που πρέπει να εξορυχτούν, αν η ανθρωπότητα θελήσει κάποτε να αντιληφθεί τα μυστήρια των κυττάρων. [58]

Κατά τα τελευταία χρόνια, τεράστια πρόοδος έχει επιτευχθεί, αλλά εξακολουθούν να υπάρχουν μια σειρά από θεμελιώδη προβλήματα στην βιο-πληροφορική, όπως η δυσκολία στην πρόβλεψη πρωτεϊνικών δομών και στην εξεύρεση γονιδίων. Η εξόρυξη δεδομένων θα διαδραματίσει θεμελιώδη ρόλο στην κατανόηση της γονιδιακής έκφρασης, στην ανάπτυξη φαρμάκων και επίλυση άλλων προβλημάτων στον τομέα της γονιδιωματικής και πρωτεϊνωματικής. Επιπλέον, η εξόρυξη κειμένου θα είναι σημαντική καθώς θα φιλτράρει τις γνώσεις από την αυξανόμενη προσφορά της βιβλιογραφίας σχετικά με τη βιο-πληροφορική. [58]

3.11 Εφαρμογή της εξόρυξης δεδομένων στην ιατρική

Ο χώρος της ιατρικής παράγει έναν συνεχώς αυξανόμενο όγκο δεδομένων και συνεπώς όλο και περισσότερη κρυμμένη πληροφορία υπάρχει σε αυτά. Είναι κοινά αποδεκτό από όλους τους φορείς υγείας ότι η αποτελεσματική διαχείριση των ιατρικών δεδομένων παίζει καθοριστικό ρόλο στην παροχή υγείας υψηλού επιπέδου [59]. Η ανάλυση των ιατρικών δεδομένων μπορεί να συνεισφέρει στη βελτίωση της αποτελεσματικότητας με ταυτόχρονη μείωση του χρόνου και του κόστους. Η αποκτώμενη γνώση με τις μεθόδους εξόρυξης δεδομένων μπορεί να αξιοποιηθεί από την ιατρική έρευνα, τόσο στο επίπεδο της διάγνωσης όσο και της θεραπείας [60].

Τα ιατρικά δεδομένα είναι από τη φύση τους ετερογενή, περιλαμβάνουν αποτελέσματα απεικονιστικά, γραφήματα, κείμενα από την κλινική εξέταση, εργαστηριακές μετρήσεις σε αριθμούς καθώς και δεδομένα σε άλλες μορφές. Έτσι, η εξόρυξη της κρυμμένης γνώσης από αυτά πρέπει να γίνει από συνδυασμό εικόνων, σχημάτων, κειμένων, αριθμών, το οποίο είναι δυσκολότερο από τις κλασσικές περιπτώσεις επεξεργασίας δεδομένων σε αριθμούς και κατηγορίες. Οι σύγχρονες τεχνολογίες εξόρυξης δεδομένων επιτρέπουν πλέον τη διαχείριση της ετερογενούς φύσης των ιατρικών δεδομένων. Πρώτιστα, η δυνατότητα επεξεργασίας της φυσικής γλώσσας και οι τεχνικές εξόρυξης δεδομένων από κείμενα επιτρέπουν την εξαγωγή πληροφορίας και γνώσης από τις ιατρικές σημειώσεις και τις κλινικές εξετάσεις [61],[62]. Ιατρική οντολογία και ορολογία μπορούν να ανιχνευθούν με τη χρήση μεθόδων εξόρυξης δεδομένων του παγκόσμιου ιστού και με τεχνικές εκμάθησης οντολογίας [62], [63].

Η έμφαση που δίνεται σήμερα σε ιατρικές πράξεις στηριζόμενες σε τεκμήρια (Evidence-based Medicine -EBM) είναι ένας από τους κύριους λόγους που ενισχύουν την εφαρμογή μεθόδων εξόρυξης δεδομένων στην ιατρική πρακτική [60]. Αυτές οι ιατρικές πράξεις συνήθως αποτελούν οδηγίες κλινικής πρακτικής ή κανόνες κλινικών αποφάσεων [64]. Η διασύνδεση της εξόρυξης δεδομένων με την ιατρική πρακτική υλοποιείται με την υπάρχουσα προσπάθεια ανάπτυξης αυτόματων συστημάτων υποστήριξης απόφασης. Η ύπαρξη συγκεκριμένης οντολογίας επιτρέπει να ξεκαθαριστεί ποιοί κανόνες αποφάσεων επιβεβαιώνουν την ιατρική γνώση [65]. Τα ιατρικά συστήματα υποστήριξης απόφασης υποστηρίζουν τους ιατρούς στη λήψη ιατρικών αποφάσεων [65].

Ένα πρόβλημα στο οποίο προσκρούει η εφαρμογή της εξόρυξης δεδομένων στην ιατρική είναι η έλλειψη στοιχειώδους οργάνωσης και αρχειοθέτησης των ιατρικών δεδομένων. Ο όγκος, η πολυπλοκότητα, η πολυσύνθετη δομή της ιατρικής πληροφορίας δυσχεραίνουν την προσπάθεια σε αυτό το επίπεδο. Επίσης, η απουσία

ενιαίου ιατρικού φακέλου, η έλλειψη και η ετερογένεια των πληροφοριακών συστημάτων και η έλλειψη διαλειτουργικότητας, η ανομοιομορφία στην κωδικοποίηση των δεδομένων είναι στοιχεία που πρέπει να ξεπεραστούν, ώστε να ανοίξει ο δρόμος προς την αποτελεσματική εφαρμογή των μεθόδων εξόρυξης δεδομένων στο χώρο της υγείας. Διάφορες σημαντικές προσπάθειες έχουν γίνει διεθνώς για να συνδέσουν τους κλινικούς και ερευνητικούς στόχους με τους διοικητικούς, όπως το πρόγραμμα I2B2 στο Χάρβαρντ [66], ή, σε μικρότερη κλίμακα, το Hemostat [64] και το σύστημα René στην Ιταλία [63]. Επιπλέον, διάφορες εμπορικές λύσεις για την κοινή διαχείριση των πληροφοριών, των στοιχείων, και της γνώσης είναι διαθέσιμες στην αγορά.

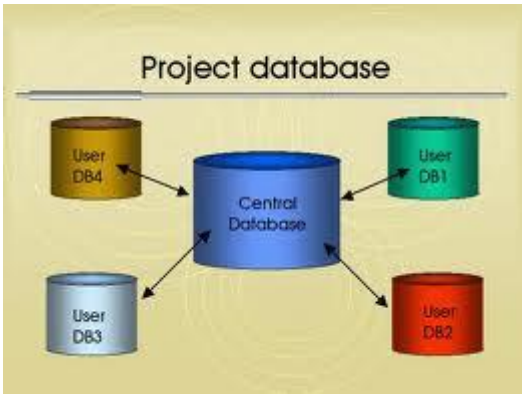
Τέλος, ένα ακόμα σημαντικό στοιχείο της ιδιαιτερότητας των ιατρικών δεδομένων αφορά το ιατρικό απόρρητο και την ασφαλή διαχείριση των δεδομένων. Τα δεδομένα συλλέγονται από ασθενείς και πρέπει να αποτραπεί κάθε πιθανότητα κακής χρήσης τους. Υπάρχει λοιπόν ηθική, νομική αλλά και κοινωνική υποχρέωση να προστατευθεί το ιατρικό απόρρητο και εγείρονται σοβαρά ζητήματα που σχετίζονται με αυτό, όπως το ζήτημα της ιδιοκτησίας των ιατρικών δεδομένων, η πιθανότητα νομικής εμπλοκής κατά τη χρήση τους καθώς και οι πιθανές συνέπειες από αυτήν [67].

Κεφάλαιο 4

4.1 Η σημασία της βάσης δεδομένων

Μια βάση δεδομένων είναι μια οργανωμένη συλλογή των δεδομένων. Τα δεδομένα οργανώνονται συνήθως για να μοντελοποιούν σχετικές πτυχές της πραγματικότητας με τρόπο που να υποστηρίζει τις διαδικασίες που χρειάζονται και χρησιμοποιούν αυτές τις πληροφορίες. Μια Βάση Δεδομένων είναι μια λογική συνεκτική συλλογή δεδομένων που έχει κάποια εγγενή σημασία. Μια τυχαία διευθέτηση δεδομένων για παράδειγμα δεν είναι μια Βάση Δεδομένων.

Μια Βάση Δεδομένων σχεδιάζεται, χτίζεται, και πληρούται με δεδομένα που εξυπηρετούν κάποιο συγκεκριμένο σκοπό. Προορίζεται για μια συγκεκριμένη ομάδα χρηστών και για κάποιες προκαθορισμένες εφαρμογές για τις οποίες οι χρήστες ενδιαφέρονται. Μια Βάση Δεδομένων, λοιπόν, έχει κάποια πηγή από την οποία παράγονται δεδομένα, αλληλεπιδρά σε κάποιο βαθμό με γεγονότα του πραγματικού κόσμου και απευθύνεται σε ένα ακροατήριο που ενδιαφέρεται ενεργά για το περιεχόμενό της.



Πολλοί χρήστες που έχουν πρόσβαση σε μια βάση δεδομένων

Οι παραδοσιακές βάσεις δεδομένων οργανώνονται από πεδία, εγγραφές και αρχεία. Ένα πεδίο είναι ένα μόνο κομμάτι των πληροφοριών. Μια εγγραφή είναι ένα πλήρες σύνολο πεδίων, και ένα αρχείο είναι μια συλλογή από εγγραφές. Για παράδειγμα, ένας τηλεφωνικός κατάλογος αναλογεί σε ένα αρχείο. Περιέχει μια λίστα με τις

εγγραφές, καθεμία από τις οποίες αποτελείται από τρεις τομείς: το όνομα, τη διεύθυνση και τον αριθμό τηλεφώνου.[13]

Μια εναλλακτική ιδέα στο σχεδιασμό βάσεων δεδομένων είναι το Hypertext. Σε μια βάση δεδομένων Hypertext, οποιοδήποτε αντικείμενο, είτε πρόκειται για ένα κομμάτι του κειμένου, μια εικόνα ή μια ταινία, μπορεί να είναι συνδεδεμένο με οποιοδήποτε άλλο αντικείμενο. Οι Hypertext βάσεις δεδομένων είναι ιδιαίτερα χρήσιμες για την οργάνωση μεγάλων ποσοτήτων διαφορετικών πληροφοριών, αλλά δεν έχουν σχεδιαστεί για αριθμητική ανάλυση. [13]

Για να υπάρχει πρόσβαση σε πληροφορίες από μια βάση δεδομένων, θα πρέπει να υπάρχει ένα σύστημα διαχείρισης βάσεων δεδομένων (DBMS). Αυτή είναι μια συλλογή από προγράμματα που δίνει τη δυνατότητα να εισαχθούν, να οργανωθούν και να επιλεγούν τα δεδομένα σε μια βάση δεδομένων.[13]

4.2 Σύστημα διαχείρισης βάσεων δεδομένων

Ένα **σύστημα διαχείρισης βάσεων δεδομένων** (database-management system - DBMS) είναι ένα σύνολο από δεδομένα που σχετίζονται και ένα σύνολο από προγράμματα για πρόσβαση σε αυτά τα δεδομένα. Η συλλογή των δεδομένων, που συνήθως αναφέρεται ως **βάση δεδομένων,** περιέχει πληροφορίες σχετικές με μια επιχείρηση. Ο βασικός στόχος ενός DBMS είναι να παρέχει ένα τρόπο να αποθηκεύονται και να ανακαλούνται οι πληροφορίες των βάσεων δεδομένων, που να είναι *βολικός* και *αποτελεσματικός*.**[71]**

Τα συστήματα βάσεων δεδομένων σχεδιάζονται για τον χειρισμό μεγάλων τμημάτων πληροφοριών. Η διαχείριση των δεδομένων περιλαμβάνει τόσο τον ορισμό των δομών για την αποθήκευση των πληροφοριών, όσο και την παροχή μηχανισμών για τον χειρισμό των πληροφοριών. Επιπλέον, τα συστήματα βάσεων δεδομένων πρέπει να διασφαλίζουν την ασφάλεια των πληροφοριών, που αποθηκεύονται, παρ' όλα τα παγώματα του συστήματος ή τις προσπάθειες μη πιστοποιημένης πρόσβασης. Αν τα δεδομένα είναι κοινόχρηστα μεταξύ διαφόρων χρηστών, το σύστημα θα πρέπει να αποφεύγει πιθανά λανθασμένα αποτελέσματα. **[71]**

Επειδή οι πληροφορίες είναι τόσο σημαντικές για τις περισσότερες εταιρείες, οι επιστήμονες της πληροφορικής έχουν αναπτύξει ένα μεγάλο σύνολο από ιδέες και τεχνικές για την διαχείριση των δεδομένων. **[71]**

Εφαρμογές Συστημάτων Βάσεων Δεδομένων [71]

Οι βάσεις δεδομένων χρησιμοποιούνται ευρέως. Εδώ αναφέρονται μερικές αντιπροσωπευτικές εφαρμογές:

Τράπεζες: Για πληροφορίες πελατών, λογαριασμών και δανείων και τραπεζικών συναλλαγών.

Αεροπορικές εταιρείες: Για κρατήσεις θέσεων και πληροφορίες πτήσεων. Οι αεροπορικές εταιρείες ήταν μεταξύ των πρώτων που χρησιμοποίησαν βάσεις δεδομένων με ένα γεωγραφικά κατανεμημένο τρόπο, δηλαδή με τερματικά, που βρίσκονταν σε όλο τον κόσμο, μπορούσαν να έχουν πρόσβαση στην κεντρική βάση δεδομένων μέσω τηλεφωνικών γραμμών και άλλων δικτύων δεδομένων.

Πανεπιστήμια: Για πληροφορίες φοιτητών, εγγραφές σε μαθήματα και βαθμούς.

Συναλλαγές πιστωτικών καρτών: Για αγορές μέσω πιστωτικών καρτών και δημιουργία μηνιαίων κινήσεων.

Τηλεπικοινωνίες: Για διατήρηση των κλήσεων, δημιουργία μηνιαίων λογαριασμών, διατήρηση του υπολοίπου για τις προπληρωμένες κάρτες κλήσης και αποθήκευση πληροφοριών για τα δίκτυα επικοινωνιών.

Χρηματοδοτήσεις: Για αποθήκευση πληροφοριών σχετικά με πωλήσεις και αγορές οικονομικών στοιχείων, όπως μετοχών και ομολόγων.

Πωλήσεις: Για πληροφορίες πελατών, προϊόντων και πωλήσεων.

Βιομηχανία: Για διαχείριση της αλυσίδας προμηθειών και την παρακολούθηση της παραγωγής των προϊόντων σε εργοστάσια, των προϊόντων σε μεγάλες αποθήκες και σε καταστήματα και των παραγγελιών των προϊόντων.

Ανθρώπινοι πόροι: Για πληροφορίες για εργαζόμενους, μισθούς, φόρους μισθοδοσίας και παροχές και για πληρωμές μισθών.

Όπως δείχνει η λίστα, οι βάσεις δεδομένων αποτελούν ένα απαραίτητο μέρος σχεδόν κάθε επιχείρησης σήμερα.

Τις τελευταίες τέσσερεις δεκαετίες του 20ου αιώνα, η χρήση των βάσεων δεδομένων αυξήθηκε σε όλες τις εταιρείες. Τον πρώτο καιρό, πολύ λίγοι άνθρωποι συνδιαλέγονταν κατευθείαν με συστήματα βάσεων δεδομένων, αν και χωρίς να το συνειδητοποιούν συνδιαλέγονταν με βάσεις δεδομένων έμμεσα, μέσω έντυπων αναφορών, όπως είναι η κίνηση των πιστωτικών καρτών, ή μέσω πρακτόρων, όπως μέσω ενός ταμία σε μια τράπεζα και ενός πράκτορα κράτησης μιας αεροπορικής εταιρείας. Μετά εμφανίστηκαν τα αυτοματοποιημένα μηχανήματα που επέτρεπαν στους χρήστες να συνδιαλέγονται κατευθείαν με βάσεις δεδομένων. Οι διασυνδέσεις τηλεφώνων με υπολογιστές (έμμεσα συστήματα απόκρισης με φωνή), επέτρεπαν στους χρήστες να επικοινωνούν κατευθείαν με βάσεις δεδομένων - αυτός που καλεί ένα αριθμό, μπορεί να πατήσει πλήκτρα *στο* τηλέφωνο για να εισάγει πληροφορίες ή για να επιλέξει εναλλακτικές επιλογές, ή για να μάθει την ώρα άφιξης και αναχώρησης, για παράδειγμα, ή για να εγγραφεί σε μαθήματα ενός πανεπιστημίου.

Η επανάσταση του Internet στα τέλη της δεκαετίας του 1990 αύξησε σαφώς την άμεση πρόσβαση του χρήστη στις βάσεις δεδομένων. Οι εταιρείες μετέτρεψαν πολλές από τις διασυνδέσεις τηλεφώνων με βάσεις δεδομένων, σε Web διασυνδέσεις και έκαναν διαθέσιμες πολλές υπηρεσίες και πληροφορίες online. Για παράδειγμα, όταν έχετε πρόσβαση σε ένα online βιβλιοπωλείο και κοιτάζετε μια συλλογή βιβλίων ή μουσικής, προσπελάζετε δεδομένα που είναι αποθηκευμένα σε μια βάση δεδομένων. Όταν δίνετε μια παραγγελία online, η παραγγελία σας αποθηκεύεται σε μια βάση δεδομένων. [71]

Η σημασία των συστημάτων βάσεων δεδομένων μπορεί να κριθεί με ένα άλλο τρόπο. Σήμερα, οι προμηθευτές συστημάτων βάσεων δεδομένων, όπως η Oracle, είναι μεταξύ των μεγαλύτερων εταιρειών λογισμικού *στον* κόσμο και τα συστήματα βάσεων δεδομένων αποτελούν ένα σημαντικό μέρος της γραμμής

προϊόντων πολλών διαφορετικών εταιρειών, όπως της Microsoft και της IBM. [71]

4.3 Οι Βάσεις Δεδομένων και τα υπολογιστικά συστήματα υποβοηθούμενης διάγνωσης

Τα τελευταία χρόνια, πολλές ερευνητικές ομάδες έχουν εστιάσει το ενδιαφέρον τους στην ανάπτυξη υπολογιστικών συστημάτων, τα οποία είναι σε θέση να αναλύουν διάφορους τύπους ιατρικών εικόνων και να εξάγουν χρήσιμες πληροφορίες για τους ιατρούς. Οι βάσεις δεδομένων αποτελούν το συνδετικό κρίκο μεταξύ ιατρικής και υπολογιστικών συστημάτων υποβοηθούμενης διάγνωσης. Οι ιατρικές εικόνες προκύπτουν από μια ποικιλία μεθόδων και μηχανημάτων και αποθηκεύονται σε βάσεις δεδομένων.
Έχουν χρησιμοποιηθεί ευρέως πολλές υπολογιστικές διατάξεις για την ανάλυση ιατρικών σημάτων που απεικονίζονται σε μία διάσταση, όπως το Ηλεκτροκαρδιογράφημα (ECG) και το Ηλεκτρομυογράφημα (EMG). Εντούτοις, η πλειοψηφία των ιατρικών σημάτων απεικονίζεται σε δύο διαστάσεις. Για την αυτοματοποιημένη ανίχνευση των χαρακτηριστικών των ανωμαλιών των εικόνων αυτών, έχουν σχεδιαστεί διάφορα υπολογιστικά συστήματα που έχουν την ικανότητα να εξασφαλίζουν στους ιατρούς χρήσιμα δεδομένα. Τα συστήματα αυτά καλούνται συνήθως υπολογιστικά συστήματα υποβοηθούμενης διάγνωσης ή υπολογιστικοί ταξινομητές (Computer-aided detection/diagnosis system - CAD).[4]

Ένα υπολογιστικό σύστημα υποβοηθούμενης διάγνωσης (CAD) ορίζεται ως ένας συνδυασμός τεχνικών επεξεργασίας εικόνας και ευφυών μεθόδων, που μπορούν να χρησιμοποιηθούν για την ενίσχυση της διακριτικής διαδικασίας, με αποτέλεσμα την αποδοτικότερη διακριτική ικανότητα. Το αποτέλεσμα, που δίνει αυτό το υπολογιστικό σύστημα, βοηθάει τον ακτινολόγο να αναλύσει καλύτερα την εικόνα και να κάνει πιο εύκολα τη διάγνωσή του. Επιπρόσθετα, το υπολογιστικό σύστημα αυτό, μπορεί να επιστήσει την προσοχή του ακτινολόγου στην περιοχή, όπου η πιθανότητα εμφάνισης της ασθένειας είναι μεγαλύτερη. Ένα υπολογιστικό σύστημα υποβοηθούμενης διάγνωσης παρέχει αναπαραγωγίσιμα και αρκετά ρεαλιστικά αποτελέσματα [14].

Οι περισσότερες από τις αυτοματοποιημένες προσεγγίσεις των υπολογιστικών συστημάτων υποβοηθούμενης διάγνωσης περιλαμβάνουν διαδικασίες εξαγωγής χαρακτηριστικών γνωρισμάτων. Παρόλα αυτά, έχουν περιγραφεί πολλές έρευνες ημι-αυτοματοποιημένων προσεγγίσεων, όπου οι ακτινολόγοι εκτελούν με το χέρι τις διαδικασίες εξαγωγής χαρακτηριστικών γνωρισμάτων με τη βοήθεια διάφορων μεθόδων [15]. Τα υπολογιστικά συστήματα υποβοηθούμενης διάγνωσης χωρίζονται σε δύο κατηγορίες, ανάλογα με το σκοπό τους: (α) σ’ αυτά που χρησιμοποιούνται για την αναγνώριση των περιοχών της παθολογίας και (β) σ’ αυτά που χρησιμοποιούνται για την ταξινόμηση των ευρημάτων ανάλογα με τα χαρακτηριστικά τους, τα οποία υποδηλώνουν την ιστολογία τους.[4]

Ο ρόλος αυτών των υπολογιστικών συστημάτων είναι να βελτιώνουν την ευαισθησία (sensitivity) και την ειδικότητα (specificity) των διαγνωστικών διαδικασιών και όχι για να παίρνουν αποφάσεις για την κατάσταση της υγείας του ασθενούς.[4]

Με την χρήση των βάσεων δεδομένων και την σωστή συντήρηση και ενημέρωσή τους καθώς και με την ορθολογική χρήση των υπολογιστικών συστημάτων υποβοηθούμενης διάγνωσης οι ακτινολόγοι έχουν στα χέρια τους ένα ισχυρό εργαλείο για να κάνουν μια σωστή και έγκυρη διάγνωση.

4.4 Περιγραφή των Μαστογραφικών Βάσεων Δεδομένων

Στα πλαίσια αυτής της διπλωματικής εργασίας επελέγησαν δυο βάσεις δεδομένων που περιείχαν μαστογραφικά δεδομένα απαραίτητα για να εκτελεστούν οι αλγόριθμοι εξόρυξης δεδομένων. Οι βάσεις είναι ελεύθερες στο διαδίκτυο.

Οι βάσεις δημιουργήθηκαν χρησιμοποιώντας το πρόγραμμα visual studio 2013 πρόγραμμα που παρέχεται δωρεάν στους φοιτητές της σχολής μας. Οι βάσεις είναι σε server της Microsoft.

1η Βάση Δεδομένων (http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass)

Αρχικά παραθέτονται κάποιες πληροφορίες που περιγράφουν τα στοιχεία της βάσης:

Η μαστογραφία είναι η πιο αποτελεσματική μέθοδος για την ανίχνευση του καρκίνου του μαστού μέχρι σήμερα. Ωστόσο, η χαμηλή θετική προγνωστική αξία του καρκίνου του μαστού από την βιοψία που προκύπτει από την ερμηνεία της μαστογραφία οδηγεί σε περίπου 70 % περιττές βιοψίες με καλοήθη αποτελέσματα. Για να μειωθεί ο υψηλός αριθμός των περιττών βιοψιών του μαστού, πολλά computer-aided διαγνωστικά (CAD) συστήματα έχουν προταθεί τα τελευταία χρόνια. Αυτά τα συστήματα βοηθούν τους γιατρούς στην απόφασή τους στο να εκτελέσουν μια βιοψία μαστού σε μια ύποπτη αλλοίωση, που έχει εμφανιστεί στην μαστογραφία ή να εκτελέσουν μια βραχυπρόθεσμη παρακολούθηση -εξέταση. Αυτό το σετ δεδομένων μπορεί να χρησιμοποιηθεί για να προβλέψει τη σοβαρότητα (καλοήθης ή κακοήθης) μιας αλλοίωσης (όγκου) της μαστογραφίας από BI - RADS χαρακτηριστικά και την ηλικία του ασθενούς .[6]
Περιέχει μια BI- RADS αξιολόγηση, την ηλικία του ασθενούς και τρεις BI- RADS ιδιότητες μαζί με την αλήθεια (το πεδίο σοβαρότητας) επί 516 καλοήθεις και 445 κακοήθεις μάζες που έχουν εντοπιστεί στον πλήρη τομέα με ψηφιακές μαστογραφίες, που συλλέγονται στο Ινστιτούτο Ακτινολογίας του Πανεπιστήμιο Ερλάνγκεν - Νυρεμβέργης μεταξύ 2003 και 2006 .
Κάθε περίπτωση έχει μια σχετική BI - RADS εκτίμηση που κυμαίνεται από 1 (σίγουρα καλοήθης) έως 5 (ισχυρές ενδείξεις κακοήθειας), που αποδίδεται σε μια

διπλή διαδικασία αναθεώρησης από τους γιατρούς. Υποθέτοντας, ότι όλες οι περιπτώσεις με BI - RADS εκτιμήσεις, μεγαλύτερη ή ίση μιας δεδομένης τιμής (που ποικίλλει από 1 έως 5), είναι κακοήθεις και οι άλλες περιπτώσεις καλοήθεις, ευαισθησίες και εξειδικεύσεις των συναφών μπορούν να υπολογιστούν. Αυτά μπορεί να είναι μια ένδειξη του πόσο καλά ένα σύστημα CAD αποδίδει σε σύγκριση με τους ακτινολόγους.[6]

Η βάση αυτή λοιπόν περιέχει 6 στήλες συν μιας που προστέθηκε ως πρωτεύον κλειδί.

Το πρωτεύον κλειδί σε έναν πίνακα προσδιορίζει μοναδικά κάθε εγγραφή στον πίνακα. Οι εγγραφές είναι 961.

Η πρώτη στήλη είναι το πρωτεύον κλειδί (primary key) με όνομα : id.

Η δεύτερη στήλη είναι η αξιολόγηση BI-RADS πεδίο ακέραιο με εύρος 1 έως 5.

Η τρίτη στήλη είναι η ηλικία του ασθενή.

Η τέταρτη στήλη είναι το σχήμα του όγκου με εύρος 1 έως 4 όπου :

1= σχήμα στρογγυλό

2= σχήμα οβάλ

3= σχήμα λοβωτό

4= σχήμα ακανόνιστο

Η πέμπτη στήλη είναι τα όρια που χαρακτηρίζουν την μάζα με εύρος 1 έως 5 όπου:

1= περιγεγραμμένο

2=ομαλό

3= εν μέρη αποκρυπτόμενο

4=ασαφές

5=με προεκβολές προς τους γύρω ιστούς

Η έκτη στήλη είναι η πυκνότητα της μάζας με εύρος 1 έως 4 όπου:

1=υψηλή

2=μέτρια

3=χαμηλή

4= περιέχει λίπος

Η έβδομη στήλη είναι η σοβαρότητα όπου χαρακτηρίζεται με 1 ή 0 όπου το 0 είναι καλοήθης όγκος και το 1 κακοήθης όγκος.

Στην παρακάτω φωτογραφία απεικονίζεται ένα κομμάτι με τιμές και πεδία της βάση δεδομένων:

Id	BIRADS	Age	Shape	Margin	Density	Severity
964	5	67	3	5	3	1
965	4	43	1	1	3	1
966	5	58	4	5	3	1
967	4	28	1	1	3	0
968	5	74	1	5	3	1
969	4	65	1	4	3	0
970	4	70	1	4	3	0
971	5	42	1	2	3	0
972	5	57	1	5	3	1
973	5	60	2	5	1	1
974	5	76	1	4	3	1
975	3	42	2	1	3	1
976	4	64	1	1	3	0
977	4	36	3	1	2	0
978	4	60	2	1	2	0
979	4	54	1	1	3	0
980	3	52	3	4	3	0
981	4	59	2	1	3	1
982	4	54	1	1	3	1
983	4	40	1	1	3	0
984	4	66	1	2	1	1
985	5	56	4	3	1	1

Εικόνα 1ης Βάσης Δεδομένων

Και για τις δυο βάσεις δημιουργήθηκαν διαδικασίες (procedures) INSERT και DELETE σε γλώσσα sql στο visual studio οι οποίες φαίνονται παρακάτω:

```
CREATE PROCEDURE dbo.InsertRow
        (
        @BIRADS int,
        @Age int,
        @Shape int,
        @Margin int,
        @Density int,
        @Severity int,
        @Id int OUTPUT
        )
AS
SET NOCOUNT ON;
--Insert a row of data
INSERT INTO [Mammographic data]
        ([BIRADS],Age,Shape,Margin,Density,Severity)
        VALUES
        (@BIRADS,@Age,@Shape,@Margin,@Density,@Severity)
```

```
CREATE PROCEDURE dbo.DeleteRow

(
	@Id int
)
AS
  SET NOCOUNT ON;
	DELETE FROM [Mammographic data] WHERE (Id=@Id);
```

2η Βάση Δεδομένων [10]

Η δεύτερη βάση δεδομένων περιέχει λιγότερες πληροφορίες καθώς αποτελείται από 4 στήλες.

Η πρώτη στήλη που είναι και το πρωτεύον κλειδί είναι ο αριθμός αναφοράς της βάσης.

Η δεύτερη στήλη είναι ο χαρακτηρισμός του υπόβαθρου του ιστού και χαρακτηρίζεται από 3 γράμματα F,G,D όπου:

F fatty (λιπαρό)

G Fatty-glandular (Λιπαρό-αδενικό)

D Dense-glandular (Πυκνό αδενικό)

Η τρίτη στήλη είναι η κατηγοριοποίηση της τρέχουσας ανωμαλίας και χαρακτηρίζεται από τους εξής χαρακτηρισμούς :

CALC Calcification (αποτιτάνωση)

CIRC Well-defined/circumscribed masses (καλά καθορισμένη/οριοθετημένη μάζα)

SPIC Spiculated masses (ακανθωτές μάζες)

MISC Other, ill-defined masses (άλλες, ασαφείς μάζες)

ARCH Architectural distortion (Αρχιτεκτονικά στρεβλωμένη)

ASYM Asymmetry (ασύμμετρη)

NORM Normal (κανονική)

Η τέταρτη στήλη είναι ο χαρακτηρισμός της σοβαρότητας της ανώμαλης μάζας και χαρακτηρίζεται ως B benign (καλοήθης) και ως M malignant (κακοήθης).

Στην παρακάτω φωτογραφία απεικονίζεται ένα κομμάτι με τιμές και πεδία της βάσης δεδομένων:

Reference nu...	Background ti...	Abnormality	Severity
mdb001	G	CIRC	B
mdb002	G	CIRC	B
mdb003	D	NORM	M
mdb004	D	NORM	B
mdb005	F	CIRC	B
mdb006	F	NORM	M
mdb007	G	NORM	B
mdb008	G	NORM	M
mdb009	F	NORM	B
mdb010	F	CIRC	B
mdb011	F	NORM	M
mdb012	F	CIRC	B
mdb013	G	MISC	B
mdb014	G	NORM	B
mdb015	G	CIRC	B
mdb016	G	NORM	B
mdb017	G	CIRC	B
mdb018	G	NORM	M
mdb019	G	CIRC	B
mdb020	G	NORM	M
mdb021	G	CIRC	B
mdb022	G	NORM	B

Εικόνα 2ης Βάσης Δεδομένων

Κεφάλαιο 5

5.1 Προετοιμασία των Βάσεων για εξόρυξη

Το λογισμικό, που χρησιμοποιήθηκε για την προετοιμασία των βάσεων καθώς και την δημιουργία των μοντέλων εξόρυξης είναι ο Microsoft sql server 2012. Ο Microsoft SQL Server είναι ένα σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων που αναπτύχθηκε από τη Microsoft. Ως βάση δεδομένων, είναι ένα προϊόν λογισμικού του οποίου η κύρια λειτουργία είναι να αποθηκεύει και να ανακτά τα δεδομένα, όπως του ζητείται από άλλες εφαρμογές λογισμικού, είτε πρόκειται για εφαρμογές που βρίσκονται στον ίδιο υπολογιστή είτε λειτουργούν σε έναν άλλο υπολογιστή σε ένα δίκτυο (συμπεριλαμβανομένου του Διαδικτύου). Υπάρχουν αρκετές διαφορετικές εκδόσεις του Microsoft SQL Server, που στοχεύουν σε διαφορετικά ακροατήρια και για διάφορα φορτία (που κυμαίνονται από μικρές εφαρμογές, που αποθηκεύουν και ανακτούν δεδομένα στον ίδιο υπολογιστή, μέχρι εκατομμύρια χρήστες και υπολογιστές, που έχουν πρόσβαση σε τεράστιους όγκους δεδομένων από το Internet την ίδια στιγμή). Η έκδοση, που χρησιμοποιήθηκε για την διπλωματική εργασία είναι η enterprise edition. Κύριες γλώσσες επερωτήσεων είναι T-SQL και ANSI SQL.

Περιγράφεται στην συνέχεια η διαδικασία προετοιμασίας των δυο βάσεων για εξόρυξη:

Αρχικά δημιουργήθηκε ένα analysis services multidimensional and data mining project, όπου συνδέθηκε το data sources με τον τοπικό server του laptop για να βλέπει και να μπορεί να διαβάσει τις βάσεις, που είχαν δημιουργηθεί από το visual studio.

Το data source είναι μια σύνδεση δεδομένων, που είναι αποθηκευμένη και διαχειρίσιμη από το project και φορτώνεται στον Microsoft SQL Server Analysis Services database. Το data source περιέχει τα ονόματα του server και της βάσης δεδομένων, όπου βρίσκονται τα δεδομένα προέλευσης, και επιπλέον άλλες απαιτούμενες ιδιότητες της σύνδεσης.

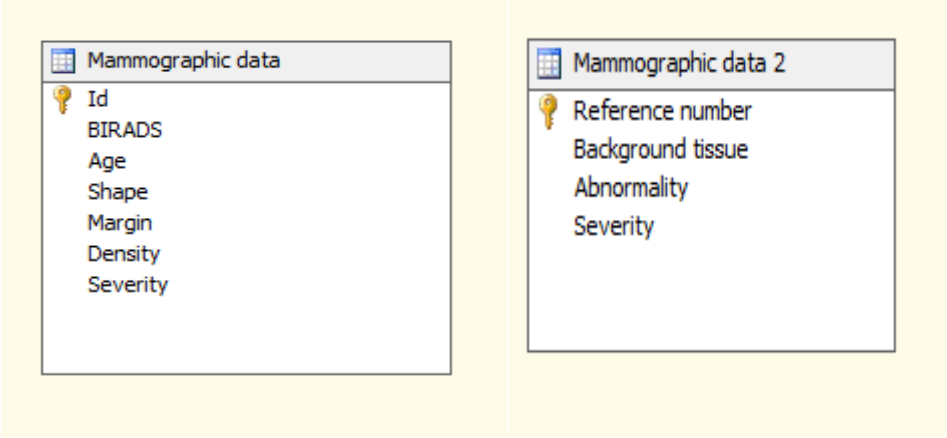
Στην συνέχεια δημιουργήθηκαν 2 όψεις, μια για κάθε βάση δεδομένων για να μπορούν να τύχουν επεξεργασίας από το μοντέλο εξόρυξης.

Μια όψη δεδομένων (data source view DSV) είναι η αφαίρεση μιας σχεσιακής data source, που γίνεται η βάση των διαστάσεων, που δημιουργούνται σε ένα πολυδιάστατο έργο. Ο σκοπός της DSV είναι να δώσει τον έλεγχο των δομών

δεδομένων, που χρησιμοποιούνται στο έργο, όπου δίνεται η δυνατότητα να εργαστεί ο αναλυτής ανεξάρτητα από τις υποκείμενες πηγές δεδομένων (όπως είναι για παράδειγμα, η δυνατότητα μετονομασίας ή συνένωσης στηλών χωρίς άμεση τροποποίηση της αρχικής προέλευσης δεδομένων).

Μια όψη δεδομένων είναι χτισμένη σε μια data source (πηγή δεδομένων) και ορίζει ένα υποσύνολο των δεδομένων, τα οποία στη συνέχεια μπορεί να χρησιμοποιήσει ο αναλυτής σε δομές εξόρυξης. Μπορεί επίσης να χρησιμοποιήσει την προβολή προέλευσης δεδομένων για να προσθέσει στήλες, να δημιουργήσει υπολογιζόμενες στήλες και αδρανή υλικά και να προσθέσει ονοματισμένες όψεις. Με τη χρήση των όψεων, μπορεί να επιλέξει τα δεδομένα, που σχετίζονται με το έργο, να δημιουργήσει σχέσεις μεταξύ πινάκων και να τροποποιήσει τη δομή των δεδομένων, χωρίς να τροποποιηθεί η αρχική πηγή δεδομένων.

Οι δυο όψεις φαίνονται στις παρακάτω εικόνες:



Όψη(view) της 1ης Βάσης Δεδομένων και Όψη(view) της 2ης Βάσης Δεδομένων

5.2 Δημιουργία των μοντέλων εξόρυξης

Για την 1η Βάση Δεδομένων

Αρχικά επιλέγεται ο αλγόριθμος, που θα χρησιμοποιηθεί για το μοντέλο και στην προκειμένη περίπτωση επελέγη ο αλγόριθμος «δένδρα αποφάσεων». Για την δημιουργία του μοντέλου εξόρυξης ορίσθηκαν οι στήλες, που θα χρησιμοποιηθούν ως στήλες εισόδου (input) και η μια στήλη, που θα χρησιμοποιηθεί για την πρόβλεψη

(prediction). Ως στήλη πρόβλεψης ορίσθηκε η σοβαρότητα (severity) του όγκου : αναλόγως του εάν είναι καλοήθης ή κακοήθης. Όλες οι υπόλοιπες στήλες ορίσθηκαν ως στήλες εισόδου για να χρησιμοποιηθούν από το μοντέλο εξόρυξης. Στην συνέχεια ορίζεται η πρώτη σημαντική παράμετρος του μοντέλου:

Percentage of data for testing: Τα δεδομένα εισόδου χωρίζονται σε δύο κατηγορίες, σε ένα σετ, που αφορά την εκπαίδευση του μοντέλου εξόρυξης και σε ένα δεύτερο σετ που αφορά τα τεστ. Η προκαθορισμένη τιμή είναι 30% για testing και 70% για training. Το μοντέλο εξόρυξης χρησιμοποιεί ένα ποσοστό από τα δεδομένα για να διαβάσει και να αντιληφθεί, πώς προκύπτει από τα δεδομένα εισόδου η στήλη πρόβλεψης, δηλαδή βλέπει ένα ποσοστό των δεδομένων (70% ετέθη στην προκειμένη περίπτωση) και την τελική απάντηση. Η στήλη πρόβλεψης στην προκειμένη περίπτωση, είναι η στήλη σοβαρότητας του όγκου (severity). Οπότε, το τεθέν ποσοστό των δεδομένων (70%) δεν χρησιμοποιείται για την εξαγωγή των αποτελεσμάτων (το μοντέλο έχει «δει» την απάντηση σε αυτά τα δεδομένα). Δεν υπάρχει τέλειο νούμερο, που πρέπει να τεθεί, καθώς είναι στην κρίση του αναλυτή να ορίσει το ιδανικό ποσοστό για κάθε τύπο δεδομένων, και τούτο εξαρτάται από διάφορες παραμέτρους όπως:

- όγκο των δεδομένων: ανάλογα με τον όγκο των δεδομένων που είναι διαθέσιμα για τον αναλυτή (αν έχει για παράδειγμα 100000 εγγραφές διαθέσιμες, δεν χρειάζεται για την εκπαίδευση του μοντέλου εξόρυξης να θέσει ως ποσοστό το 70%, που είναι υψηλό, αλλά ένα μικρότερο ποσοστό).
- αριθμό στηλών των δεδομένων εισόδου: εάν ο αριθμός των στηλών των δεδομένων εισόδου είναι μεγάλος, αν δηλαδή έχουμε διαθέσιμες περισσότερες από 10 στήλες και ταυτόχρονα μεγάλο αριθμό γραμμών θα χρειάζονται περισσότερες πληροφορίες-δεδομένα για την εκπαίδευση του μοντέλου εξόρυξης, ώστε να μπορέσει επιτυχώς να βρει τους κατάλληλους συσχετισμούς μεταξύ των δεδομένων).
- ιδιαιτερότητα των δεδομένων εισόδου: οι αριθμοί είναι πλέον δόκιμοι για να μπορούν να αναγνωρίσουν τα συστήματα και τους συσχετισμούς των δεδομένων. Όμως τα δεδομένα (όπως στην δεύτερη βάση δεδομένων που έγινε η επεξεργασία στα πλαίσια αυτής της διπλωματικής) μπορεί να είναι χαρακτήρες, οπότε δυσκολεύεται η αναγνώριση και η εκπαίδευση του μοντέλου εξόρυξης.

Στην συνέχεια δίνεται η δυνατότητα να μπορεί ο αναλυτής να κάνει drill through (γεώτρηση).

Η γεώτρηση μπορεί να ενεργοποιηθεί σε μοντέλα και δομές. Αφότου το μοντέλο έχει υποβληθεί σε επεξεργασία, θα μπορεί ο αναλυτής να ανακτήσει λεπτομερείς πληροφορίες από τα δεδομένα εκπαίδευσης, που χρησιμοποιούνται για τη δημιουργία του μοντέλου.[16]

Εάν η υποκείμενη δομή εξόρυξης έχει ρυθμιστεί ώστε να επιτρέπει την γεώτρηση, τότε μπορεί να ανακτήσει ο αναλυτής λεπτομερείς πληροφορίες τόσο από τις εγγραφές του μοντέλου όσο και από τη δομή εξόρυξης, συμπεριλαμβανομένων των στηλών που δεν είχαν συμπεριληφθεί στο μοντέλο εξόρυξης.[16]

Η γεώτρηση είναι χρήσιμη αν θέλει ο αναλυτής να δει τις εγγραφές που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου, σε σχέση με τις εγγραφές που χρησιμοποιούνται για τη δοκιμή του, ή αν θέλει, μπορεί επίσης να δει περισσότερες λεπτομέρειες από τα δεδομένα εγγραφής.[16]

Οι Υπηρεσίες Εξόρυξης Ανάλυσης Δεδομένων προσφέρουν δύο διαφορετικές επιλογές για γεώτρηση:[16]

Γεώτρηση στις εγγραφές του μοντέλου

Η γεώτρηση χρησιμοποιείται σε εγγραφές του μοντέλου, όταν θέλει ο αναλυτής να κατευθυνθεί από ένα συγκεκριμένο μοτίβο του μοντέλου-όπως από ένα σύμπλεγμα ή από παρακλάδι του δένδρου απόφασης-και να δει λεπτομέρειες σχετικά με τις μεμονωμένες περιπτώσεις.

Γεώτρηση στις δομές του μοντέλου

Η γεώτρηση στις δομές του μοντέλου χρησιμοποιείται, όταν η δομή περιέχει πληροφορίες, που ενδέχεται να μην είναι διαθέσιμες στο μοντέλο. Για παράδειγμα, παίρνουμε την περίπτωση κατά την οποία ο αναλυτής δεν θέλει να χρησιμοποιήσει τα στοιχεία επικοινωνίας πελατών σε ένα μοντέλο- ομαδοποίηση, ακόμη και αν τα δεδομένα περιλαμβάνονται στη δομή. Ωστόσο, μετά από την δημιουργία του μοντέλου, ίσως να θέλει να ανακτήσει τα στοιχεία επικοινωνίας για τους πελάτες που έχουν ομαδοποιηθεί σε ένα συγκεκριμένο σύμπλεγμα. Η γεώτρηση του δίνει την δυνατότητα αυτή.

Μετά από αυτές τις επιλογές δημιουργείται το μοντέλο εξόρυξης.

Η ίδια διαδικασία ακολουθήθηκε για την δημιουργία μοντέλου για την ομαδοποίηση (clustering) για την $1^{η}$ Βάση Δεδομένων.

Για την $2^{η}$ Βάση Δεδομένων

Οι ίδιοι αλγόριθμοι χρησιμοποιήθηκαν για την βάση αυτή, οπότε ακολουθήθηκε η ίδια διαδικασία. Στην περίπτωση της στήλης, που αφορά την πρόβλεψη επελέγη η σοβαρότητα (severity) και όλες οι άλλες στήλες, εκτός του πρωτεύοντος κλειδιού, ως στήλες εισόδου.

Εξαιτίας του μικρότερου όγκου δεδομένων (329 εγγραφές) οι παράμετροι που ορίστηκαν είναι διαφορετικοί ως προς το training data αλλά θα περιγραφούν στα αποτελέσματα ώστε να αντιληφθεί ο αναγνώστης την σημασία σωστής επιλογής παραμέτρων.

5.3 Εξόρυξη 1ης Βάσης Δεδομένων

Αλγόριθμος Decision trees

Υπάρχουν 7 τουλάχιστον παράμετροι του αλγορίθμου που μπορούν να τροποποιηθούν. Αυτοί είναι:

- COMPLEXITY_PENALTY : Ελέγχει την ανάπτυξη του δέντρου απόφασης. Μια χαμηλή τιμή αυξάνει τον αριθμό των διασπάσεων, και μια υψηλή τιμή μειώνει τον αριθμό των διασπάσεων. Η προεπιλεγμένη τιμή με βάση τον αριθμό των χαρακτηριστικών για ένα συγκεκριμένο μοντέλο, όπως περιγράφεται στον ακόλουθο κατάλογο:

  Για 1 έως 9 χαρακτηριστικά, η προεπιλεγμένη τιμή είναι 0,5.

  Για 10 έως 99 χαρακτηριστικά, η προεπιλεγμένη τιμή είναι 0,9.

  Για 100 ή περισσότερα χαρακτηριστικά, η προεπιλεγμένη τιμή είναι 0,99.

- FORCED_REGRESSOR : Αναγκάζει τον αλγόριθμο να χρησιμοποιήσει τις αναφερόμενες στήλες ως παλινδρομικές, ανεξάρτητα από τη σημασία των στηλών, όπως υπολογίζεται από τον αλγόριθμο.
- MAXIMUM_INPUT_ATTRIBUTES : Καθορίζει τον αριθμό των χαρακτηριστικών εισόδου που ο αλγόριθμος μπορεί να χειριστεί πριν επικαλεστεί την επιλογή χαρακτηριστικών. Ορίζοντας την τιμή σε 0 απενεργοποιείται η επιλογή χαρακτηριστικών.
  Η προεπιλογή είναι 255.

- MAXIMUM_OUTPUT_ATTRIBUTES : Καθορίζει τον αριθμό των χαρακτηριστικών εξόδου που ο αλγόριθμος μπορεί να χειριστεί πριν επικαλεστεί την επιλογή χαρακτηριστικών. Ορίζοντας την τιμή σε 0 απενεργοποιείται η επιλογή χαρακτηριστικών.
  Η προεπιλογή είναι 255.

- MINIMUM_SUPPORT : Καθορίζει τον ελάχιστο αριθμό των περιπτώσεων φύλλων που απαιτείται για να δημιουργήσει μια διάσπαση στο δέντρο απόφασης.
  Η προεπιλογή είναι 10.

- SCORE_METHOD : Καθορίζει τη μέθοδο που χρησιμοποιείται για τον υπολογισμό της βαθμολογίας διαχωρισμού. Οι ακόλουθες επιλογές είναι διαθέσιμες: (1) Εντροπία (2) Bayesian με K2 Πριν, ή (3) Bayesian Dirichlet Ισοδύναμο (BDE) Πριν.
  Η προεπιλεγμένη τιμή είναι 3.

- SPLIT_METHOD : Καθορίζει τη μέθοδο που χρησιμοποιείται για να χωρίσει τον κόμβο. Οι ακόλουθες επιλογές είναι διαθέσιμες: Binary (1), Complete (2), ή και τα δύο (3).

  Η προεπιλεγμένη τιμή είναι 3.

Το testing data ήταν ίσο με 30%.

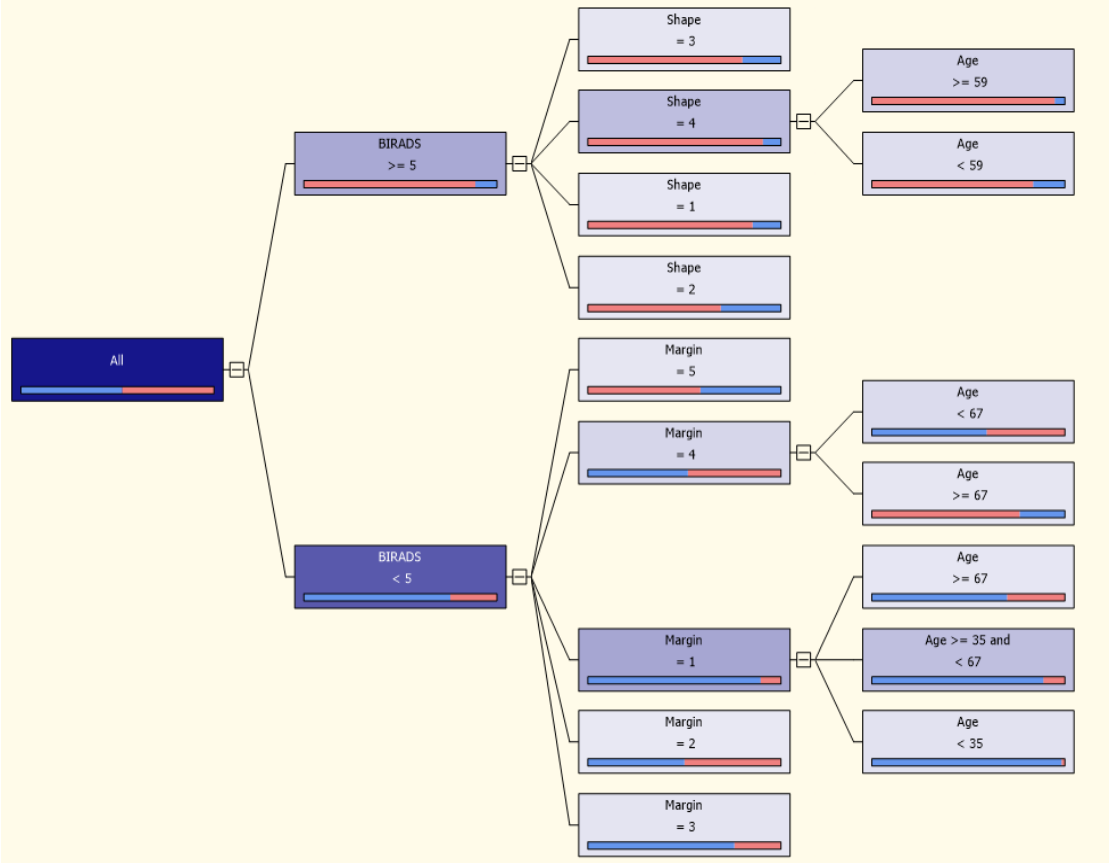
Οι τιμές των παραμέτρων που χρησιμοποιήθηκαν για το πιο επιτυχημένο μοντέλο της συγκεκριμένης βάσης με κριτήριο το ποσοστό πρόβλεψης είναι:

COMPLEXITY_PENALTY =0.3

SCORE_METHOD=1

SPLIT_METHOD=2

Όλες οι υπόλοιπες παράμετροι σε προεπιλεγμένες τιμές. Τα αποτελέσματα από την μέθοδο δένδρα αποφάσεων φαίνονται παρακάτω:



Δένδρο Απόφασης 1ης Βάσης Δεδομένων

Το μπλε είναι καλοήθης όγκος και το κόκκινο ο κακοήθης όγκος.

Αλγόριθμος Clustering

Υπάρχουν 9 τουλάχιστον παράμετροι του αλγορίθμου που μπορούν να τροποποιηθούν. Αυτοί είναι:

- CLUSTER_COUNT : Καθορίζει τον κατά προσέγγιση αριθμό των συστάδων που θα κατασκευαστούν από τον αλγόριθμο. Εάν ο κατά προσέγγιση αριθμός των συστάδων δεν μπορεί να κατασκευαστεί από τα δεδομένα, ο αλγόριθμος χτίζει όσο το δυνατόν περισσότερες συστάδες.
  Η προεπιλογή είναι 10.

- CLUSTER_SEED : Καθορίζει τον αριθμό των σπόρων που χρησιμοποιείται για να δημιουργήσει τυχαία συστάδες για το αρχικό στάδιο του μοντέλου οικοδόμησης.
  Η προεπιλογή είναι 0.

- CLUSTERING_METHOD : Καθορίζει τη μέθοδο ομαδοποίησης για τον αλγόριθμο που θα χρησιμοποιήσει. Οι ακόλουθες μέθοδοι ομαδοποίησης είναι διαθέσιμες: κλιμακούμενες EM (1), μη κλιμακούμενες EM (2), επεκτάσιμη K-Means (3), και μη επεκτάσιμη K-Means (4).
  Η προεπιλογή είναι 1.

- MAXIMUM_INPUT_ATTRIBUTES : Καθορίζει τον αριθμό των χαρακτηριστικών εισόδου που ο αλγόριθμος μπορεί να χειριστεί πριν επικαλεστεί την επιλογή χαρακτηριστικών. Ορίζοντας την τιμή σε 0 απενεργοποιείται η επιλογή χαρακτηριστικών.
  Η προεπιλογή είναι 255.

- MAXIMUM_STATES : Καθορίζει το μέγιστο αριθμό των καταστάσεων των χαρακτηριστικών που ο αλγόριθμος υποστηρίζει. Εάν ο αριθμός των καταστάσεων που ένα χαρακτηριστικό έχει είναι μεγαλύτερος από το μέγιστο αριθμό των καταστάσεων, ο αλγόριθμος χρησιμοποιεί τις καταστάσεις του πιο δημοφιλούς χαρακτηριστικού και αγνοεί τις υπόλοιπες καταστάσεις.
  Η προεπιλογή είναι 100.

- MINIMUM_SUPPORT : Καθορίζει τον ελάχιστο αριθμό των εγγραφών σε κάθε συστάδα.
  Η προεπιλογή είναι 1.

- MODELLING_CARDINALITY : Καθορίζει τον αριθμό των μοντέλων δείγματος που κατασκευάστηκαν κατά τη διάρκεια της διαδικασίας ομαδοποίησης.

  Η προεπιλογή είναι 10.

- SAMPLE_SIZE: Καθορίζει τον αριθμό των εγγραφών που ο αλγόριθμος χρησιμοποιεί σε κάθε πέρασμα, εάν η παράμετρος CLUSTERING_METHOD είναι ίση με 1 από τις κλιμακούμενες μεθόδους ομαδοποίησης. Ρύθμιση του SAMPLE_SIZE παράμετρο στο 0 θα προκαλέσει όλο το σύνολο δεδομένων να συγκεντρωθεί σε ένα μόνο πέρασμα. Αυτό μπορεί να προκαλέσει θέματα μνήμης και απόδοσης..

  Η προεπιλεγμένη τιμή είναι 50000.

- STOPPING_TOLERANCE : Συγκεκριμενοποιεί την τιμή που χρησιμοποιείται για να καθορίσει πότε έχει επιτευχθεί σύγκλιση και πότε ο αλγόριθμος έχει τελειώσει την κατασκευή του μοντέλου. Σύγκλιση επιτυγχάνεται όταν η συνολική μεταβολή των πιθανοτήτων του συμπλέγματος είναι μικρότερη από την αναλογία της παραμέτρου STOPPING_TOLERANCE διαιρούμενο με το μέγεθος του μοντέλου.

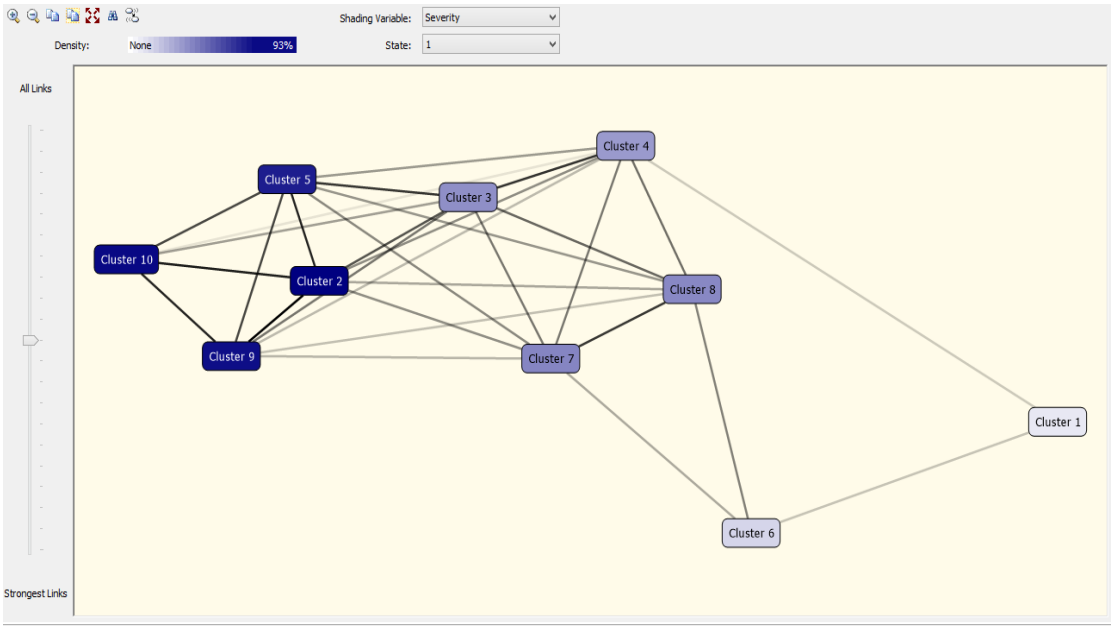
  Η προεπιλογή είναι 10.

Το testing data ήταν ίσο με 30%.

Οι τιμές των παραμέτρων που κατέληξα για το πιο επιτυχημένο μοντέλο για την συγκεκριμένη βάση με κριτήριο το ποσοστό πρόβλεψης είναι:

CLUSTERING_METHOD=2

Όλες οι υπόλοιπες παράμετροι σε προεπιλεγμένες τιμές. Τα αποτελέσματα από την μέθοδο ομαδοποίηση φαίνονται παρακάτω:

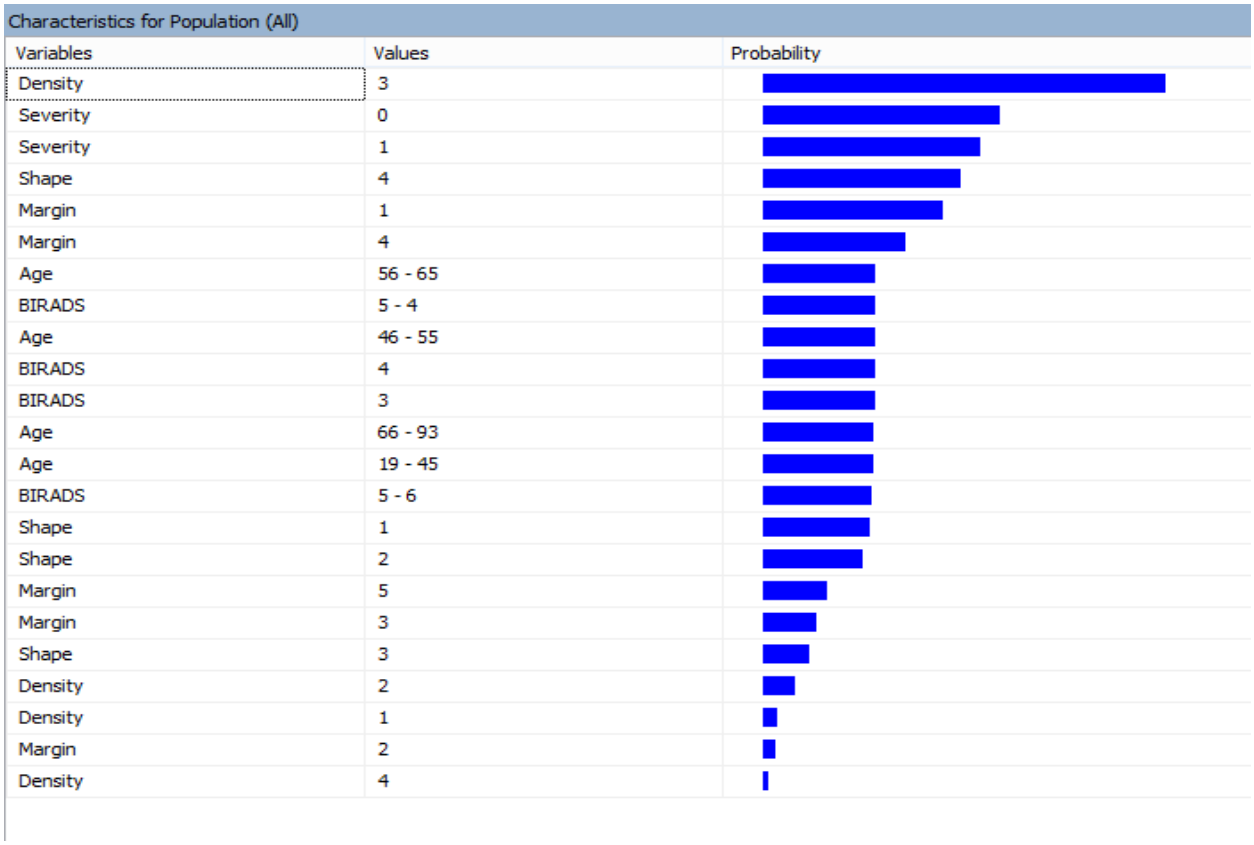


Clusters 1ης Βάσης Δεδομένων

Με έντονο μπλε είναι οι ομάδες-συστάδες οι οποίες έχουν μαζέψει κακοήθη όγκο. Στις υπόλοιπες ομάδες-συστάδες όσο πιο αχνό είναι το μπλε τόσο περισσότερο καλοήθη όγκο έχουν στην ομάδα τους. Τα κριτήρια και τα χαρακτηριστικά στα οποία διαχώρισε τις εγγραφές σε 10 ομάδες ο αλγόριθμος φαίνονται παρακάτω:

Attributes		Cluster profiles										
Variables	States	Populatio... Size: 673	Cluster 1 Size: 175	Cluster 2 Size: 135	Cluster 3 Size: 114	Cluster 4 Size: 97	Cluster 5 Size: 86	Cluster 6 Size: 22	Cluster 7 Size: 21	Cluster 8 Size: 11	Cluster 9 Size: 7	Cluster 10 Size: 5
Age	93,00 56,00 19,00											
BIRADS	6,00 4,00 1,84											
Density	3 2 1 4 Other											
Margin	1 4 5 3 Other											
Severity	0 1 missing											
Shape	4 1 2 3 Other											

Πληροφορίες διαχωρισμού των clusters της 1ης Βάσης Δεδομένων



Characteristics for Population (All)

Variables	Values	Probability
Density	3	
Severity	0	
Severity	1	
Shape	4	
Margin	1	
Margin	4	
Age	56 - 65	
BIRADS	5 - 4	
Age	46 - 55	
BIRADS	4	
BIRADS	3	
Age	66 - 93	
Age	19 - 45	
BIRADS	5 - 6	
Shape	1	
Shape	2	
Margin	5	
Margin	3	
Shape	3	
Density	2	
Density	1	
Margin	2	
Density	4	

Πληροφορίες των clusters της 1ης Βάσης Δεδομένων

5.4 Σύγκριση αποτελεσμάτων 1ης Βάσης Δεδομένων

Στην συνέχεια έγινε σύγκριση των δύο τεχνικών για το ποια είναι καλύτερη.

5.4.1 Lift Chart

Το πρώτο κριτήριο είναι η ευστοχία των προβλέψεων. Χρησιμοποιήθηκε το Lift γράφημα. Ένα γράφημα Lift αναπαριστά γραφικά τη βελτίωση που ένα μοντέλο εξόρυξης παρέχει, όταν γίνεται σύγκριση με μια τυχαία εικασία, και μετρά τη μεταβολή σε όρους μιας βαθμολογία lift. Με τη σύγκριση των lift βαθμολογιών σε διάφορα τμήματα του σετ δεδομένων και για τα διάφορα μοντέλα, μπορούμε να προσδιορίσουμε ποιο μοντέλο είναι καλύτερο, και ποιό ποσοστό των εγγραφών του συνόλου δεδομένων θα επωφεληθεί από την εφαρμογή των προβλέψεων του μοντέλου.

Με ένα γράφημα lift, μπορούμε να συγκρίνουμε την ακρίβεια των προβλέψεων για πολλαπλά μοντέλα που έχουν το ίδιο προβλέψιμο χαρακτηριστικό. Μπορούμε να αξιολογήσουμε επίσης την ακρίβεια της πρόβλεψης, είτε για ένα μόνο αποτέλεσμα

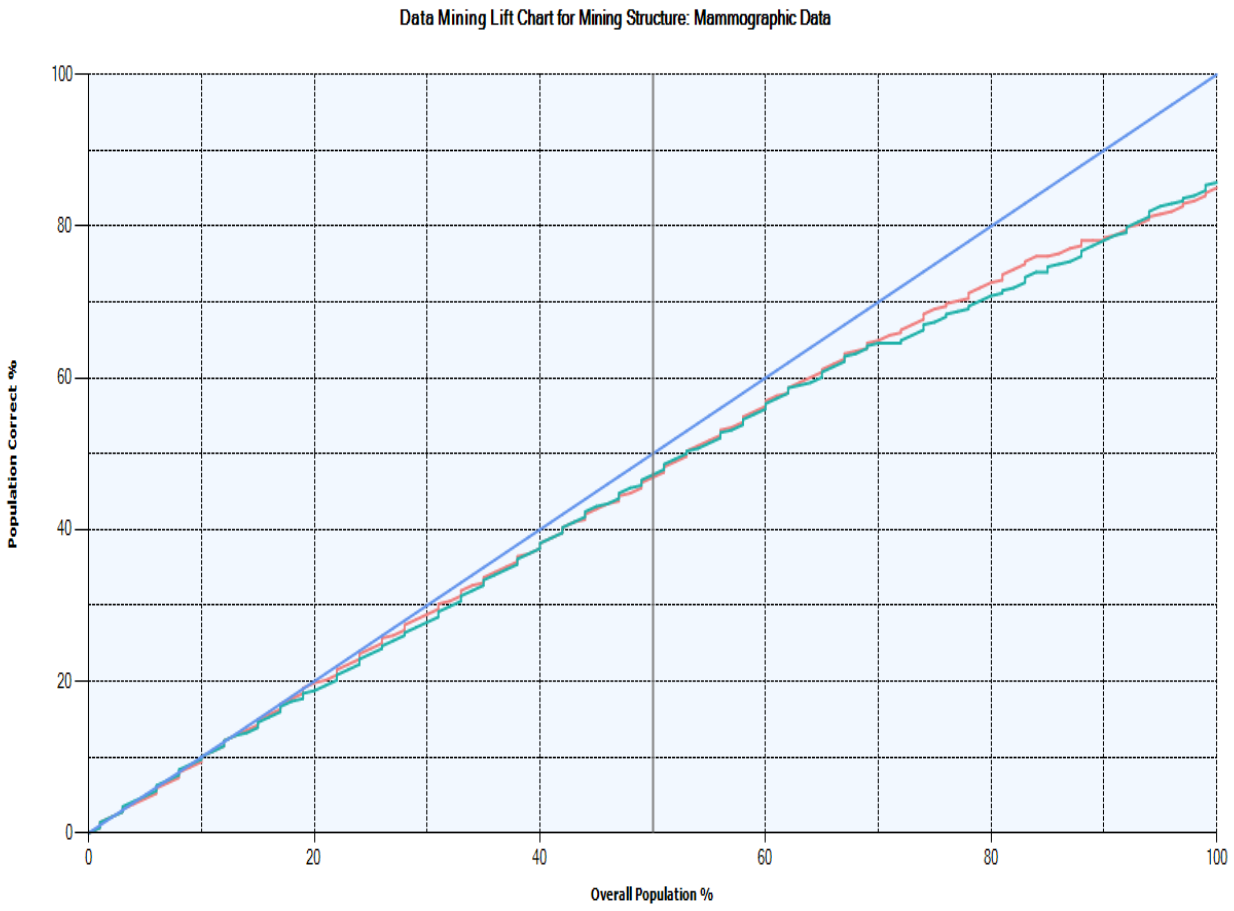
(μία μόνο τιμή του χαρακτηριστικού πρόβλεψης), ή για όλα τα αποτελέσματα (όλες οι τιμές του καθορισμένου χαρακτηριστικού).

Ο x-άξονας του γραφήματος αντιπροσωπεύει το ποσοστό του test dataset που χρησιμοποιείται για να συγκρίνει τις προβλέψεις. Ο άξονας y του γραφήματος αντιπροσωπεύει το ποσοστό των προβλέψεων που είναι σωστές.

Η διαγώνια ευθεία γραμμή, που φαίνεται παρακάτω σε μπλε, αντιπροσωπεύει το αποτέλεσμα της τυχαίας εικασίας, και είναι η βάση με την οποία αξιολογείται το lift. Η μπλε γραμμή δείχνει τα ιδανικά αποτελέσματα, που δημιουργούν ένα μοντέλο το οποίο προβλέπει πάντα σωστά. Η κόκκινη γραμμή είναι η γραμμή πρόβλεψης του μοντέλου ομαδοποίησης και η πράσινη του μοντέλου «δένδρα αποφάσεων».

Overall Score των δύο αλγορίθμων 1ης Βάσης Δεδομένων

Series Model	Score
TM_Clustering	0.92
TM_Decision_Tress	0.91
Ideal Model	1



Διάγραμμα lift chart σύγκρισης των δύο αλγορίθμων 1ης Βάσης Δεδομένων

Μπλε γραμμή : Ιδανικό μοντέλο

Κόκκινη γραμμή: Clustering μοντέλο

Πράσινη γραμμή :Decision Trees μοντέλο

5.4.2 Classification or confusion matrix

Ένα δεύτερο κριτήριο είναι ο λεγόμενος classification matrix.

Ο classification matrix κατατάσσει όλες τις περιπτώσεις του μοντέλου σε κατηγορίες, προσδιορίζοντας αν η προβλεπόμενη τιμή ταίριαξε με την πραγματική αξία. Όλες οι περιπτώσεις σε κάθε κατηγορία στη συνέχεια μετρούνται, και τα σύνολα εμφανίζονται στη μήτρα. Ο classification matrix είναι ένα πρότυπο εργαλείο για την αξιολόγηση των στατιστικών μοντέλων και μερικές φορές αναφέρεται ως confusion matrix.

Στο διάγραμμα, που δημιουργείται, όταν επιλεγεί ο εν λόγω πίνακας συγκρίνονται οι πραγματικές τιμές με τις προβλεπόμενες τιμές για κάθε προβλεπόμενη κατάσταση που καθορίζει ο αναλυτής. Οι σειρές του πίνακα αντιπροσωπεύουν τις προβλεπόμενες τιμές για το μοντέλο, ενώ οι στήλες αντιπροσωπεύουν τις πραγματικές αξίες. Οι κατηγορίες που χρησιμοποιούνται στην ανάλυση είναι ψευδώς θετικές, αληθώς θετικές, ψευδώς αρνητικές, και αληθώς αρνητικές.

Ένας πίνακας κατάταξης είναι ένα σημαντικό εργαλείο για την αξιολόγηση των αποτελεσμάτων της πρόβλεψης, διότι καθιστά εύκολο να κατανοήσει κάποιος και να καταλογίσει τις επιπτώσεις των λανθασμένων προβλέψεων. Με την προβολή της ποσότητας και των ποσοστών σε κάθε κελί του πίνακα αυτού, μπορεί ο αναλυτής να δει γρήγορα πόσο συχνά το μοντέλο προέβλεψε με ακρίβεια.

Για το Clustering:

Classification matrix clustering 1^{ης} Βάσης Δεδομένων

Predicted	0(Actual)	1(Actual)
0	151	29
1	14	94

Για να καταλάβουμε τι σημαίνει ο πίνακας: Το νούμερο 151 στην 1^{η} γραμμή σημαίνει ότι το μοντέλο προέβλεψε ως 0 την στήλη πρόβλεψης (severity) και ήταν όντως 0 (καλοήθης όγκος) αληθώς αρνητικά (προέβλεψε σωστά ότι δεν είναι κακοήθης όγκος), ενώ το νούμερο 29 στην ίδια γραμμή σημαίνει ότι προέβλεψε 29 επίσης ως 0 (καλοήθης όγκος) ενώ ήταν 1 η σωστή πρόβλεψη-απάντηση (ψευδώς αρνητικά καθώς προέβλεψε ως καλοήθη όγκο περιπτώσεις κακοήθους). Αντίστοιχα στην δεύτερη γραμμή το νούμερο 94 μας δείχνει τις σωστές προβλέψεις που πραγματοποίησε ο

αλγόριθμος (αληθώς θετικά, ότι δηλαδή υπάρχει κακοήθης όγκος και ήταν σωστή η πρόβλεψη) ενώ το νούμερο 14 μας δείχνει τις λάθος προβλέψεις που έκανε (θεώρησε 1, κακοήθη όγκο, ενώ στην πραγματικότητα η σωστή απάντηση ήταν 0-πρόβλεψη δηλαδή καλοήθης όγκος, ψευδώς θετικά).

Για το Decision Trees:

Classification matrix decision trees 1ης Βάσης Δεδομένων

Predicted	0(Actual)	1(Actual)
0	142	18
1	23	105

5.4.3 Cross Validation

Ένα τρίτο κριτήριο είναι το cross validation.[17]

Το cross validation είναι ένα πρότυπο εργαλείο ανάλυσης και είναι ένα σημαντικό χαρακτηριστικό που βοηθά στην ανάπτυξη και στην τελειοποίηση των μοντέλων εξόρυξης δεδομένων. Χρησιμοποιείται το εργαλείο αυτό για να εξακριβωθεί η εγκυρότητα του μοντέλου. Το cross validation έχει τις ακόλουθες εφαρμογές:

- Επικύρωση της ευρωστίας ενός συγκεκριμένου μοντέλου εξόρυξης.
- Αξιολόγηση πολλαπλών μοντέλων από μια ενιαία δήλωση.
- Κτίσιμο πολλαπλών μοντέλων και στη συνέχεια προσδιορισμός των καλύτερων μοντέλων που βασίζεται σε στατιστικά στοιχεία.

Μια cross validation έκθεση είναι θεμελιωδώς διαφορετική από ένα γράφημα ακρίβεια, όπως ένα γράφημα lift ή από τον classification matrix .

- Το cross validation αξιολογεί τη συνολική κατανομή των δεδομένων που χρησιμοποιούνται σε ένα μοντέλο ή δομή. Ως εκ τούτου, δεν καθορίζεται από τον αναλυτή ένα testing σύνολο δεδομένων. Το cross validation αξιοποιεί πάντα μόνο τα αρχικά δεδομένα που χρησιμοποιούνται για να εκπαιδεύσουν το μοντέλο ή τη δομή της εξόρυξης .
- Το cross validation μπορεί να πραγματοποιηθεί μόνο σε σχέση με ένα προβλέψιμο αποτέλεσμα.
- Μόνο τα μοντέλα που σχετίζονται με την επιλεγμένη δομή είναι διαθέσιμα για cross validation.

Το αποτέλεσμα που επιστρέφει το cross validation για τους δύο αλγόριθμους που χρησιμοποιήθηκαν στην τρέχουσα βάση φαίνονται παρακάτω:

Cross validation αποτελέσματα των δύο αλγορίθμων 1ης Βάσης Δεδομένων

TM_Clustering				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	67	Classification	Pass	54
2	68	Classification	Pass	55
3	68	Classification	Pass	52
4	68	Classification	Pass	54
5	67	Classification	Pass	53
6	67	Classification	Pass	58
7	67	Classification	Pass	58
8	67	Classification	Pass	54
9	67	Classification	Pass	49
10	67	Classification	Pass	53
			Average	53,9985
			Standard Deviation	2,5256
1	67	Classification	Fail	13
2	68	Classification	Fail	13
3	68	Classification	Fail	16
4	68	Classification	Fail	14
5	67	Classification	Fail	14
6	67	Classification	Fail	9
7	67	Classification	Fail	9
8	67	Classification	Fail	13
9	67	Classification	Fail	18
10	67	Classification	Fail	14
			Average	13,3046
			Standard Deviation	2,606
1	67	Likelihood	Log Score	-0,4216
2	68	Likelihood	Log Score	-0,402

3	68	Likelihood	Log Score	-0,483
4	68	Likelihood	Log Score	-0,4877
5	67	Likelihood	Log Score	-0,4606
6	67	Likelihood	Log Score	-0,3425
7	67	Likelihood	Log Score	-0,3589
8	67	Likelihood	Log Score	-0,4126
9	67	Likelihood	Log Score	-0,549
10	67	Likelihood	Log Score	-0,4326
			Average	-0,4352
			Standard Deviation	0,059
1	67	Likelihood	Lift	0,2705
2	68	Likelihood	Lift	0,2895
3	68	Likelihood	Lift	0,2097
4	68	Likelihood	Lift	0,205
5	67	Likelihood	Lift	0,2316
6	67	Likelihood	Lift	0,3496
7	67	Likelihood	Lift	0,3333
8	67	Likelihood	Lift	0,2796
9	67	Likelihood	Lift	0,1432
10	67	Likelihood	Lift	0,2595
			Average	0,257
			Standard Deviation	0,0588
1	67	Likelihood	Root Mean Square Error	0,2302
2	68	Likelihood	Root Mean Square Error	0,2601
3	68	Likelihood	Root Mean Square Error	0,2314
4	68	Likelihood	Root Mean Square Error	0,2197
5	67	Likelihood	Root Mean Square Error	0,2446
6	67	Likelihood	Root Mean Square Error	0,2385
7	67	Likelihood	Root Mean Square Error	0,2644
8	67	Likelihood	Root Mean Square Error	0,2298

9	67	Likelihood	Root Mean Square Error	0,248
10	67	Likelihood	Root Mean Square Error	0,2473
			Average	0,2414
			Standard Deviation	0,0135
TM_Decision_Trees				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	67	Classification	Pass	51
2	68	Classification	Pass	54
3	68	Classification	Pass	51
4	68	Classification	Pass	53
5	67	Classification	Pass	56
6	67	Classification	Pass	59
7	67	Classification	Pass	60
8	67	Classification	Pass	58
9	67	Classification	Pass	51
10	67	Classification	Pass	54
			Average	54,6909
			Standard Deviation	3,2232
1	67	Classification	Fail	16
2	68	Classification	Fail	14
3	68	Classification	Fail	17
4	68	Classification	Fail	15
5	67	Classification	Fail	11
6	67	Classification	Fail	8
7	67	Classification	Fail	7
8	67	Classification	Fail	9
9	67	Classification	Fail	16
10	67	Classification	Fail	13
			Average	12,6122
			Standard Deviation	3,4391
1	67	Likelihood	Log Score	-0,4432
2	68	Likelihood	Log Score	-0,4119
3	68	Likelihood	Log Score	-0,5508
4	68	Likelihood	Log Score	-0,5059
5	67	Likelihood	Log Score	-0,4821
6	67	Likelihood	Log Score	-0,31

7	67	Likelihood	Log Score	-0,3775
8	67	Likelihood	Log Score	-0,3795
9	67	Likelihood	Log Score	-0,533
10	67	Likelihood	Log Score	-0,4258
			Average	-0,4422
			Standard Deviation	0,0724
1	67	Likelihood	Lift	0,249
2	68	Likelihood	Lift	0,2795
3	68	Likelihood	Lift	0,1419
4	68	Likelihood	Lift	0,1868
5	67	Likelihood	Lift	0,2101
6	67	Likelihood	Lift	0,3821
7	67	Likelihood	Lift	0,3147
8	67	Likelihood	Lift	0,3126
9	67	Likelihood	Lift	0,1592
10	67	Likelihood	Lift	0,2663
			Average	0,25
			Standard Deviation	0,0723
1	67	Likelihood	Root Mean Square Error	0,223
2	68	Likelihood	Root Mean Square Error	0,23
3	68	Likelihood	Root Mean Square Error	0,221
4	68	Likelihood	Root Mean Square Error	0,196
5	67	Likelihood	Root Mean Square Error	0,2389
6	67	Likelihood	Root Mean Square Error	0,1947
7	67	Likelihood	Root Mean Square Error	0,2429
8	67	Likelihood	Root Mean Square Error	0,2594
9	67	Likelihood	Root Mean Square Error	0,2328

10	67	Likelihood	Root Mean Square Error	0,2203
			Average	0,2259
			Standard Deviation	0,0189

Κατανόηση των αποτελεσμάτων:

Η πρώτη στήλη Partition Index είναι μια ένας δείκτης με βάση το 1 που προσδιορίζει σε ποια διαμέριση ισχύουν τα αποτελέσματα.

Η δεύτερη στήλη Partition Size είναι ένας ακέραιος που δείχνει πόσες εγγραφές συμπεριελήφθησαν σε κάθε διαμέριση.

Η τρίτη στήλη Test είναι μια κατηγορία που περιγράφει το τεστ που εκτελέστηκε.

- Classification test: Τα μέτρα που ισχύουν για τα μοντέλα ταξινόμησης

  True Positive
  True Negative
  False Positive
  False Positive

  Καταμέτρηση των γραμμών ή τιμών στην διαμέριση όπου η προβλεπόμενη κατάσταση ταιριάζει με την κατάσταση στόχο, και η πιθανότητα πρόβλεψης είναι μεγαλύτερη από το καθορισμένο όριο.
  Οι εγγραφές που έχουν τιμές που λείπουν για το χαρακτηριστικό προορισμού αποκλείονται, δηλαδή οι μετρήσεις όλων των τιμών μπορεί να μην αθροίζονται.
  Χαρακτηρισμός Πέρασε/Απέτυχε:
  Καταμέτρηση των γραμμών ή αξιών στην διαμέριση όπου η προβλεπόμενη κατάσταση ταιριάζει με την κατάσταση στόχο, και όπου η τιμή πρόβλεψης πιθανότητας είναι μεγαλύτερη από 0.
- Likelihood test: έχει 3 τρόπους-μεθόδους μέτρησης.

  Lift
  Η αναλογία της πραγματικής πιθανότητας πρόβλεψης με την οριακή πιθανότητα στις test εγγραφές. Εξαιρούνται οι γραμμές που δεν έχουν τιμή για το χαρακτηριστικό προορισμού .

  Το μέτρο αυτό γενικά δείχνει πόσο η πιθανότητα του αποτελέσματος στόχου βελτιώνεται όταν το μοντέλο χρησιμοποιείται.

Root Mean Square Error
Η Τετραγωνική ρίζα του μέσου σφάλματος για όλες τις περιπτώσεις διαμέρισης, διαιρείται με τον αριθμό των εγγραφών στη διαμέριση, εκτός από τις γραμμές που δεν έχουν τιμή για το χαρακτηριστικό προορισμού.

RMSE είναι ένα δημοφιλής εκτιμητής για μοντέλα πρόβλεψης. Η βαθμολογία βάζει το μέσο όρο για κάθε περίπτωση, για να δώσει ένα ενιαίο δείκτη του σφάλματος του μοντέλου .

Log score
Ο λογάριθμος της πραγματικής πιθανότητας για κάθε εγγραφή, αθροίζεται, και στη συνέχεια διαιρείται με τον αριθμό των γραμμών στο σύνολο δεδομένων εισόδου, εξαιρουμένων των γραμμών, που δεν έχουν τιμή για το χαρακτηριστικό στόχο.
Επειδή η πιθανότητα αναπαρίσταται ως δεκαδικό κλάσμα, οι βαθμολογίες του λογάριθμου είναι πάντα αρνητικοί αριθμοί. Ένας αριθμός πιο κοντά στο μηδέν είναι μια καλύτερη βαθμολογία. Εκτιμώντας ακατέργαστες βαθμολογίες, οι οποίες μπορεί να έχουν πολύ ακανόνιστη ή στραβές διανομές, μια λογαριθμική βαθμολόγηση είναι παρόμοια με ένα ποσοστό.

Η τέταρτη στήλη περιέχει το όνομα της μέτρησης που επιστρέφεται από το τεστ της τρίτης στήλης.

Η πέμπτη στήλη περιέχει την τιμή του καθορισμένου τεστ.

Βλέπουμε ότι και οι δύο αλγόριθμοι με τις τροποποιήσεις που έγιναν συγκλίνουν αρκετά καλά προς το ιδανικό μοντέλο. Παρόλα αυτά και στα δύο κριτήρια ο αλγόριθμος άρα και η τεχνική ομαδοποίησης (clustering) έχει λίγο καλύτερη απόδοση στην πρόβλεψη σε σχέση με τον αλγόριθμο, άρα και την τεχνική «δένδρα αποφάσεων».

5.5 Εξόρυξη 2ης Βάσης Δεδομένων

Για την δεύτερη βάση δεδομένων χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι αλλά με διαφορετικές παραμέτρους.

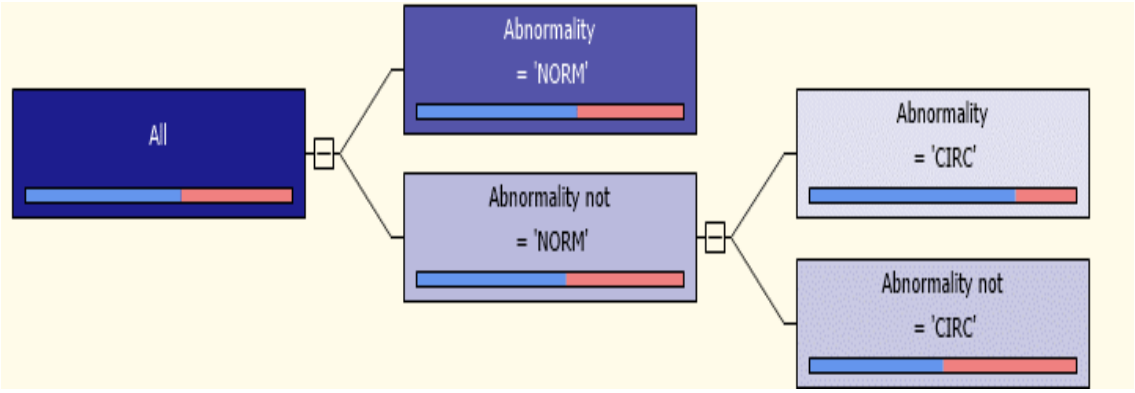
Αλγόριθμος Decision trees

Το testing data ήταν ίσο με 50%.

Οι τιμές των παραμέτρων, που διαμορφώθηκαν για το πιο επιτυχημένο μοντέλο για την συγκεκριμένη βάση με κριτήριο το ποσοστό πρόβλεψης είναι:

COMPLEXITY_PENALTY =0.1

Όλες οι υπόλοιπες παράμετροι σε προεπιλεγμένες τιμές. Επειδή οι τιμές στα πεδία αυτής της βάσης ήταν χαρακτήρες και όχι νούμερα θα παρατηρηθεί η όχι τόσο αποτελεσματική πρόβλεψη και των δύο μοντέλων. Σε αυτό επίσης οφείλεται και ο μικρός όγκος δεδομένων (329 εγγραφές). Τα αποτελέσματα από την μέθοδο «δένδρα αποφάσεων» φαίνονται παρακάτω:



Δένδρο Απόφασης 2ης Βάσης Δεδομένων

Το μπλε είναι καλοήθης όγκος και το κόκκινο ο κακοήθης όγκος.

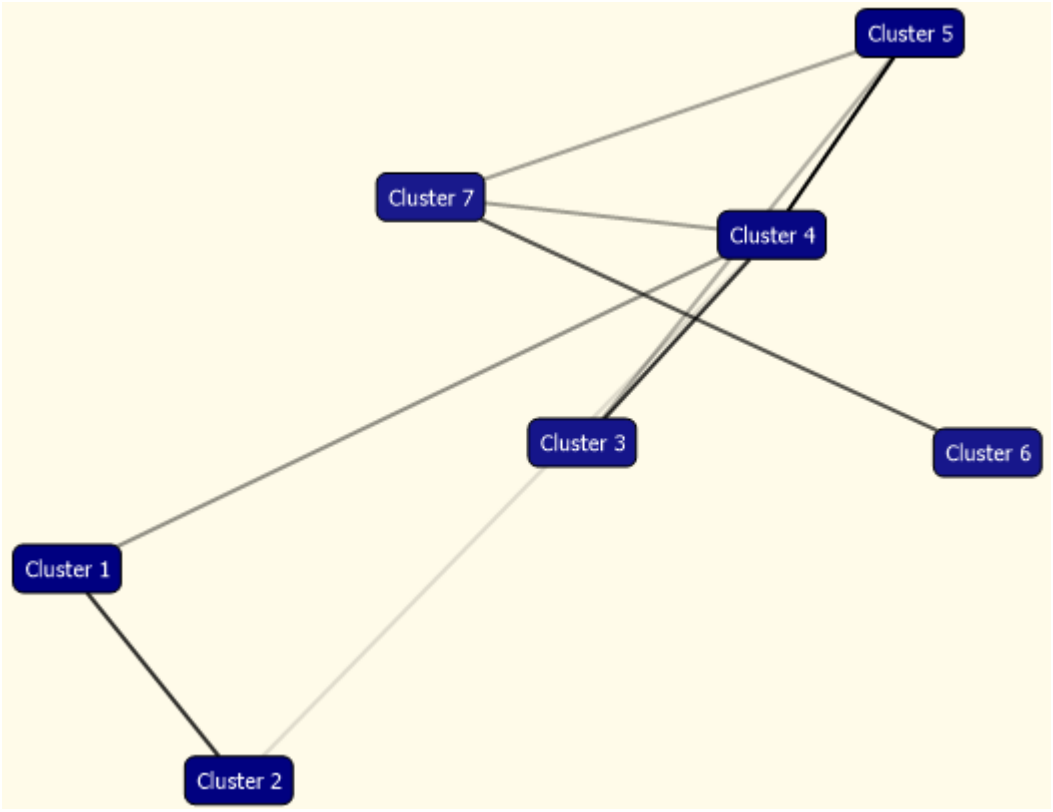
Αλγόριθμος Clustering

Το testing data ήταν ίσο με 50%.

Οι τιμές των παραμέτρων, που διαμορφώθηκαν για το πιο επιτυχημένο μοντέλο για την συγκεκριμένη βάση με κριτήριο το ποσοστό πρόβλεψης είναι:

CLUSTERING_METHOD=2

Όλες οι υπόλοιπες παράμετροι σε προεπιλεγμένες τιμές. Τα αποτελέσματα από την μέθοδο ομαδοποίηση φαίνονται παρακάτω:

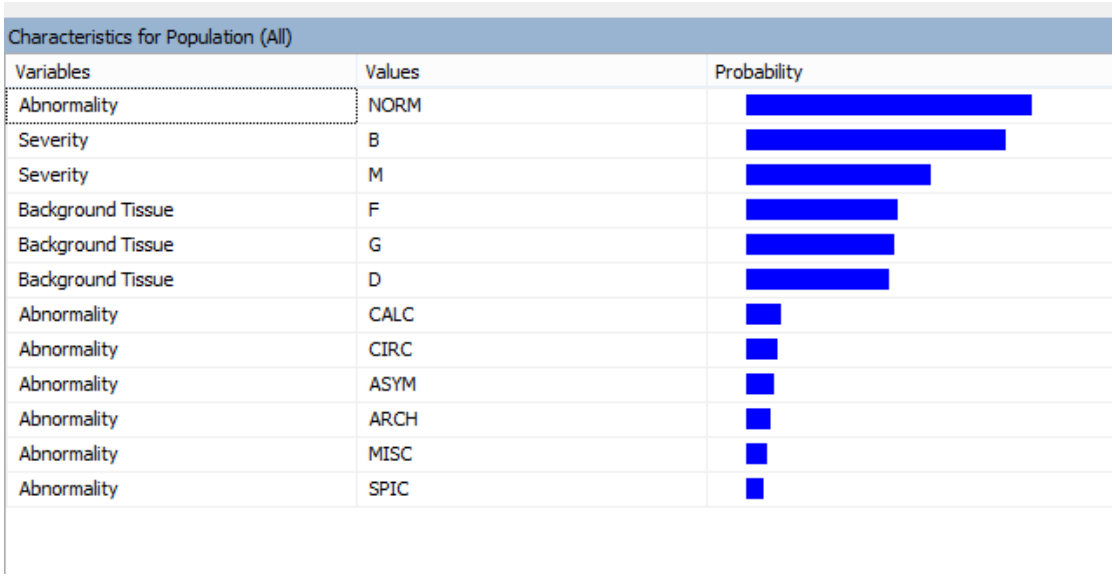


Clusters 2ης Βάσης Δεδομένων

Με έντονο μπλε είναι οι ομάδες-συστάδες οι οποίες έχουν μαζέψει κακοήθη όγκο. Στις υπόλοιπες ομάδες-συστάδες, όσο πιο αχνό είναι το μπλε τόσο περισσότερο καλοήθη όγκο έχουν στην ομάδα τους (Δυστυχώς στην φωτογραφία δεν διακρίνεται εύκολα ο διαχωρισμός αυτός). Τα κριτήρια και τα χαρακτηριστικά στα οποία διαχώρισε τις εγγραφές σε 8 ομάδες ο αλγόριθμος φαίνονται παρακάτω:

Attributes		Cluster profiles							
Variables	States	Populatio... Size: 161	Cluster 2 Size: 42	Cluster 6 Size: 39	Cluster 1 Size: 39	Cluster 5 Size: 14	Cluster 3 Size: 14	Cluster 7 Size: 8	Cluster 4 Size: 5
Abnormality	NORM CALC CIRC ASYM Other								
Background Tissue	F G D missing								
Severity	B M missing								

Πληροφορίες διαχωρισμού των clusters της 2ης Βάσης Δεδομένων



Characteristics for Population (All)

Variables	Values	Probability
Abnormality	NORM	
Severity	B	
Severity	M	
Background Tissue	F	
Background Tissue	G	
Background Tissue	D	
Abnormality	CALC	
Abnormality	CIRC	
Abnormality	ASYM	
Abnormality	ARCH	
Abnormality	MISC	
Abnormality	SPIC	

Πληροφορίες των clusters της 2ης Βάσης Δεδομένων

5.6 Σύγκριση αποτελεσμάτων 2ης Βάσης Δεδομένων

Στην συνέχεια έγινε σύγκριση των δύο τεχνικών για το ποια είναι καλύτερη.

Όπως έγινε η ανάλυση και για την 1η Βάση Δεδομένων ομοίως και για την 2η Βάση Δεδομένων θα έχουμε τα αντίστοιχα κριτήρια.

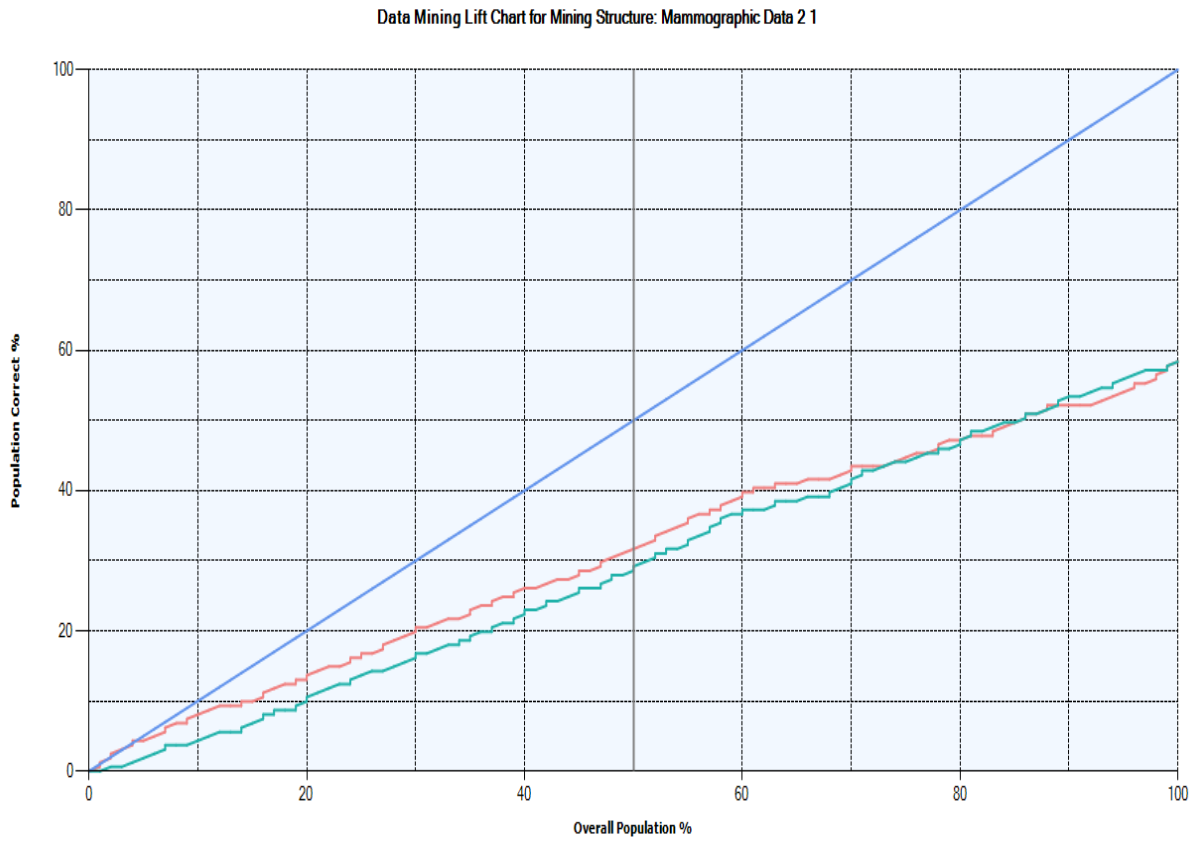
5.6.1 Lift Chart

Ο x-άξονας του γραφήματος αντιπροσωπεύει το ποσοστό του test dataset που χρησιμοποιείται για να συγκρίνει τις προβλέψεις. Ο άξονας y του γραφήματος αντιπροσωπεύει το ποσοστό των προβλέψεων που είναι σωστές.

Η διαγώνια ευθεία γραμμή, που φαίνεται παρακάτω σε μπλε, αντιπροσωπεύει το αποτέλεσμα της τυχαίας εικασίας, και είναι η βάση με την οποία αξιολογεί το lift. Η μπλε γραμμή δείχνει τα ιδανικά αποτελέσματα όπου δημιουργούν ένα μοντέλο το οποίο προβλέπει πάντα σωστά. Η κόκκινη γραμμή είναι η γραμμή πρόβλεψης του μοντέλου ομαδοποίησης και η πράσινη του μοντέλου δένδρα αποφάσεων.

Overall Score των δύο αλγορίθμων 2ης Βάσης Δεδομένων

Series Model	Score
TM_Clustering	0.58
TM_Decision_Tress	0.62
Ideal Model	1



Διάγραμμα lift chart σύγκρισης των δύο αλγορίθμων 2ης Βάσης Δεδομένων

Μπλε γραμμή : Ιδανικό μοντέλο

Κόκκινη γραμμή: Clustering μοντέλο

Πράσινη γραμμή :Decision Trees μοντέλο

Παρατηρούμε όπως ήταν αναμενόμενο, ότι η ευστοχία και των δύο μοντέλων δεν είναι τόσο καλή όσο στην προηγούμενη βάση εξαιτίας του μικρού όγκου δεδομένων (το $\frac{1}{3}$ σε σχέση με πριν) αλλά και των τιμών των πεδίων που είναι χαρακτήρες και όχι διακριτές τιμές.

5.6.2 Classification or confusion matrix

Στην συνέχεια παρατίθεται ο classification matrix των δύο αλγορίθμων:

Για το Clustering:

Classification matrix clustering 2ης Βάσης Δεδομένων

Predicted	B(Actual)	M(Actual)

B	94	67
M	0	0

Για το Decision Trees:

Classification matrix decision trees 2ης Βάσης Δεδομένων

Predicted	B(Actual)	M(Actual)
B	94	67
M	0	0

5.6.3 Cross Validation

Το τρίτο κριτήριο είναι το cross validation tool του οποίου τα αποτελέσματα φαίνονται παρακάτω:

Cross validation αποτελέσματα των δύο αλγορίθμων 2ης Βάσης Δεδομένων

Decision Trees				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	15	Classification	Pass	6
2	16	Classification	Pass	10
3	17	Classification	Pass	10
4	17	Classification	Pass	10
5	17	Classification	Pass	8
6	17	Classification	Pass	10
7	16	Classification	Pass	9
8	16	Classification	Pass	9
9	15	Classification	Pass	9
10	15	Classification	Pass	9
			Average	9,0311
			Standard Deviation	1,1659

1	15	Classification	Fail	9
2	16	Classification	Fail	6
3	17	Classification	Fail	7
4	17	Classification	Fail	7
5	17	Classification	Fail	9
6	17	Classification	Fail	7
7	16	Classification	Fail	7
8	16	Classification	Fail	7
9	15	Classification	Fail	6
10	15	Classification	Fail	6
			Average	7,1118
			Standard Deviation	1,0336
1	15	Likelihood	Log Score	-0,7132
2	16	Likelihood	Log Score	-0,664
3	17	Likelihood	Log Score	-0,6811
4	17	Likelihood	Log Score	-0,675
5	17	Likelihood	Log Score	-0,6947
6	17	Likelihood	Log Score	-0,6763
7	16	Likelihood	Log Score	-0,69
8	16	Likelihood	Log Score	-0,7086
9	15	Likelihood	Log Score	-0,7206
10	15	Likelihood	Log Score	-0,6702
			Average	-0,689

			Standard Deviation	0,0181
1	15	Likelihood	Lift	-0,0402
2	16	Likelihood	Lift	-0,0024
3	17	Likelihood	Lift	-0,0036
4	17	Likelihood	Lift	0,0025
5	17	Likelihood	Lift	-0,0172
6	17	Likelihood	Lift	0,0012
7	16	Likelihood	Lift	-0,0047
8	16	Likelihood	Lift	-0,0233
9	15	Likelihood	Lift	-0,0476
10	15	Likelihood	Lift	0,0028
			Average	-0,0128
			Standard Deviation	0,017
1	15	Likelihood	Root Mean Square Error	0,3982
2	16	Likelihood	Root Mean Square Error	0,4261
3	17	Likelihood	Root Mean Square Error	0,4093
4	17	Likelihood	Root Mean Square Error	0,4144

5	17	Likelihood	Root Mean Square Error	0,3949
6	17	Likelihood	Root Mean Square Error	0,4187
7	16	Likelihood	Root Mean Square Error	0,4144
8	16	Likelihood	Root Mean Square Error	0,4244
9	15	Likelihood	Root Mean Square Error	0,4424
10	15	Likelihood	Root Mean Square Error	0,4172
			Average	0,4158
			Standard Deviation	0,0128
Clustering				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	15	Classification	Pass	9
2	16	Classification	Pass	10
3	17	Classification	Pass	10
4	17	Classification	Pass	10
5	17	Classification	Pass	10
6	17	Classification	Pass	10

7	16	Classification	Pass	9
8	16	Classification	Pass	9
9	15	Classification	Pass	9
10	15	Classification	Pass	9
			Average	9,5217
			Standard Deviation	0,4995
1	15	Classification	Fail	6
2	16	Classification	Fail	6
3	17	Classification	Fail	7
4	17	Classification	Fail	7
5	17	Classification	Fail	7
6	17	Classification	Fail	7
7	16	Classification	Fail	7
8	16	Classification	Fail	7
9	15	Classification	Fail	6
10	15	Classification	Fail	6
			Average	6,6211
			Standard Deviation	0,4851
1	15	Likelihood	Log Score	-0,6942
2	16	Likelihood	Log Score	-0,665
3	17	Likelihood	Log Score	-0,6974
4	17	Likelihood	Log Score	-0,6713
5	17	Likelihood	Log Score	-0,6798
6	17	Likelihood	Log Score	-0,6768
7	16	Likelihood	Log Score	-0,6889
8	16	Likelihood	Log Score	-0,6945
9	15	Likelihood	Log Score	-0,6738
10	15	Likelihood	Log Score	-0,6706
			Average	-0,6813
			Standard Deviation	0,011
1	15	Likelihood	Lift	-0,0212
2	16	Likelihood	Lift	-0,0034
3	17	Likelihood	Lift	-0,0199
4	17	Likelihood	Lift	0,0062
5	17	Likelihood	Lift	-0,0023
6	17	Likelihood	Lift	0,0007
7	16	Likelihood	Lift	-0,0036

8	16	Likelihood	Lift	-0,0092
9	15	Likelihood	Lift	-0,0008
10	15	Likelihood	Lift	0,0024
			Average	-0,0051
			Standard Deviation	0,0086
1	15	Likelihood	Root Mean Square Error	0,4093
2	16	Likelihood	Root Mean Square Error	0,4138
3	17	Likelihood	Root Mean Square Error	0,4135
4	17	Likelihood	Root Mean Square Error	0,4089
5	17	Likelihood	Root Mean Square Error	0,4183
6	17	Likelihood	Root Mean Square Error	0,414
7	16	Likelihood	Root Mean Square Error	0,4024
8	16	Likelihood	Root Mean Square Error	0,4057
9	15	Likelihood	Root Mean Square Error	0,4131
10	15	Likelihood	Root Mean Square Error	0,4075
			Average	0,4107
			Standard Deviation	0,0045

Παρατηρούμε ότι αντίθετα με πριν ο αλγόριθμος «decision trees» είναι καλύτερος στην πρόβλεψη αποτελέσματος για την μικρή βάση δεδομένων. Ακόμα και στην μεγάλη είχε πολύ μικρή διαφορά σε σχέση με τον αλγόριθμο clustering (0.91 score decision trees, 0.92 score clustering). Αντίθετα σε μικρό δείγμα δεδομένων ο αλγόριθμος clustering έμεινε σε αρκετά πιο χαμηλό επίπεδο (0.58 score) σε σχέση με τον αλγόριθμο decision trees (0.62 score).

Κεφάλαιο 6

6.1 Στόχοι που επετεύχθησαν

Οι στόχοι οι οποίοι επετεύχθησαν στην παρούσα διπλωματική εργασία εκτείνονται σε όλο το φάσμα του διεπιστημονικού αντικειμένου της ιατρικής πληροφορικής. Πρώτος στόχος που επετεύχθη ήταν η αξιοποίηση του υλικού που βρισκόταν στο διαδίκτυο, από το οποίο δημιουργήθηκαν και στήθηκαν οι βάσεις δεδομένων. Επίσης, με την μεθοδολογική ανάλυση, που παρουσιάστηκε στην παρούσα διπλωματική εργασία έγινε πλήρης ανάλυση και εύρεση των καλύτερων χαρακτηριστικών παραμέτρων για τα εγκυρότερα αποτελέσματα πρόβλεψης των δύο αλγορίθμων που χρησιμοποιήθηκαν. Εφαρμόστηκαν ανεξάρτητα δύο γνωστές μέθοδοι εξόρυξης δεδομένων μεγάλης αξιοπιστίας, Δέντρα αποφάσεων (Decision Trees) και ομαδοποίηση (clustering). Έγινε εκτενής ανάλυση των χαρακτηριστικών και του τρόπου εκτέλεσης και λειτουργίας των δύο αλγορίθμων (decision trees και clustering) για την βαθύτερη κατανόησή τους. Τέλος εφαρμόστηκαν τεχνικές πρόβλεψης των δύο αλγορίθμων σε αληθινά ιατρικά στοιχεία του καρκίνου του μαστού και έγινε σύγκριση της επίδοσης και των αποτελεσμάτων των δύο αλγορίθμων με πολλαπλά κριτήρια. Με τον πειραματισμό και τις δοκιμές εξήχθη ένα αξιόπιστο αποτέλεσμα χαρακτηριστικών παραμέτρων για τον κάθε αλγόριθμο, καθώς επίσης και για την απόδοσή και την αξιοπιστία τους σε σχέση με τις ιατρικές εκτιμήσεις των ακτινολόγων.

6.2 Μελλοντικές επεκτάσεις

Μια πρώτη προοπτική επέκτασης θα μπορούσε να είναι η ανάπτυξη μιας βάσης δεδομένων στην οποία θα συλλέγονται όχι μόνο στοιχεία του καρκίνου του μαστού αλλά και άλλων ειδών καρκίνου. Έτσι οι αλγόριθμοι θα μπορούσαν να παρουσιάσουν πιο εκτενή και έγκυρα αποτελέσματα καθώς θα υπήρχε η δυνατότητα διασταύρωσης των συμπερασμάτων.

Μια δεύτερη προοπτική επέκτασης αποτελεί σίγουρα η παροχή περισσότερων ιατρικών δεδομένων. Εάν οι βάσεις είχαν περισσότερες εγγραφές θα μπορούσαμε να δούμε και την αποτελεσματικότητα των αλγορίθμων με την πρόβλεψή τους να πλησιάζει το ιδανικό μοντέλο ακόμα καλύτερα. Η προοπτική αυτή φάνηκε και στα αποτελέσματα της $2^{ης}$ Βάσης Δεδομένων, όπου, επειδή οι εγγραφές αντιπροσώπευαν ποσοστό $\frac{1}{3}$ σε σχέση με την $1^{η}$ Βάση Δεδομένων οι αλγόριθμοι δεν είχαν αρκετά δεδομένα για να εκπαιδευτούν και να μάθουν τις ιδιαιτερότητες του καρκίνου με αποτέλεσμα τα ποσοστά πρόβλεψης να είναι πολύ χαμηλότερα σε σχέση με την $1^{η}$ Βάση Δεδομένων. Όταν παρέχεται στον αλγόριθμο ένας ικανός αριθμός δεδομένων, ώστε να μπορεί να «δει» την πρόβλεψη για να «εκπαιδευτεί» και να «μάθει» τους συσχετισμούς της Βάσης, είναι πολύ πιο αξιόπιστος στις προβλέψεις του.

Μια άλλη προοπτική μελλοντικής επέκτασης αποτελεί και η περαιτέρω ανάλυση των δεδομένων με άλλες μεθόδους και αλγόριθμους, πέραν αυτών που παρουσιάζονται

στην παρούσα διπλωματική εργασία για την εύρεση των πιθανών πηγών λανθασμένης διάγνωσης. Ένα ψευδώς αρνητικό (όπως, η πρόβλεψη ύπαρξης καλοήθη όγκου ενώ στην πραγματικότητα πρόκειται για κακοήθη) αποτέλεσμα μπορεί να έχει θανατηφόρες συνέπειες ενώ τα ψευδώς θετικά (όπως, πρόβλεψη περί κακοήθειας ενώ στην πραγματικότητα είναι καλοήθης ο όγκος), οδηγούν τόσο σε ανώφελη επιβάρυνση του ασθενούς θέτοντας σε κίνδυνο την υγεία του, όσο και σε ανώφελο ψυχολογικό βάρος αλλά και σε οικονομική επιβάρυνσή του. Επίσης θα μπορούσαν να εξεταστούν και οι περιπτώσεις, οι οποίες χαρακτηρίζονται, ως ύποπτες και τυγχάνουν διαχείρισης, ως κακοήθειες καταλήγοντας σε χειρουργικές επεμβάσεις.

Μία άλλη τέλος προοπτική μελλοντικής επέκτασης θα μπορούσε να είναι μια πιο πολύπλοκη αναζήτηση του αλγορίθμου ή ένας τελείως διαφορετικός τρόπος εξαγωγής προβλέψεων, όπου στο μέλλον θα μπορούσαν να δημιουργηθούν καινούργιοι αλγόριθμοι, ακόμα πιο εύστοχοι στις προβλέψεις τους εξαιτίας του μαθηματικού τους υπόβαθρου και του διαφορετικού τρόπου εύρεσης και δημιουργίας των συσχετισμών.

Βιβλιογραφία

[1] http://el.Wikipedia.org/wiki/Μαστός

[2] http://www.eurocytology.eu/static/eurocytology/gre/breast/mod5contA.html

[3] http://www.karkinos24.gr/index.php/karkinostoumastou

[4] http://artemis-new.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/5118/1/DT2010-0004.pdf

[5]http://el.wikipedia.org/wiki/Μαστογραφία

[6] http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass

Source:

Matthias Elter Fraunhofer Institute for Integrated Circuits (IIS) Image Processing and Medical Engineering Department (BMT) Am Wolfsmantel 33 91058 Erlangen, Germany matthias.elter'@'iis.fraunhofer.de (49)9131-7767327

Prof. Dr. Rüdiger Schulz-Wendtland Institute of Radiology, Gynaecological Radiology, University Erlangen-Nuremberg Universitätsstraße 21-23 91054 Erlangen, Germany

[7] Gulta Rahbar, Angela C. Sie, Gail C. Hansen, Jeffrey S. Prince, Michelle L. Melany, Handel E. Reynolds, Valerie P. Jackson, James W. Sayre, and Lawrence W. Bassett, Benign versus Malignant Solid Breast Masses: US Differetiation Radiology 1999; 21(3):889-94.

[8] http://www.acr.org/Quality-Safety/National-Radiology-Data-Registry/National-Mammography-DB

[9]The American College of Radiology BI-RADS® ATLAS and MQSA (http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/BIRADS/BIRADSFAQs.pdf)

[10] http://peipa.essex.ac.uk/info/mias.html
J Suckling *et al* (1994): *The Mammographic Image Analysis Society Digital Mammogram Database* Exerpta Medica. International Congress Series 1069 pp375-378.

[11] "IBM What is big data? — Bringing big data to the enterprise". www.ibm.com. Retrieved 2013-08-26.

[12]http://www.plasticsurgery-plasis.gr/index.php?option=com_content&view=article&id=96&Itemid=86&lang=el

[13] http://www.webopedia.com/TERM/D/database.html

[14] Lena Costaridou, Medical Image Analysis Methods, Taylor and Francis Group, 2005.

[15] Bin Zheng, Yuan-Hsiang Chang, Walter F. Good and David Gur, Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme, Acad Radiol 1997; 4:497-502.

[16] http://technet.microsoft.com/en-us/library/bb895170

[17] http://technet.microsoft.com/en-us/library/bb895174.aspx

[18]http://el.wikipedia.org/wiki/Καρκίνος_του_μαστού

[19] World Health Organization, http://www.who.int/mediacentre/factsheets/fs297/en/index.html

[20] American Cancer Society

[21] Breast Cancer, The Merk Manual of Medical Information.

[22] Joseph Y. Lo, Marios Gavrielides, Mia K. Markey, Jonathan L. Jesneck, Computeraided classification of breast microcalcification clusters: Merging of features from image processing and radiologists, Proceedings of SPIE 2003; 5032:882-9.

[23] Timothy W. Freer, Michael J. Ulissey, Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center, Radiology 2001; 22:781-6.

[24] Bin Zheng, Lara A. Hardesty, William R. Poller, Jules H. Sumkin, Sara Golla, Mammography with Computer-aided Detection: Reproducibility Assessment - Initial Experience, Radiology 2003; 228:58-62.

[25] Corinne Balleyguier, Salma Ayadi, Kim Van Nguyen, Daniel Vanel, Clarisse Dromain, and Robert Sigal, BIRADS(TM) classification in mammography, European Journal of Radiology 2007; 61:192-4.

[26] Leconte I, Feger C, Galant C, et al., Mammography and subsequent whole-breast sonography of nonpalpable breast cancers: the importance of radiologic breast density,American Journal of Roentgenology 2003; 180:1675-9.

[27]Orel SG, Kay N, Reynolds C, Sullivan DC, BI-RADS categorization as a predictor of malignancy, Radiology 1999; 211:845-50.

[28] Mendez A, Cabanillas F, Echenique M, Malekshamran K, Perez I, Ramos E., Evaluation of Breast Imaging Reporting and Data System Category 3 mammograms and the use of stereotactic vacuum-assisted breast biopsy in a nonacademic community practice, Cancer 2004; 100:710–4.

[29] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.

[30] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.

[31] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[32] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[33] http://en.wikipedia.org/wiki/Data_mining#Medical_data_mining

[34] David G. Savage (2011-06-24). "Pharmaceutical industry: Supreme Court sides with pharmaceutical industry in two decisions". *Los Angeles Times*. Retrieved 2012-11-07.

[35] Bao, HO Tu. Introduction to knowledge discovery and data mining, course, Institute of βInformation Technology National Center for Natural Science and Technology, http://www.ebook.edu.vn/?page=1.9&view=169 4, [Web Accessed 22th January 2008].

[36] MacLennan, ZhaoHui Tang and Jamie. Data Mining with SQL Server 2005. s.l.: Wiley Publishing, Inc., 2005.

[37] Mehmed Kantardzic. Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons © 2003.

[38] Oded Maimon and Lior Rokach. Data mining and knowledge discovery handbook. Tel-Aviv University, Israel, Springel 2005.

[39] Nada Lavrac. Selected techniques for data mining in medicine. Department of Intelligent Systems, J. Stefan Institute, 1000 Ljubljana, Slovenia.

[40] Βαζιργιάννης Μ. – Χαλκίδη Μ., Εξόρυξη γνώσης από βάσεις δεδομένων, Τυπωθήτω, Αθήνα 2003.

[41] Jiawei Han and Kamber Micheline, Data mining concepts and techniques, Simon Fraser University, Academic Press, USA 2001

[42]Advanced Data Mining Techniques Olson, D.L.; Delen, D. 2008 ISBN:978-3-540-76916-3

[43] http://technet.microsoft.com/en-us/library/ms175595.aspx

[44]Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711.

[45] Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers

[46] Barros R. C., Cerri R., Jaskowiak P. A., Carvalho, A. C. P. L. F., A bottom-up oblique decision tree induction algorithm. Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011).

[47] Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

[48] Breiman, L. (1996). Bagging Predictors. "Machine Learning, 24": pp. 123-140.

[49] Friedman, J. H. (1999). *Stochastic gradient boosting.* Stanford University.

[50] Hastie, T., Tibshirani, R., Friedman, J. H. (2001). *The elements of statistical learning : Data mining, inference, and prediction.* New York: Springer Verlag.

[51] Rodriguez, J.J. and Kuncheva, L.I. and Alonso, C.J. (2006), Rotation forest: A new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10):1619-1630.

[52] Kass, G. V. (1980). "An exploratory technique for investigating large quantities of categorical data". *Applied Statistics* **29** (2): 119–127. doi:10.2307/2986296. JSTOR 2986296

[53] Lior Rokach Department of Industrial Engineering Tel-Aviv University Chapter 9 Decision trees: http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

[54]Structure and parameter Learning for casual independence and casual interaction models by Christopher Meek and David Heckerman: http://research.microsoft.com/en-us/um/people/heckerman/lNoisyOr.pdf

[55] Microsoft Decision Trees Algorithm Technical Reference: http://technet.microsoft.com/en-us/library/cc645868.aspx

[56] Learning Bayesian Networks: The Combination of Knowledge and Statistical Data by D. Heckerman, D. Geiger, and D.M. Chickering March 1995

[57] Autoregressive Tree Models for Time-Series Analysis by C. Meek, D.M. Chickering,and D. Hecherman

[58] New Trends in Data Mining by J. HUYSMANS, B. BAESENS, D. MARTENS, K. DENYS and J. VANTHIENEN

(http://www.econ.kuleuven.be/rebel//jaargangen/2001-2010/2005/TEM%202005-4/TEM_4_05_Huysmans.pdf)

[59] Krzysztof j. Giosa, G. William Mooree, Uniqueness of medical data mining, Joornal of Artificial intelligence in medicine, 2002.

[60] Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R B., Richardson, W. S., Evidence based medicine: what it is and what it isn't. BMJ, 312 (7023), 71-2, 2004.

[61] Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., Chute, C. G., Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc, 15(1), 25-8, 2008.

[62] Cimiano, A., Hoto, A., Staab, S., Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research, 24, 305-339, 2005.

[63] Montani, S., Portinale, L., Leonardi, G., Bellazzi, R., Case-based retrieval to support the treatment of end stage renal failure patients. Artif Intell Med, 37(1), 31-42, 2006.

[64] Bellazzi, R., Larizza, C., Magni, P., Bellazzi, R., Temporal data mining for the quality assessment of hemodialysis services. Artif Intell Med, 34(1), 25-39, 2005.

[65] Raj, R., O'Connor, M. J., Das, A. K., An Ontology-Driven Method for Hierarchical Mining of Temporal Patterns: Application to HIV Drug Resistance Research. AMIA Symp., 2008.

[66] Heinze, D. T., Morsch, M. L., Potter, B. C., Sheffer, R.E Jr., Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology. J Am Med Inform Assoc, 15(1), 40-3, 2008.

[67] Petr Berka, Jan Rauch, Djamel Abdelkader Zighed, Data Mining and Medical Knowledge Management: Cases and Applications, Information Science Reference, Hershey, USA, 2009.

[68] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/clustering.htm

[69] http://technet.microsoft.com/en-us/library/ms174879.aspx

[70]Clustering SDSC Summer Institute 2012 Natasha Balac, Ph.D.

[71] Συστήματα Βάσεων Δεδομένων «η πλήρης θεωρία των Βάσεων Δεδομένων» 4η έκδοση Silberschatz ,Korth, Sudarshan