ΠΕΡΙΛΗΨΗ

Η μελέτη αυτή αφορά την αξιολόγηση κανόνων συσχέτισης και κατηγοριοποίησης βάσει πολλαπλών μέτρων. Τα μέτρα ενδιαφέροντος των κανόνων αναφέρονται στους κανόνες που είναι της μορφής $A \rightarrow B$. Όλα τα μέτρα μπορούν να παίξουν σημαντικό ρόλο στην επιλογή των κανόνων που έχουν ενδιαφέρον από ένα σύνολο δεδομένων. Γι' αυτό είναι σημαντικό στο στάδιο της εξαγωγής των κανόνων, να λαμβάνεται υπόψη ένας συνδυασμός από αυτά τα μέτρα.

Έχουν χρησιμοποιηθεί κάποια αντικειμενικά μέτρα όπως η Υποστήριξη, η Εμπιστοσύνη, η Κάλυψη, η Πεποίθηση κτλ. καθώς επίσης και κάποια στατιστικά μέτρα όπως το χ^2 (Chi-Square), p-Value, Framingham Event Risk και το ποσοστό σωστής ταξινόμησης των κανόνων (Correct Classified Instances), τα οποία μας παρέχουν πληροφορίες σχετικά με τη σημαντικότητα του κάθε κανόνα καθώς και τον κίνδυνο για τυχόν νέο επεισόδιο του ασθενή.

Για τη μελέτη αυτή έχει χρησιμοποιηθεί μία βάση δεδομένων με ασθενείς με καρδιαγγειακά επεισόδια. Οι τιμές των μέτρων ενδιαφέροντος των κανόνων που εξάχθηκαν από τις μεθόδους συσχέτισης και κατηγοριοποίησης έχουν μελετηθεί και κωδικοποιηθεί έτσι ώστε να δημιουργήσουμε ένα δέντρο απόφασης βασισμένο στα μέτρα. Με τη δημιουργία του δέντρου λαμβάνουμε γνώση σχετικά με το ποια μέτρα παρουσιάζουν περισσότερο ενδιαφέρον. Στη συνέχεια φιλτράρουμε τους αρχικούς κανόνες που έχουν εξαχθεί λαμβάνοντας υπόψη αυτά τα μέτρα.

Ο κύριος στόχος μας είναι εξάγοντας τους φιλτραρισμένους κανόνες παρέχοντας επιπρόσθετα πληροφορίες σχετικά με τη στατιστική σημαντικότητα καθώς επίσης και τον κίνδυνο που διατρέχει ο κάθε ασθενής για ενδεχόμενο επεισόδιο, να αυξήσουμε την ιατρική γνώση του ιατρού με απώτερο σκοπό τη μείωση του αριθμού των θανάτων ασθενών με τέτοιου είδους καρδιολογικά προβλήματα.

ΑΞΙΟΛΟΓΗΣΗ ΚΑΝΟΝΩΝ ΒΑΣΕΙ ΠΟΛΛΑΠΛΩΝ ΜΕΤΡΩΝ

Κουμπάρου Νικόλας

Η Διατριβή αυτή

Υποβλήθηκε προς Μερική Εκπλήρωση των

Απαιτήσεων για την Απόκτηση

Τίτλου Σπουδών Master

σε Προηγμένες Τεχνολογίες Πληροφορικής

στο

Πανεπιστήμιο Κύπρου

Συστήνεται προς Αποδοχή

από το Τμήμα Πληροφορικής

Ιούνιος, 2010

ΣΕΛΙΔΑ ΕΓΚΡΙΣΗΣ

Διατριβή Master

ΑΞΙΟΛΟΓΗΣΗ ΚΑΝΟΝΩΝ ΒΑΣΕΙ ΠΟΛΛΑΠΛΩΝ ΜΕΤΡΩΝ

Παρουσιάστηκε από

Κουμπάρου Νικόλα

Ερευνητικός Σύμβουλος	Κωνσταντίνος Σ. Παττίχης
	Όνομα Ερευνητικού Συμβούλου
Μέλος Επιτροπής	Χρίστος Ν. Σχίζας
	Όνομα Μέλους Επιτροπής
Μέλος Επιτροπής	Χρίστος Χριστοδούλου
	Όνομα Μέλους Επιτροπής

Πανεπιστήμιο Κύπρου

Ιούνιος, 2010

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω όσους έχουν συμβάλει άμεσα ή έμμεσα στην διεκπεραίωση της διατριβής μου, που έχω αναλάβει προς εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου master. Αρχικά ευχαριστώ τον επιβλέποντα καθηγητή μου Δρ. Κωνσταντίνο Παττίχη όπως επίσης και τον κύριο Μηνά Καραολή για την βοήθεια και τη σωστή καθοδήγηση που μου πρόσφεραν οποιαδήποτε στιγμή την χρειάστηκα καθ' όλη την διάρκεια της εκπόνησης αυτής της μελέτης. Θα ήθελα να ευχαριστήσω επίσης και τον καρδιολόγο Δρ. Ιωσήφ Μαντίρη, από το Καρδιολογικό Τμήμα του Γενικού Νοσοκομείου Πάφου, από τον οποίο έχουμε λάβει την βάση δεδομένων.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω και τις κυρίες Λουκία Παπακωνσταντίνου και Δήμητρα Χατζηπαναγή για την άδεια που μου έδωσαν να χρησιμοποιήσω τις εφαρμογές που έχουν υλοποιήσει για τους αλγόριθμους Συσχέτισης και Κατηγοριοποίησης αντίστοιχα.

Ένα μεγάλο ευχαριστώ στην αρραβωνιαστικιά μου για την υποστήριξη και τη δύναμη που μου έδινε κατά την εκπόνηση της διπλωματικής μου εργασίας.

Τελευταίους άφησα τους γονείς και την οικογένεια μου που στάθηκαν στο πλευρό μου καθ' όλη τη διάρκεια της εκπόνησης αυτής της διπλωματικής εργασίας. Ένα μεγάλο ευχαριστώ για όλη τη βοήθεια, εμψύχωση και συμπαράσταση που μου παρέχουν. Αφιερωμένο.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΜΕ ΠΙΝΑΚΕΣ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Οι καρδιακές παθήσεις είναι μια από τις πιο συχνές αιτίες θανάτου στον κόσμο και μία σημαντική αύξηση έχει παρατηρηθεί και στην Κύπρο [1]. Το έμφραγμα του μυοκαρδίου, κοινώς γνωστό ως καρδιακό επεισόδιο, συμβαίνει όταν μπλοκαριστεί η ροή του αίματος σε ένα τμήμα του καρδιακού μυός λόγω δημιουργίας θρόμβου σε μια στεφανιαία αρτηρία. Αν η ροή του αίματος στο τμήμα του καρδιακού μυός δεν μπορεί να αποκατασταθεί γρήγορα, το τμήμα του καρδιακού μυός δεν τροφοδοτείται με οξυγόνο και αρχίζει να πεθάνει [2].

Κλασικά συμπτώματα ενός καρδιακού επεισοδίου είναι ο ξαφνικός πόνος στο στήθος, η δύσπνοια, η ναυτία, ο εμετός, η ταχυπαλμία, η εφίδρωση, και η ανησυχία. Οι γυναίκες μπορεί να εμφανίσουν λιγότερα τυπικά συμπτώματα από τους άνδρες, όπως δυσκολίες στην αναπνοή, αδυναμίες, αισθήματα δυσπεψίας και κόπωσης. Περίπου το ένα τέταρτο των ασθενών με έμφραγμα του μυοκαρδίου δεν παρουσιάζουν κάποια συμπτώματα όπως πόνο στο στήθος ή άλλα συμπτώματα που αναφέρθηκαν. Μια καρδιακή προσβολή είναι μια επείγουσα ιατρική κατάσταση, και οι άνθρωποι που νιώθουν πόνο στο στήθος πρέπει να επισκεφτούν επειγόντως καρδιολόγο διότι η άμεση θεραπεία μπορεί να είναι ζωτικής σημασίας για τη επιβίωση του ασθενή [3].

Στην Κύπρο, ο αριθμός των ασθενών με καρδιακές παθήσεις αυξάνεται κάθε χρόνο. Το 2007, περίπου 64 χιλιάδες άτομα επισκέφθηκαν τα δημόσια νοσοκομεία και το 14% αυτών των ασθενών παρουσίασαν πρόβλημα στην κυκλοφορία του αίματος. Περίπου 8 χιλιάδες από

αυτούς εμφάνισαν προβλήματα ροής του αίματος σε κάποιο από τα τμήματα της καρδίας. Οι περισσότεροι από αυτούς τους ασθενείς, χωρίς χειρουργική επέμβαση αλλά με φαρμακευτική αγωγή, πέτυχαν βελτίωση της κατάστασής τους. Για ένα μικρό ποσοστό των περιπτώσεων η λειτουργία της καρδιάς έχει αποκατασταθεί πλήρως, ενώ ένα άλλο ποσοστό δεν είχε καμία αλλαγή στην κατάστασή τους. Δυστυχώς, το 2007, 560 ασθενείς απεβίωσαν λόγω της καρδιακής νόσου, επτά από τους οποίους έχασαν τη ζωή τους κατά τη διάρκεια της εγχείρησης [4].

Στην προσπάθεια μας για μείωση του μεγάλου αριθμού θυμάτων που πεθαίνουν λόγω καρδιακής ανακοπής, θα επικεντρωθούμε στη μελέτη και των εντοπισμό των κύριων παραγόντων που προκαλούν καρδιακό έμφραγμα αλλά κυρίως των λόγων που οδηγούν σε έμφραγμα του μυοκαρδίου. Αυτό γίνεται με τη μελέτη των καρδιακών περιπτώσεων που παρατηρήθηκαν στο Γενικό Νοσοκομείο Πάφου.

1.2 Εξόρυξη Δεδομένων (Data Mining)

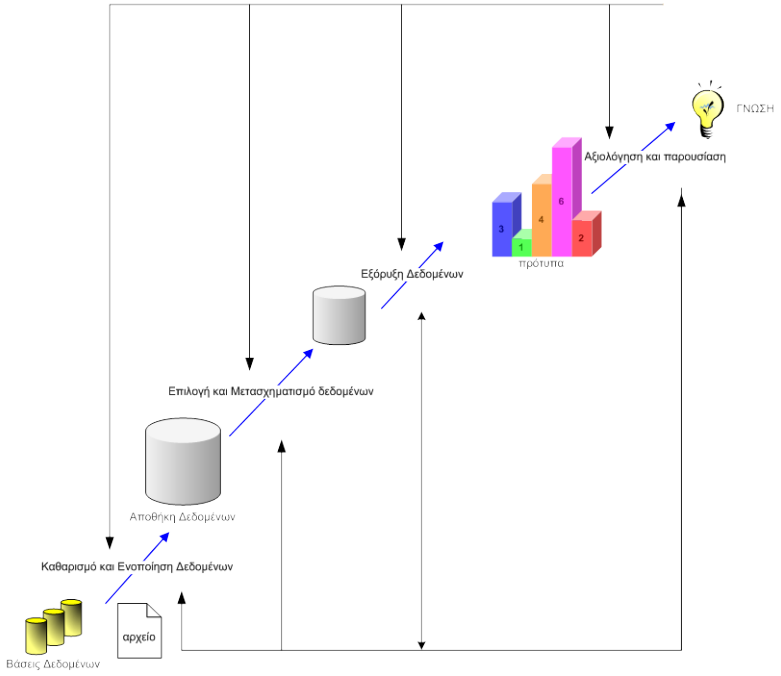
Η εξόρυξη δεδομένων είναι η μέθοδος για εξαγωγή ή «εξόρυξη» γνώσης από μεγάλο όγκο δεδομένων. Η εξόρυξη δεδομένων έχει προσελκύσει το ενδιαφέρον στη βιομηχανία της πληροφορίας και στην κοινωνία γενικά τα τελευταία χρόνια, λόγω της ευρείας διαθεσιμότητας τεράστιων ποσών δεδομένων και της επικείμενης ανάγκης για μετατροπή τους σε χρήσιμες πληροφορίες και γνώση. Οι πληροφορίες και η γνώση που εξάγονται μπορούν να χρησιμοποιηθούν για διάφορες εφαρμογές, όπως ανάλυση αγοράς, ανίχνευση απάτης, διατήρηση πελατών, έλεγχο παραγωγής και άλλα.

Για την εξόρυξη δεδομένων από μια τεράστια βάση δεδομένων, θα πρέπει να ακολουθηθούν κάποια βήματα, που είναι ευρέως γνωστά με τον όρο, *Ανακάλυψη Γνώσης από Βάση Δεδομένων (Knowledge Discovery from Data ή KDD)* [5], τα οποία θα προετοιμάσουν τη βάση δεδομένων για την εξόρυξη χρήσιμης γνώσης. Η διαδικασία KDD είναι επαναληπτική. Από κάθε βήμα μπορεί κανείς να μεταπηδήσει σε οποιοδήποτε προγενέστερο

βήμα. Η ροή των βημάτων είναι απεικονισμένη στο Σχήμα 1.1. Παρόλο που το κάθε βήμα της εξόρυξης δεδομένων αποτελεί μια κύρια εργασία στη διαδικασία εξόρυξης γνώσης, όλα τα βήματα είναι εξίσου σημαντικά για τη σωστή και επιτυχή εφαρμογή της τεχνικής KDD.

Τα επαναληπτικά βήματα της τεχνικής KDD είναι τα ακόλουθα:

1. *Καθαρισμός Δεδομένων (Data Cleaning)*: Αφαίρεση θορύβου, δεδομένων με ελλιπείς τιμές και ασυνεπών δεδομένων.
2. *Ολοκλήρωση Δεδομένων (Data Intergation)*: Διάφορες πηγές δεδομένων μπορούν να ενωθούν μαζί.
3. *Επιλογή Δεδομένων (Data Selection)*: Ανακτώνται δεδομένα από τη βάση δεδομένων, που θεωρούνται χρήσιμα, σχετικά με το στόχο ανάλυσης
4. *Μετασχηματισμός Δεδομένων (Data Transformation)*: Τα δεδομένα μετασχηματίζονται ή μορφοποιούνται σε κατάλληλες μορφές για την εξόρυξη δεδομένων με την εκτέλεση διάφορων διαδικασιών όπως μέθοδος συνάθροισης (aggregation), σύνοψης (summary) κτλ.
5. *Εξόρυξη Δεδομένων (Data Mining)*: Μια ουσιαστική διαδικασία κατά την οποία εφαρμόζονται ευφυείς μέθοδοι, αλγόριθμοι, έτσι ώστε να εξαχθούν πρότυπα δεδομένων.
6. *Αξιολόγηση Προτύπων (Pattern Evaluation)*: Η διαδικασία η οποία θα καθορίσει τα πρότυπα με πραγματικό ενδιαφέρον που απεικονίζουν την γνώση βάσει διάφορων μέτρων.
7. *Παρουσίαση Γνώσης (Knowledge Representation)*: Χρησιμοποιούνται διάφορες τεχνικές απεικόνισης και παρουσίασης γνώσης για να παρουσιάσουν στο χρήστη την γνώση που εξάχθηκε [6].



Σχήμα 1.1: Τα βήματα της Διαδικασίας KDD (από J. Han, M. Kamber, 'Data Mining, Concepts and Techniques', Morgan Kaufman Publishers, Academic Press, 2001).

Το είδος των προτύπων του στόχου της εξόρυξης δεδομένων καθορίζεται από την διαδικασία - αλγόριθμο εξόρυξης δεδομένων που χρησιμοποιείται. Γενικά, τα πρότυπα εξόρυξης δεδομένων μπορούν να ταξινομηθούν σε δύο κατηγορίες: τα *περιγραφικά (descriptive)* και τα *προβλεπτικά (predictive)*. Τα περιγραφικά πρότυπα χαρακτηρίζουν τις γενικές ιδιότητες των δεδομένων της βάσης δεδομένων, ενώ τα προβλεπτικά πρότυπα βγάζουν κάποιο συμπέρασμα από τα υπάρχοντα δεδομένα προκειμένου να κάνουν προβλέψεις.

Για εξόρυξη δεδομένων έχουν προταθεί διάφορες τεχνικές ανάλυσης δεδομένων, όπως η εξόρυξη συχνών προτύπων (frequent pattern), η κατηγοριοποίηση (classification), η συσταδοποίηση (clustering), και η εξόρυξη outlier. Η *κατηγοριοποίηση* (*classification*) είναι η διαδικασία για να βρεθούν μοντέλα – λειτουργίες που να περιγράφουν και να διακρίνουν τις κλάσεις των δεδομένων, με σκοπό να χρησιμοποιηθούν τα μοντέλα αυτά για να προβλέψουν την άγνωστη κλάση κάποιου αντικειμένου.

Αντίθετα από την ταξινόμηση, η *συσταδοποίηση* (*clustering*) αναλύει τα δεδομένα χωρίς να λαμβάνει υπόψη μια γνωστή ετικέτα κλάσης. Οι ετικέτες κλάσης, στην ομαδοποίησης, δεν είναι γνωστές στα δεδομένα κατάρτισης κι έτσι η ομαδοποίηση μπορεί να χρησιμοποιηθεί για να παραγάγει τέτοιες ετικέτες.

Πολλές φορές, μια βάση δεδομένων μπορεί να περιέχει δεδομένα που να μην συμμορφώνονται σε ένα γενικό μοντέλο συμπεριφοράς ή κάποιο πρότυπο δεδομένων. Αυτά τα δεδομένα είναι ιδιαίτερα (outliers). Οι περισσότερες μέθοδοι εξόρυξης δεδομένων απορρίπτουν τα outliers ως θόρυβο ή εξαιρέσεις. Εντούτοις, σε μερικές εφαρμογές όπως η ανίχνευση απάτης, τα σπάνια γεγονότα μπορούν να είναι πιο ενδιαφέροντα από τα συχνά εμφανιζόμενα. Η ανάλυση των outlier δεδομένων ονομάζεται εξόρυξη outlier.

Τα πιο *συχνά πρότυπα (frequent pattern)*, είναι τα πρότυπα που εμφανίζονται περισσότερες φορές στα δεδομένα. Υπάρχουν διάφορα είδη συχνών προτύπων, τα συχνά σύνολα αντικειμένων (frequent itemset), τα ακολουθιακά πρότυπα (sequential patterns) και οι υποδομές (substructure). Το συχνό σύνολο αντικειμένων (frequent itemset) αναφέρεται σε ένα σύνολο από αντικείμενα που εμφανίζονται συχνά μαζί σε ένα σύνολο δεδομένων δοσοληψίας, παραδείγματος χάρη κάποιος που αγοράζει γάλα αγοράζει και ψωμί. Η συχνά εμφανιζόμενη ακολουθία, όπως το πρότυπο όπου ο πελάτης τείνει να αγοράζει πρώτα έναν υπολογιστή, μετά μία ψηφιακή φωτογραφική, και μετά μια κάρτα μνήμης, είναι ένα (συχνό) ακολουθιακό πρότυπο (sequential pattern). Μία υποδομή μπορεί να αναφερθεί σε διάφορες μορφές δομών, όπως γράφους, δέντρα που μπορούν να συνδυαστούν με τα σύνολα αντικειμένων ή με ακολουθίες. Εάν μια υποδομή εμφανίζεται συχνά, τότε ονομάζεται (συχνό) πρότυπο δομής. Η εξόρυξη συχνών προτύπων οδηγεί στην ανακάλυψη ενδιαφέρων σχέσεων και συσχετίσεων στα δεδομένα [6].

1.3 Κανόνες Συσχέτισης (Association Rules)

Η εξόρυξη συχνών προτύπων οδηγεί στην ανακάλυψη ενδιαφέρων σχέσεων και συσχετίσεων στα δεδομένα. Συνεχώς, τεράστιες ποσότητες δεδομένων συλλέγονται και αποθηκεύονται, και οι κανόνες συσχέτισης παρέχουν ένα συνοπτικό τρόπο για να

εκφραστούν χρήσιμες πληροφορίες, που γίνονται εύκολα κατανοητές από τους χρήστες. Έτσι πολλές βιομηχανίες ενδιαφέρονται να εξαγάγουν κανόνες συσχέτισης από τις βάσεις δεδομένων τους. Η ανακάλυψη ενδιαφέρον συσχετίσεων μεταξύ τεράστιων ποσών από εγγραφές επιχειρησιακών δοσοληψιών μπορούν να βοηθήσουν σε πολλές διαδικασίες λήψης επιχειρησιακών αποφάσεων, όπως σχεδιασμό καταλόγου και ανάλυση της συμπεριφοράς αγορών των πελατών [7].

1.4 Αλγόριθμοι Κατηγοριοποίησης Δέντρων Απόφασης

Κατά την εκτέλεση του αλγορίθμου των Δέντρων Απόφασης (Decision Tree), δημιουργείται το δέντρο από τη ρίζα και συνεχίζει προς τα κάτω (top-down), επιλέγοντας ένα πεδίο ή χαρακτηριστικό (attribute) από όλο το σύνολο των χαρακτηριστικών στη ρίζα του δέντρου.

Το δέντρο αποφάσεων έχει κόμβους απόφασης και κόμβους φύλλα. Ο κόμβος απόφασης περιέχει το χαρακτηριστικό που επιλέχθηκε να χωρίσει το δέντρο στο επόμενο επίπεδο. Ο κόμβος φύλλο περιέχει την τιμή της τάξης (class) που παρατηρήθηκε.

Ο αλγόριθμος ξεκινά με τη δημιουργία ενός συνόλου δεδομένων για εκπαίδευση (training) του αλγορίθμου. Το σύνολο των δεδομένων για εκπαίδευση είναι ένας πίνακας που αποτελείται από τις τιμές που έχουν κάποια ανύσματα για μία σειρά χαρακτηριστικών καθώς επίσης και την τιμή της τάξης του κάθε ανύσματος. Στη επόμενο βήμα του αλγορίθμου, το σύνολο δεδομένων για εκπαίδευση, η λίστα με τα χαρακτηριστικά και κριτήρια διαχωρισμού (splitting criteria) χρησιμοποιούνται επαναληπτικά για το χτίσιμο του δέντρου απόφασης [6].

Ο αλγόριθμος των δέντρων απόφασης είναι δημοφιλής λόγω της απλότητας και της διαφάνειας του [8].

1.5 Μέτρα αξιολόγησης κανόνων

Τα μέτρα ενδιαφέροντος των κανόνων αναφέρονται στους κανόνες που είναι της μορφής $A \rightarrow B$. Όλα τα μέτρα μπορούν να παίξουν σημαντικό ρόλο στην επιλογή των κανόνων που έχουν ενδιαφέρον από ένα σύνολο δεδομένων. Γι' αυτό είναι σημαντικό στο στάδιο της εξαγωγής των κανόνων, να λαμβάνεται υπόψη ένας συνδυασμός από αυτά τα μέτρα.

Ένας επιπρόσθετος λόγος για τον οποίο χρησιμοποιούνται τα μέτρα ενδιαφέροντος των κανόνων έγκειται στο ότι κατά την εκτέλεση των αλγορίθμων συσχέτισης και κατηγοριοποίησης εξάγεται ένας υπέρογκος αριθμός κανόνων που δε βοηθούν το χρήστη να λάβει κάποια σημαντική γνώση. Γι' αυτό, για τη αξιολόγηση των κανόνων μπορούν να χρησιμοποιηθούν κάποια *αντικειμενικά* μέτρα για να φιλτράρουμε τους εξαγόμενους κανόνες και να ανακτήσουμε τους πιο σημαντικούς.

Στατιστικά μέτρα όπως το μέτρο χ^2 (chi-square) [9], το οποίο είναι καθοριστικό για να υπολογιστεί το μέτρο p-value [10] μπορούν να χρησιμοποιηθούν έτσι ώστε να συμπεραίνουμε κατά πόσο ένας κανόνας είναι στατιστικά σημαντικός ή όχι. Ένα άλλο σημαντικό *στατιστικό* μέτρο είναι η εξίσωση του Framingham [11] η οποία υπολογίζει το ποσοστό του κινδύνου ένας ασθενής να πάθει καρδιακό επεισόδιο.

Περισσότερες πληροφορίες σχετικά με τα μέτρα αξιολόγησης κανόνων καθώς και το πώς θα χρησιμοποιηθούν για να μειώσουν το πλήθος των εξαγόμενων κανόνων παρουσιάζονται στο Κεφάλαιο 3.

1.6 Στόχος

Σκοπός της διπλωματικής μας εργασίας είναι η αξιολόγηση των κανόνων που εξάγονται από τους αλγόριθμους συσχέτισης και κατηγοριοποίησης με τη χρήση πολλαπλών μέτρων αξιολόγησης. Στόχος μας είναι να ελαχιστοποιήσουμε όσο το δυνατόν – αξιολογώντας τα μέτρα αξιολόγησης του κάθε κανόνα – το πλήθος των κανόνων που εξάγονται και να λάβουμε τους πιο σημαντικούς από αυτούς.

Επιπλέον, θέλουμε να υπολογίσουμε τη στατιστική σημαντικότητα του κάθε κανόνα καθώς επίσης και το ποσοστό κινδύνου που έχει ο κάθε ασθενής να πάθει καρδιακό επεισόδιο.

1.7 Δομή Μελέτης

Κεφάλαιο 2: Γίνεται μία περιγραφή των αλγορίθμων συσχέτισης και ταξινόμησης

Κεφάλαιο 3: Παρουσιάζονται τα στατιστικά και αντικειμενικά μέτρα αξιολόγησης κανόνων που έχουν χρησιμοποιηθεί και υλοποιηθεί καθώς και η κωδικοποίηση που έγινε σε κάθε ένα από αυτό.

Κεφάλαιο 4: Παρουσιάζεται μία σύντομη περιγραφή της Βάσης δεδομένων που έχει χρησιμοποιηθεί. Επιπρόσθετα, περιγράφεται η μεθοδολογία που έχει ακολουθηθεί για την κωδικοποίηση, την αξιολόγηση και την εξαγωγή γνώσης με βάση τα μέτρα.

Κεφάλαιο 5: Παρουσίαση αποτελεσμάτων

Κεφάλαιο 6: Συζήτηση

Κεφάλαιο 7: Αξιολόγηση και Μελλοντική Εργασία

Κεφάλαιο 2

Αλγόριθμοι κανόνων Συσχέτισης και Δέντρα Αποφάσεων

2.1 Αλγόριθμοι κανόνων Συσχέτισης (Association Algorithms)

2.1.1 Αλγόριθμος Apriori

Ο αλγόριθμος Apriori έχει προταθεί από τους R. Agrawal R. Srikant το 1994[12]. Ο αλγόριθμος χρησιμοποιείται για ανόρυξη συχνών συνόλων αντικειμένων (itemsets) για εξόρυξη κανόνων συσχέτισης. Ο αλγόριθμος έχει πάρει το όνομα του από την προγενέστερη γνώση (prior knowledge) των χαρακτηριστικών των συχνών συνόλων αντικειμένων, που χρησιμοποιεί. Ο Apriori υιοθετεί την τεχνική αναζήτησης level-wise, η οποία είναι μια επαναλαμβανόμενη τεχνική που χρησιμοποιεί τα k-itemesets για να κτίσει τα (k+1)-itemsets.

Στην αρχή, ο αλγόριθμος βρίσκει τα συχνά εμφανιζόμενα 1-itemsets (το σύνολο αντικειμένων με 1 μόνο χαρακτηριστικό). Ο αλγόριθμος αναζητά και συναθροίζει τον αριθμό που εμφανίζεται κάθε αντικείμενο – χαρακτηριστικό στη βάση δεδομένων, και μετά συλλέγει τα αντικείμενα που ικανοποιούν το ελάχιστο support, στο σύνολο L1. Κατόπιν, χρησιμοποιώντας το σύνολο L1, χτίζεται το σύνολο L2 το οποίο περιλαμβάνει όλα τα συχνά σύνολα αντικειμένων με 2 χαρακτηριστικά (2-itemsets), το οποίο κι αυτό χρησιμοποιείται για να χτιστεί το L3, και ούτω κάθε εξής μέχρις ότου να μην μπορεί βρεθεί άλλο σύνολο με k-itemesets, δηλαδή το Lk να είναι κενό. Για να βρεθεί κάθε Lk απαιτείται μία αναζήτηση της βάσης δεδομένων.

Για την δημιουργία κάθε επιπέδου με τα συχνά σύνολα αντικειμένων, χρησιμοποιείται η ιδιότητα Apriori (Apriori Property) η οποία μειώνει τον χώρο αναζήτησης και έτσι βελτιώνεται σημαντικά η αποδοτικότητα του αλγορίθμου. Η ιδιότητα Apriori αναφέρει ότι: ***όλα τα μη κενά υποσύνολα των συχνών συνόλων αντικειμένων πρέπει να είναι επίσης συχνά.***

Η ιδιότητα Apriori βασίζεται στο ότι: εάν ένα σύνολο αντικειμένων I δεν ικανοποιεί το ελάχιστο όριο support (min_sup,) τότε το I δεν είναι συχνό, P(I) < min_sup .Εάν το αντικείμενο A προστίθεται στο σύνολο αντικείμενων I, τότε το καινούργιο σύνολο I (IUA) δεν μπορεί να εμφανίζεται πιο συχνά από το I. Επομένως, ούτε το σύνολο IUA είναι συχνό, επειδή P(IUA) < min_sup.

Η ιδιότητα Apriori χρησιμοποιείται για την παραγωγή του Lk από το Lk-1, για k≥2, και ακολουθείται μια διαδικασία δύο βημάτων, που αποτελείται από την διαδικασία ένωσης (join) και κλαδέματος (prune):

Διαδικασία Ένωσης: Για να βρεθεί το σύνολο L_k, παράγετε ένα σύνολο από υποψήφια σύνολα με k αντικείμενα (k-itemsets) από την ένωση του συνόλου L_{k-1} με τον εαυτό του. Το σύνολο με τα υποψήφια σύνολα αντικειμένων καλείται C_k. Εάν το l_i είναι μέλος του L_{k-1}, τότε το $l_i[j]$ αναφέρεται στο αντικείμενο j του συνόλου αντικειμένων l_i. Ο Apriori θεωρεί ότι τα αντικείμενα στα σύνολα είναι ταξινομημένα σε αλφαβητική σειρά. Για κάποιο σύνολο αντικειμένων l_i με (k-1) αντικείμενα, τα αντικείμενα είναι ταξινομημένα σε $l_i[1] < l_i[2] < l_i[3] < ... < l_i[k-1]$. Όταν η ένωση $L_{k-1} \bowtie L_{k-1}$ εκτελείται, τα μέλη του L_{k-1} μπορούν να ενωθούν εάν τα πρώτα (k-2) αντικείμενα είναι τα ίδια. Για παράδειγμα το l_1 και l_2 itemsets που ανήκουν στο σύνολο L_{k-1} μπορούν να ενωθούν εάν $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$. Ο έλεγχος $(l_1[k-1] < l_2[k-2])$ γίνεται για να εξασφαλιστεί ότι δεν θα παραχθεί κανένα αντίγραφο του ίδιου itemset στο C_k. Το αποτέλεσμα της ένωσης των l1 και l2 itemsets είναι $l_1[1], l_1[2], l_1[3], ..., l_1[k-1], l_2[k-1]$.

Διαδικασία Κλαδέματος (prune): Κάποια από τα σύνολα αντικειμένων που ανήκουν στο Ck, μπορεί να είναι συχνά εμφανιζόμενα κι άλλα όχι, όμως όλα τα συχνά εμφανιζόμενα σύνολα k αντικειμένων (k-itemsets) συμπεριλαμβάνονται στο Ck. Θα πρέπει να γίνει μία

αναζήτηση στη βάση δεδομένων για να μετρηθεί ο αριθμός όπου κάθε υποψήφιο σύνολο στο Ck, εμφανίζεται στη βάση δεδομένων. Όλα τα σύνολα αντικειμένων που περιλαμβάνονται στο Ck, και εμφανίζονται στη βάση δεδομένων όχι λιγότερό αριθμό από το ελάχιστο support, προστίθενται στο Lk. Αυτό γίνεται επειδή, όπως αναφέρει η Apriori ιδιότητα, οποιοδήποτε (k-1)-itemset σύνολο αντικειμένων δεν είναι συχνό τότε δεν μπορεί να είναι υποσύνολο κάποιου k-itemset σύνολο αντικειμένων. Έτσι επειδή το σύνολο Ck, μπορεί να γίνει αρκετά μεγάλο, τα σύνολα αντικειμένων που δεν είναι συχνά αφαιρούνται.

2.1.2 Περιγραφή Ψευδοκώδικα Αλγόριθμου Apriori

Στο Σχήμα 2.1 παρουσιάζεται ο ψευδοκώδικας του αλγόριθμου Apriori και οι σχετικές διαδικασίες:

1. Καταρχάς ο αλγόριθμος δέχεται ως είσοδο μία βάση δεδομένων με δοσοληψίες. Η βάση αυτή αποτελείται από ένα αρχείο και κάθε εγγραφή του αρχείου αντιπροσωπεύει μία δοσοληψία. Η δοσοληψία συνήθως περιλαμβάνει ένα μοναδικό αριθμό ταυτότητας και μία λίστα από αντικείμενα (items) – χαρακτηριστικά όπου συνθέτουν την δοσοληψία.
2. Στο πρώτο βήμα βρίσκονται όλα τα συχνά σύνολα αντικειμένων με 1 χαρακτηριστικό και φυλάγονται στο σύνολο L1
3. Στο βήμα 3 είναι η διαδικασία όπου το σύνολο υποψηφίων Ck παράγεται από την ένωση του Lk-1 με τον εαυτό του. Η διαδικασία apriori_gen παράγει τα υποψήφια σύνολα αντικειμένων και μετά χρησιμοποιεί την ιδιότητα Apriori για να αφαιρέσει αυτά που δεν είναι συχνά.
4. Στα βήματα 4 – 10 γίνεται μία αναζήτηση στη βάση δεδομένων, για να βρεθεί ο αριθμός που τα σύνολα αντικειμένων εμφανίζονται στην βάση. Στο βήμα 9 βρίσκει τα υποψήφια σύνολα αντικειμένων που έχουν μεγαλύτερο από το ελάχιστο support και τα προσθέτει στο σύνολο Lk.

5. Στο τελικό βήμα 11 γίνεται μια ένωση όλων των συχνών συνόλων αντικειμένων των συνόλων Lk στο L. Έτσι μετά από μια διαδικασία για εξαγωγή κανόνων συσχέτισης μπορεί να χρησιμοποιήσει το σύνολο L.

Αλγόριθμος: Apriori. Εύρεση των συχνών συνόλων αντικειμένων (itemsets) χρησιμοποιώντας την επαναλαμβανόμενη τεχνική level-wise βασισμένη στα παραγωγή υποψηφίων.

Είσοδος:
- D, βάση δεδομένων με δοσοληψίες
- min-sup, το ελάχιστο αριθμός support.

Έξοδος: L, σύνολο με όλα τα συχνά σύνολα αντικειμένων που ανήκουν στο D

Μέθοδος:

```
(1) L1 =find_frequent_l-itemsets(D);
(2) for(k = 2;Lk-l ≠ 0;k++){
(3)     Ck = apriori_gen(Lk-l);
(4)     for each transaction t∈D { //scan D for counts
(5)             Ct = subset(Ck,t);  //get the subsets of t that are candidates
(6)             for each candidate c ∈Ct
(7)                     c.count++;
(8)     }
(9)     Lk = {c ∈ Ck| c.count ≥ min-sup}
(10) }
(11) return L = UkLk;

procedure apriori_gen(Lk-1 :frequent (k-1)-itemsets)
(1) for each itemset l1∈Lk-1
(2)     for each itemset l2 ∈ Lk-1
(3)             if((l1[1]=l2[1])∧(l1[2 =l2[2])∧…∧( l1 [k-2]= l2 [k-2])∧
                   (l1 [k-1]]< l2[k-1])) then {
(4)                     c = l1 ⊳⊲ l2; //join step: generate candidates
(5)                     if has_infrequent_subset(c, Lk-1) then
(6)                             delete c;  //prune step: remove unfruitful candidate
(7)                     else add c to Ck;
(8)             }
(9) return Ck;

procedure has_infrequent_subset(c: candidate k-itemset;
                                Lk-1: frequent (k - 1)-itemsets); //use prior knowledge
(1) for each (k - 1)-subset s of c
(2)     if s ∈ Lk-1 then
(3)             return TRUE;
(4)     return FALSE;
```

Σχήμα 2.1: Ψευδοκώδικας Αλγόριθμου Apriori

2.1.3 Παράδειγμα Εκτέλεσης αλγόριθμου Apriori

Χρησιμοποιήσαμε μία μικρή βάση δεδομένων με 20 ασθενείς με έμφραγμα του μυοκαρδίου που παρουσιάζεται στο Πίνακα 2.1. Σε αυτή τη βάση εφαρμόζουμε τον αλγόριθμο Apriori για να βρούμε τα συχνά σύνολα αντικειμένων. Θεωρούμε το ελάχιστο

support 0,4 (40%), που αυτό σημαίνει ότι θα πρέπει ένα σύνολο αντικειμένων να εμφανίζεται τουλάχιστο 8 φορές .

Πίνακας 2.1: Βάση δεδομένων δοσοληψιών για ασθενείς με καρδιαγγειακά επεισόδια (File: example_mi.arff)

Αρι.	SEX (Φύλο)	SM BEF (Καπνιστής)	HDL (Λιποπρωτεΐνες Υψηλής Πυκνότητας)	GLU (Γλυκόζη)	HT (Υπέρταση)	MI (Στεφανιαία Νόσος)
1.	M	Y	M	N	N	Y
2.	M	Y	M	N	Y	Y
3.	M	Y	L	H	N	Y
4.	M	Y	M	N	N	Y
5.	F	Y	L	H	N	N
6.	M	N	M	N	Y	N
7.	F	N	M	N	Y	Y
8.	M	Y	M	H	Y	N
9.	M	Y	L	N	N	Y
10.	M	Y	L	N	N	Y
11.	M	Y	M	H	N	Y
12.	M	Y	M	H	N	Y
13.	M	Y	M	N	N	Y
14.	M	Y	M	N	Y	Y
15.	M	Y	M	H	N	Y
16.	M	Y	H	H	N	Y
17.	M	Y	L	N	N	N
18.	M	Y	M	N	N	Y
19.	M	Y	L	N	N	Y
20.	M	Y	M	N	Y	Y

Βήμα 1: Καταρχάς γίνεται μια αναζήτηση στη βάση δεδομένων για να βρεθούν όλα τα σύνολα αντικειμένων με 1 χαρακτηριστικό και πόσες φορές αυτά εμφανίζονται στην βάση δεδομένων. Όλα τα σύνολα αντικειμένων αποθηκεύονται στο σύνολο C1 . Σημειώνουμε ότι, το χαρακτηριστικό κλάσης δεν εισάγεται στα υποψήφια σύνολα αντικειμένων, θα προστεθεί στο τέλος της διαδικασίας για να βεβαιωθεί ότι δεν θα εξαχθούν κανόνες συσχέτισης που να μην συμπεριλαμβάνουν το χαρακτηριστικό για την κλάση. Τα αποτελέσματα του συνόλου C1 παρουσιάζονται στο πιο κάτω Πίνακα 2.2.

Πίνακας 2.2: Παραγόμενο σύνολο C1 (Παράδειγμα αλγόριθμου Apriori)

C1			
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M	18	3	15
SEX = F	2	1	1
SMBEF = N	2	1	1
SMBEF = Y	18	3	15
HDL = H	1	0	1
HDL = L	6	2	4
HDL = M	13	2	11
GLU = N	13	2	11
GLU = H	7	2	5
HT = N	14	2	12
HT = Y	6	2	4

Βήμα 2: Από το C1 επιλέγουμε τα σύνολα αντικειμένων, που όταν ενωθούν και με κάποιο αντικείμενο της κλάσης έχουν support μεγαλύτερο από το ελάχιστο support και τα αποθηκεύουμε στο σύνολο L1. Δηλαδή θα αφαιρεθούν τα σύνολα αντικειμένων που εμφανίζονται λιγότερο από 8 φορές. Τα αποτελέσματα του συνόλου L1 παρουσιάζονται στο πιο κάτω Πίνακα 2.3.

Πίνακας 2.3: Παραγόμενο σύνολο L1 (Παράδειγμα αλγόριθμου Apriori)

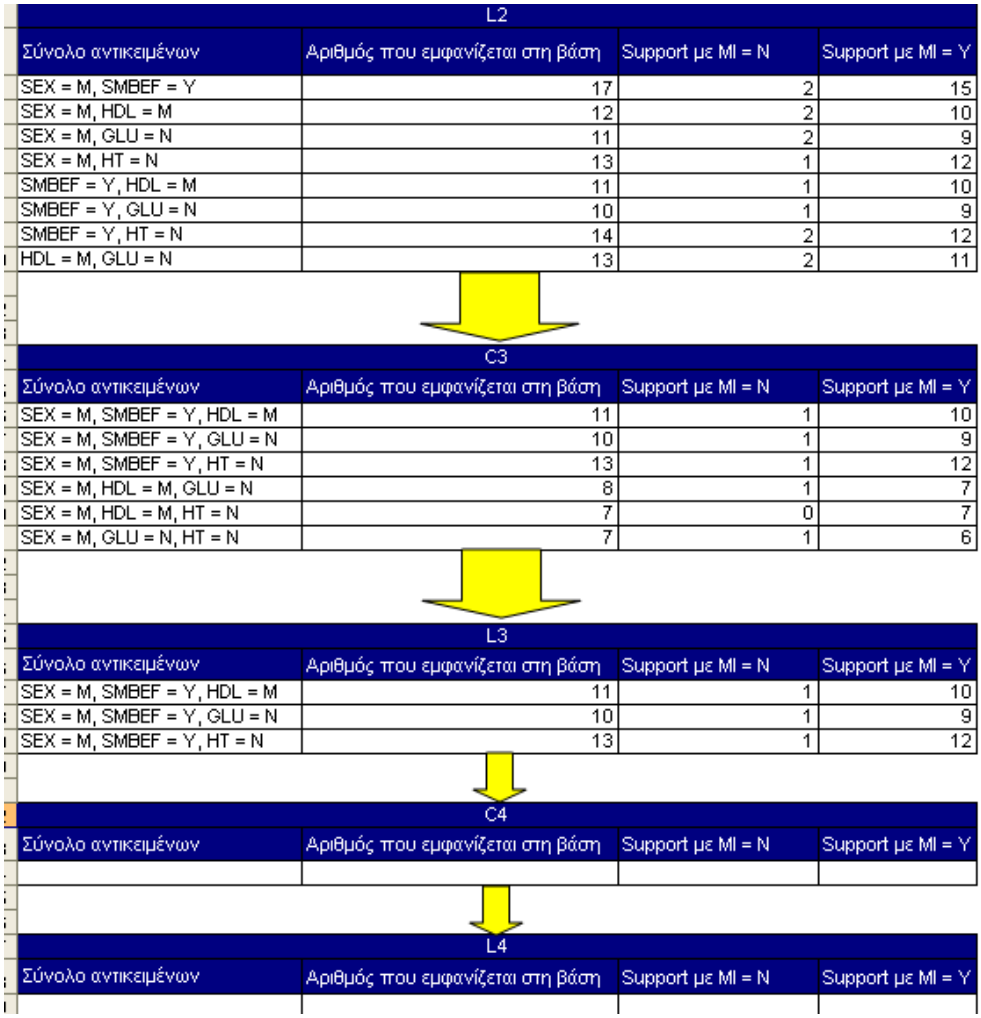
L1			
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M	18	3	15
SMBEF = Y	18	3	15
HDL = M	13	2	11
GLU = N	13	2	11
HT = N	14	2	12

Βήμα 3: Σε αυτό το βήμα θα γίνει η ένωση του L1 με τον εαυτό του για να κτιστεί το σύνολο C2. Για κάθε σύνολο αντικειμένων του L1 γίνεται συνένωση με τα υπόλοιπα. Τα αποτελέσματα του συνόλου C2 παρουσιάζονται στο πιο κάτω Πίνακα 2.4.

Πίνακας 2.4: Παραγόμενο σύνολο C2 (Παράδειγμα αλγόριθμου Apriori)

C2			
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M, SMBEF = Y	17	2	15
SEX = M, HDL = M	12	2	10
SEX = M, GLU = N	11	2	9
SEX = M, HT = N	13	1	12
SMBEF = Y ,HDL = M	11	1	10
SMBEF = Y ,GLU = N	10	1	9
SMBEF = Y ,HT = N	14	2	12
HDL = M ,GLU = N	13	2	11
HDL = M ,HT = N	7	0	7
GLU = N ,HT = N	7	1	6

Βήμα 4: Επανάληψη των Βημάτων 2 και 3 μέχρι το Lk να είναι κενό, σε αυτή τη περίπτωση, οι επαναλήψεις γίνονται μέχρι το L4 να είναι κενό. Στο Σχήμα 2.2, παρουσιάζονται τα υπόλοιπα βήματα.



L2

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M, SMBEF = Y	17	2	15
SEX = M, HDL = M	12	2	10
SEX = M, GLU = N	11	2	9
SEX = M, HT = N	13	1	12
SMBEF = Y, HDL = M	11	1	10
SMBEF = Y, GLU = N	10	1	9
SMBEF = Y, HT = N	14	2	12
HDL = M, GLU = N	13	2	11

C3

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M, SMBEF = Y, HDL = M	11	1	10
SEX = M, SMBEF = Y, GLU = N	10	1	9
SEX = M, SMBEF = Y, HT = N	13	1	12
SEX = M, HDL = M, GLU = N	8	1	7
SEX = M, HDL = M, HT = N	7	0	7
SEX = M, GLU = N, HT = N	7	1	6

L3

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M, SMBEF = Y, HDL = M	11	1	10
SEX = M, SMBEF = Y, GLU = N	10	1	9
SEX = M, SMBEF = Y, HT = N	13	1	12

C4

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y

L4

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y

Σχήμα 2.2: Παραγόμενα σύνολα αντικειμένων (Παράδειγμα εφαρμογής αλγόριθμου Apriori)

2.1.4 Διαδικασία εξόρυξης κανόνων συσχέτισης από τα εξαγόμενα συχνά σύνολα αντικειμένων

Αφού ολοκληρωθεί ο αλγόριθμος Apriori, το σύνολο L, με όλα τα συχνά σύνολα αντικειμένων, αποστέλλονται στην διαδικασία για δημιουργία των κανόνων συσχέτισης. Η διαδικασία αυτή λαμβάνει επίσης, από τον χρήστη, και ένα μέτρο αξιολόγησης των κανόνων. Για κάθε σύνολο αντικείμενων l που ανήκει στο σύνολο L, γίνεται ένωση του με κάποιο χαρακτηριστικό κλάσης, και ελέγχεται εάν το μέτρο ικανοποιείται. Για την αξιολόγηση των κανόνων χρησιμοποιούμε το μέτρο αξιολόγησης κανόνων, confidence (εμπιστοσύνη). Σε ένα κανόνα $A \rightarrow B$, το confidence, είναι η πιθανότητα να ισχύει το B νοουμένου ότι ισχύει το A, και υπολογίζεται από την Εξίσωση 3.2 (Κεφάλαιο 3.1). (Το support_count είναι ο αριθμός εμφάνισης στην βάση δεδομένων) .

2.2 Δέντρα Αποφάσεων (Decision Tree)

Οι αλγόριθμοι για την παραγωγή δέντρων αποφάσεων ακολουθούν συνήθως αναλυτική προσέγγιση. Δημιουργούν δηλαδή το δέντρο από τη ρίζα και συνεχίζουν προς τα κάτω (top-down), επιλέγοντας ένα πεδίο ή χαρακτηριστικό (attribute) από όλο το σύνολο των χαρακτηριστικών στη ρίζα του δέντρου. Στη συνέχεια, για κάθε τιμή (ή διάστημα) του χαρακτηριστικού αυτού ορίζεται ένα υποσύνολο εγγραφών, οι οποίες έχουν στο συγκεκριμένο χαρακτηριστικό τη συγκεκριμένη τιμή (ή διάστημα). Αφού ολοκληρωθεί το βήμα αυτό και ο αλγόριθμος έχει κάνει την πρώτη διακλάδωση αναζητά για κάθε υποσύνολο ένα υποδέντρο αποφάσεων (subtree). Όταν βρει ένα υποσύνολο, το οποίο ανήκει αποκλειστικά σε μία μόνο τάξη, τότε η διαδικασία σταματά, η διακλάδωση προς τα κάτω τελειώνει και παίρνει φύλλο με την τάξη στην οποία ανήκει το υποσύνολο. Αξίζει να σημειωθεί ότι έχουν προταθεί και άλλες προσεγγίσεις για το σχεδιασμό ταξινομητών δέντρων αποφάσεων, όπως η προσέγγιση bottom-up [15], όπου υπολογίζονται οι αποστάσεις μεταξύ προταξινομημένων τάξεων και σε κάθε βήμα οι δύο τάξεις με τη μικρότερη απόσταση ενώνονται, ώστε να δημιουργήσουν μία νέα ομάδα, μέχρις ότου μείνει ένας κόμβος, ο οποίος περιέχει όλες τις τάξεις, δηλαδή η ρίζα του δέντρου. Επίσης, έχει προταθεί μία υβριδική (hybrid) μέθοδος [16], η οποία συνδυάζει τόσο αναλυτική (top-down) προσέγγιση, όσο και προσέγγιση bottom-up διαδοχικά. Παρ' όλα αυτά η πιο διαδεδομένη μέθοδος σχεδιασμού ταξινομητών δέντρων αποφάσεων είναι η αναλυτική προσέγγιση.

Αξίζει να σημειωθεί ότι το μεγαλύτερο μέρος της έρευνας σχετικά με τους ταξινομητές δέντρων αποφάσεων έχει επικεντρωθεί στην εύρεση κανόνων διαίρεσης (splitting rules) [17]. Αυτό συμπεριλαμβάνει και την απόφαση για τους τερματικούς κόμβους. Οι τερματικοί κόμβοι συνδέονται με τις τάξεις εκείνες, οι οποίες έχουν τη μεγαλύτερη πιθανότητα, προκειμένου να ελαχιστοποιηθεί το ποσοστό των λανθασμένα ταξινομημένων εγγραφών. Μία εγγραφή ταξινομείται αφού περάσει από το δέντρο ξεκινώντας από τη ρίζα. Ο έλεγχος σε κάθε ενδιάμεσο κόμβο εφαρμόζεται στα χαρακτηριστικά της εγγραφής, προκειμένου να καθορισθεί το επόμενο τόξο (arc), στο οποίο η εγγραφή πρέπει να προχωρήσει. Η τιμή στον τερματικό κόμβο, στον οποίο καταλήγει η εγγραφή είναι και η ταξινόμησή της. Μία εγγραφή

ταξινομείται λάθος (misclassified) από το δέντρο, εάν η ταξινόμησή της δεν είναι η ίδια από τη σωστή τάξη της εγγραφής. Το ποσοστό των εγγραφών που ταξινομείται σωστά από ένα δέντρο αποφάσεων ονομάζεται ακρίβεια (accuracy), ενώ το ποσοστό των λανθασμένων ταξινομημένων εγγραφών αναφέρεται ως λάθος (error) [18].

Ένας από τους αρχικούς και βασικότερους αλγόριθμους ταξινόμησης δέντρων αποφάσεων είναι ο ID3 [19]. Ο αλγόριθμος αυτός ακολουθεί την αναλυτική προσέγγιση και δέχεται πλειάδες που είναι ήδη σε προταξινομημένες τάξεις. Ο αλγόριθμος επιλύει δυαδικά προβλήματα, δηλαδή θεωρεί δύο τάξεις (οι οποίες συμβολίζονται ως P (Positive) και N (Negative), μπορεί όμως να επεκταθεί και σε προβλήματα με περισσότερες τιμές τάξης. Το δέντρο αποφάσεων παράγεται από ένα υποσύνολο πλειάδων και βάσει τούτου ταξινομείται όλο το σύνολο εκπαίδευσης. Στη συνέχεια ελέγχεται η ακρίβεια της ταξινόμησης. Έτσι, αν όλες οι πλειάδες έχουν ταξινομηθεί σωστά, ο αλγόριθμος τερματίζει, διαφορετικά προστίθενται και άλλες πλειάδες και η διαδικασία επαναλαμβάνεται, μέχρις ότου όλες οι πλειάδες να ταξινομηθούν σωστά από το δέντρο. Βασική παράμετρος του αλγορίθμου είναι το ποσοστό των πλειάδων που θα λαμβάνεται υπόψη και με ποιο ρυθμό θα μεγαλώνει, εφόσον δεν είναι επαρκές. Σημαντικότερη παράμετρο στον αλγόριθμο αποτελεί το κριτήριο επιλογής του πεδίου για κάθε κόμβο, βάσει του οποίου θα γίνει η διακλάδωση. Ο αλγόριθμος αυτός χρησιμοποιεί σαν κριτήριο επιλογής την εντροπία, η οποία παρέχει μία εκτίμηση όσον αφορά το βαθμό του σφάλματος που επιτελείται κάθε φορά κατά το χωρισμό του συνόλου εκπαίδευσης, βάσει του συγκεκριμένου πεδίου. Η εντροπία είναι ένα μέγεθος που χρησιμοποιείται στη Θεωρία της Πληροφορίας και έχει αρχικά προταθεί από τον Shannon [20]. Η εντροπία μπορεί να δοθεί από την εξίσωση:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

(Εξίσωση 2.1)

όπου b είναι η βάση του λογάριθμου. Οι τιμές που παίρνει το b είναι 2, ο αριθμός Euler e και 10, και η εντροπία είναι σε bits για b = 2, nat για b = e και dit (or digit) για b = 10 [7], όπου οι τιμές a1, a2,…., am ανήκουν σε ένα πεδίο A. Η δεσμευμένη πιθανότητα P(ci/aj)

αντιπροσωπεύει την πιθανότητα να συμβαίνει το ci, δεδομένου ότι συμβαίνει το aj. Έτσι, το πεδίο με τη μικρότερη εντροπία χωρίζει καλύτερα το σύνολο εκπαίδευσης. Αναλυτικά, τα βήματα του αλγορίθμου ID3 έχουν ως εξής:

Διάλεξε ένα πεδίο ως ρίζα του δέντρου, βάσει της μικρότερης εντροπίας και σχημάτισε διακλαδώσεις για κάθε διαφορετική τιμή (ή διάστημα) του πεδίου αυτού.

Το δέντρο απόφασης που έχει κατασκευαστεί μέχρι στιγμής χρησιμοποιείται για ταξινόμηση του συνόλου εκπαίδευσης. Εάν όλες οι εγγραφές που ταξινομούνται σε ένα συγκεκριμένο φύλλο ανήκουν στην ίδια τάξη, ονόμασε το φύλλο με αυτήν την τάξη. Αν όλα τα φύλλα έχουν ονομασθεί σε κάποια τάξη, ο αλγόριθμος τελειώνει.

Διαφορετικά, για κάθε φύλλο που δεν έχει ονομασθεί με κάποια τάξη, επέλεξε ένα πεδίο που δεν έχει επιλεγεί στο μονοπάτι από το φύλλο έως τη ρίζα, βάσει της μικρότερης εντροπίας. Ονόμασε τον κόμβο με αυτό το πεδίο και σχημάτισε διακλάδωση με ένα φύλλο για κάθε διαφορετική τιμή (ή διάστημα) αυτού του πεδίου. Επανέλαβε το βήμα 2.

2.2.1 Αλγόριθμος C 4.5

Ο αλγόριθμος C4.5 αναπτύχθηκε από τον Quinlan [13] και αποτελεί εξέλιξη του αλγορίθμου ID3. Ο καινούργιος αλγόριθμος σε σχέση με τον προκάτοχό του έχει τα εξής βασικά πλεονεκτήματα:

- Δέχεται αριθμητικά δεδομένα
- Υποστηρίζει το χειρισμό άγνωστων τιμών
- Παρέχει ανθεκτικότητα στην ύπαρξη θορύβου
- Αποφυγή της μεγάλης προσαρμογής στα δεδομένα του δείγματος εκμάθησης

Ο αλγόριθμος ξεκινά με τη δημιουργία δεδομένων εκπαίδευσης. Το σύνολο των στοιχείων εκπαίδευσης είναι ένας πίνακας των παρατηρήσεων. Στη συνέχεια, τα στοιχεία της εκπαίδευσης, η λίστα των χαρακτηριστικών και τα κριτήρια διαχωρισμού περνούν από μια

επαναληπτική μέθοδο η οποία θα κτίσει το δέντρο απόφασης [14]. Στο Σχήμα 2 3 περιγράφονται τα βήματα εκτέλεσης του αλγορίθμου.

Αλγόριθμος δημιουργίας δέντρου απόφασης: C4.5 βασισμένος στο [13]

Είσοδος:

- Σύνολο εκπαίδευσης D, το οποίο είναι ένα σύνολο παρατηρήσεων και η σχετική τιμή της τάξης
- Λίστα των χαρακτηριστικών A, ένα σύνολο από υποψήφια χαρακτηριστικά
- Επιλεγόμενο κριτήριο διαχωρισμού (Information Gain, Gain Ratio, Gini Index, Distance Measure, Likelihood Ratio Chi-squared statistics)

Έξοδος: Δέντρο απόφασης

Μέθοδος:

(1) *Δημιουργία κόμβου N*

(2) *Αν όλες οι περιπτώσεις του συνόλου εκπαίδευσης έχουν την ίδια τιμή της τάξης C, τότε επέστρεψε το N σαν φύλλο με την ετικέτα C*

(3) *Αν η λίστα των χαρακτηριστικών είναι άδεια, τότε επέστρεψε N σαν φύλλο με την ετικέτα με τη μεγαλύτερη τιμή της τάξης στην έξοδο στο σύνολο της εκπαίδευσης*

(4) *Εφάρμοσε το επιλεγμένο κριτήριο διαχωρισμού στο σύνολο της εκπαίδευσης για να βρεθεί το καλύτερο χαρακτηριστικό για διαχωρισμό*

(5) *Ετικέτα κόμβου N με το χαρακτηριστικό του κριτηρίου διαχωρισμού*

(6) *Αφαίρεση του χαρακτηριστικού του κριτηρίου διαχωρισμού από τη λίστα των χαρακτηριστικών*

(7) ***Για κάθε τιμή** j στο χαρακτηριστικό του κριτηρίου διαχωρισμού*

- *Ας είναι D_j οι περιπτώσεις στο σύνολο εκπαίδευσης που ικανοποιούν το χαρακτηριστικό με τιμή j*
- *Αν D_j είναι άδειο (καμιά περίπτωση), τότε πάρε σαν φύλλο με την ετικέτα με τη μεγαλύτερη τιμή της τάξης στην έξοδο στον κόμβο N*
- ***διαφορετικά** πάρε τον κόμβο που έδωσε το **Generate Decision Tree** (D_j, λίστα χαρακτηριστικών, επιλεγμένο κριτήριο διαχωρισμού) στον κόμβο N*

(8) ***Τέλος για (for)***

(9) *Δώσε τον κόμβο N*

Σχήμα 2.3: Ψευδοκώδικας αλγόριθμου δημιουργίας δέντρου απόφασης

2.2.2 Παράδειγμα Εκτέλεσης αλγόριθμου C4.5

χρησιμοποιούμε μία μικρή βάση δεδομένων με 20 ασθενείς με έμφραγμα του μυοκαρδίου που παρουσιάζεται στο Πίνακα 2.5. Σε αυτή τη βάση εφαρμόζουμε τον αλγόριθμο C4.5 για να χτίσουμε το δέντρο απόφασης.

Πίνακας 2.5: Παράδειγμα βάσης δεδομένων δοσοληψιών για ασθενείς με καρδιαγγειακά επεισόδια (File: example2_mi.arff)

Αρι.	SEX (Φύλο)	SM BEF (Καπνιστής)	HDL (Λιποπρωτεΐνες Υψηλής Πυκνότητας)	AGE (Ηλικία)	MI (Στεφανιαία Νόσος)
1.	M	Y	H	1	N
2.	M	N	H	1	N
3.	M	Y	H	2	Y
4.	M	Y	M	3	Y
5.	F	Y	L	3	Y
6.	F	N	L	3	N
7.	F	N	L	2	Y
8.	M	Y	M	1	N
9.	F	Y	L	1	Y
10.	F	Y	M	3	Y
11.	F	N	M	1	Y
12.	M	N	M	2	Y
13.	F	Y	H	2	Y
14.	M	N	M	3	N

1η Επανάληψη

(1) Δημιουργία κόμβου N

(2) Όλες οι περιπτώσεις του συνόλου εκπαίδευσης έχουν την ίδια τιμή της τάξης C, τότε επέστρεψε το N σαν φύλλο με την ετικέτα C ?

Όχι → Συνέχεια στο Βήμα 3

(3) Η λίστα των χαρακτηριστικών είναι άδεια ;

Όχι → Συνέχεια στο Βήμα 4

(4) Εφάρμοσε το επιλεγμένο κριτήριο διαχωρισμού στο σύνολο της εκπαίδευσης για να βρεθεί το καλύτερο χαρακτηριστικό για διαχωρισμό

Πίνακας 2.6: Τιμές κριτηρίου διαχωρισμού για όλα τα χαρακτηριστικά στην 1η επανάληψη

Χαρακτηριστικό	Κριτήριο Διαχωρισμού (Information Gain)
SEX	0.151
SMBEF	0.048
HDL	0.029
AGE	**0.69**

- Το χαρακτηριστικό με τη μεγαλύτερη τιμή κατά τη μέθοδο του διαχωρισμού είναι το AGE.

(5) Ετικέτα κόμβου N με το χαρακτηριστικό του κριτηρίου διαχωρισμού AGE

(6) Αφαίρεση του χαρακτηριστικού AGE από τη λίστα των χαρακτηριστικών

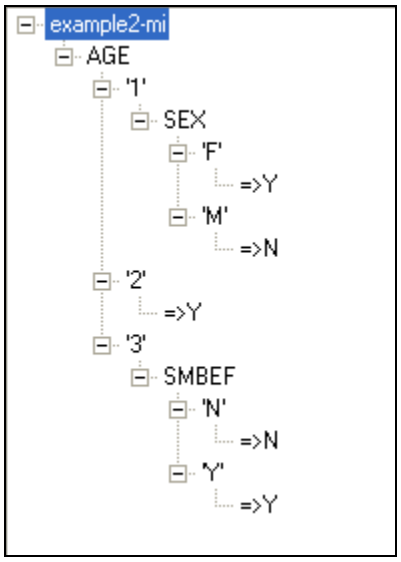
(7) ***Για κάθε τιμή*** *j στο χαρακτηριστικό του κριτηρίου διαχωρισμού*

- Ας είναι D_j οι περιπτώσεις στο σύνολο εκπαίδευσης που ικανοποιούν το χαρακτηριστικό με τιμή j
- Αν D_j είναι άδειο (καμιά περίπτωση), τότε πάρε σαν φύλλο με την ετικέτα με τη μεγαλύτερη τιμή της τάξης στην έξοδο στον κόμβο N
- ***διαφορετικά*** πάρε τον κόμβο που έδωσε το ***Generate Decision Tree*** (D_j, λίστα χαρακτηριστικών, επιλεγμένο κριτήριο διαχωρισμού) στον κόμβο N

(8) ***Τέλος για (for)***

(9) Δώσε τον κόμβο N

Με το πέρας όλων των επαναλήψεων δημιουργείται το δέντρο στο Σχήμα 2.4



Σχήμα 2.4: Δέντρο Απόφασης Παραδείγματος αλγορίθμου C4.5

Κεφάλαιο 3

Περιγραφή Μέτρων Αξιολόγησης

3.1 Αξιολόγηση Κανόνων

Χρησιμοποιώντας τους αλγόριθμους κανόνων Συσχέτισης και Κατηγοριοποίησης χωρίς να χρησιμοποιήσουμε κάποιο μέτρο αξιολόγησης των υποψήφιων κανόνων, τότε ο αλγόριθμος θα εξαγάγει εξαιρετικά μεγάλο αριθμό κανόνων, που δεν θα βοηθήσουν το χρήστη να λάβει κάποια σημαντική γνώση χρήστη. Έτσι θα πρέπει να γίνει αξιολόγηση των κανόνων, με κάποια μέτρα. Αυτά τα μέτρα καλούνται αντικειμενικά μέτρα και είναι βασισμένα στις πιθανότητες. Είναι συνήθως λειτουργίες από 2×2 πίνακα ενδεχομένων. Ένας πίνακας ενδεχομένων αποθηκεύει τις συχνότητας που ικανοποιούν τους δεδομένους όρους. Ο Πίνακας 9 είναι ένας πίνακας ενδεχομένων για τον κανόνα A → B, όπου το n(AB) δείχνει τον αριθμό εγγραφών που ικανοποιούν και το A και το B, ενώ το N δείχνει τον συνολικό αριθμό εγγραφών [21].

Πίνακας 3.1: Πίνακας ενδεχομένων 2 ×2για τον κανόνα A → B [21]

	B	$\overline{B}$	
A	$n(AB)$	$n(A\overline{B})$	$n(A)$
$\overline{A}$	$n(\overline{A}B)$	$n(\overline{A}\overline{B})$	$n(\overline{A})$
	$n(B)$	$n(\overline{B})$	N

Κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων, τα μέτρα αξιολόγησης μπορούν να χρησιμοποιηθούν με τρεις τρόπους:

1. Τα μέτρα μπορούν να χρησιμοποιηθούν για να κλαδέψουν τους κανόνες, που δεν είναι ενδιαφέρον, κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων.
2. Μπορεί επίσης να καθοριστεί ένα κατώτατο όριο για μέτρα βασισμένα στην χρησιμότητα, για περικοπή κανόνων σύμφωνα με τη σειρά των αποτελεσμάτων.
3. Και τρίτον, τα μέτρα μπορούν να χρησιμοποιηθούν επίσης κατά τη διάρκεια επιλογής ενδιαφέρον κανόνων.

Η πρώτη προσέγγιση, χρησιμοποιείται από τους αλγόριθμους για να αφαιρεθούν σύνολα αντικειμένων που δεν είναι συχνά εμφανιζόμενα. Το μέτρο που χρησιμοποιείται είναι το Support [14], το οποίο, για ένα κανόνα συσχέτισης $A \to B$, δείχνει την πιθανότητα εμφάνισης του Α και του Β μαζί στα δεδομένα (Εξίσωση 3.1).

$$\sup port = P(AB) = \frac{n(AB)}{N}$$

(Εξίσωση 3.1)

Διάφορα άλλα αντικειμενικά μέτρα αξιολόγησης κανόνων που χρησιμοποιήθηκαν για φιλτράρισμα των εξαγόμενων κανόνων ή ταξινόμησης τους είναι:

1. *Εμπιστοσύνη (Confidence) [14]*: η πιθανότητα να ισχύει το Β αφού ισχύει το Α. Δηλαδή το ποσοστό των περιπτώσεων που ισχύει το Β, στις περιπτώσεις που ισχύει το Α. Δυνατές τιμές που μπορεί να πάρει, είναι τιμές από μεταξύ 0 και 1. Όσο πιο κοντά στο 1 είναι η τιμή, τόσο πιο ενδιαφέρον είναι ο κανόνας.

$$confidence = P(B/A) = \frac{P(AB)}{P(A)} = \frac{n(AB)}{n(A)}$$

(Εξίσωση 3.2)

2. *Κάλυψη (Coverage)[22]*: η πιθανότητα να ισχύει το Α, δηλαδή το ποσοστό των περιπτώσεων όπου ισχύει το Α, και παίρνει τιμές μεταξύ 0 και 1.

$$\cov erage = P(A) = \frac{n(A)}{N}$$

(Εξίσωση 3.3)

3. *Επικράτηση (Prevalence)[22]:* η πιθανότητα να ισχύει το Β και παίρνει τιμές μεταξύ 0 και 1.

$$prevalance = P(B) = \frac{n(B)}{N}$$

(Εξίσωση 3.4)

4. *Ανάκληση (Recall)[22]*: είναι η πιθανότητα να ισχύει το Α, δεδομένου ότι ισχύει το Β. Παίρνει τιμές μεταξύ 0 και 1. Σημαντικοί κανόνες μπορεί να θεωρηθούν οι κανόνες που έχουν τιμή κοντά στο 1.

$$recall = P(A/B) = \frac{P(AB)}{P(B)} = \frac{n(AB)}{n(B)}$$

(Εξίσωση 3.5)

5. *Ιδιομορφία (Specificity)[22]:* είναι η πιθανότητα να μην ισχύει το Β, δεδομένου ότι δεν ισχύει ούτε το Α. Το specificity δείχνει πόσο καλά μπορεί ο αλγόριθμος να αναγνωρίσει τα αρνητικά αποτελέσματα. Παίρνει τιμές μεταξύ 0 και 1. Σημαντικοί κανόνες μπορεί να θεωρηθούν οι κανόνες που έχουν τιμή κοντά στο 1.

$$specificity = P(\overline{B}/\overline{A}) = \frac{P(\overline{A}\overline{B})}{P(\overline{A})} = \frac{n(\overline{A}\overline{B})}{n(\overline{A})}$$

(Εξίσωση 3.6)

6. *Ακρίβεια (Accuracy)[24]:* είναι η πιθανότητα να ισχύουν το Α και το Β συν την πιθανότητα να μην ισχύει ούτε το Α αλλά ούτε και το Β μαζί. Βασικά είναι ο βαθμός του πόσο κοντινές είναι οι μετρήσεις ή οι υπολογισμένες ποσότητες των πραγματικών τιμών (των σωστών τιμών). Παίρνει τιμές μεταξύ 0 και 1. Σημαντικοί κανόνες μπορεί να θεωρηθούν οι κανόνες που έχουν τιμή κοντά στο 1.

$$accuracy = P(AB) + P(\overline{A}\overline{B})$$

(Εξίσωση 3.7)

7. Lift [22]: Είναι ένα μέτρο της απόδοσης του προτύπου, και υπολογίζει το ποσοστό της προβλεπόμενης απάντησης. Είναι η αναλογία του μέτρου αξιολόγησης confidence με την πιθανότητα να ισχύει το B.

$$Lift = \frac{P(B/A)}{P(B)}$$

(Εξίσωση 3.7)

Παίρνει πραγματικές θετικές τιμές. Όταν η τιμή τείνει στο ένα, τα δύο μέρη είναι ανεξάρτητα και έτσι ο κανόνας δεν παρουσιάζει κάποιο ενδιαφέρον. Όταν η τιμή τείνει στο $+\infty$ αυτό σημαίνει ότι το P(B) τείνει στο μηδέν δείχνει ότι ο κανόνας δεν είναι σημαντικός ή το P(B|A) τείνει στο ένα τότε δείχνει ότι ο κανόνας είναι ενδιαφέρον. Όταν όμως το lift=0 σημαίνει ότι ο κανόνας δεν είναι σημαντικός.

8. *Δύναμη (Leverage)[22]*: Είναι η ανάλυση οπισθοδρόμησης και ειδικότερα υπολογισμό εκείνων των περιπτώσεων που έχουν επιδρούν αρνητικά [11]. Υπολογίζει το ποσοστό των ακραίων περιπτώσεων, των πρόσθετων περιπτώσεων που καλύπτονται και από το αριστερό και το δεξιό μέρος του κανόνα, πάνω από εκείνο που αναμένονται αν τα δύο μέρη ήταν ανεξάρτητα. Ορίζεται από την Εξίσωση 9.

$$Leverage = P(B/A) - P(A)P(B)$$

(Εξίσωση 3.8)

Παίρνει τιμές από το -1 μέχρι το 1 . Τιμές ίσες ή μικρότερες του μηδέν δείχνουν μια ισχυρή ανεξαρτησία μεταξύ των δύο μερών. Τιμές κοντά στο ένα δείχνουν ότι ο κανόνας είναι σημαντικός.

9. *Προστιθέμενη Αξία (Added Value)[22]*: Ορίζει την διαφορά μεταξύ της τελικής απάντησης με την άμεση και έμμεση απάντηση. Και είναι η διαφορά του confidence με την πιθανότητα να ισχύει το B.

$$AddedValue = P(B/A) - P(B)$$

(Εξίσωση 3.9)

Μπορεί να πάρει τιμές από το -1 μέχρι το 1. Τιμές ίσες ή μικρότερες του μηδέν δείχνουν μια ισχυρή εξάρτηση μεταξύ των δύο μερών.

10. *Σχετικός κίνδυνος (Relative Risk)[22]:* Είναι ο υπολογισμός του κινδύνου γεγονότος (πχ μίας ασθένειας) στη βάση δεδομένων. Υπολογίζεται από την αναλογία της πιθανότητας να ισχύει το Β δεδομένου ότι ισχύει το Α (εκτεθειμένης ομάδας), με την πιθανότητα να ισχύει το Β δεδομένου ότι δεν ισχύει το Α (μη εκτεθειμένης ομάδας).

$$\text{Relative Risk} = \frac{P(B/A)}{P(B/\overline{A})}$$

(Εξίσωση 3.10)

Παίρνει πραγματικές θετικές τιμές (>0). Όταν η τιμή τείνει στο ένα, τα δύο μέρη είναι ανεξάρτητα και έτσι ο κανόνας δεν παρουσιάζει κάποιο ενδιαφέρον. Όταν η τιμή τείνει στο $+\infty$ αυτό σημαίνει ότι το $P(B/\overline{A})$ τείνει στο μηδέν ή το P(B| A) τείνει στο ένα και δείχνει ότι ο κανόνας είναι σημαντικός.

11. *Αναλογία Πιθανοτήτων (Odds Ratio)[22]*: Είναι η αναλογία της πιθανότητας ενός γεγονότος να εμφανίζεται σε μια ομάδα, με τις πιθανότητες να εμφανίζεται σε μια άλλη ομάδα. Υπολογίζεται από την πιο κάτω εξίσωση.

$$OddsRatio = \frac{P(AB)P(\overline{A}\overline{B})}{P(A\overline{B})P(\overline{A}B)}$$

(Εξίσωση 3.11)

Παίρνει πραγματικές θετικές τιμές (>0). Όταν η τιμή τείνει στο ένα, τα δύο μέρη είναι ανεξάρτητα και έτσι ο κανόνας δεν παρουσιάζει κάποιο ενδιαφέρον. Όταν η τιμή τείνει στο $+\infty$ αυτό σημαίνει ότι τα δύο μέρη είναι εξαρτώμενα και δείχνει ότι ο κανόνας είναι σημαντικός.

12. *Πεποίθηση (Conviction) [23]*

$$\text{Conviction}=\frac{n-p(b)}{(1-Confidence)}$$

(Εξίσωση 3.13)

Προτάθηκε από τον Brin [23] και το n συμβολίζει τον αριθμό των συναλλαγών στη βάση δεδομένων.

Τόσο το Lift όσο και η πεποίθηση είναι μονότονα σε σχέση με την εμπιστοσύνη.

Εκτός από τα παραπάνω μέτρα αξιολόγησης κανόνων έχουν υλοποιηθεί το Chi-Square Statistic, η πιθανότητα P-Value καθώς και το Framingham Event Risk τα οποία παρέχουν ένα στατιστικό έλεγχο για κάθε κανόνα.

3.2 Chi-Square Test – Chi-Square Statistic [9]

Το Chi-Square test είναι ένας έλεγχος κάθε στατιστικής υπόθεσης κατά την οποία η δειγματοληπτική κατανομή του στατιστικού αποτελέσματος της δοκιμής είναι μια κατανομή chi-square (χ^2). Αυτό ισχύει όταν η μηδενική υπόθεση είναι αληθής ή ασυμπτωτικά αληθής. Σ' αυτή την περίπτωση η δειγματική κατανομή μπορεί να προσδιοριστεί από μία chi-square κατανομή με αποτέλεσμα να γενικεύσουμε το αποτέλεσμα για όσο πιο μεγάλο δείγμα.

Στην περίπτωση της μηδενικής υπόθεσης, το chi-square test προσδιορίζει αν η κατανομή των συμβάντων που παρατηρήθηκαν σε ένα δείγμα είναι σύμφωνο με τη θεωρητική κατανομή chi-square (χ^2). Τα συμβάντα που λαμβάνονται υπόψη πρέπει να είναι αποκλειστικά και να έχουν συνολική πιθανότητα 1.

Το Chi-Square test χρησιμοποιείται για την αξιολόγηση δύο ειδών σύγκρισης:

- τον έλεγχο καλής προσαρμογής (test of Goodness of fit) και
- τον έλεγχο της ανεξαρτησίας (test of independence)

Κατά τον έλεγχο καλής προσαρμογής ελέγχεται κατά πόσο ή όχι μια συχνότητα κατανομής που παρατηρήθηκε διαφέρει από μία θεωρητική κατανομή.

Κατά τον έλεγχο της ανεξαρτησίας ελέγχεται αν ζεύγη παρατηρήσεων που αναπαριστούνται σε ένα πίνακα είναι ανεξάρτητα μεταξύ τους (π.χ. εάν οι άνθρωποι που έχουν διαβήτη και είναι άνω των 60 ετών διαφέρουν όσο αφορά τη συχνότητα με την οποία έπαθαν έμφραγμα.

Το Chi-Square statistic υπολογίζεται από την πιο κάτω εξίσωση.

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

(Εξίσωση 3.14)

3.3 p-Value [10]

Στον έλεγχο στατιστικών υποθέσεων, η τιμή p-Value προσδιορίζει την πιθανότητα ένας στατιστικός έλεγχος (κανόνας) που παρατηρήθηκε να είναι σημαντικός (significant) ή όχι αν η μηδενική υπόθεση είναι αληθής.

Όσο μικρότερη είναι η πιθανότητα p-Value, τόσο απίθανο είναι το αποτέλεσμα αν η μηδενική υπόθεση είναι αληθής. Κατά συνέπεια τόσο πιο σημαντικός είναι ο κανόνας.

Η πιθανότητα p-Value δηλώνει το επίπεδο σημαντικότητας που παρατηρήθηκε. Αλλά το επίπεδο σημαντικότητας που παρατηρήθηκε αντιπροσωπεύει την πιθανότητα να υπάρχει κάποιο λάθος στον έλεγχο. Κατά συνέπεια όσο πιο μικρή είναι η τιμή του p-Value τόσο πιο μικρή είναι η πιθανότητα να υπάρχει κάποιο λάθος.

Σημαντικοί κανόνες θεωρούνται οι κανόνες που η πιθανότητα p-Value είναι μικρότερη η ίση από 0.05. Σε τέτοια περίπτωση, η μηδενική υπόθεση απορρίπτεται και ο κανόνας είναι σημαντικός.

Η πιθανότητα p-Value υπολογίζεται από τη εξίσωση:

$$p-Value = \frac{(1/2)^{k/2}}{\Gamma(k/2)} \chi^{k/2-1} e^{-\chi/2}$$

(Εξίσωση 3.15)

όπου

χ = η τιμή του Chi-Square,

k = ο βαθμός ελευθερίας (degrees of freedom) και

$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$ η συνάρτηση Gamma.

3.4 Framingham Event Risk [11]

Η πιθανότητα Framingham Event Risk παρουσιάζει το ποσοστό ένας ασθενής να παρουσιάσει ένα καρδιακό επεισόδιο ή όχι. Η εξίσωση προέκυψε μετά από μελέτες και κάποια μεθοδολογία που εφάρμοσαν οι ιατροί προσπαθώντας να βρουν ποιοι παράγοντες επηρεάζουν τα καρδιακά επεισόδια αλλά ταυτόχρονα και πόση βαρύτητα έχει ο κάθε παράγοντας.

Σύμφωνα με τους ιατρούς, εάν ένας ασθενής παρουσιάζει ποσοστό 0-5% τότε έχει χαμηλό κίνδυνο (low risk) να πάθει καρδιακό επεισόδιο, από 5-15% τότε βρίσκεται στη μεσαία κατηγορία κινδύνου (medium risk) ενώ από 15 %και άνω ο κίνδυνος να πάθει καρδιακό επεισόδιο ο ασθενής είναι αρκετά μεγάλο (high risk).

Η πιθανότητα Framingham Event Risk υπολογίζεται στα παρακάτω βήματα:

Βήμα 1:

$$\mu = 15.5303 - 0.9119x^{\log_{10}(SBP)} - 0.2767x(SMBF) - 0.7181x^{\log_{10}(TC/HDL)} - 0.5865x(ELVH) - 1.4792x^{\log_{10}(AGE)} - 0.1759x(DM)$$

όπου

SBP = Systolic Blood Pressure
SMBF = Smoking Before
TC = Total Cholesterol
HDL = High Density Lipoprotein Cholesterol
ELVH = Electrocardigraphic Left Ventricular Hypertrophy
AGE = η ηλικία του ασθενή
DM = Διαβήτης
Οι μεταβλητές SMBF, ELVH και DM παίρνουν τιμή 1 αν παρουσιάζονται και 0 όταν δεν είναι διαθέσιμες.

$$\sigma = e^{(-0.3155-0.2784x(\mu-4.4181))}$$

Βήμα 2: $u = (\log_{10}10 - \mu)/\sigma$

Βήμα 3: $p = 1 - e^{-e^u}$

(Εξίσωση 3.16)

Κεφάλαιο 4

Μεθοδολογία

4.1 Γενικά

Όπως αναφέραμε και στο Κεφάλαιο 1, σκοπός της διπλωματικής μας εργασίας είναι η αξιολόγηση των κανόνων που εξάγονται από τους αλγόριθμους συσχέτισης και κατηγοριοποίησης με τη χρήση των μέτρων αξιολόγησης που παρουσιάστηκαν στο Κεφάλαιο 3. Στόχος μας είναι να ελαχιστοποιήσουμε όσο το δυνατόν – με τη χρήση των μέτρων αξιολόγησης – το πλήθος των κανόνων που εξάγονται και να λάβουμε τους πιο σημαντικούς από αυτούς.

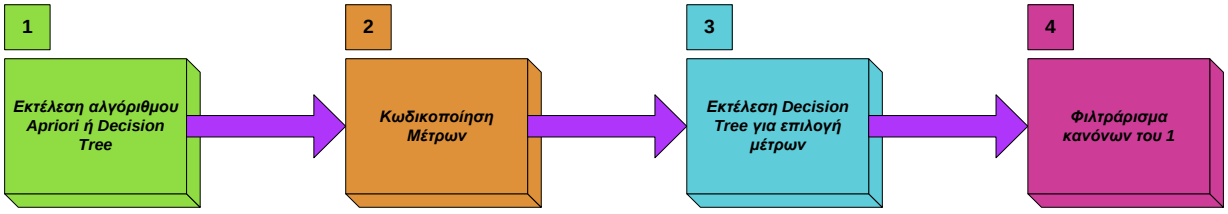
Στα πλαίσια της εκπόνησης της διπλωματικής εργασίας έχουμε χρησιμοποιήσει μία βάση δεδομένων του Γενικού Νοσοκομείου Πάφου [25] για ασθενείς με τριών ειδών καρδιοαγγειακά νοσήματα: α) Έμφραγμα μυοκαρδίου (MI), β) Αγγειοπλαστική (PCI) και γ) Στεφανιαία Παράκαμψη (bypass) (CABG).

Για την επίτευξη του στόχου μας ακολουθήθηκε η εξής μεθοδολογία:

- Εφαρμογή των σταδίων εξόρυξης δεδομένων
- Δημιουργία μοντέλων για την εξαγωγή των κανόνων
- Κωδικοποίηση των μέτρων αξιολόγησης
- Υπολογισμός του μέτρου Chi-Square (χ^2) το οποίο χρησιμοποιήθηκε για τον υπολογισμό της στατιστικής σημαντικότητας του κάθε κανόνα (μέτρο p-value)
- Υπολογισμός για κάθε ασθενή του ποσοστού του κινδύνου να πάθει ένα επεισόδιο με την εξίσωση του Framingham.

- Εξαγωγή κανόνων από τους αλγόριθμους Συσχέτισης [26] και Κατηγοριοποίησης [27] (Δέντρα Απόφασης) από την αρχική βάση.
- Εισαγωγή των κωδικοποιημένων κανόνων στον αλγόριθμο Κατηγοριοποίησης και δημιουργία δέντρου απόφασης βασισμένο στα μέτρα αξιολόγησης. Ανάκτηση από το δέντρο απόφασης των τιμών των σημαντικότερων μέτρων αξιολόγησης. Για την εξαγωγή αυτών των κανόνων χρησιμοποιήθηκε η μέθοδος της 10-πτυχης διασταυρούμενης επικύρωσης (10-fold cross validation).
- Φιλτράρισμα αρχικών κανόνων βάσει των αποτελεσμάτων του Δέντρου Απόφασης.
- Υπολογισμός για κάθε μοντέλο του ποσοστού της σωστής ταξινόμησης (Correct Classification) του.

Στο Σχήμα 4.1 παρουσιάζεται σχηματικά η μεθοδολογία που έχει ακολουθηθεί.



Σχήμα 4.1: Μεθοδολογία που ακολουθήθηκε για εξαγωγή φιλτραρισμένων κανόνων

4.2 Σύντομη περιγραφή Βάσης Δεδομένων

Η βάση δεδομένων προέκυψε από ένα πρωτόκολλο που χρησιμοποιήθηκε στο Γενικό Νοσοκομείο Πάφου. Για τέσσερα χρόνια οι γιατροί μάζευαν τριακόσιους ασθενείς κάθε χρόνο. Στη βάση δεδομένων υπήρχαν κάποια πεδία που είτε δεν είχαν καθόλου τιμές, είτε είχαν σε πολύ λίγες πλειάδες τιμές. Πέραν τούτου υπήρχαν κάποια πεδία που δεν χρειάζονταν στην ανάλυση, γιατί δεν θα πρόσφεραν καινούργια γνώση, όπως για παράδειγμα το πεδίο κάπνισμα μετά από το επεισόδιο. Η αρχική βάση δεδομένων περιείχε 1200 ασθενείς με τριών ειδών καρδιοαγγειακά νοσήματα: α) Έμφραγμα μυοκαρδίου (MI), β) Αγγειοπλαστική (PCI)

και γ) Στεφανιαία Παράκαμψη (bypass) (CABG). Υπάρχουν ασθενείς που έχουν ένα, δύο ή ακόμη και τρία από τα αναφερθέντα νοσήματα. Επειδή το κάθε νόσημα το εξετάζουμε ξεχωριστά, υπάρχουν ασθενείς που εμφανίζονται στη μια, ή και στις δύο, ή και στις τρεις ομάδες. Για την επιλογή των ασθενών που είναι στη βάση δεδομένων δεν υπήρχε κανένα κριτήριο, παρά μόνο να είχε τουλάχιστο ένα από τα πιο πάνω νοσήματα.

Για την επιλογή των χαρακτηριστικών είχαν ληφθεί υπόψη οι ακόλουθες προϋποθέσεις:

- Δόθηκαν από τους ειδικούς γιατρούς οι κατευθυντήριες γραμμές για το τι θα μελετηθεί στην έρευνα αυτή. Έτσι, έγινε η επιλογή των παραγόντων που έπρεπε να μελετηθούν.
- Παράγοντες που περικλείονταν σε άλλους παράγοντες δεν λήφθηκαν υπόψη, για παράδειγμα το ύψος και το βάρος του ασθενή που περιέχονται στον παράγοντα δείκτη μάζας σώματος (BMI).
- Παράγοντες που είχαν πολλές ελλιπείς τιμές και δεν υπήρχε η ευχέρεια ανάκτησης αυτών των τιμών, έχουν αφαιρεθεί.

Στον Πίνακα 4.1 που ακολουθεί, παρουσιάζεται μια γενική αναφορά και περιγραφή των πεδίων της Βάσης Δεδομένων.

Πίνακας 4.1: Πεδία Βάσης Δεδομένων

Όνομα Πεδίου	Συντομογραφία	Περιγραφή	Τιμές
Έμφραγμα Μυοκαρδίου	MI	Ένδειξη εάν ο ασθενής έχει υποστεί έμφραγμα μυοκαρδίου.	Y/N
Αγγειοπλαστική	PCI	Δείχνει κατά πόσο ο ασθενής έχει υποστεί σε αγγειοπλαστική εγχείρηση.	Y/N
Στεφανιαία Παράκαμψη (bypass)	CABG	Δείχνει κατά πόσο ο ασθενής έχει κάνει στεφανιαία παράκαμψη (bypass).	Y/N
Ηλικία	AGE	Αντιπροσωπεύει την ηλικία του ασθενή.	Αριθμός
Φύλο	SEX	Δείχνει το φύλο του ασθενή. Παίρνει τιμές M (MALE) και F (FEMALE).	M/F
Βάρος	W	Αντιπροσωπεύει το βάρος του ασθενή.	Αριθμός
Ύψος	H	Αντιπροσωπεύει το ύψος του ασθενή.	Αριθμός
Δείκτης Μάζας Σώματος	BMI	Ο δείκτης μάζας σώματος υπολογίζει το βάρος ενός ασθενή βάσει του ύψους του.	Αριθμός
Ενεργός Καπνιστής	AS	Δείχνει εάν ο ασθενής είναι ενεργός καπνιστής.	Y/N
Παθητικός Καπνιστής	PS	Δείχνει εάν ο ασθενής είναι παθητικός καπνιστής.	Y/N
Σταμάτημα-Ξεκίνημα καπνίσματος	S-R	Δείχνει εάν ο ασθενής είχε σταματήσει το κάπνισμα και μετά το ξεκίνησε ξανά.	Y/N
Πρώην Καπνιστής	EX-SM	Δείχνει εάν ο ασθενής είναι πρώην καπνιστής.	Y/N
Ιστορικό Οικογένειας	POS FH	Παρουσιάζει το ιστορικό της οικογένειας του ασθενή σε καρδιακά επεισόδια.	Y/N
Υπέρταση	HT	Η υπέρταση είναι η υψηλή πίεση αίματος. Δείχνει αν ο ασθενής πάσχει από υπέρταση.	Y/N
Διαβήτης	DM	Ο διαβήτης χαρακτηρίζεται από υψηλά επίπεδα ζάχαρης στο αίμα. Μπορεί να προκαλείται από πολύ λίγη ινσουλίνη (ορμόνη που παράγεται από το πάγκρεας για να ρυθμίζει την ζάχαρη αίματος), αντίσταση στην ινσουλίνη, ή και τα δύο. Δείχνει αν ο ασθενής πάσχει από διαβήτη.	Y/N
Άγχος	STAT	Δείχνει κατά πόσο ο ασθενής καταβάλλεται από άγχος.	Y/N
Άσκηση	EXER	Δείχνει κατά πόσο ο ασθενής γυμνάζεται ή όχι.	Y/N
Παλμοί καρδίας	HR	Δείχνει τους παλμούς της καρδίας του ασθενούς.	Αριθμός

…/συνεχίζεται

…/συνέχεια Πίνακα 6.1

Όνομα Πεδίου	**Συντομογραφία**	**Περιγραφή**	**Τιμές**
Ψηλή Πίεση (Συστολική Πίεση)	SBP	Η συστολική πίεση αναπαριστά τη μέγιστη πίεση που εξασκείται όταν η καρδιά συστέλλεται.	Αριθμός
Χαμηλή Πίεση (Διαστολική Πίεση)	DBP	Η διαστολική πίεση αναπαριστά την πίεση στις αρτηρίες όταν η καρδιά είναι ξεκούραστη.	Αριθμός
Ολική Χοληστερόλη	TC	Δείχνει την ολική ποσότητα χοληστερόλης στο αίμα.	Αριθμός
Λιποπρωτεΐνες Υψηλής Πυκνότητας	HDL	Δείχνει την ποσότητα των λιποπρωτεϊνών υψηλής πυκνότητας στο αίμα.	Αριθμός
Λιποπρωτεΐνες Χαμηλής Πυκνότητας	LDL	Δείχνει την ποσότητα των λιποπρωτεϊνών χαμηλής πυκνότητας στο αίμα.	Αριθμός
Τριγλυκερίδια	TG	Δείχνει την ποσότητα των τριγλυκεριδίων στο αίμα.	Αριθμός
Γλυκόζη	GLU	Δείχνει την ποσότητα της γλυκόζης στο αίμα.	Αριθμός
Ουρία – Ουρικό Οξύ	UA	Δείχνει την ποσότητα της ουρίας στο αίμα.	Αριθμός
Fibrinogen	FIBR	Ινοδογόνο	Αριθμός

Στη βάση δεδομένων υπήρχαν πολλές πλειάδες που είχαν ελλιπείς τιμές. Σαν πρώτο βήμα έγινε έλεγχος των γραπτών αναφορών των ασθενών. Συμπληρώθηκαν κάποιες τιμές που δεν είχαν περαστεί στη βάση δεδομένων όπως επίσης διορθώθηκαν τιμές που δεν ήταν σωστές. Μετά εφαρμόστηκαν οι τύποι που έχουν να κάνουν με το δείκτη μάζας σώματος, το ύψος και το βάρος, τα τριγλυκερίδια και τη χοληστερόλη.

Αφού τελείωσε αυτή η διαδικασία, όσες πλειάδες είχαν ακόμη ελλιπείς τιμές αγνοήθηκαν. Παρόλο που υπάρχουν διάφορες τεχνικές συμπλήρωσης ελλιπών τιμών όπως με τον υπολογισμό του μέσου όρου ή της μέσης τιμής, σε αυτή την περίπτωση θεωρήσαμε ότι αυτό θα αλλοίωνε τα αποτελέσματά μας και έτσι δεν εφαρμόσαμε καμιά από αυτές τις τεχνικές. Έτσι καταλήξαμε σε μια βάση δεδομένων με 528 περιπτώσεις, όπου είχαμε 358 περιπτώσεις με έμφραγμα μυοκαρδίου, 213 με αγγειοπλαστική και 215 με στεφανιαία παράκαμψη. Ο Πίνακας 4.2 παρουσιάζει τη βάση δεδομένων κατά μοντέλο.

Πίνακας 4.2: Κατανομή περιπτώσεων ανά τάξη

	MI		PCI		CABG	
	Y	N	Y	N	Y	N
Περιπτώσεις	358	170	213	315	215	313

Το επόμενο στάδιο ήταν η κωδικοποίηση των τιμών των παραγόντων. Λαμβάνοντας υπόψη τόσο τις διεθνείς προδιαγραφές, όσο και τις υποδείξεις των ειδικών γιατρών προχωρήσαμε στην κωδικοποίηση των παραγόντων. Οι παράγοντες που επιλέξαμε για να μελετήσουμε είναι δύο ειδών: 1. κλινικοί και 2. βιοχημικοί. Επίσης οι παράγοντες μπορούν να διαχωριστούν σε δύο κατηγορίες: α. στους μεταβαλλόμενους και β. στους μη μεταβαλλόμενους. Μεταβαλλόμενοι είναι οι παράγοντες που μπορούν να αλλάξουν, για παράδειγμα η χοληστερόλη. Μη μεταβαλλόμενοι είναι οι παράγοντες που δεν μπορούμε να τους αλλάξουμε, για παράδειγμα η ηλικία.

Οι παράγοντες που έχουν επιλεγεί μπορούν να ομαδοποιηθούν χρονολογικά σε δύο κατηγορίες: τους παράγοντες που έχουμε πριν από το επεισόδιο και αυτούς που καταγράφηκαν μετά το επεισόδιο. Για το λόγο αυτό έχουμε δημιουργήσει εννέα μοντέλα:

- Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά πριν το επεισόδιο (B)
- Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά μετά το επεισόδιο (A)
- Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά πριν και μετά το επεισόδιο (B+A)
- Μοντέλο Αγγειπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά πριν από το επεισόδιο (B)
- Μοντέλο Αγγειπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά μετά από το επεισόδιο (A)
- Μοντέλο Αγγειπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)
- Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν από το επεισόδιο (B)
- Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά μετά από το επεισόδιο (A)
- Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

Τα παραπάνω μοντέλα χρησιμοποιήθηκαν για την εξαγωγή κανόνων και των μέτρων αξιολόγησής τους.

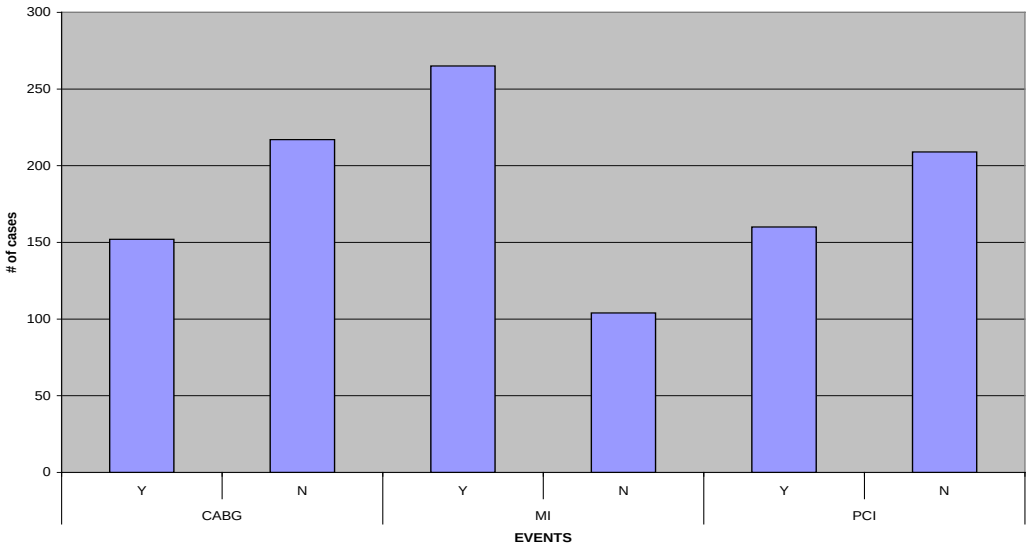
Ο Πίνακας 4.3 παρουσιάζει την κωδικοποίηση των χαρακτηριστικών που επιλέγηκαν για να μελετηθούν.

Πίνακας 4.3: Κωδικοποίηση χαρακτηριστικών

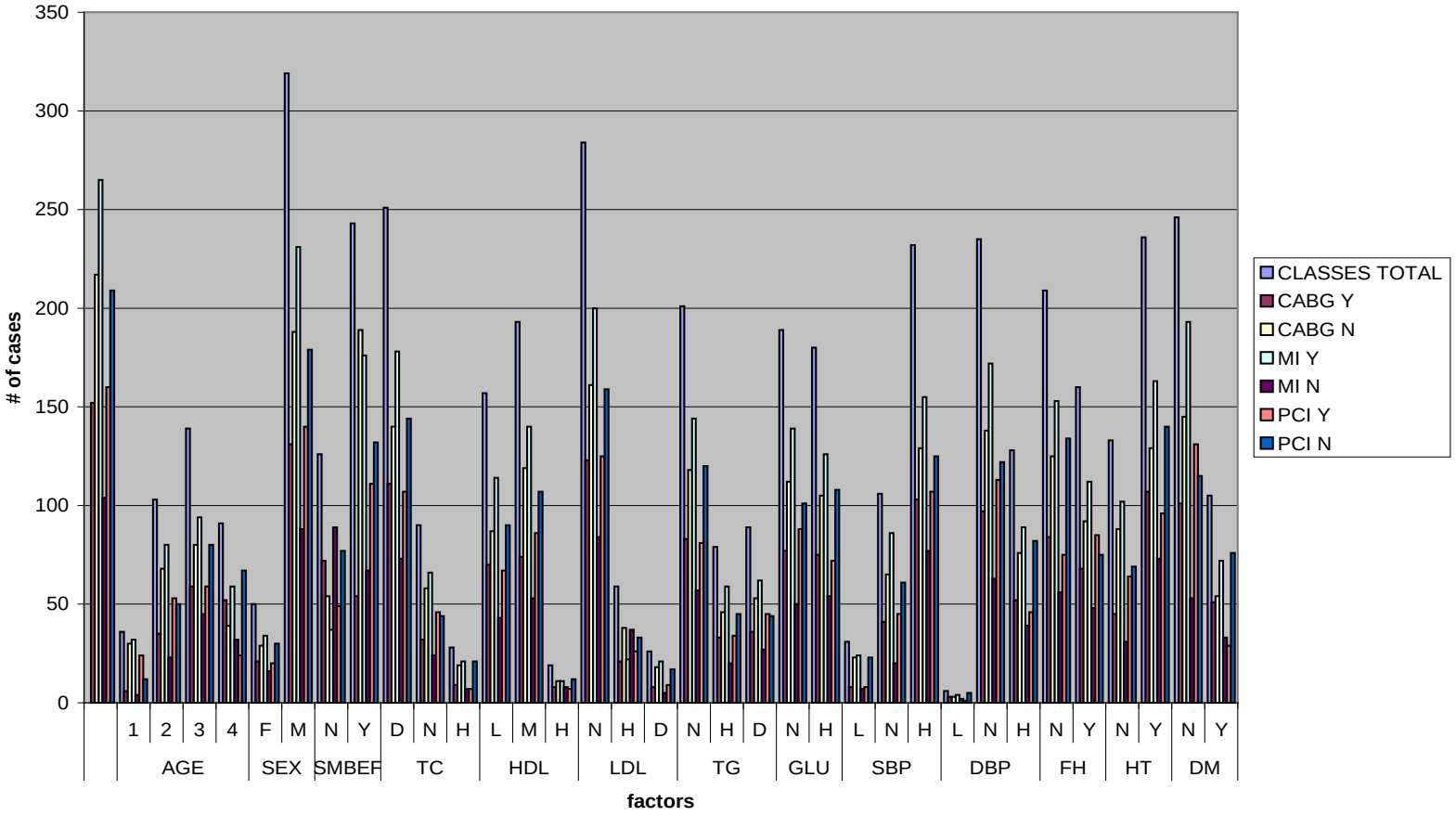
	Παράγοντες	**Κωδικ. 1**	**Κωδικ. 2**	**Κωδικ. 3**	**Κωδικ. 4**
Κλινικοί παράγοντες					
1	AGE	1: 34-50	2: 51-60	3:61-70	4: 71-85
2	SEX	M: MALE	F:FEMALE		
3	SMBEF	Y: YES	N: NO		
4	SBP*	L<90	N:90-120	H>20	
5	DBP *	L<60	N:60-80	H>80	
6	FH	Y: YES	N: NO		
7	HT	Y: YES	N: NO		
8	DM	Y: YES	N: NO		
Βιοχημικοί παράγοντες					
9	TC **	D <200	N:201 –240	H>240	
10	HDL** Women Men	 L<50 L<40	 M:50-60 M:40-60	 H>60	
11	LDL**	N<130	H:131-160	D>60	
12	TG**	N<150	H:151-200	D>200	
13	GLU**	H>110	N <110		

* in mmHg ** in mg/dL

Το Σχήμα 4 2 αναπαριστά τον αριθμό ασθενών σε σχέση με τις διάφορες τιμές των τάξεων, MI, CABG και PCI. Παρατηρούμε ότι για τις τάξεις CABG και PCI, ο αριθμός των ασθενών για τις τιμές 'Y' και 'N' των τάξεων είναι περίπου ο ίδιος ενώ για την τάξη MI παρουσιάζονται περισσότερες δοσοληψίες με τιμή 'Y' παρά με 'N'. Ενώ το Σχήμα 4 3 παρουσιάζει τον αριθμό των ασθενών για κάθε χαρακτηριστικό. Παρατηρούμε ότι παρουσιάζονται περισσότερες περιπτώσεις να είναι άνδρας (SEX = M), και πολλές περιπτώσεις με κανονική ποσότητα των λιποπρωτεϊνών χαμηλής πυκνότητας στο αίμα (LDL = N).



Σχήμα 4.2: Κατανομή περιστατικών για τις τάξεις CABG, MI και PCI



Σχήμα 4.3: Αριθμός περιστατικών έναντι κωδικοποιημένων χαρακτηριστικών για τις τάξεις CABG, MI και PCI

4.3 Κωδικοποίηση μέτρων αξιολόγησης

Η κωδικοποίηση των τιμών των μέτρων αξιολόγησης που εξάγονται από την εκτέλεση του αλγόριθμου Apriori και Δέντρων Απόφασης είναι *αναγκαία* έτσι ώστε να μπορούμε να τρέξουμε τον αλγόριθμο Δέντρων Απόφασης με είσοδο τα κωδικοποιημένα μέτρα. Η έξοδος του δέντρου μας παρουσιάζει τα πιο σημαντικά μέτρα.

4.3.1 Κωδικοποίηση αντικειμενικών μέτρων

Χρησιμοποιώντας τα αρχικά σύνολα δεδομένων, εκτελέσαμε τον αλγόριθμο συσχέτισης και Decisio n Tree ξεχωριστά με μοναδικό περιορισμό το Support να είναι μεγαλύτερο του 0.01 (1%). Επιλέξαμε αυτό το κατώφλι (threshold) έτσι ώστε να ανακτηθούν όσο το δυνατόν περισσότερο ι κανόνες. Το πλήθος των κανόνων που έχουν εξαχθεί από την εκτέλεση του αλγόριθμου συσχέτισης και ταξινόμησης παρουσιάζεται στον Πίνακα 5.3.

Οι δύο αλγόριθμοι εκτός από τα χαρακτηριστικά που συμμετέχουν στον κανόνα, εξάγουν και την τιμή που παίρνει κάθε μέτρο αξιολόγησης για το συγκεκριμένο κανόνα.

Για κάθε μέτρο αξιολόγησης ξεχωριστά, συγκεντρώσαμε τις τιμές που παίρνει σε όλους τους κανόνες για όλα τα σύνολα δεδομένων.

Βασισμένοι σε στατιστικά δεδομένα (ελάχιστη / μέγιστη τιμή και την τάση των τιμών) και λαμβάνοντας υπόψη την τιμή του κάθε μέτρου αξιολόγησης που παρουσιάζει ενδιαφέρον, κωδικοποιήσαμε την τιμή που έπαιρνε το κάθε μέτρο αξιολόγησης σε όλους τους κανόνες.

Το κάθε μέτρο αξιολόγησης κωδικοποιήθηκε σε τρεις κατηγορίες 1 , 2 και 3 . Για να βρούμε την κωδικοποίηση της κάθε τιμής του κάθε μέτρου βασιστήκαμε στον ακόλουθο πίνακα 4.4:

Πίνακας 4.4: Κωδικοποίηση μέτρων αξιολόγησης και καθορισμός του μέγιστου αριθμού κανόνων σε κάθε κατηγορία

Κατηγορία	Μέγιστος αριθμός κανόνων	Σημαντικότητα
3	50	Πιο Σημαντικοί
2	100	Σημαντικοί
1	Υπόλοιποι	Λιγότερο Σημαντικοί

Με τον τρόπο αυτό όσοι κανόνες ανήκουν στην Κατηγορία 3 για το συγκεκριμένο μέτρο αξιολόγησης, θεωρούνται σημαντικοί ενώ οι κανόνες που μετά την κωδικοποίηση του κάθε μέτρου αξιολόγησης που έγινε ανήκουν στην Κατηγορία 1 για το συγκεκριμένο μέτρο αξιολόγησης, δεν αξιολογούνται ως σημαντικοί σύμφωνα με το μέτρο αξιολόγησης αυτό.

Επιλέχθηκε σε κάθε κατηγορία να ανήκει ο συγκεκριμένος αριθμός των κανόνων έτσι ώστε όταν εισάγουμε τα κωδικοποιημένα μέτρα αξιολόγησης στο δέντρο απόφασης, να δημιουργηθεί ένα δέντρο στο οποίο να παρουσιάζονται τα μέτρα αξιολόγησης με τιμές 1, 2 και 3. Οπότε όσο πιο λίγοι είναι κανόνες που ανήκουν στην κατηγορία 3 για ένα μέτρο αξιολόγησης, τόσο πιο λίγοι θα είναι και οι κανόνες που θα εξαχθούν από το δέντρο απόφασης με τιμή 3 (Κατηγορία 3) αλλά ταυτόχρονα οι κανόνες που θα εξαχθούν με τιμή του μέτρου αξιολόγησης = 3, θα είναι και οι πιο σημαντικοί.

Στον παρακάτω πίνακα 4.5 παρουσιάζουμε την κωδικοποίηση που έτυχαν τα μέτρα αξιολόγησης [25].

Πίνακας 4.5: Κωδικοποίηση μέτρων αξιολόγησης

	Μέτρα	Κωδικ. 1	Κωδικ. 2	Κωδικ. 3
1	Support	<0.2	>=0.2 AND <0.3	>=0.3
2	Confidence	<0.82	>=0.82 AND <0.84	>=0.84
3	Coverage	<0.4	>=0.4 AND <0.5	>=0.5
4	Recall	<0.4	>=0.4 AND <0.5	>=0.5
5	Specificity	<0.7	>=0.7 AND <0.73	>=0.73
6	Accuracy	<0.4	>=0.4 AND <0.5	>=0.5
7	Lift/Interest	<0.8	>=0.8 AND <0.83	>=0.83
8	Leverage	<0.7	>=0.7 AND <0.73	>=0.73
9	AddedValue	>=0.2 AND <=1	>-0.2 AND <0.2	>=-1 AND <=-0.2
10	Relative Risk	<0.8	>=0.8 AND <0.83	>=0.83
11	Odds Ratio	>=0.68	>=0.64 AND <0.67	>=0.63
12	Conviction	>=0.82	>=0.79 AND <0.81	>=0.78

Σ' αυτό το σημείο πρέπει να αναφέρουμε ότι τα μέτρα Chi-Square, P-Value και Framingham Event Risk δεν έχουν κωδικοποιηθεί για το λόγο ότι ως στατιστικά μέτρα που είναι θα χρησιμοποιηθούν στο τέλος της μεθοδολογίας μας για την αξιολόγηση του κάθε εξαγόμενου κανόνα.

4.3.2 Αλγόριθμος Υλοποίησης Chi-Square statistic

Παρακάτω παρουσιάζεται ο αλγόριθμος [25] που χρησιμοποιήθηκε για τον υπολογισμό του Chi-Square statistic.

Input:
Rules dataset
Events Dataset

For Each Rule
ChiSquareValue = 0
Find the number of attributes in the rule
Find the value of each attribute
Create the Chi Square ***Observed*** Table (Class[2] x NumberOfAttributes^2)
For Each Row in Events Dataset
a. Check if the attribute's value(s) is satisfied
b. Increase the number of the appropriate cell in the Chi Square Observed Table
Create the Chi Square ***Expected*** Table
For Each Cell in Chi Square Observed Table

a. $$ExpectedTableCell_{(i,j)} = \frac{(Total\#ofValues\ in\ Row_{(i)})x(Total\#ofValues\ in\ Column_{(j)})}{TotalNumberofEvents}$$

b. $$ChiSquareValue = \sum_{i=0,\,j=0}^{i=2,\,j=NumberOfAttributes^2} \frac{(Observed_{(i,j)} - Expected_{(i,j)})^2}{Expected_{(i,j)}}$$

Return ***ChiSquareValue***

End Loop

Σχήμα 4.4: Αλγόριθμος Chi-Square Statistic

Περιγραφή του αλγόριθμου για την εύρεση του Chi-Square Statistic για ένα κανόνα

Η εύρεση του Chi-Square Statistic για ένα κανόνα υπολογίζεται από το άθροισμα όλων των τετραγώνων της διαφοράς της συχνότητας των περιστατικών που παρατηρήθηκαν (Observed) σε όλα συμβάντα – της αναμενόμενης συχνότητας των αποτελεσμάτων (Expected), δια της αναμενόμενης συχνότητας των αποτελεσμάτων (Expected).

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

όπου

X^2 = το Chi-Square statistic,

O_i = η συχνότητα με την οποία παρατηρήθηκε το συμβάν,

E_i = η αναμενόμενη συχνότητα των συμβάντων βασισμένη στη μηδενική υπόθεση,

n = ο συνολικός αριθμός των περιστατικών.

Η τιμή του Chi-Square statistic θα χρησιμοποιηθεί στη συνέχεια για τον υπολογισμό του p-Value έτσι ώστε να αποφανθούμε αν ο κανόνας είναι σημαντικός (significant) ή όχι.

Παράδειγμα υπολογισμού του Chi-Square statistic

Ως παράδειγμα θα ελέγξουμε και θα υπολογίσουμε το Chi-Square statistic του κανόνα αν SEX = “Male” $\Rightarrow$ MI = “Yes”.

Από μία βάση με 100 περιστατικά παρατηρήθηκαν (observed) τα εξής αποτελέσματα για τον παραπάνω κανόνα:

Πίνακας 4.6: Παρουσίαση αποτελεσμάτων που παρατηρήθηκαν (Observed)

Rule: SEX = “M” $\Rightarrow$ MI = “Y”.		SEX		
		M	***F***	Row Total
Class	***Y***	50	20	70
	N	20	10	30
	Column Total	70	30	100

Για κάθε περίσταση υπολογίζουμε το αναμενόμενο αποτέλεσμα από τη σχέση:

$$E_{i,j} = \frac{Row_i x Column_j}{Total\ \ Events}$$

Οπότε για την περίσταση SEX = “M” AND Class = “Y” το αναμενόμενο αποτέλεσμα είναι : $E_{Y,M} = \frac{70x70}{100} = 49$

Με τον ίδιο τρόπο υπολογίζονται και το αναμενόμενο αποτέλεσμα για όλες τις άλλες σχέσεις

$$E_{N,M} = \frac{30x70}{100} = 21 \text{ , } E_{Y,F} = \frac{70x30}{100} = 21 \text{ και } E_{N,F} = \frac{30x30}{100} = 9 \text{ .}$$

Το Chi-Square statistic υπολογίζεται από τη εξίσωση 3.1.3 και για αυτό τον κανόνα ισούται με

$$X^2 = \frac{(50-49)^2}{49} + \frac{(20-21)^2}{21} + \frac{(20-21)^2}{21} + \frac{(10-9)^2}{9} = 0.226757.$$

4.3.3 Αλγόριθμος Υλοποίησης πιθανότητας p-Value

Παρακάτω παρουσιάζεται ο αλγόριθμος [25] που χρησιμοποιήθηκε για τον υπολογισμό του p-Value.

Input:
Rules dataset
Chi Square Test values

For Each Rule

1. P-Value = 0
2. Find the Number of Attributes in the Rule
3. Calculate Degrees of Freedom

   Degrees of Freedom = (c – 1) x (r-1)
   = [(#of Attributes)^2 – 1] x [(ClassValues) – 1]

4. Calculate P-Value

$$\text{P-Value} = \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{k/2-1} e^{-x/2}$$

   Where: k = Degrees of Freedom
   x = Chi Square Value
   Γ() = Gamma Function

5. Return ***P-Value***

End Loop

Σχήμα 4.5: Αλγόριθμος p-Value.

<u>Παράδειγμα υπολογισμού του p-Value</u>

Χρησιμοποιώντας το ίδιο δείγμα περιστατικών που χρησιμοποιήθηκε για τον υπολογισμό του Chi-Square statistic βρίσκουμε ότι το $\chi^2 = 0.226757$.

Στη συνέχεια υπολογίσουμε το βαθμό ελευθερίας k από τη σχέση:

k = (r – 1) x (c – 1) όπου

r = το πλήθος των γραμμών του πίνακα X

c = το πλήθος των στηλών του πίνακα X

Στο παράδειγμα μας το r = 2 και το c = 2 $\Rightarrow$ k = 1

Οπότε η τιμή του P-Value σε αυτό το παράδειγμα είναι P-Value = 0.633982 άρα ο κανόνας SEX = ”Male” $\Rightarrow$ MI = “Yes” **δεν** είναι σημαντικός (significant).

4.3.4 Αλγόριθμος Υλοποίησης Framingham Event Risk

Παρακάτω παρουσιάζεται ο αλγόριθμος [25] που χρησιμοποιήθηκε για τον υπολογισμό του Framingham Event Risk.

Input:
Rules dataset
Events Dataset
Non Coded Events Dataset

For each Row on Non Coded Events Dataset
1. Calculate Event Risk (ER) of each event based on equations:

$\mu = 15.5303 - 0.9119 \times \log_{10}(SBP) - 0.2767 \times (Smoking) - 0.7181 \times \log_{10}(TC/HDL) - 0.5865 \times (ELVH) - 1.4792 \times \log_{10}(AGE) - 0.1759 \times (DM)$

Where: SBP = Systolic Blood Pressure
TC = Total Cholesterol
HDL = High Density Lipoprotein Cholesterol
ELVH = Electrocardiographic Left Ventricular Hypertrophy
DM = Diabetes
Variables smoking, electrocardiographic left ventricular hypertrophy, and diabetes are set to 1 whenpresent and 0 when absent.

$\sigma = e^{(-0.3155-0.2784x(\mu-4.4181))}$

ER = $1 - e^{-e^{\sigma}}$ for each event

2. Add ER value into a new Column(ER) in Events dataset

End Loop

For Each Rule
3. EventRiskValue=0, Counter=0
4. Find the Number of Attributes in the Rule
5. Find the value of each attribute
6. For Each Row in Events Dataset
   a. Check if the attribute's value(s) is/are satisfied
   b. IF (Yes)
      i. EventRiskValue += ER
      ii. Counter ++
   c. ENDIF

   End Loop
7. Event Risk = EventRiskValue / Counter(ER for each rule)
8. Return ***Event Risk***

End Loop

Σχήμα 4.6: Αλγόριθμος Framingham Event Risk

<u>Εξήγηση αλγορίθμου και παράδειγμα υπολογισμού του Framingham Event Risk</u>

Κατά τον υπολογισμό του Framingham Event Risk για κάθε κανόνα, εκτός από την κωδικοποιημένη βάση, χρησιμοποιούμε και την αρχική βάση ανακτώντας τις τιμές SBP,SMBF, TC, HDL, ELVH, AGE και DM για κάθε πλειάδα. Το Framingham Event Risk υπολογίζεται από την εξίσωση 3.15.

Για τον κανόνα SEX = M, AGE = 4 , CLASS = Y παρουσιάζονται 5 περιστατικά. Ανατρέχοντας στην αρχική βάση ανακτούμε τα περιστατικά και υπολογίζουμε για κάθε περιστατικό το event risk.

Πίνακας 4.7: Πλειάδες που ικανοποιούν τον κανόνα από την μη κωδικοποιημένη βάση

AGE	SEX	SMBF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	CLASS	risk
82	M	Y	159	33	101	124	117	120	80	Y	Y	Y	Y	0.148122
82	M	N	208	33	148	136	127	150	90	N	Y	Y	Y	0.144195
77	M	Y	182	30	107	205	160	130	90	Y	Y	Y	Y	0.15054
77	M	N	134	41	49	219	106	120	75	N	N	N	Y	0.126389
83	M	N	109	47	45	85	103	150	90	N	Y	Y	Y	0.133311

Στην τελευταία στήλη του πίνακα παρουσιάζεται το Framingham Event Risk έτσι όπως υπολογίστηκε από τις παραπάνω εξισώσεις.

Αφού υπολογίσουμε για κάθε πλειάδα το Event Risk, το Framingham Event Risk του κανόνα υπολογίζεται από το μέσο όρο των Event risks των συμβάντων.

Οπότε το Framingham Even Risk του κανόνα SEX = M, AGE = 4, CLASS = Y είναι:

Event Risk = 0.140511 = 14.05% δηλαδή συμβάντα που ικανοποιούν τον παραπάνω κανόνα βρίσκονται στη μεσαία κατηγορία κινδύνου (medium risk) να έχουν καρδιακό επεισόδιο.

4.4 Δημιουργία Δέντρου Απόφασης βασισμένο στα μέτρα

Για τη δημιουργία του δέντρου απόφασης με τα κωδικοποιημένα μέτρα ακολουθήσαμε τα παρακάτω βήματα:

- Εισαγωγή κανόνων με τις κωδικοποιημένες τιμές των μέτρων
- Καθορισμός της μεθόδου διαχωρισμού
- Επιλογή για κλάδεμα του δέντρου
- Καθορισμός του ποσοστού των δεδομένων εκπαίδευσης
- Εκτέλεση αλγορίθμου
- Ανάκτηση από το δέντρο απόφασης των σημαντικών μέτρων

Μετά την κωδικοποίηση των μέτρων που έγινε στο προηγούμενο βήμα, έχουμε δημιουργήσει ένα .arff αρχείο που περιέχει το νέο σύνολο των δεδομένων (dataset) μας. Ένα δείγμα των κωδικοποιημένων μέτρων παρουσιάζεται στο Πίνακα 4.8. Πλέον σε αυτό το σύνολο δεδομένων δεν υπάρχουν τα αρχικά χαρακτηριστικά (attributes) της βάσης μας (π.χ. Sex, Age, SMBEF κτλ.) και τη θέση τους παίρνουν τα κωδικοποιημένα μέτρα.

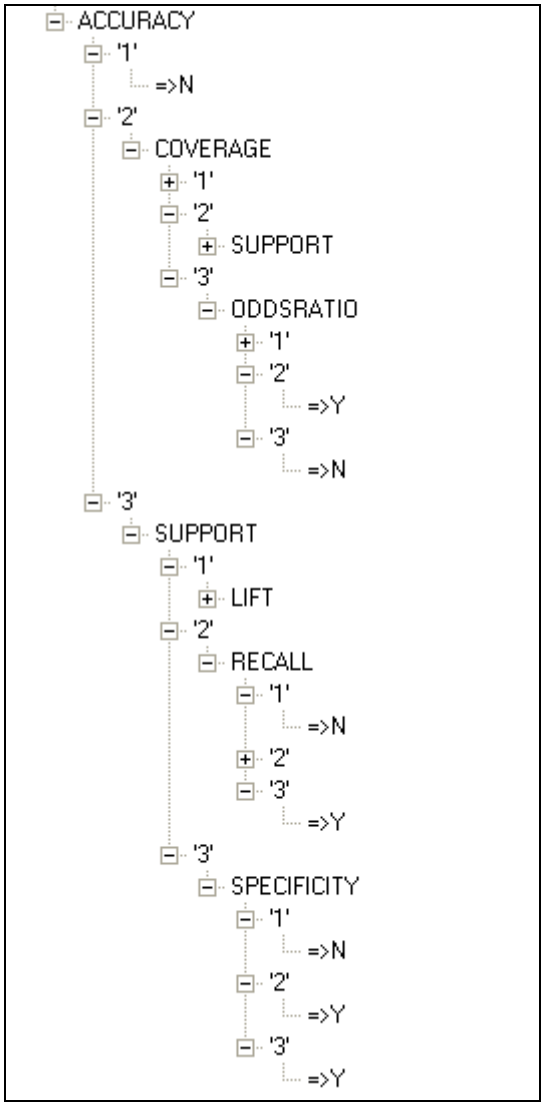
Μία πλειάδα στο νέο σύνολο δεδομένων μας αντιστοιχεί σε ένα κανόνα που έχει ανακτηθεί είτε με τη χρήση της μεθόδου Συσχέτισης είτε της Κατηγοριοποίησης.

Πίνακας 4.8: Μέτρα αξιολόγησης μετά την κωδικοποίηση

Support	Confidence	Coverage	Recall	Specificity	Accuracy	Lift	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	CLASS
1	1	1	1	1	2	1	1	1	1	1	1	N
3	1	2	3	1	3	1	1	1	1	1	1	N
1	1	2	2	1	2	1	1	1	1	2	1	Y
2	1	3	2	1	2	1	1	1	1	2	1	N
2	1	3	3	1	3	1	1	1	1	1	1	Y
3	1	3	3	1	3	1	1	1	1	3	1	N
3	1	3	3	3	2	1	1	1	1	1	1	Y
1	1	1	1	1	2	1	1	1	1	1	1	N
1	1	1	1	1	2	1	1	1	1	1	1	N

Το νέο σύνολο δεδομένων που αναπαριστά τους κανόνες και ως χαρακτηριστικά τα μέτρα, εισήχθηκε στο Δέντρο Απόφασης έτσι ώστε να δημιουργηθεί το δέντρο αποτελούμενο από τα μέτρα. Το δέντρο απόφασης που δημιουργείται παρουσιάζει τα μέτρα ως κόμβους του δέντρου και την τάξη του κάθε κανόνα ως φύλλα του δέντρου.

Στο Σχήμα 4.7 παρουσιάζεται ένα στιγμιότυπο του δέντρου που δημιουργήθηκε για το σύνολο δεδομένων PCI πριν + μετά.



Σχήμα 4.7: Παράδειγμα Δέντρου Απόφασης

Μελετώντας το δέντρο μας μπορούμε εύκολα να προσδιορίσουμε για ποιους κανόνες η τάξη μας είναι "Yes" και να βρούμε για ποιες κατηγορίες των μέτρων ισχύει αυτό.

Παραδείγματος χάριν βασισμένοι στο παραπάνω δέντρο απόφασης παρατηρούμε ότι για Accuracy = 3, Support = 3 και Specificity = 3 η τάξη των κανόνων είναι “Yes”.

Πλέον η έξοδος του δέντρου απόφασης παρουσιάζει ένα μειωμένο πλήθος κανόνων κατηγοριοποιημένο ανάλογα με τις τιμές που έχουν πάρει τα μέτρα.

Για να λάβουμε τους κατηγοριοποιημένους κανόνες από το δέντρο απόφασης έχουμε εκτελέσει για κάθε σύνολο δεδομένων 10 φορές (10-fold cross validation) τον αλγόριθμο και καταγράψαμε για ποιες τιμές των μέτρων μας έχουμε τους πιο σημαντικούς κανόνες και το ποσοστό των ορθών ταξινομημένων περιστατικών.

Επιπρόσθετα πρέπει να αναφέρουμε ότι για τη δημιουργία του δέντρου χρησιμοποιήθηκαν 30% των κανόνων για έλεγχο του δέντρου και 70% για εκπαίδευση του αλγορίθμου. Τα αποτελέσματα παρουσιάζονται αναλυτικά στο επόμενο κεφάλαιο.

Οπότε στο επόμενο βήμα, βασισμένοι στους πιο σημαντικούς κανόνες που έχουν εξαχθεί από το δέντρο απόφασης και λαμβάνοντας υπόψη την τιμή των μέτρων που παίρνουν σε κάθε κανόνα, θα έρθουμε να εκτελέσουμε ξανά τους αλγόριθμους συσχέτισης και κατηγοριοποίησης αλλά με είσοδο:

- Την αρχική μας βάση και
- Το κατώφλι των μέτρων του κάθε κανόνα

4.5 Φιλτράρισμα Αρχικών Κανόνων

Για το φιλτράρισμα των αρχικών κανόνων και την ανάκτηση των σημαντικότερων από αυτούς ακολουθήσαμε τα παρακάτω βήματα:

- Εισαγωγή στον αλγόριθμο Apriori (αντιστοίχως στο C4.5) το αρχικό μας σύνολο δεδομένων της καρδιαγγειακής βάσης
- Καθορισμός του ποσοστού των δεδομένων εκπαίδευσης
- Επιλογή σημαντικών μέτρων αξιολόγησης
- Εισαγωγή ανάλογου κατωφλιού σε όλα τα σημαντικά μέτρα που ανακτήθηκαν από το δέντρο απόφασης.
- Εκτέλεση αλγορίθμου

Στο τελευταίο βήμα της μελέτης μας, χρησιμοποιούμε τις κωδικοποιημένες τιμές των μέτρων αξιολόγησης που έχουμε λάβει από τη μελέτη του δέντρου απόφασης. Αυτές τις τιμές των μέτρων αξιολόγησης τις εισάγουμε τόσο στον αλγόριθμο Apriori όσο και στον αλγόριθμο Δέντρων Απόφασης έτσι ώστε να παρατηρήσουμε πόσοι από τους αρχικούς κανόνες εξάγονται.

Μελετώντας το κάθε δέντρο απόφασης, βρίσκουμε τους πιο σημαντικούς κανόνες για τα 9 διαφορετικά σύνολα δεδομένων. Βασισμένοι στην κωδικοποίηση των μέτρων αξιολόγησης του Πίνακα 4.5, εκτελούμε τον αλγόριθμο συσχέτισης με είσοδο την αρχική βάση μας. Για κάθε σημαντικό κανόνα από το δέντρο απόφασης του κάθε συνόλου δεδομένων, εισάγουμε την κατώτερη τιμή του μέτρου αξιολόγησης του κανόνα που αντιστοιχεί στην κωδικοποιημένη τιμή (1,2,3) του δέντρου απόφασης ως κατώφλι του μέτρου αξιολόγησης και καταγράφουμε τους κανόνες που εξάγονται.

Τους κανόνες αυτούς τους αξιολογούμε σύμφωνα με τις τιμές των στατιστικών μέτρων p-Value και Framingham Event Risk έτσι ώστε να δούμε ποιοι από αυτούς είναι σημαντικοί (significant) καθώς επίσης και την πιθανότητα να συμβεί επεισόδιο ξανά.

Τα αποτελέσματα παρουσιάζονται στο κεφάλαιο 5.4.

Κεφάλαιο 5

Αποτελέσματα

Σ' αυτό το κεφάλαιο παρουσιάζουμε τα αποτελέσματα που εξήχθηκαν από την εκτέλεση της μεθοδολογίας μας. Για τον έλεγχο των αποτελεσμάτων έχουν χρησιμοποιηθεί και τα 9 μοντέλα που έχουμε αναφέρει στο Κεφάλαιο 4.2. Οι κανόνες συσχέτισης έχουν εξαχθεί με τον αλγόριθμο Apriori μέσω του εργαλείου που έχει υλοποιηθεί στο [26] ενώ για τη δημιουργία των Δέντρων Απόφασης χρησιμοποιήθηκε το εργαλείο που υλοποιήθηκε στο [27]. Επιπρόσθετα, τα στατιστικά μέτρα p-value, EventRisk και το ποσοστό CC του κάθε μοντέλου έχουν υλοποιηθεί στο κάθε εργαλείο έτσι ώστε να αποφανθούμε αν ο κάθε κανόνας είναι σημαντικός (p-value), πόσος είναι ο κίνδυνος ένας ασθενής να παρουσιάσει επεισόδιο (EventRisk) και πόσο σωστά έχουν ταξινομηθεί οι κανόνες στο κάθε μοντέλο (CC). Τα εξαγόμενα αποτελέσματα παρουσιάζονται στα επόμενα υποκεφάλαια.

5.1 Εξαγωγή κανόνων και μέτρων αξιολόγησης από αλγόριθμους Apriori και Δέντρων Απόφασης

Έχουν εξαχθεί κανόνες από τα μοντέλα MI (Εμφραγμα Μυοκαρδίου), PCI (Αγγειοπλαστική) και CABG (Στεφανιαία Παράκαμψη) με χαρακτηριστικά πριν (B), μετά (A) και πριν+μετά (B+A) το επεισόδιο. Στιγμιότυπα των κανόνων καθώς επίσης και των τιμών των μέτρων αξιολόγησης από την εκτέλεση των αλγόριθμων Apriori (Πίνακας 5.1) και Δέντρων Απόφασης (Πίνακας 5.2) παρουσιάζονται παρακάτω. Το σύνολο των εξαγόμενων κανόνων από κάθε μοντέλο παρουσιάζεται στον Πίνακα 5.3

Πίνακας 5.1: Επιλεγμένοι κανόνες για το μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν από το επεισόδιο αλγόριθμου Apriori

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	p-Value	Significant	Event Risk	Risk	CC-Rule
1.1	M						Y	0.59	0.68	0.86	0.67	0.88	0.40	0.64	1.02	0.10	0.01	1.14	1.42	1.03	1.29169	0.2557	NS	0.1396	M	68%
1.2		4		Y			Y	0.08	0.70	0.11	0.67	0.11	0.33	0.37	1.04	0.63	0.03	1.05	1.17	1.10	6.56518	0.0871	NS	0.1802	H	70%
1.3	F			N			Y	0.08	0.60	0.13	0.67	0.11	0.32	0.35	0.89	0.51	-0.07	0.87	0.69	0.82	6.38101	0.0945	NS	0.1227	M	60%
1.4				Y		Y	Y	0.09	0.60	0.14	0.67	0.13	0.32	0.36	0.90	0.51	-0.07	0.89	0.71	0.83	13.8507	0.0031	S	0.1496	M	60%
1.5	M			N			N	0.12	0.40	0.29	0.33	0.35	0.70	0.61	1.22	0.31	0.07	1.34	1.56	1.12	6.38101	0.0945	NS	0.1065	M	40%
1.6	M	3					Y	0.22	0.65	0.33	0.67	0.32	0.32	0.43	0.97	0.43	-0.02	0.96	0.87	0.94	2.51368	0.4728	NS	0.1341	M	65%
1.7	M				Y		Y	0.29	0.69	0.42	0.67	0.44	0.35	0.49	1.03	0.41	0.02	1.06	1.19	1.07	3.41193	0.3324	NS	0.1568	H	69%
1.8	F			N		Y	Y	0.03	0.69	0.04	0.67	0.04	0.33	0.35	1.03	0.66	0.02	1.03	1.09	1.06	15.2603	0.0328	S	0.1004	M	69%
1.9		3		Y		Y	Y	0.03	0.59	0.05	0.67	0.04	0.33	0.34	0.88	0.56	-0.08	0.87	0.69	0.80	16.9202	0.0179	S	0.1151	M	59%
1.10	M	4				Y	N	0.02	0.28	0.07	0.33	0.06	0.67	0.64	0.85	0.26	-0.05	0.84	0.78	0.93	12.6125	0.0821	NS	0.1987	H	28%
1.11			N	N		Y	Y	0.05	0.68	0.08	0.67	0.08	0.33	0.36	1.01	0.63	0.01	1.01	1.04	1.03	17.9212	0.0123	S	0.1096	M	68%
1.12			N	Y		Y	Y	0.06	0.67	0.10	0.67	0.10	0.33	0.36	0.99	0.60	0.00	0.99	0.98	0.99	17.9212	0.0123	S	0.1388	M	67%
1.13				Y	N	N	Y	0.18	0.76	0.23	0.67	0.26	0.36	0.45	1.14	0.61	0.09	1.19	1.81	1.40	17.9038	0.0124	S	0.124	M	76%
1.14	F	4	Y			N	Y	0.01	0.80	0.01	0.67	0.02	0.33	0.34	1.19	0.79	0.13	1.20	1.98	1.65	17.6663	0.2806	NS	0.2368	H	80%
1.15	F	4	Y	N			Y	0.01	0.83	0.02	0.67	0.02	0.33	0.34	1.24	0.82	0.16	1.25	2.49	1.98	12.8342	0.6151	NS	0.2293	H	83%
1.16		4	Y	Y		N	Y	0.01	0.50	0.02	0.67	0.02	0.33	0.33	0.75	0.49	-0.17	0.74	0.48	0.66	25.378	0.0451	S	0.1697	H	50%
1.17		1	Y		N	N	Y	0.03	1.00	0.03	0.67	0.04	0.34	0.36	1.49	0.98	0.33	1.51	Infinity	Infinity	24.6096	0.0554	NS	0.0856	M	100%
1.18		2		Y	Y	Y	Y	0.01	0.40	0.03	0.67	0.02	0.32	0.32	0.60	0.38	-0.27	0.59	0.32	0.55	26.3276	0.0347	S	0.197	H	40%
1.19	M	1	Y		N		Y	0.03	0.91	0.03	0.67	0.04	0.34	0.35	1.36	0.89	0.24	1.37	5.08	3.63	21.0972	0.1338	NS	0.0856	M	91%
1.20		4	N	N		Y	N	0.01	0.45	0.03	0.33	0.04	0.67	0.67	1.38	0.44	0.12	1.39	1.72	1.23	25.378	0.0451	S	0.1386	M	45%
1.21		2		N	N	N	Y	0.01	0.36	0.04	0.67	0.02	0.32	0.32	0.53	0.33	-0.31	0.52	0.26	0.51	26.3276	0.0347	S	0.1953	H	36%
1.22		1	N	Y	N		Y	0.05	0.94	0.05	0.67	0.07	0.34	0.37	1.41	0.91	0.27	1.44	8.90	5.94	25.3579	0.0453	S	0.1297	M	94%
1.23		3		Y	Y	N	Y	0.07	0.75	0.10	0.67	0.11	0.34	0.38	1.12	0.68	0.08	1.13	1.53	1.32	25.5851	0.0426	S	0.136	M	75%
1.24	M	4	N		Y	N	Y	0.02	0.75	0.03	0.67	0.04	0.33	0.35	1.12	0.73	0.08	1.12	1.49	1.32	26.4947	0.6973	NS	0.2527	H	75%
1.25	M	3	N		Y	Y	Y	0.02	0.67	0.03	0.67	0.03	0.33	0.34	0.99	0.64	0.00	0.99	0.98	0.99	23.4461	0.8325	NS	0.1045	M	67%

Πίνακας 5.2: Επιλεγμένοι κανόνες για το μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν από το επεισόδιο αλγόριθμου Δέντρων Απόφασης

	SEX	AGE	FH	SMBEF	HXHTN	HXDM	CLASS	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift	Levarage	AddedValue	Conviction	OddsRatio	Chi_Square	P-Value	Significant	EventRisk	Risk
1.1		1					Y	0.09	0.88	0.11	0.68	0.14	0.35	0.40	1.29	0.80	0.20	2.57	3.69	7.9468	0.0048	S	15.6%	H
1.2		2	N	N	N	N	N	0.03	0.67	0.04	0.32	0.08	0.58	0.05	2.07	0.65	0.35	2.04	2.00	44.4481	0.0558	NS	15.7%	H
1.3		2		Y		N	Y	0.12	0.73	0.16	0.68	0.17	0.32	0.14	1.07	0.62	0.05	1.19	0.47	13.3903	0.0632	NS	14.1%	M
1.4		2			N	Y	Y	0.01	0.60	0.01	0.68	0.01	0.28	0.07	0.88	0.59	-0.08	0.80	0.27	2.9185	0.8924	NS	19.0%	H
1.5		2	N		Y	Y	N	0.01	0.33	0.02	0.32	0.03	0.81	0.07	1.04	0.33	0.01	1.02	0.29	14.2622	0.5057	NS	18.0%	H
1.6	M	2	Y	Y	Y	Y	N	0.01	0.25	0.02	0.32	0.02	0.33	0.01	0.78	0.24	-0.07	0.90	0.10	62.3463	0.4996	NS	23.3%	H
1.7	F	3	N	N		N	N	0.02	0.58	0.03	0.32	0.06	0.67	0.05	1.81	0.57	0.26	1.63	1.27	40.5309	0.1175	NS	10.7%	M
1.8	F	3	Y		Y	Y	Y	0.01	0.50	0.01	0.68	0.01	0.32	0.06	0.74	0.49	-0.18	0.64	0.33	34.7944	0.2920	NS	32.7%	H
1.9	M	3	N	Y	N	N	Y	0.02	0.60	0.04	0.68	0.04	1.00	0.03	0.88	0.57	-0.08	0.80	0.75	67.1431	0.3371	NS	14.4%	M
1.10	M	3		Y	Y	N	Y	0.06	0.79	0.08	0.68	0.09	0.00	0.06	1.16	0.73	0.11	1.50	0.00	40.2913	0.1226	NS	17.4%	H
1.11	M	3	N	Y	N	Y	N	0.01	0.43	0.02	0.32	0.03	0.63	0.02	1.33	0.42	0.11	1.19	1.25	67.1431	0.3371	NS	17.6%	H
1.12	M	3	Y	Y		Y	N	0.02	0.55	0.03	0.32	0.05	0.50	0.03	1.70	0.54	0.22	1.49	0.50	40.5309	0.1175	NS	15.1%	H
1.13	F	4					Y	0.02	0.56	0.04	0.68	0.04	0.28	0.22	0.83	0.53	-0.12	0.74	0.42	5.7009	0.1271	NS	15.6%	H
1.14	M	4	Y		N	N	N	0.01	0.50	0.01	0.32	0.02	1.00	0.02	1.55	0.50	0.18	1.36	1.67	40.8420	0.1111	NS	19.7%	H
1.15	M	4			Y	N	Y	0.04	0.63	0.06	0.68	0.06	0.50	0.04	0.92	0.58	-0.05	0.86	1.67	26.4038	0.0340	S	16.7%	H
1.16	M	4		N	Y	Y	Y	0.01	0.83	0.02	0.68	0.02	0.00	0.01	1.23	0.82	0.15	1.93	0.00	45.2284	0.0476	S	15.6%	H
1.17	M	4	N	Y	Y	Y	Y	0.01	0.80	0.01	0.68	0.02	0.00	0.01	1.18	0.79	0.12	1.61	0.00	76.1197	0.1241	NS	17.6%	H
1.18	M	4	Y	Y	Y	Y	Y	0.01	0.50	0.01	0.68	0.01	0.75	0.03	0.74	0.49	-0.18	0.64	3.00	76.1197	0.1241	NS	19.2%	H

Πίνακας 5.3: Πίνακας παρουσίασης του πλήθους των εξαγόμενων κανόνων μετά από 1 εκτέλεση, των σημαντικών καθώς επίσης και του κινδύνου για νέο επεισόδιο από αρχική βάση με Support 0.0 1 για τα μοντέλα Πριν και Μετά και Support 0.1 για τα μοντέλα Πριν+Μετά. Ποσοστό δεδομένων εκπαίδευσης 50%.

Μοντέλα		**Μέθοδος Συσχέτισης**						**Μέθοδος Δέντρων Απόφασης**					
		Κανόνες	Σημαντικοί Κανόνες	Κίνδυνος Επεισοδίου			% CC	Κανόνες	Σημαντικοί Κανόνες	Κίνδυνος Επεισοδίου			% CC
				L	M	H				L	M	H	
Πριν	MI	1161	313	0	441	720	56.1	64	12	0	17	47	66.8
	PCI	1167	1043	0	397	770	55.3	39	29	0	16	23	62.5
	CABG	1192	965	1	450	741	53.4	52	10	0	18	34	58.5
Μετά	MI	5188	1309	23	1744	3380	57.6	74	5	0	34	40	56.0
	PCI	5288	350	31	1796	3391	54.0	75	2	0	28	47	62.4
	CABG	5365	1612	10	1812	3543	56.0	73	7	0	29	44	57.1
Πριν + Μετά	MI	3406	1022	0	797	2609	64.5	229	16	0	86	143	61.4
	PCI	2902	1424	0	1034	1868	57.8	242	38	0	82	160	60.4
	CABG	2758	902	0	673	2085	55.7	270	23	0	96	174	54.4

Όπως παρατηρούμε από τους Πίνακες 5.1 και 5.2, τα μέτρα λαμβάνουν συνεχόμενες τιμές. Αυτό συνεπάγεται ότι αν εισάγουμε τις τιμές των μέτρων με τη μορφή που βρίσκονται τώρα στο Δέντρο Απόφασης, δε θα έχουμε το επιθυμητό αποτέλεσμα. Αντιθέτως θα δημιουργηθεί ένα δέντρο με χιλιάδες κανόνες από το οποίο θα είναι δύσκολο να εξάγουμε οποιαδήποτε γνώση σχετικά με τα μέτρα που τα σημαντικά. Γι' αυτό αποφασίσαμε να κωδικοποιήσουμε τις τιμές των μέτρων σε 1, 2 και 3 έτσι ώστε να είναι δυνατή η εισαγωγή τους στο δέντρο απόφασης.

Από το Πίνακα 5.3 μπορούμε να παρατηρήσουμε ότι το πλήθος των κανόνων που εξάγονται από τον αλγόριθμο Apriori σε κάθε μοντέλο είναι πολύ περισσότερες από τον αριθμό των κανόνων που εξάγονται από τον αλγόριθμο Δέντρων Απόφασης. Αυτό οφείλεται στο ότι το δέντρο απόφασης κτίζει τους κανόνες με όσο το δυνατόν περισσότερα χαρακτηριστικά να συμμετέχουν στον κανόνα ενώ ο αλγόριθμος Apriori βγάζει κανόνες που να ικανοποιείται το ελάχιστο support με αποτέλεσμα σε πολλούς κανόνες να συμμετέχουν πολύ λίγα χαρακτηριστικά.

Επιπρόσθετα, παρατηρούμε ότι με τον αλγόριθμο Apriori, το ποσοστό των σημαντικών (μέτρο p-value) κανόνων - έναντι του συνόλου των κανόνων σε κάθε μοντέλο - που εξάγονται

από τον αλγόριθμο Apriori είναι μεγαλύτερο από αυτό που παράγονται από τον αλγόριθμο Δέντρων Απόφασης.

Όσο αφορά τον κίνδυνο να συμβεί επεισόδιο σε ένα ασθενή (Framingham Event Risk), ακολουθήσαμε τις οδηγίες των ιατρών για να αποφανθούμε αν ένας ασθενής έχει χαμηλό (5%), μέτριο (5%-15%) ή υψηλό (15% και πάνω) κίνδυνο να έχει ένα επεισόδιο. Και στους δύο αλγόριθμους παρατηρούμε ότι ο κίνδυνος επεισοδίου είναι μέτριο-υψηλό στην πλειονότητα των κανόνων σε όλα τα μοντέλα.

Τέλος, σημαντική διαφορά παρουσιάζεται και στο ποσοστό των σωστών ταξινομήσεων (*Correct Classification - CC*) ανάμεσα στους δύο αλγόριθμους. Παρατηρούμε ότι σχεδόν σε όλα τα μοντέλα το ποσοστό των σωστών ταξινομημένων περιπτώσεων της μεθόδου Κατηγοριοποίησης είναι μεγαλύτερο απ' ότι της μεθόδου Συσχέτισης. Αυτό ίσος μας προσφέρει και ένα μέτρο σύγκρισης των δύο μεθόδων για να αποφανθούμε ποια μέθοδο θα προτιμήσουμε.

5.2 Κωδικοποίηση Μέτρων

Για την κωδικοποίηση των μέτρων τρέξαμε όλα τα μοντέλα στους δύο αλγορίθμους. Για κάθε μέτρο ξεχωριστά, συγκεντρώσαμε τις τιμές που παίρνει σε κάθε κανόνα. Για να αποφανθούμε το διάστημα τιμών της κάθε κατηγορίας βασιστήκαμε στα παρακάτω κριτήρια:

- Την τάση τιμών του κάθε μέτρου, δηλαδή για ποιες τιμές το μέτρο αξιολόγησης είναι ενδιαφέρον,
- Την ελάχιστη και μέγιστη τιμή του μέτρου,
- Τον Πίνακα 4.4.

Ο Πίνακας 4.5 παρουσιάζει τα αποτελέσματα της διαδικασίας κωδικοποίησης που έλαβε μέρος για κάθε μέτρο.

Στιγμιότυπα των κανόνων καθώς επίσης και των κωδικοποιημένων τιμών των μέτρων αξιολόγησης από την εκτέλεση των αλγόριθμων Apriori (Πίνακας 5.4) και Δέντρων Απόφασης (Πίνακας 5.5) παρουσιάζονται παρακάτω (βασισμένοι στους Πίνακες 5.1 και 5.2).

Πίνακας 5.4: Κωδικοποιημένοι κανόνες για το μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν από το επεισόδιο αλγόριθμου Apriori

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	CLASS	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	p-Value	Significant	EventRisk	Risk	CC-Rule
1.1	M						Y	3	1	3	0.67	3	1	3	3	1	1	3	1	1	1.29169	0.25574	NS	0.13961	M	68%
1.2		4		Y			Y	1	1	1	0.67	1	1	1	3	1	1	3	1	1	6.56518	0.08713	NS	0.18019	H	70%
1.3	F			N			Y	1	1	1	0.67	1	1	1	3	1	2	3	1	1	6.38101	0.09448	NS	0.1227	M	60%
1.4				Y		Y	Y	1	1	1	0.67	1	1	1	3	1	2	3	1	1	13.85065	0.00312	S	0.14958	M	60%
1.5	M			N			N	1	1	1	0.33	1	2	3	3	1	1	3	1	1	6.38101	0.09448	NS	0.10651	M	40%
1.6	M	3					Y	2	1	1	0.67	1	1	2	3	1	1	3	1	1	2.51368	0.47282	NS	0.13414	M	65%
1.7	M				Y		Y	3	1	2	0.67	2	1	2	3	1	1	3	1	1	3.41193	0.33237	NS	0.15675	H	69%
1.8	F			N		Y	Y	1	1	1	0.67	1	1	1	3	1	1	3	1	1	15.26028	0.0328	S	0.10043	M	69%
1.9		3		Y		Y	Y	1	1	1	0.67	1	1	1	2	1	2	3	1	2	16.92019	0.01792	S	0.1151	M	59%
1.10	M	4				Y	N	1	1	1	0.33	1	1	3	2	1	1	2	1	1	12.6125	0.08213	NS	0.19868	H	28%
1.11			N	N		Y	Y	1	1	1	0.67	1	1	1	3	1	1	3	1	1	17.92117	0.01233	S	0.10957	M	68%
1.12			N	Y		Y	Y	1	1	1	0.67	1	1	1	3	1	1	3	1	1	17.92117	0.01233	S	0.13878	M	67%
1.13				Y	N	N	Y	2	1	1	0.67	1	1	2	3	1	1	3	1	1	17.90377	0.01241	S	0.12398	M	76%
1.14	F	4	Y			N	Y	1	1	1	0.67	1	1	1	3	3	1	3	1	1	17.66634	0.28061	NS	0.23676	H	80%
1.15	F	4	Y	N			Y	1	2	1	0.67	1	1	1	3	3	1	3	1	1	12.8342	0.6151	NS	0.22933	H	83%
1.16		4	Y	Y		N	Y	1	1	1	0.67	1	1	1	2	1	3	1	3	3	25.37798	0.04509	S	0.16973	H	50%
1.17		1	Y		N	N	Y	1	3	1	0.67	1	1	1	3	3	1	3	1	1	24.60962	0.05544	NS	0.08564	M	100%
1.18		2		Y	Y	Y	Y	1	1	1	0.67	1	1	1	1	1	3	1	3	3	26.32758	0.03472	S	0.19702	H	40%
1.19	M	1	Y		N		Y	1	3	1	0.67	1	1	1	3	3	1	3	1	1	21.09717	0.13375	NS	0.08564	M	91%
1.20		4	N	N		Y	N	1	1	1	0.33	1	1	3	3	1	1	3	1	1	25.37798	0.04509	S	0.13856	M	45%
1.21		2		N	N	N	Y	1	1	1	0.67	1	1	1	1	1	3	1	3	3	26.32758	0.03472	S	0.19526	H	36%
1.22		1	N	Y	N		Y	1	3	1	0.67	1	1	1	3	3	1	3	1	1	25.35785	0.04534	S	0.12965	M	94%
1.23		3		Y	Y	N	Y	1	1	1	0.67	1	1	1	3	1	1	3	1	1	25.58514	0.04262	S	0.13597	M	75%
1.24	M	4	N		Y	N	Y	1	1	1	0.67	1	1	1	3	2	1	3	1	1	26.49466	0.6973	NS	0.25266	H	75%
1.25	M	3	N		Y	Y	Y	1	1	1	0.67	1	1	1	3	1	1	3	1	1	23.44611	0.83248	NS	0.10448	M	67%

Πίνακας 5.5: Κωδικοποιημένοι κανόνες για το μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν από το επεισόδιο αλγόριθμου Δέντρων Απόφασης

	SEX	AGE	FH	SMBEF	HXHTN	HXDM	CLASS	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift	Levarage	Added_Value	Conviction	Odds_Ratio	Chi_Square	p-Value	Significant	Event_Risk	Risk
1.1		1					Y	1	3	1	0.68	1	1	2	3	3	1	1	1	7.9468	0.0048	S	15.6%	H
1.2		2	N	N	N	N	N	1	1	1	0.32	1	1	1	3	1	1	1	1	44.4481	0.0558	NS	15.7%	H
1.3		2		Y		N	Y	1	1	1	0.68	1	1	1	3	1	1	1	3	13.3903	0.0632	NS	14.1%	M
1.4		2			N	Y	Y	1	1	1	0.68	1	1	1	3	1	2	2	3	2.9185	0.8924	NS	19.0%	H
1.5		2	N		Y	Y	N	1	1	1	0.32	1	3	1	3	1	1	1	3	14.2622	0.5057	NS	18.0%	H
1.6	M	2	Y	Y	Y	Y	N	1	1	1	0.32	1	1	1	1	1	2	1	3	62.3463	0.4996	NS	23.3%	H
1.7	F	3	N	N		N	N	1	1	1	0.32	1	1	1	3	1	1	1	1	40.5309	0.1175	NS	10.7%	M
1.8	F	3	Y		Y	Y	Y	1	1	1	0.68	1	1	1	1	1	3	3	3	34.7944	0.2920	NS	32.7%	H
1.9	M	3	N	Y	N	N	Y	1	1	1	0.68	1	3	1	3	1	2	2	3	67.1431	0.3371	NS	14.4%	M
1.10	M	3		Y	Y	N	Y	1	1	1	0.68	1	1	1	3	2	1	1	3	40.2913	0.1226	NS	17.4%	H
1.11	M	3	N	Y	N	Y	N	1	1	1	0.32	1	1	1	3	1	1	1	1	67.1431	0.3371	NS	17.6%	H
1.12	M	3	Y	Y		Y	N	1	1	1	0.32	1	1	1	3	1	1	1	3	40.5309	0.1175	NS	15.1%	H
1.13	F	4					Y	1	1	1	0.68	1	1	1	2	1	3	3	3	5.7009	0.1271	NS	15.6%	H
1.14	M	4	Y		N	N	N	1	1	1	0.32	1	3	1	3	1	1	1	1	40.8420	0.1111	NS	19.7%	H
1.15	M	4			Y	N	Y	1	1	1	0.68	1	1	1	3	1	1	1	1	26.4038	0.0340	S	16.7%	H
1.16	M	4		N	Y	Y	Y	1	2	1	0.68	1	1	1	3	3	1	1	3	45.2284	0.0476	S	15.6%	H
1.17	M	4	N	Y	Y	Y	Y	1	1	1	0.68	1	1	1	3	3	1	1	3	76.1197	0.1241	NS	17.6%	H
1.18	M	4	Y	Y	Y	Y	Y	1	1	1	0.68	1	3	1	1	1	3	3	1	76.1197	0.1241	NS	19.2%	H

Πρέπει να αναφέρουμε ότι το μέτρο Prevalence δεν έχει κωδικοποιηθεί και δεν έχει χρησιμοποιηθεί στα επόμενα στάδια της μελέτης μας για το λόγο ότι δε μας προσφέρει οποιαδήποτε γνώση (παρουσιάζει τη πιθανότητα να συμβεί το B, *P(B)*).

Όπως αναφέραμε και στον Πίνακα 4.4, στο κάθε μέτρο παρουσιάζεται περίπου 50 φορές η Κατηγορία 3, 100 φορές η Κατηγορία 2 και τα υπόλοιπα Κατηγορία 1.

5.3 Αποτελέσματα από Δέντρο Απόφασης

Μετά την εκτέλεση του αλγόριθμου Δέντρων Απόφασης με είσοδο τις κωδικοποιημένες τιμές των μέτρων, ανακτήσαμε τις τιμές των μέτρων για τους κανόνες που είναι οι πιο σημαντικοί. Οι παρακάτω πίνακες (Πίνακες 5.6 – 5.11) παρουσιάζουν τα σημαντικά μέτρα με τη χρήση της μεθόδου 10-πτυχης διασταυρούμενης επικύρωσης (10-fold cross validation).

5.3.1 Είσοδος: Κωδικοποιημένα μέτρα από αλγόριθμο Apriori

Πίνακας 5.6: Σημαντικά μέτρα από Δέντρο απόφασης για το μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG

Μοντέλο	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Αρ. Εκτέλεσης
MI vs PCI ή CABG (B)	Accuracy	Support	-	-	10	1
MI vs PCI ή CABG (A)	Accuracy	Support	-	-	15	1
MI vs PCI ή CABG (B+A)	Accuracy	Support	Recall	-	8	2

Πίνακας 5.7: Σημαντικά μέτρα από Δέντρο απόφασης για το μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG

Μοντέλο	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Αρ. Εκτέλεσης
PCI vs MI ή CABG (B)	Accuracy	Recall	-	-	12	3
PCI vs MI ή CABG (A)	Accuracy	Coverage	Recall	-	9	4
PCI vs MI ή CABG (B+A)	Accuracy	Support	Recall	Coverage	16	1

Πίνακας 5.8: Σημαντικά μέτρα από Δέντρο απόφασης για το μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI

Μοντέλο	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Αρ. Εκτέλεσης
CABG vs MI ή PCI (B)	Accuracy	Recall	-	-	7	2
CABG vs MI ή PCI (A)	Accuracy	Support	Recall	-	8	2
CABG vs MI ή PCI (B+A)	Accuracy	Support	p-value	-	8	10

5.3.2 Είσοδος: Κωδικοποιημένα μέτρα από αλγόριθμο Δέντρων Απόφασης

Πίνακας 5.9: Σημαντικά μέτρα από Δέντρο απόφασης για το μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG

Μοντέλο	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Αρ. Εκτέλεσης
MI vs PCI ή CABG (B)	Specificity	Confidence	-	-	6	4
MI vs PCI ή CABG (A)	Added Value	Odds Ratio	-	-	7	3
MI vs PCI ή CABG (B+A)	Specificity	Added Value	-	-	13	5

Πίνακας 5.10: Σημαντικά μέτρα από Δέντρο απόφασης για το μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG

Μοντέλο	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Αρ. Εκτέλεσης
PCI vs MI ή CABG (B)	Conviction	Specificity	-	-	8	5
PCI vs MI ή CABG (A)	Leverage	Confidence	Specificity	-	15	4
PCI vs MI ή CABG (B+A)	Specificity	Confidence	-	-	8	10

Πίνακας 5.11: Σημαντικά μέτρα από Δέντρο απόφασης για το μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI

Μοντέλο	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Αρ. Εκτέλεσης
CABG vs MI ή PCI (B)	Leverage	Confidence	Specificity	-	8	3
CABG vs MI ή PCI (A)	Specificity	Confidence	-	-	14	8
CABG vs MI ή PCI (B+A)	Specificity	Leverage	Confidence	-	9	4

Ως ρίζα του δέντρου αναφέρεται το Μέτρο 1 και στη συνέχεια διακλαδώνεται στις τιμές του Μέτρου 2 κτλ.

5.4 Φιλτραρισμένοι Κανόνες

Βασισμένοι στα αποτελέσματα των Δέντρων Απόφασης που ανακτήθηκαν στο 5.3 (Πίνακες 5.6 – 5.11) και χρησιμοποιώντας ως κριτήριο την αντίστοιχη μη κωδικοποιημένη τιμή του μέτρου, τρέξαμε ξανά τα αρχικά μοντέλα δεδομένων και στους δύο αλγόριθμους. Το πλήθος των εξαγόμενων κανόνων εμφανίζεται στους παραπάνω πίνακες.

Η χρήση των μέτρων στην αξιολόγηση κανόνων έχει ως αποτέλεσμα την ανάκτηση των πιο σημαντικών κανόνων καθώς και τη μείωση το υ πλήθους των κανόνων από μερικές χιλιάδες σε μερικές δεκάδες.

Ένα παράδειγμα παρουσιάζεται στο Πίνακα 5.12 όπου μετά από το φιλτράρισμα των κανόνων του μοντέλου MI Πριν σύμφωνα με τις τιμές των μέτρων που ανακτήθηκαν από το Δέντρο Απόφασης, το πλήθος των εξαγόμενων κανόνων έχει μειωθεί στους 10.

Οι αρχικοί κανόνες έχουν φιλτραριστεί σύμφωνα με τα δεδομένα του Πίνακα 5.6. Σ’ αυτή την περίπτωση μόνο 2 μέτρα θεωρούνται σημαντικά από το Δέντρο Απόφασης, το μέτρο “Accuracy” κατηγορίας 3 και το μέτρο ‘Support” κατηγορίας 3.

Πίνακας 5.6: Φιλτραρισμένοι κανόνες του μοντέλου Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG Πριν

SEX	AGE	FH	SMBEF	HxHTN	HxDM	CLASS	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	Significant	EventRisk	Risk
				Y		Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	1.04399	0.3069	NS	0.1808	H
			Y			Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	6.18087	0.01291	S	0.1646	H
		N				Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	0.438	0.50809	NS	0.1531	H
					N	Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	0.02542	0.87334	NS	0.1557	H
M						Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	1.6761	0.19544	NS	0.1666	H
			Y		N	Y	3	1	2	0.7	2	1	3	1	1	1	1	1	1	15.7117	0.0013	S	0.1601	H
M		N				Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	2.46397	0.48184	NS	0.1546	H
M			Y			Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	6.29737	0.09801	NS	0.1653	H
M					N	Y	3	1	3	0.7	3	1	3	1	1	1	1	1	1	2.45389	0.48368	NS	0.1573	H
M			Y		N	Y	3	1	2	0.7	2	1	3	1	1	1	1	1	1	16.4314	0.02145	S	0.1611	H

Κεφάλαιο 6

Συζήτηση

Στόχος της εργασίας ήταν η αξιολόγηση των κανόνων που εξάγονται από τις μεθόδους Συσχέτισης και Κατηγοριοποίησης με τη χρήση πολλαπλών μέτρων αξιολόγησης. Με αυτή τη μέθοδο θέλαμε να εξάγουμε όχι μόνο ένα περιορισμένο αριθμό κανόνων αξιολογώντας τα μέτρα του κάθε κανόνα αλλά ταυτόχρονα να εξάγουμε στατιστικά σημαντικούς κανόνες που να παρέχουν και μία πρόγνωση στους ιατρούς για τον κίνδυνο που έχει ένας ασθενής για μελλοντικό επεισόδιο.

Στα πλαίσια της εργασίας μελετήσαμε μία ιατρική βάση η οποία παρείχε δεδομένα ασθενών με τριών ειδών καρδιοαγγειακά νοσήματα: α) Έμφραγμα μυοκαρδίου (MI), β) Αγγειοπλαστική (PCI) και γ) Στεφανιαία Παράκαμψη (bypass) (CABG). Στη βάση υπήρχαν πολλές πλειάδες που είχαν ελλιπείς τιμές. Ως πρώτο βήμα έγινε έλεγχος των γραπτών αναφορών των ασθενών. Συμπληρώθηκαν κάποιες τιμές που δεν είχαν περαστεί στη βάση δεδομένων όπως επίσης διορθώθηκαν τιμές που δεν ήταν σωστές. Μετά εφαρμόστηκαν οι τύποι που έχουν να κάνουν με το δείκτη μάζας σώματος, το ύψος και το βάρος, τα τριγλυκερίδια και τη χοληστερόλη. Αφού τελείωσε αυτή η διαδικασία, όσες πλειάδες είχαν ακόμη ελλιπείς τιμές αγνοήθηκαν. Έτσι καταλήξαμε σε μια βάση δεδομένων με 528 περιπτώσεις, όπου είχαμε 358 περιπτώσεις με έμφραγμα μυοκαρδίου, 213 με αγγειοπλαστική και 215 με στεφανιαία παράκαμψη.

Μετά την κωδικοποίηση των τιμών των παραγόντων που έχουν επιλεγεί βάσει των υποδείξεων των ιατρών, ομαδοποιήσαμε χρονολογικά τους παράγοντες σε παράγοντες που

παρατηρήθηκαν πριν από το επεισόδιο και παράγοντες που καταγράφηκαν μετά το επεισόδιο. Για το λόγο αυτό δημιουργήθηκαν εννέα μοντέλα:

- Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά πριν το επεισόδιο (B)
- Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά μετά το επεισόδιο (A)
- Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά πριν και μετά το επεισόδιο (B+A)
- Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά πριν από το επεισόδιο (B)
- Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά μετά από το επεισόδιο (A)
- Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)
- Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν από το επεισόδιο (B)
- Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά μετά από το επεισόδιο (A)
- Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

Για τη μελέτη των παραπάνω μοντέλων τρέξαμε τους αλγόριθμους Συσχέτισης (Apriori) και Δέντρων Απόφασης (Decision Tree) έτσι ώστε να εξάγουμε τους σημαντικούς κανόνες. Με την εκτέλεση των δύο αλγορίθμων, ο αριθμός των εξαγόμενων κανόνων ήταν πολύ μεγάλος με αποτέλεσμα να αναζητήσουμε κάποια άλλη μέθοδο με την οποία θα περιορίζαμε το πλήθος των κανόνων αλλά ταυτόχρονα να ανακτούσαμε και τους πιο σημαντικούς.

Γι' αυτό το λόγο αποφασίσαμε να χρησιμοποιήσουμε τα μέτρα αξιολόγησης του κάθε κανόνα για την επίτευξη του στόχου μας. Μελετήσαμε τη διακύμανση των τιμών του κάθε μέτρου σε όλα τα μοντέλα και για ποιες τιμές του ο κανόνας θεωρείται σημαντικός. Κωδικοποιήσαμε τις τιμές του κάθε μέτρου για να μπορούμε να εισάγουμε στον αλγόριθμο Δέντρων Απόφασης τις κωδικοποιημένες τιμές των μέτρων έτσι ώστε να βρούμε για ποια μέτρα οι κανόνες παρουσιάζουν ενδιαφέρον. Αφού ανακτήσαμε τα σημαντικά μέτρα ανατρέξαμε στους αρχικούς κανόνες των μοντέλων. Χρησιμοποιώντας τις τιμές των μέτρων φιλτράραμε τους αρχικούς κανόνες. Το αποτέλεσμα ήταν να ανακτήσουμε ένα πολύ μικρό αριθμό κανόνων για κάθε μοντέλο. Τους κανόνες αυτούς μπορούμε να τους μελετήσουμε έτσι ώστε να αποφανθούμε ποιοι παράγοντες προκαλούν αυτού του είδους τις παθήσεις. Οι κανόνες μπορούν να χρησιμοποιηθούν είτε σε νέους ασθενείς για πρόβλεψη καρδιοαγγειακών νοσημάτων είτε σε ασθενείς που έχουν παρουσιάσει αυτού του είδους τις παθήσεις.

Επιπρόσθετα, στα πλαίσια της μελέτης έχουν υλοποιηθεί και τα στατιστικά μέτρα αξιολόγησης Chi-Square (χ^2), p-Value, Framingham Event Risk και Correct Classification (CC).

Το στατιστικό μέτρο Chi-Square υπολογίζει τον έλεγχο κάθε στατιστικής υπόθεσης κατά την οποία η δειγματοληπτική κατανομή του στατιστικού αποτελέσματος της δοκιμής είναι μια κατανομή chi-square (χ2). Αυτό ισχύει όταν η μηδενική υπόθεση είναι αληθής ή ασυμπτωτικά αληθής. Σ' αυτή την περίπτωση η δειγματική κατανομή μπορεί να προσδιοριστεί από μία chi-square κατανομή με αποτέλεσμα να γενικεύσουμε το αποτέλεσμα για όσο πιο μεγάλο δείγμα. Η τιμή του Chi-Square statistic χρησιμοποιήθηκε στη συνέχεια για τον υπολογισμό του p-Value έτσι ώστε να αποφανθούμε αν ο κανόνας είναι σημαντικός (significant) ή όχι.

Στον έλεγχο στατιστικών υποθέσεων, η τιμή p-Value προσδιορίζει την πιθανότητα ένας στατιστικός έλεγχος (κανόνας) που παρατηρήθηκε να είναι σημαντικός (significant) ή όχι αν η μηδενική υπόθεση είναι αληθής. Η πιθανότητα p-Value δηλώνει το επίπεδο σημαντικότητας

που παρατηρήθηκε. Αλλά το επίπεδο σημαντικότητας που παρατηρήθηκε αντιπροσωπεύει την πιθανότητα να υπάρχει κάποιο λάθος στον έλεγχο. Κατά συνέπεια όσο πιο μικρή είναι η τιμή του p-Value τόσο πιο μικρή είναι η πιθανότητα να υπάρχει κάποιο λάθος. Σημαντικοί κανόνες θεωρούνται οι κανόνες που η πιθανότητα p-Value είναι μικρότερη η ίση από 0.05.

Υλοποιώντας τα μέτρα Chi-Square και p-Value στους αλγόριθμους Apriori και Δέντρων Απόφασης καταφέραμε να προσδιορίσουμε αν κάποιος κανόνας θεωρείται στατιστικά σημαντικός (statistically significant) ή όχι.

Η εξίσωση Framingham παρουσιάζει το ποσοστό ένας ασθενής να παρουσιάσει ένα καρδιακό επεισόδιο ή όχι. Η εξίσωση προέκυψε μετά από μελέτες και κάποια μεθοδολογία που εφάρμοσαν οι ιατροί προσπαθώντας να βρουν ποιοι παράγοντες επηρεάζουν τα καρδιακά επεισόδια αλλά ταυτόχρονα και πόση βαρύτητα έχει ο κάθε παράγοντας. Σύμφωνα με τους ιατρούς, εάν ένας ασθενής παρουσιάζει ποσοστό 0-5% τότε έχει χαμηλό κίνδυνο (low risk) να πάθει καρδιακό επεισόδιο, από 5-1 5 %τότε βρίσκεται στη μεσαία κατηγορία κινδύνου (medium risk) ενώ από 15% και άνω ο κίνδυνος να πάθει καρδιακό επεισόδιο ο ασθενής είναι αρκετά μεγάλος (high risk).

Υλοποιώντας το μέτρο Framingham Event Risk για κάθε ασθενή και για κάθε κανόνα, οι ιατροί μπορούν να αποφανθούν για κάθε ασθενή τον κίνδυνο που διατρέχει να πάθει καρδιακό επεισόδιο.

Το ποσοστό σωστής ταξινόμησης (Correct Classified Instances) παρουσιάζει πόσες από τις ταξινομημένες πλειάδες έχουν ταξινομηθεί σωστά στους κανόνες έναντι του συνολικού αριθμού ταξινομήσεων. Με το μέτρο αυτό μπορούμε να παρατηρήσουμε το ποσοστό των σωστών ταξινομήσεων του κάθε αλγόριθμου αλλά ταυτόχρονα και των πλειάδων σε κάθε κανόνα.

Τα αποτελέσματα των παραπάνω στατιστικών μέτρων για τους δύο αλγορίθμους παρουσιάζονται στον Πίνακα 5.3.

Δοκιμάζοντας να τρέξουμε τους αλγόριθμους με μόνο στατιστικά σημαντικούς (significant) κανόνες (σύμφωνα με την τιμή του p-Value) παρατηρήσαμε ότι το ποσοστό της σωστής ταξινόμησης των αλγόρίθμων είχε βελτιωθεί σε μεγάλο βαθμό σε σύγκριση με το ποσοστό που πήραμε κατά την εκτέλεση με όλους τους κανόνες.

Κεφάλαιο 7

Συμπεράσματα και Μελλοντική Εργασία

7.1 Συμπεράσματα

Για την αξιολόγηση κανόνων βάσει πολλαπλών μέτρων έχουν χρησιμοποιηθεί εννέα μοντέλα από τη βάση δεδομένων με καρδιαγγειακά νοσήματα. Τα μοντέλα που παρουσιάζουν περισσότερο ενδιαφέρον είναι τα μοντέλα Πριν+Μετά καθώς επίσης και τα μοντέλα Πριν.

Ο λόγος που τα μοντέλα Πριν+Μετά παρουσιάζουν ενδιαφέρον και προτείνονται για περισσότερη μελέτη έγκειται στο ότι σ' αυτά τα μοντέλα έχουν συμπεριληφθεί όλοι οι παράγοντες που συνέστησαν οι ιατροί. Εκτελώντας είτε τη μέθοδο Συσχέτισης είτε τη μέθοδο Κατηγοριοποίησης μπορούν να εξαχθούν περισσότεροι κανόνες στους οποίους να συμμετέχουν σημαντικοί παράγοντες που παρουσιάζονται πριν αλλά και μετά το επεισόδιο. Με αυτό τον τρόπο μετά την αξιολόγηση των κανόνων από τους ιατρούς, οι ιατροί θα έχουν μία πιο σφαιρική και ολοκληρωμένη εικόνα για τους παράγοντες που επηρεάζουν τέτοιων ειδών επεισόδια με αποτέλεσμα την άμεση πρόγνωση και πρόβλεψη μελλοντικών περιπτώσεων.

Τα μοντέλα Πριν μπορούν να μελετηθούν περισσότερο για το λόγο ότι το ποσοστό των σωστών ταξινομήσεων (CC) που παρουσιάστηκαν στους δύο αλγόριθμους για τα μοντέλα Πριν, είναι καλύτερο από το ποσοστό στα άλλα μοντέλα. Πιο σωστή ταξινόμηση όμως συνεπάγεται και πιο αξιόπιστους κανόνες με αποτέλεσμα να δίνουν στους ιατρούς περισσότερη και πιο ακριβή γνώση για την πρόβλεψη νέων συμβάντων.

Για την αξιολόγηση κανόνων βάσει πολλαπλών μέτρων έχουν χρησιμοποιηθεί δύο μέθόδοι, η μέθοδος Συσχέτισης (Association) με τη χρήση του αλγόριθμου Apriori καθώς επίσης και η μέθοδος Κατηγοριοποίησης (Classification) με τη χρήση του αλγόριθμου Δέντρων Απόφασης.

Και στις δύο περιπτώσεις οι εξαγόμενοι κανόνες παρουσίασαν ενδιαφέρον.

Στην περίπτωση της μεθόδου Συσχέτισης, ο αριθμός των εξαγόμενων κανόνων ήταν κατά πολύ μεγαλύτερος σε σύγκριση με τους εξαγόμενους κανόνες από τη μέθοδο Κατηγοριοποίησης. Αυτό είναι ένα σημαντικό πλεονέκτημα της μεθόδου αφού περισσότεροι κανόνες $\Rightarrow$ κάλυψη περισσότερων πιθανών περιστατικών με αποτέλεσμα οι ιατροί να έχουν μεγαλύτερη ευχέρεια αντιμετώπισης ενός νέου περιστατικού.

Στην περίπτωση της μεθόδου Κατηγοριοποίησης, το πλήθος των εξαγόμενων κανόνων περιορίζεται σε μερικές εκατοντάδες. Αυτό οφείλεται στο ότι κατά το κτίσιμο του δέντρου απόφασης, σε ένα κανόνα συμμετέχουν πολλά χαρακτηριστικά του μοντέλου μας. Περισσότερα χαρακτηριστικά σε ένα κανόνα $\Rightarrow$ λιγότερα κλαδιά στο δέντρο μας άρα και πιο λίγοι κανόνες. Από τη στιγμή όμως που σε ένα κανόνα συμμετέχουν πολλά χαρακτηριστικά αυτό καθιστά τον κανόνα και πιο συγκεκριμένο. Τέτοιοι κανόνες όμως μπορούν να δώσουν πιο εξειδικευμένη γνώση. Επιπρόσθετα τα Δέντρα Απόφασης μπορούν να χρησιμοποιηθούν και σε βάσεις δεδομένων που αποτελούνται από πολλά χαρακτηριστικά. Επίσης, κάτι που παρατηρήσαμε και από την εκτέλεση και των δύο αλγορίθμων ήταν το ότι στο αλγόριθμο Δέντρων Απόφασης το ποσοστό σωστών ταξινομήσεων (CC) ήταν πιο υψηλό απ' ότι στη μέθοδο Συσχέτισης. Αυτό καθιστά και τους κανόνες από τον αλγόριθμο Δέντρων Απόφασης πιο αξιόπιστους.

Παρατηρήσαμε ότι διαφορετικές μεθοδολογίες παράγουν διαφορετικά αποτελέσματα, εφαρμόζοντάς τις στις ίδιες βάσεις δεδομένων. Επομένως ανάλογα με τη βάση δεδομένων θα πρέπει να εφαρμόζεται και η κατάλληλη μεθοδολογία.

7.2 Μελλοντική Εργασία

Για το σκοπό της μελέτης έχουν χρησιμοποιηθεί δύο εργαλεία εξόρυξης δεδομένων τα οποία στηρίζονται σε αλγόριθμους εξόρυξης κανόνων συσχέτισης και κατηγοριοποίησης. Μέσω αυτών των εργαλείων μελετήθηκαν αντικειμενικά μέτρα καθώς επίσης και υλοποιήθηκαν κάποια στατιστικά μέτρα που παρέχουν πληροφορίες σχετικά με τη σημαντικότητα του κάθε κανόνα (χ^2 και p-Value), το ποσοστό του κινδύνου που έχει ένας ασθενής για τυχόν επεισόδιο (Framingham Event Risk) και του ποσοστού σωστής ταξινόμησης των κανόνων (Correct Classification). Οι τιμές των μέτρων κωδικοποιήθηκαν και εισήχθηκαν στον αλγόριθμο Δέντρων Απόφασης για να εξάγουμε τα σημαντικά μέτρα. Με την ανάκτηση των σημαντικών μέτρων, ανατρέξαμε πίσω στους αρχικούς κανόνες φιλτράροντάς τους για να αποφανθούμε ποιοι από αυτούς μας δίνουν περισσότερη γνώση.

Στο μέλλον θα μπορούσε να γίνει αξιολόγηση των φιλτραρισμένων κανόνων. Για την επίτευξη του στόχου αυτού επιβάλλεται η συνεργασία με τους ιατρούς που είναι οι πλέον αρμόδιοι για την αξιολόγηση των εξαγόμενων κανόνων.

Επιπρόσθετα στα εργαλεία εξόρυξης κανόνων που έχουμε χρησιμοποιήσει θα μπορούσαν να γίνουν κάποιες αναβαθμίσεις έτσι ώστε να δίνεται η δυνατότητα στο χρήστη να επιλέγει τα χαρακτηριστικά της βάσης δεδομένων που θα ήθελε να τρέξει. Με την αναβάθμιση αυτή δε θα χρειαζόταν να σπάσουμε την αρχική μας βάση σε 9 μοντέλα παρά μόνο να επιλέγουμε τα χαρακτηριστικά που θα θέλαμε να τρέξουμε σε κάθε περίπτωση.

Τέλος, ένα νέο Γραφικό Περιβάλλον (GUI) - εργαλείο μπορεί να αναπτυχθεί που να συνδυάζει τις δύο μεθόδους εξόρυξης κανόνων. Με το περιβάλλον αυτό θα δίνεται η δυνατότητα στο χρήστη να επιλέγει ποια μέθοδο θέλει να ακολουθήσει για την εξόρυξη κανόνων.

Βιβλιογραφία

[1] M. Russell, 'The Biggest Cause of Death in the Western World', Ezine @rticles®.

[2] M Wijesinghe, K Perrin, A Ranchord, M Simmonds, M Weatherall and R Beasley, 'Routine use of oxygen in the treatment of myocardial infarction: systematic review', Heart, Vol. 95, p. 198 – 202, 2009.

[3] J. G. Canto, R. J. Goldberg, M. M. Hand, R. O. Bonow, G. Sopko, C. J. Pepine, T. Long, 'Symptom Presentation of Women With Acute Coronary Syndromes: Myth vs Reality', Arch Intern Med, Vol. 167, p. 2405 – 2413, 2007.

[4] Ministry of Health, Cyprus: Annual Report 2007, 2008

[5] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, Ph. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D, Steinberg, 'Top 10 algorithms in data mining,' Knowl Inf Syst 14:1–37, 2008.

[6] J. Han, M. Kamber, 'Data Mining, Concepts and Techniques', Morgan Kaufman Publishers, Academic Press, 2001.

[7] Μ. Βαζιργιάννης, Μ. Χαλκίδη, 'Εξόρυξη γνώσης από βάσεις δεδομένων', Τυπωθήτω, 2003.

[8] http://www.wikipedia.org (Keywords: Myocardial Infarction, Percutaneous coronary intervention, Coronary artery bypass surgery, Data Mining)

[9] P.E. Greenwood, M.S. Nikulin, 'A guide to chi-squared testing', Wiley, New York, 1996

[10] M. J. Schervish, 'P Values: What They Are and What They Are Not', The American Statistician 50 (3): 203–206, 1996

[11] Z. Wang and W. E. Hoy, 'Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people?', The Medical Journal of Australia, vol. 182, no. 2, pp. 66-69, 2005.

[12] R. Agrawal, R. Srikant, 'Fast Algorithms for Mining Association Rules', IBM Almaden Research Center 650 Harry Road, San Jose, CA 9512, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994

[13] K. C. You, and K. S. Fu, 'An approach to the design of a linear binary tree classifier', Proc. of the 3rd Symposium on Machine Processing of Remotely Sensed Data, pp.: 3-10, 1976.

[14] M. F. Usama, G. Piatesky-Shapiro, P. Smuth, R. Uthurusamy, 'Advances in knowledge discovery and data mining,' AAAI Press, 1996.

[15] M. W. Gillo, 'MAID: A Honeywell 600 program for an automatised survey analysis,' Behavioral Science. 17, pp.: 251-252, 1972.

[16] J. N. Morgan, R. C. Messenger, 'THAID: a sequential search program for the analysis of nominal scale dependent variables', Technical report, Institute for Social Research, University of Michigan, Ann Arbor, MI, 1973

[17] M. Mehta, J. Rissanen, R. Agrawal, 'MDL-based decision tree pruning', In Proc. of KDD, 1995.

[18] I. H. Witten, E. Frank, 'Data mining: Practical machine Learning Tools and Techniques', (The Morgan Kaufmann Series in Data Management Systems), 2nd edition. Publ: Hanser Fachbuch, 2005.

[19] Β. Βουτσινάς, 'Θέματα επιχειρηματικής νοημοσύνης: Θεωρητική θεμελίωση και εφαρμογές', Εκδόσεις Κωσταράκη, 2003.

[20] G. V. Kass, 'An exploratory technique for investigating large quantities of categorical data', Applied Statistics. 29 (2), pp.: 119-127, 1980.

[21] L. Geng, H. J. Hamilton, 'Interestingness Measures for Data Mining: A Survey', ACM Computing Surveys, Vol. 38, No. 3, Article 9, 2006.

[22] J. S. Park, M.-S. Chen, P. S. Yu, 'Efficient Parallel Data Mining for Association Rules', Proc. of the International Conference on Information and Knowledge Management, pp.: 31-36, 1995b.

[23] D. W-L. Cheung, J. Han, V. Ng, A. W-C Fu, Y. Fu, 'A Fast Distributed Algorithm for Mining Association Rules,' Proceedings of PDIS, pp.: 31-43, 1996.

[24] L. Harada, N. Akaboshi, K. Ogihara, R. Take, 'Dynamic Skew Handling in Parallel Mining of Association Rules,' Proc. of the 7th International Conference on Information and Knowledge management, pp.: 76-85, 1998.

[25] Μ. Καραολής, 'Εξόρυξη γνώσης με εξαγωγή κανόνων σε Καρδιαγγειακές Βάσεις Δεδομένων', Διδακτορική Εργασία στο Τμήμα Πληροφορικής Πανεπιστημίου Κύπρου, 2010.

[26] Λ. Παπακωνσταντίνου, 'Εξόρυξη κανόνων από καρδιαγγειακή βάση με τη χρήση αλγόριθμων συσχέτισης', Μεταπτυχιακή Εργασία στο Τμήμα Πληροφορικής Πανεπιστημίου Κύπρου, 2009.

[27] Δ. Χατξηπαναγή, 'Rule extraction of cardiovascular database using decision trees', Μεταπτυχιακή Εργασία στο Τμήμα Πληροφορικής Πανεπιστημίου Κύπρου, 2009.