ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΜΕ ΕΞΑΓΩΓΗ ΚΑΝΟΝΩΝ ΣΕ ΚΑΡΔΙΑΓΓΕΙΑΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Μηνάς Καραολής

Πανεπιστήμιο Κύπρου, 2010

Εκτιμήσεις του παγκόσμιου οργανισμού υγείας δείχνουν ότι οι καρδιακές παθήσεις είναι και θα παραμείνουν στις επόμενες δεκαετίες η κύρια αιτία θανάτου παγκόσμια. Οι παράγοντες κινδύνου πάθησης ενός καρδιακού επεισοδίου είναι γνωστοί, αλλά όχι τόσο ο συνδυασμός αυτών, και πως μπορεί η καλύτερη διαχείριση των πολλαπλών παραγόντων κινδύνου να βοηθήσει στην μείωση των περιστατικών.

Στόχος της διατριβής αυτής είναι η ανάπτυξη ενός καινοτόμου ολοκληρωμένου συστήματος που θα υποστηρίζει την αξιολόγηση των παραγόντων κινδύνου σε καρδιαγγειακές βάσεις δεδομένων και την εξόρυξη κανόνων εκτίμησης κινδύνου βασισμένων σε αλγόριθμους δένδρων αποφάσεων και κανόνων συσχέτισης.

Η πρωτοτυπία της διατριβής εστιάζεται στα ακόλουθα: i. Ανάπτυξη αλγορίθμων εξαγωγής κανόνων από δέντρα απόφασης βασισμένων σε διαφορετικά κριτήρια διαχωρισμού. ii. Ανάπτυξη ενός νέου αλγορίθμου εξαγωγής κανόνων συσχέτισης, που ονομάζεται AKAMAS, που με μια σάρωση της βάσης δεδομένων δημιουργεί κανόνες, και οι οποίοι υπολογίζονται με διαφορετικά μέτρα. iii. Ανάπτυξη μίας καινοτόμας μεθοδολογίας αξιολόγησης των στατιστικά σημαντικών κανόνων παραγόντων κινδύνου για εκτίμηση κινδύνου για ανεύρεση των καλύτερων κανόνων βασισμένη σε πολλαπλά μέτρα.

Το προτεινόμενο σύστημα έχει εφαρμοστεί για την εξόρυξη κανόνων παραγόντων κινδύνου σε βάσεις δεδομένων για καρδιακά επεισόδια με έμφραγμα μυοκαρδίου, αγγειοπλαστικής και αρτηριακής παράκαμψης με πολύ ικανοποιητικά αποτελέσματα.

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΜΕ ΕΞΑΓΩΓΗ ΚΑΝΟΝΩΝ ΣΕ ΚΑΡΔΙΑΓΓΕΙΑΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Μηνάς Καραολής

Η Διατριβή αυτή

Υποβλήθηκε προς Μερική Εκπλήρωση των

Απαιτήσεων για την Απόκτηση

Διδακτορικού Τίτλου Σπουδών

στο

Πανεπιστήμιο Κύπρου

Συστήνεται προς Αποδοχή

από το Τμήμα Πληροφορικής

Ιούνης, 2010

2010

ΣΕΛΙΔΑ ΕΓΚΡΙΣΗΣ

Διδακτορική Διατριβή

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΜΕ ΕΞΑΓΩΓΗ ΚΑΝΟΝΩΝ ΣΕ ΚΑΡΔΙΑΓΓΕΙΑΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Παρουσιάστηκε από

Μηνά Καραολή

Ερευνητικός Σύμβουλος Κωνσταντίνος Σ. Παττίχης

Όνομα Ερευνητικού Συμβούλου

Μέλος Επιτροπής Χρίστος Ν. Σχίζας

Όνομα Μέλους Επιτροπής

Μέλος Επιτροπής Χρίστος Χριστοδούλου

Όνομα Μέλους Επιτροπής

Μέλος Επιτροπής Δημήτρης Φωτιάδης

Όνομα Μέλους Επιτροπής

Μέλος Επιτροπής Γιώργος Κονταξάκης

Όνομα Μέλους Επιτροπής

Πανεπιστήμιο Κύπρου

Ιούνης, 2010

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ όλους εκείνους που με βοήθησαν για την αποπεράτωση αυτής της διατριβής.

Ο γιατρός, Δρ. Ιωσήφ Μουτήρης μου έδωσε το πολύτιμο υλικό και τις ιατρικές συμβουλές για να μπορέσω να διατρίψω σε αυτό το θέμα.

Ο γιός μου Ανδρέας που ποτέ δεν παραπονέθηκε ότι του λείπω, όταν αυτός είχε την ανάγκη της συντροφιάς του πατέρα του.

Η σύζυγος μου Κυπρούλα που εκτός από την κατανόηση της, με την εμπειρία της με βοήθησε και στην συγγραφή αυτής της διατριβής.

Τέλος, ο Διδακτορικός μου πατέρας, Καθηγητής Δρ. Κωνσταντίνος Σ. Παττίχης, ο εμπνευστής, ο υπομονετικός, ο σύμβουλος, πάντα κοντά μου να βρίσκει λύσεις σε οποιοδήποτε πρόβλημα και εμπόδιο. ’Στον πατέρα μου χρωστώ το ζην, στον δάσκαλο μου το ευ ζην’.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΜΕ ΠΙΝΑΚΕΣ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Κεφάλαιο 1: Εισαγωγή

1.1 Γενικά

1.1.1 Εισαγωγή στην εξόρυξη δεδομένων

Η εξόρυξη γνώσης προέκυψε από την ανάγκη επεξεργασίας των τεράστιων αποθηκών δεδομένων και έχει εξελιχθεί σε ένα από τα βασικότερα ερευνητικά θέματα στην περιοχή των βάσεων δεδομένων. Οι τεχνικές που χρησιμοποιούνταν μέχρι πρότινος για την επεξεργασία των δεδομένων δεν ικανοποιούν και δεν ανταποκρίνονται λόγω των τεράστιων όγκων δεδομένων που έχουν μαζευτεί. Η ανάγκη βελτίωσης της ποιότητας των παρεχόμενων πληροφοριών όπως επίσης και της εξαγωγής χρήσιμων συμπερασμάτων, επισπεύσανε την εξεύρεση τεχνικών διαχείρισης και επεξεργασίας δεδομένων. Όλο και περισσότερο αναγνωρίζεται η συνεισφορά των διαφόρων αυτών τεχνικών στην ανάλυση μεγάλων βάσεων δεδομένων σε σχέση με τις κλασικές στατιστικές μεθόδους. Επομένως, ενώ οι στατιστικές δοκιμές απαιτούν έλεγχο στατιστικών υποθέσεων αναφορικά με έναν πληθυσμό, οι τεχνικές εξόρυξης από δεδομένα παρέχουν αυτόματα επιβεβαιωμένες σχέσεις με τη μορφή κανόνων σε μεγάλες βάσεις από δεδομένα [1]. Πολλές από τις τεχνικές που χρησιμοποιούνται σήμερα ως τεχνικές εξόρυξης από δεδομένα υπήρχαν εδώ και αρκετά χρόνια, μέσα στα πλαίσια της επιστήμης της τεχνητής νοημοσύνης και αναπτύχθηκαν στα τέλη της δεκαετίας του 1980. Παρ' όλα αυτά, μόλις τα τελευταία χρόνια οι τεχνικές αυτές εφαρμόζονται σε μεγάλες βάσεις δεδομένων [2].

Η ανάπτυξη της τεχνολογίας των βάσεων δεδομένων που άρχισε την προηγούμενη εικοσαετία συντέλεσε στη γρήγορη αύξηση της παραγωγής και συλλογής δεδομένων. Οι διάφορες τεχνικές για ποσοτική και αποτελεσματική αποθήκευση, συλλογή και επεξεργασία των δεδομένων είχε σαν αποτέλεσμα να παράγονται κάθε χρόνο τεράστιοι όγκοι δεδομένων [3]. Μεγάλες εταιρείες, οργανισμοί, δημόσιες υπηρεσίες, νοσοκομεία και πανεπιστήμια, μπορούν σήμερα να συλλέγουν παντός είδους πληροφορίες, και να τις συντηρούν με πολύ μικρό

κόστος. Φυσικά σε αυτό συντέλεσε σε μεγάλο βαθμό και η τρομακτική μείωση του κόστους των αποθηκευτικών χώρων.

Με τους τεράστιους όγκους δεδομένων που συλλέγονται προέκυψαν δύο προβλήματα. Το πρώτο είναι ότι υπάρχουν δεδομένα που είναι πολύ περίπλοκα και το δεύτερο ότι υπάρχουν δεδομένα που είναι αχρείαστα. Στη δεύτερη περίπτωση μπορούμε, έχοντας πλήρη γνώση και επίγνωση, να αγνοήσουμε αυτά τα δεδομένα αλλά όχι να τα εξαλείψουμε. Η πρώτη περίπτωση επέφερε την ανάγκη δημιουργίας νέων εργαλείων και τεχνικών για ευφυή ανάλυση βάσεων δεδομένων, έτσι ώστε να γίνεται εξαγωγή «χρήσιμης» γνώσης. Αυτή την ανάγκη την αντιλήφθηκαν οι ερευνητές από διάφορες περιοχές, όπως των βάσεων δεδομένων, της τεχνητής νοημοσύνης και της στατιστικής, με αποτέλεσμα να δημιουργηθεί ένας νέος ερευνητικός τομέας, γνωστός ως εξόρυξη δεδομένων και γνώσης.

Το βασικό πλεονέκτημα των τεχνικών αυτών σε σχέση με τις στατιστικές μεθόδους είναι ότι μπορούν να αναγνωρίσουν σχέσεις μεταξύ μεταβλητών εκ των οποίων είναι πιθανό είτε να μην είναι γνωστές από την αρχή σε ένα σύνολο δεδομένων, είτε ένα πολύ πιο μικρό ποσοστό του συνόλου αυτού να είναι γνωστές εξαρχής. Αυτό σημαίνει ότι ορισμένα πρότυπα τα οποία μπορεί να ενδιαφέρουν έναν αναλυτή ίσως να μην αντιπροσωπεύουν συνολικές τάσεις των δεδομένων και κατά συνέπεια να μην μπορούν να αποκαλυφθούν από μια στατιστική δοκιμή.

Οι τεχνικές εξόρυξης από δεδομένα περιλαμβάνουν δύο βασικά ζητήματα: την προεπεξεργασία δεδομένων και την αναγνώριση προτύπων στα δεδομένα αυτά. Κατά το στάδιο της προεπεξεργασίας των δεδομένων αναγνωρίζονται τα σχετικά με το πρόβλημα χαρακτηριστικά. Το στάδιο αυτό μπορεί να περιλαμβάνει και διάφορα άλλα υποστάδια, όπως διαχείριση κενών τιμών στα πεδία, κανονικοποίηση των δεδομένων, μείωση δεδομένων κ.ο.κ.. Το δεύτερο στάδιο περιλαμβάνει αναγνώριση προτύπων βάσει των χαρακτηριστικών που είχαν καθοριστεί στο προηγούμενο στάδιο και την αξιολόγησή τους. Η αναγνώριση των προτύπων αυτών γίνεται αρχικά με την επιλογή της κατάλληλης μεθόδου εξόρυξης από δεδομένα και στη συνέχεια με τη χρήση αλγορίθμων εξόρυξης από δεδομένα. Παρόλο που κατά τη διαδικασία εξόρυξης από δεδομένα δίνεται ιδιαίτερη έμφαση στο στάδιο της αναγνώρισης προτύπων, αξίζει να σημειωθεί ότι πολλοί έχουν αναγνωρίσει ότι το στάδιο της

προεπεξεργασίας δεδομένων είναι αυτό που επηρεάζει περισσότερο την επιτυχία της ανεύρεσης γνώσης από δεδομένα [4]. Τα βήματα εξόρυξης δεδομένων παρατίθενται αναλυτικότερα στο κεφάλαιο 1.1.1.2.

Οι σημαντικότερες τεχνικές εξόρυξης από δεδομένα, οι οποίες αναφέρονται ως τύποι γνώσης που εξάγονται αυτόματα με τη μορφή κανόνων βασίζονται στους:

i. Αλγόριθμους ταξινόμησης βασισμένους σε δέντρα απόφασης.
ii. Αλγόριθμους συσχέτισης.

Οι Wu *et al.* [5] έχουν επιλέξει τους 10 σημαντικότερους και διαδεδομένους αλγόριθμους στην περιοχή εξόρυξης δεδομένων. Σε αυτή τη δημοσίευση κατέληξαν στο συμπέρασμα ότι στη λίστα αυτή συμπεριλαμβάνονται οι C4.5 [6] και CART [7] για την πρώτη κατηγορία, όπως και ο αλγόριθμος Apriori [8] για την δεύτερη κατηγορία. Στα κεφάλαια 2 και 3 που ακολουθούν θα παρουσιασθούν αναλυτικά οι δύο αυτές τεχνικές εξόρυξης από δεδομένα.

1.1.1.1 Εξόρυξη γνώσης δεδομένων

Η εξόρυξη γνώσης δεδομένων συμπεριλαμβάνει δύο σημαντικά στοιχεία. Το πρώτο είναι η ανακάλυψη γνώσης από βάσεις δεδομένων και το δεύτερο είναι οι τεχνικές που χρησιμοποιούνται για την ανάλυση και εξαγωγή της γνώσης από διάφορα σύνολα δεδομένων. Η όλη διαδικασία ανάλυσης δεδομένων είναι γνωστή και σαν Knowledge Discovery in Databases (KDD) [9], ενώ για τις μεθόδους και τις τεχνικές που χρησιμοποιούνται στη διαδικασία ανάλυσης συνηθίζεται ο όρος εξόρυξη δεδομένων. Όμως, παρά το γεγονός ότι συνηθίζεται να χρησιμοποιείται ο όρος εξόρυξη δεδομένων, πιστεύω ότι η πιο σωστή ορολογία που πρέπει να χρησιμοποιείται είναι η εξόρυξη γνώσης. Περνώντας από όλα τα βήματα της διαδικασίας εξόρυξης, καταλήγουμε στην αξιολόγηση προτύπων στη γνώση. Η γνώση αυτή προέρχεται σίγουρα από τα δεδομένα, αλλά όχι με την έννοια των δεδομένων αυτών καθαυτών. Οι ερευνητές που ασχολούνται με αυτό το θέμα, έχουν την ικανότητα να εξάγουν γνώση μέσα από την πληθώρα των κανόνων και συσχετίσεων που παράγονται από τα

δεδομένα. Αυτή η γνώση προέρχεται από τα αξιολογημένα πρότυπα, τα οποία φυσικά δεν είναι απλά δεδομένα. Επειδή ο όρος εξόρυξη δεδομένων έχει επικρατήσει και χαρακτηρίζει τη διαδικασία εξεύρεσης γνώσης, θα χρησιμοποιούμε στα πλαίσια αυτή της διατριβής τον όρο αυτό.

1.1.1.2 Βήματα και προϋποθέσεις για εξόρυξη δεδομένων

Η ανακάλυψη γνώσης από δεδομένα είναι η μη τετριμμένη διαδικασία εύρεσης έγκυρων, πρωτότυπων, χρήσιμων και οπωσδήποτε κατανοητών προτύπων μέσα στα δεδομένα. Αυτή η διαδικασία είναι μια επαναληπτική διαδικασία η οποία αποτελείται από τα ακόλουθα βήματα [9] (βλέπε Σχήμα 1.1):

Την επιλογή του συνόλου των δεδομένων. Από μια αποθήκη δεδομένων γίνεται η επιλογή εκείνων των οποίων θα χρειαστούν στη διαδικασία εξόρυξης δεδομένων που πρόκειται να εφαρμοστεί.

Τον καθαρισμό και προεπεξεργασία των δεδομένων. Σε αυτό το βήμα γίνεται η συμπλήρωση των ελλειπόντων πεδίων δεδομένων με βάση τις γνωστές στρατηγικές διαχείρισης δεδομένων, η αφαίρεση θορύβου και ακραίων δεδομένων (outliers) και η συλλογή των απαραίτητων πληροφοριών για τον εντοπισμό του θορύβου.

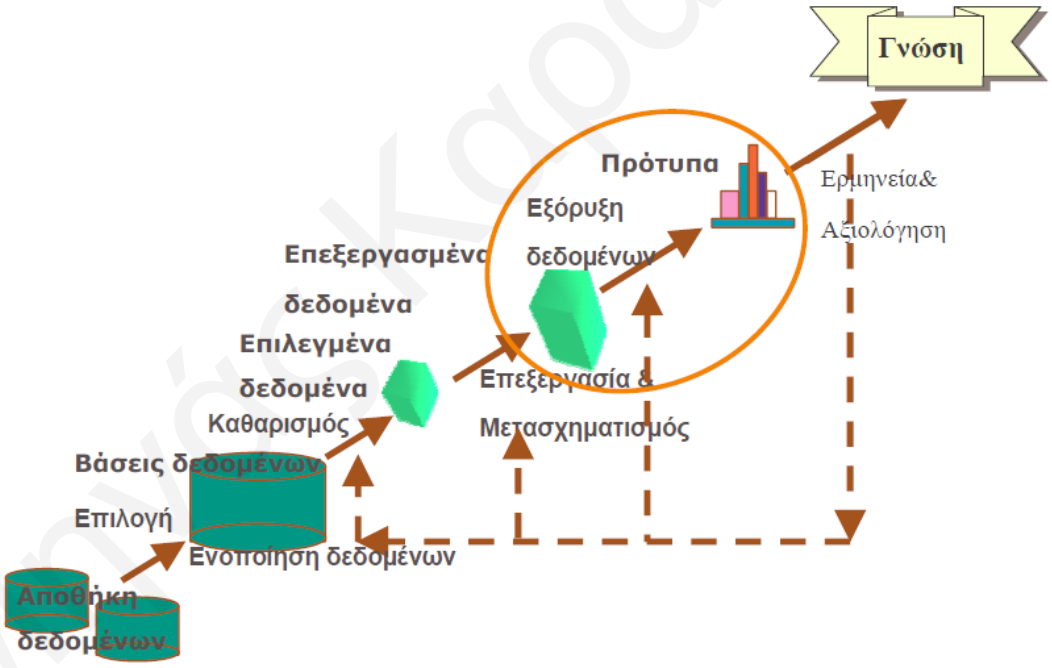
Το μετασχηματισμό των δεδομένων. Τα δεδομένα μετασχηματίζονται κατάλληλα για να γίνει η εξόρυξη. Χρησιμοποιούνται διάφορες μέθοδοι για να μειωθούν οι διαστάσεις και οι μεταβλητές. Γίνεται κωδικοποίηση των δεδομένων για να αποφευχθούν όπου είναι δυνατό οι μεταβλητές.

Την επιλογή αλγορίθμων εξόρυξης δεδομένων. Γίνεται η επιλογή των μεθόδων που θα χρησιμοποιηθούν λαμβάνοντας υπόψη τους στόχους που θέλουμε να επιτύχουμε. Αποτέλεσμα αυτού του βήματος είναι η παραγωγή προτύπων που επιδεικνύουν ενδιαφέρον και θα συμβάλουν στην αποκόμιση νέων πληροφοριών και γνώσης.

Την αξιολόγηση των προτύπων. Τα πρότυπα που έχουν εξαχθεί στο προηγούμενο βήμα αξιολογούνται με κάποια κριτήρια και αγνοούνται αυτά τα οποία δεν προσφέρουν νέα γνώση. Έτσι παραμένουν μόνο τα πρότυπα που έχουν ενδιαφέρον και προσφέρουν κάτι το καινούργιο.

Την παρουσίαση της γνώσης. Σε αυτό το βήμα η νέα γνώση που έχει εξορυχτεί απεικονίζεται με κάποιες τεχνικές στο χρήστη.

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 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.



Σχήμα 1.1: Τα βήματα της διαδικασίας KDD (από Μ. Βαζιργιάννης και Μ. Χαλκίδη, 'Εξόρυξη γνώσης από βάσεις δεδομένων,' Τυπωθήτω, 2003)

Για την καλή και επιτυχημένη εξόρυξη δεδομένων πρέπει να ληφθούν υπόψη κάποιες προϋποθέσεις και απαιτήσεις που πρέπει να έχει το σύστημα καθώς επίσης και οι τεχνικές που πρέπει να χρησιμοποιηθούν.

Οι αλγόριθμοι εξόρυξης δεδομένων πρέπει να προσαρμοστούν κατάλληλα στα μεγάλα σύνολα δεδομένων έτσι ώστε να έχουν καλή απόδοση. Μπορούν να δοκιμαστούν διάφοροι αλγόριθμοι και να γίνει μέτρηση του χρόνου εκτέλεσής τους έτσι ώστε να γίνει σωστή επιλογή όσον αφορά τον χρόνο. Αλγόριθμοι που δεν έχουν αναμενόμενο ή αποδεκτό χρόνο εκτέλεσης δεν είναι κατάλληλα προσαρμοσμένοι στα δεδομένα μας.

Τα αποτελέσματα που παίρνουμε από την εξόρυξη δεδομένων, δηλαδή η γνώση, πρέπει να είναι ακριβή. Η ακρίβεια των αποτελεσμάτων μπορεί να διαπιστωθεί χρησιμοποιώντας είτε άλλα εργαλεία εξόρυξης δεδομένων, είτε άλλους αλγόριθμους, ή ακόμη συγκρίνοντας τα με το περιεχόμενο της βάσης δεδομένων. Αυτά τα αποτελέσματα πρέπει να εκφράζονται με διάφορους τρόπους όπως για παράδειγμα με γραφικές διεπαφές, για να μπορούν οι απλοί χρήστες να κατανοούν και να χρησιμοποιούν αυτήν τη γνώση. Πρέπει να ληφθεί υπόψη ότι τα δεδομένα είναι δυνατό να βρίσκονται σε διάφορες βάσεις δεδομένων. Για το λόγο αυτό θα ήταν καλό να χρησιμοποιηθούν παράλληλοι και κατανεμημένοι αλγόριθμοι εξόρυξης δεδομένων.

1.1.1.3 Ανάλυση δεδομένων OnLine Analytical Processing (OLAP)

Η τεχνική OLAP [9] είναι μια από τις κύριες τεχνικές της Επιχειρηματικής Νοημοσύνης. Συνδέεται πολύ συχνά με την ανάπτυξη αποθήκευσης δεδομένων και μαζί αποτελούν βασικά στοιχεία σύγχρονων τεχνικών για την υποστήριξη αποφάσεων. Οι μεγάλες αποθήκες δεδομένων (datawarehouses) προορίζονται για υποστήριξη αποφάσεων, διατηρώντας ιστορικά, συγκεντρωτικά και ολοκληρωμένα δεδομένα, ενδεχομένως από πολλές επιχειρησιακές βάσεις δεδομένων, ή και από εξωτερικές πηγές, κάνοντάς τα έτσι πιο σημαντικά, αλλά και πολύ μεγαλύτερα σε μέγεθος. Υποστηρίζουν την ανάπτυξη OLAP

εφαρμογών, των οποίων οι λειτουργικές απαιτήσεις είναι διαφορετικές από αυτές της συναλλακτικής διαδικασίας η οποία υποστηρίζεται παραδοσιακά από τις επιχειρησιακές βάσεις δεδομένων. Για πολλά χρόνια οι σχεσιακές βάσεις δεδομένων ήταν το σημείο αναφοράς στη διαχείριση δεδομένων. Το σχεσιακό μοντέλο αναπτύχθηκε προκειμένου να λύσει ορισμένα προβλήματα που υπήρχαν στο χώρο της διαχείρισης δεδομένων. Τα συστήματα που υπήρχαν μέχρι τα τέλη της δεκαετίας του 1960 σχεδιάζονταν για να λύσουν συγκεκριμένα προβλήματα και μόνο αργότερα επεκτάθηκαν με στόχο να δώσουν μια γενικότερη και πιο ολοκληρωμένη λύση. Το σχεσιακό μοντέλο έκανε τα προϊόντα συστημάτων διαχείρισης βάσεων δεδομένων πιο ελκυστικά για όλους τους τύπους των χρηστών, μέσω της επεξεργασίας ερωτημάτων και αναφορών, με διάφορους τύπους δεδομένων και συνολικά πιο εύκολης σχεδίασης και λύσης προβλημάτων.

Η τεχνική OLAP εισήχθηκε το 1993 κα αναπτύχθηκε λόγω της ανάγκης για μια πιο περίπλοκη ανάλυση και ταχύτερη σύνθεση καλύτερης ποιότητας πληροφορίας από δεδομένα, καθώς επίσης λόγω της ανάγκης για πρόσβαση σε αυτήν την πληροφορία και ανάλυση της από περισσότερα άτομα μέσα σε έναν οργανισμό [10].

Η εν λόγω τεχνική χρησιμοποιείται σε διάφορους τύπους βάσεων δεδομένων και αρχείων με δεδομένα και στους διάφορους τύπους πακέτων front-end, για παράδειγμα στατιστικά πακέτα και φύλα εργασίας.

Ο κλασικός ορισμός της τεχνικής OLAP είναι ο εξής [9]: «Η τεχνική OLAP είναι μια κατηγορία τεχνολογίας λογισμικού, η οποία δίνει τη δυνατότητα στους αναλυτές, στους διευθυντές και στα στελέχη να αντιλαμβάνονται τα δεδομένα της επιχείρησης μέσω της ταχείας, συστηματικής και συνολικής πρόσβασης σε αυτά από διάφορες πιθανές οπτικές γωνίες της πληροφορίας, που έχει μετασχηματισθεί σε πληροφορία από δεδομένα, προκειμένου να αντικατοπτρίσει την πραγματική διάσταση της επιχείρησης έτσι όπως την αντιλαμβάνεται ο χρήστης».

Έχουν ορισθεί ως κανόνες για την τεχνική OLAP τα πιο κάτω [11]: πολυδιάστατο μοντέλο, διαφάνεια του εξυπηρετητή (server), προσβασιμότητα, σταθερή αποτελεσματικότητα πρόσβασης, αρχιτεκτονική του πελάτη εξυπηρετητή (client server), συνολική διάσταση,

διαχείριση της διασκόρπισης των δεδομένων, πολυχρηστική, λειτουργικότητα στις διαστάσεις, διαισθητικός χειρισμός των δεδομένων, ελαστική και εύκολη προετοιμασία και ενημέρωση, πολλαπλές διαστάσεις και επίπεδα.

1.1.1.4 Αρχιτεκτονική αποθήκευσης δεδομένων

Στην εφαρμογή OLAP υπάρχουν δύο βασικοί τύποι αποθήκευσης δεδομένων. Είναι η αρχιτεκτονική σχεσιακής OLAP και η αρχιτεκτονική πολυδιάστατης OLAP. Η αρχιτεκτονική Relational Online Analytical Processing (ROLAP) [12] βασίζεται σε εξυπηρετητές σχεσιακών βάσεων δεδομένων. Έχει όμως και τη δυνατότητα να συγκεντρώνει ή να διαιρεί δεδομένα, απλά ή πολυδιάστατα. Τα δεδομένα στην αρχιτεκτονική ROLAP έχουν τη μορφή αστεριού ή νιφάδας. Η αρχιτεκτονική Multidimensional Online Analytical processing (MOLAP) [13] βασίζεται σε καθαρά πολυδιάστατες βάσεις δεδομένων. Ο συνδυασμός των δύο πιο πάνω αρχιτεκτονικών είναι η υβριδική αρχιτεκτονική Hybrid Online Analytical Processing (HOLAP) [13]. Τα πολυδιάστατα συστήματα διαχείρισης βάσεων δεδομένων αποθηκεύουν τα δεδομένα σε ν-διάστατη παράταξη. Κάθε διάσταση αντιπροσωπεύει την αντίστοιχη διάσταση του κύβου. Ένας κύβος είναι μια ομάδα από κελιά δεδομένων, που οργανώνονται βάσει των διαστάσεων των δεδομένων. Μια διάσταση είναι ένα «δομημένο χαρακτηριστικό ενός κύβου, δηλαδή μία λίστα χαρακτηριστικών, όλα εκ των οποίων είναι του ίδιου τύπου, σύμφωνα με την αντίληψη του χρήστη, όσον αφορά τα δεδομένα» [13]. Κάθε διάσταση αποτελείται από μία σειρά συγκεντρωτικών επιπέδων των δεδομένων (δηλαδή μπορεί κάποιος να τα δει από διαφορετικά επίπεδα λεπτομέρειας). Συνεπώς, αν για παράδειγμα οι διαστάσεις στα δεδομένα είναι ο χρόνος, τόπος, περιοχή της πώλησης, προϊόν, καθώς και ο πωλητής, τότε ο χρόνος μπορεί να οργανωθεί σαν μέρα-μήνας-τετράμηνο-έτος (ιεραρχικά), το προϊόν σαν προϊόν-κατηγορία-βιομηχανία κ.ο.κ. Υπάρχουν διάφορες συναρτήσεις που μπορεί ο χρήστης να χρησιμοποιήσει για να παρουσιάσει τα δεδομένα του κύβου. Ως τέτοιες συναρτήσεις ορίζονται η ’ανάλυσε’, όπου γίνεται πλοήγηση ανάμεσα σε επίπεδα δεδομένων, από το

υψηλότερο στο χαμηλότερο επίπεδο, η ’συνόψισε’, που κάνει ακριβώς το αντίθετο από την ’ανάλυσε’, η ’επικέντρωσε’, στην οποία γίνεται μείωση των διαστάσεων των δεδομένων και η ’περίστρεψε’ όπου γίνεται αλλαγή στον προσανατολισμό των διαστάσεων του κύβου.

1.2 Ειδικά προβλήματα ενδιαφέροντος

Ο παγκόσμιος οργανισμός υγείας για το θέμα της καρδιαγγειακής νόσου αναφέρει τα ακόλουθα [14]:

«Η καρδιαγγειακή νόσος είναι η νούμερο ένα αιτία θανάτου σε παγκόσμιο επίπεδο: περισσότερα άτομα πεθαίνουν κάθε χρόνο από καρδιαγγειακή νόσο παρά από οποιαδήποτε άλλη αιτία. Εκτιμάται ότι 17,1 εκατομμύρια άνθρωποι πέθαναν από καρδιαγγειακή νόσο το 2004, δηλαδή το 29% του συνόλου των παγκόσμιων θανάτων. Από αυτούς τους θανάτους, περίπου 7,2 εκατομμύρια οφείλονταν σε στεφανιαία νόσο και 5,7 εκατομμύρια λόγω εγκεφαλικού επεισοδίου. Το 82% των θανάτων με καρδιαγγειακή νόσο πραγματοποιείται σε χώρες χαμηλού και μέσου εισοδήματος και εμφανίζονται σχεδόν εξίσου σε άνδρες και γυναίκες. Μέχρι το 2030, περίπου 23.6 εκατομμύρια άνθρωποι θα πεθάνουν από καρδιαγγειακή νόσο, κυρίως από καρδιακές παθήσεις και εγκεφαλικά επεισόδια. Αναμένεται ότι αυτές θα παραμείνουν κύριες αιτίες θανάτου. Επίσης αναφέρεται ότι η μεγαλύτερη ποσοστιαία αύξηση θα πραγματοποιηθεί στην Ανατολική Μεσόγειο. Η μεγαλύτερη αύξηση του αριθμού των θανάτων που θα επέλθουν θα είναι στην περιφέρεια της Νότιο-Ανατολική Ασίας.»

Στα κράτη της Ευρώπης ο απολογισμός είναι τεράστιος, αντιπροσωπεύοντας το ήμισυ περίπου των θανάτων. Η τραγικότητα συνίσταται στο γεγονός ότι οι θάνατοι, ή τουλάχιστον το μεγαλύτερο μέρος αυτών των θανάτων, θα μπορούσαν να προληφθούν και μάλιστα με μεθόδους απλές και διαδικασίες που θα μπορούσε να ακολουθήσει κάθε ένας που καθίσταται υποψήφιο θύμα των καρδιαγγειακών νοσημάτων [15].

Από διάφορες στατιστικές μελέτης που έχουν γίνει στη Κύπρο [16], έχει εκτιμηθεί ότι ετησίως συμβαίνουν 700 με 800 καρδιακοί θάνατοι, ενώ ένας στους χίλιους Κύπριους κινδυνεύει να υποστεί καρδιακή ανακοπή. Από μελέτη που έχει γίνει το 2007, στον συνολικό αριθμό κλήσεων στο νοσοκομείο της πρωτεύουσας παρατηρήθηκε ότι από τις 83 κλήσεις οι 69 αφορούσαν περιστατικά καρδιακής ανακοπής, δηλαδή ποσοστό περίπου 83%, και το οποίο αυξάνεται.

Υπάρχουν διάφορα είδη καρδιακών παθήσεων, και μία από τις πιο σημαντικές είναι το έμφραγμα του μυοκαρδίου, το οποίο συγκαταλέγεται στα οξέα στεφανιαία σύνδρομα. Το έμφραγμα του μυοκαρδίου είναι η νέκρωση του μυοκαρδίου της καρδιάς, η οποία οφείλεται σε απόφραξη λόγω δημιουργίας θρόμβου σε μια στεφανιαία αρτηρία. Ο θρόμβος διακόπτει τη κυκλοφορία του αίματος με αποτέλεσμα την νέκρωση του μυοκαρδίου. Υπάρχουν διάφοροι και πολλοί παράγοντες οι οποίοι μπορεί να προκαλέσουν καρδιακό έμφραγμα. Έτσι είναι πολύ σημαντικό να μελετήσουμε και να βρούμε τους πιο κύριους παράγοντες στους οποίους προκαλείται καρδιακό έμφραγμα και κυρίως για τους λόγους που οδηγούν σε έμφραγμα του μυοκαρδίου.

Η αγγειοπλαστική των στεφανιαίων αρτηριών αποκαλούμενη και μερικές φορές μπαλονάκι ή PTCA (Percutaneous Transluminal Coronary Angioplasty) ή PCI (Percutaneous Coronary Intervention), είναι μια θεραπευτική πράξη που εκτελείται από καρδιολόγους προκειμένου να ανοιχτεί η αποφραγμένη στεφανιαία αρτηρία και να αποκατασταθεί η ροή αίματος στο μυοκάρδιο. Η αγγειοπλαστική χρησιμοποιείται ως εναλλακτική διεργασία στη χειρουργική επέμβαση παράκαμψης των στεφανιαίων αρτηριών (by-pass). Είναι λιγότερο αιματηρή του bypass, λιγότερο ακριβή, πιο σύντομη, ενώ ο ασθενής επιστρέφει συνήθως στο σπίτι του την επόμενη ημέρα. Το κύριο μειονέκτημα της μεθόδου είναι ότι σε ποσοστό 20-30% των ασθενών, η αρτηρία μπορεί να ξανακλείσει τους επόμενους 6 μήνες, μια κατάσταση που αποκαλείται επαναστένωση. Οι νέες μεταλλικές προσθέσεις (stents) που κυκλοφορούν με απελευθέρωση διαφόρων ουσιών μειώνουν δραματικά το ποσοστό τη επαναστένωσης. Η αγγειοπλαστική των στεφανιαίων αρτηριών εκτελείται σε μη επείγουσα βάση για τη θεραπεία

των χρόνιων στεφανιαίων στενώσεων και σε επείγουσα βάση για τη θεραπεία του οξέος εμφράγματος του μυοκαρδίου [18].

Είναι γνωστό ότι η στεφανιαία νόσος οφείλεται σε στενώσεις των στεφανιαίων αρτηριών. Η εγχειρητική τεχνική παράκαμψης των στεφανιαίων στενώσεων ονομάζεται αορτοστεφανιαία παράκαμψη. Η παράκαμψη (υπερπήδηση), με φλεβικά ή αρτηριακά μοσχεύματα, γίνεται συνήθως από την αορτή προς την στεφανιαία αρτηρία μετά την στένωση (το ένα άκρο της φλέβας ή της αρτηρίας συρράπτεται στην αορτή και το άλλο στην στεφανιαία αρτηρία μετά την στένωση. Έτσι με αρκετό αίμα τροφοδοτείται η στεφανιαία κυκλοφορία από την αορτή). Η παράκαμψη (υπερπήδηση), με αρτηρίες γίνεται από τη θέση της εκφύσεώς των προς την στεφανιαία αρτηρία πάλι μετά την στένωση (το ένα άκρο της αρτηρίας παραμένει ως έχει και συρράπτεται στην στεφανιαία αρτηρία μετά την στένωση) [17].

Η εξόρυξη συχνών προτύπων οδηγεί στην ανακάλυψη ενδιαφέρων σχέσεων και συσχετίσεων στους παράγοντες πάθησης ενός καρδιακού επεισοδίου. Με την αύξηση των καρδιακών επεισοδίων, συλλέγονται και αποθηκεύονται τα δεδομένα των ασθενών, κι οι κανόνες που εξάγονται από την εξόρυξη δεδομένων παρέχουν ένα συνοπτικό τρόπο για να εκφραστούν χρήσιμες πληροφορίες, που γίνονται εύκολα κατανοητές από τους ειδικούς. Η ανακάλυψη ενδιαφέρον συσχετίσεων μεταξύ των παραγόντων πάθησης ενός επεισοδίου μπορούν να βοηθήσουν τους ειδικούς στη λήψη αποφάσεων, έτσι ώστε να δώσουν σωστή θεραπευτική αγωγή σους ασθενείς και έτσι να αποφευχθεί κάποιο επεισόδιο.

Στη βάση δεδομένων που μελετήθηκε για τα τρία επεισόδια, έμφραγμα μυοκαρδίου (myocardial infarction, MI), αγγειοπλασική (percutaneous coronary intervention, PCI) και στεφανιαία παράκαμψη (Coronary Artery Bypass Graft surgery, CABG), είχαμε σε σύνολο 528 ασθενών, 358 ασθενείς με έμφραγμα μυοκαρδίου, 213 ασθενείς με αγγειοπλαστική και 215 ασθενείς με στεφανιαία παράκαμψη.

1.3 Στόχος διατριβής

Στόχος της διατριβής αυτής είναι η ανάπτυξη ενός ολοκληρωμένου συστήματος που θα υποστηρίζει την αξιολόγηση των παραγόντων κινδύνου σε καρδιαγγειακές βάσεις δεδομένων και την εξόρυξη κανόνων εκτίμησης κινδύνου βασισμένων σε αλγόριθμους δένδρων αποφάσεων και κανόνων συσχέτισης. Οι κανόνες αυτοί θα αποτελούν τη βάση για την αξιολόγηση ενός νέου ασθενή, συγκρίνοντας τις τιμές των παραγόντων του ασθενή με αυτές των κανόνων.

1.4 Πρωτοτυπία διατριβής

I. Ανάπτυξη αλγορίθμων εξαγωγής κανόνων από δέντρα αποφάσεων βασισμένων σε διαφορετικά κριτήρια διαχωρισμού. Κάθε κριτήριο διαχωρισμού δύναται να κτίσει διαφορετικά το δέντρο. Επιλέγεται η ομάδα από δέντρα που έχουν την καλύτερη απόδοση και κατ' επέκταση οι κανόνες που εξάγονται από τα δέντρα αυτά.
II. Ανάπτυξη αλγορίθμων εξαγωγής κανόνων από αλγόριθμους κανόνων συσχέτισης βασισμένους σε διαφορετικά μέτρα. Τα μέτρα αυτά θα βοηθήσουν στα στάδιο επιλογής των καλύτερων κανόνων.
III. Ανάπτυξη ενός νέου αλγόριθμου εξαγωγής κανόνων συσχέτισης, που με μια σάρωση της βάσης δεδομένων δημιουργεί τους κανόνες. Πέραν τούτου μπορεί να χρησιμοποιήσει οποιοδήποτε υλοποιημένο μέτρο ή μέτρα σαν κριτήριο επιλογής ενός κανόνα ή μιας ομάδας από κανόνες. Επιπλέον η εξαγωγή και παρουσίαση των κανόνων γίνεται δομημένα βάσει των παραγόντων κινδύνου που υποστηρίζει την πιο εύκολη εξόρυξη γνώσης.
IV. Ανάπτυξη μεθοδολογίας αξιολόγησης κανόνων παραγόντων κινδύνου για εκτίμηση κινδύνου για ανεύρεση των στατιστικά σημαντικότερων κανόνων.

V. Ανάπτυξη αυτόματης μεθοδολογίας φιλτραρίσματος κανόνων βάσει των σημαντικότερων μέτρων κάθε μοντέλου.

VI. Εφαρμογή των πιο πάνω σε βάσεις δεδομένων με καρδιαγγειακά επεισόδια για έμφραγμα μυοκαρδίου (MI), αγγειοπλαστική (PCI) και στεφανιαία παράκαμψη (CABG), όπως επίσης και σε βάσεις δεδομένων με γενετικά δεδομένα και με δεδομένα για άτομα με παχυσαρκία.

VII. Αξιοποίηση των ευρημάτων του πιο πάνω στόχου για την αξιολόγηση νέων περιστατικών (χρησιμοποιώντας ο καρδιολόγος τα ευρήματα ενός νέου ασθενή, θα μπορεί εύκολα να αξιολογήσει την κατάστασή του βασισμένος στους κανόνες που υπάρχουν στο σύστημα).

Στο Παράρτημα 1 δίδονται οι δημοσιεύσεις που βασίζονται στα ευρήματα αυτής της διατριβής.

1.5 Ανασκόπηση διατριβής

Στο πρώτο κεφάλαιο γίνεται μια εισαγωγή στην εξόρυξη δεδομένων και την απόκτηση γνώσης. Παρουσιάζονται τα βήματα όλης της διαδικασίας εξόρυξης δεδομένων, από τη χρήση μιας βάσης δεδομένων ή αρχείου με δεδομένα μέχρι την απόκτηση γνώσης. Γίνεται ανάλυση των δεδομένων, παρουσιάζεται η αρχιτεκτονική αποθήκευσης δεδομένων και τέλος διατυπώνεται η πρωτοτυπία της διατριβής.

Ακολουθούν, στα κεφάλαια δύο και τρία, οι τεχνικές εξόρυξης δεδομένων, όπου περιγράφονται οι αλγόριθμοι ταξινόμησης, τα δέντρα αποφάσεων και το κλάδεμα τους και οι αλγόριθμοι συσχέτισης και οι κατηγορίες τους.

Στο τέταρτο κεφάλαιο γίνεται αναφορά στα μέτρα που χρησιμοποιούνται στην εξόρυξη δεδομένων και στα νέα μέτρα που ορίστηκαν για τις μεθόδους που χρησιμοποιούνται σε αυτήν τη διατριβή. Επίσης αναφέρεται ο τρόπος αξιολόγησης των μοντέλων.

Έχουν μελετηθεί εφαρμογές των τεχνικών εξόρυξης από δεδομένα σε ιατρικά προβλήματα και ιδιαίτερα σε καρδιαγγειακά. Αυτές οι εφαρμογές αναλύονται στο κεφάλαιο πέντε αυτής της εργασίας.

Στο κεφάλαιο έξι περιγράφεται το σύστημα που αναπτύχθηκε, ξεκινώντας από τη μεθοδολογία που χρησιμοποιήθηκε, την περιγραφή των αλγορίθμων, τον τρόπο που ιεραρχούνται οι παράγοντες πάθησης ενός επεισοδίου σε μια ασθένεια και τα μέτρα του κεφαλαίου τέσσερα που χρησιμοποιήθηκαν στο προτεινόμενο σύστημα. Επιπλέον, γίνεται σύγκριση και συνδυασμός αποτελεσμάτων από διαφορετικούς αλγόριθμους και περιγράφεται η εφαρμογή σε καρδιαγγειακές βάσεις.

Περιγράφονται αναλυτικά τα πρώτα στάδια της εξόρυξης δεδομένων, δηλαδή η προεπεξεργασία της βάσης δεδομένων, τα χαρακτηριστικά της και η επιλογή των χαρακτηριστικών που θα μελετηθούν, η συμπλήρωση των ελλειπουσών τιμών και η κωδικοποίηση των χαρακτηριστικών. Επίσης, παρουσιάζεται η στατιστική ανάλυση των χαρακτηριστικών που χρησιμοποιήθηκαν.

Στο κεφάλαιο επτά γίνεται η παρουσίαση των αποτελεσμάτων. Γίνεται παρουσίαση των κανόνων με δέντρα απόφασης και με τη μέθοδο της συσχέτισης. Τα αποτελέσματα εμφανίζονται κατά μοντέλο: πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά το επεισόδιο (B+A). Σε κάθε μοντέλο έχουμε τα επεισόδια έμφραγμα μυοκαρδίου (MI), αγγειοπλαστική (PCI) και στεφανιαία παράκαμψη (CABG). Παρουσιάζονται επίσης οι κυριότεροι παράγοντες πάθησης επεισοδίου σε κάθε μοντέλο, όπως επίσης και μια σύγκριση του αλγόριθμου Apriori με τον νέο αλγόριθμο AKAMAS. Στο κεφάλαιο οκτώ γίνεται συζήτηση των αποτελεσμάτων και σύγκριση με αποτελέσματα άλλων ερευνητών. Επίσης παρουσιάζεται το προτεινόμενο σύστημα που θα δοθεί στον καρδιολόγο. Στο κεφάλαιο εννέα έχουμε τα συμπεράσματα και η προτεινόμενη μελλοντική εργασία και στο παράρτημα Ι τις δημοσιεύσεις μας.

Κεφάλαιο 2: Αλγόριθμοι ταξινόμησης βασισμένοι σε δέντρα απόφασης

2.1 Γενικά

Στη ταξινόμηση έχουμε δεδομένα τα οποία έχουν εξαρχής γνωστές τάξεις. Είναι μια πολύ διαδεδομένη τεχνική εξόρυξης που ταξινομεί τα δεδομένα στις υπάρχουσες τάξεις και δημιουργεί πρότυπα [2]. Ο αλγόριθμος ταξινόμησης χρησιμοποιεί τα δεδομένα για να καθορίσει το σύνολο των παραμέτρων που χρειάζονται για περαιτέρω διάκριση (ταξινόμηση) δεδομένων. Στη συνέχεια κωδικοποιεί τα δεδομένα - χαρακτηριστικά (πρότυπα) σε ένα μοντέλο, που ονομάζεται ταξινομητής. Αφού δημιουργηθεί ένας αποτελεσματικός ταξινομητής, χρησιμοποιείται σαν πρόβλεψη, ώστε να ταξινομήσει νέα δεδομένα στις τάξεις.

Οι αλγόριθμοι ταξινόμησης διακρίνονται σε αλγόριθμους που παράγουν δέντρα απόφασης (decision trees) [19], σε λογιστική παλινδρόμηση [20], σε ταξινομητές Bayes [21], βασισμένους σε νευρωνικά δίκτυα [22] και σε ταξινομητές SVM (support vector machines) [23].

Τα δέντρα αποφάσεων παράγουν μία οπτική παρουσίαση των κανόνων, γεγονός το οποίο συμβάλλει σημαντικά στη διάδοση τους ως μέθοδο για ταξινόμηση. Τα δέντρα αποφάσεων είναι δυνατόν να χρησιμοποιηθούν στην ταξινόμηση (πρόβλεψη σε ποια τάξη ανήκουν κάποια δεδομένα), στην παλινδρόμηση (πρόβλεψη κάποιας συγκεκριμένης τιμής της εξαρτημένης μεταβλητής από ανεξάρτητες μεταβλητές), αλλά και για τη μείωση του όγκου δεδομένων μέσω του μετασχηματισμού τους σε μία πιο συμπιεσμένη μορφή, διατηρώντας όμως τα βασικά χαρακτηριστικά των δεδομένων [24]. Τα δέντρα αποφάσεων αποτελούν την πιο διαδεδομένη μέθοδο για ταξινόμηση και γι' αυτό παρουσιάζονται αναλυτικότερα παρακάτω. Το δέντρο απόφασης κατασκευάζεται από το σύνολο εκπαίδευσης (training set), δηλαδή από ένα σύνολο δεδομένων/ εγγραφών. Κάθε εγγραφή χαρακτηρίζεται από το σύνολο

χαρακτηριστικών (attributes) και την τάξη (label). Η λογική της κατασκευής ενός δέντρου αποφάσεων είναι η σωστή και ακριβής σχέση (ή αλληλεξάρτηση) των χαρακτηριστικών αυτών και της τάξης [24]. Ένα δέντρο αποφάσεων περιέχει μηδενικούς ή περισσότερους ενδιάμεσους κόμβους (internal nodes) και έναν ή περισσότερους τερματικούς (leaf) κόμβους. Κάθε ενδιάμεσος κόμβος αποτελείται από δύο ή περισσότερους κόμβους-παιδιά (child nodes). Όλοι οι ενδιάμεσοι κόμβοι περιέχουν διαιρέσεις (splits), οι οποίες ελέγχουν την τιμή της έκφρασης των χαρακτηριστικών. Τέλος, ένας τερματικός κόμβος αποτελείται από μία τιμή τάξης. Οι βασικοί αντικειμενικοί σκοποί των ταξινομητών δέντρων αποφάσεων είναι [25]:

i. Να ταξινομήσουν σωστά όσο το δυνατόν περισσότερο ποσοστό από το σύνολο εκπαίδευσης (training set).
ii. Να γενικεύσουν πέρα από το δείγμα εκπαίδευσης, έτσι ώστε ένα νέο και άγνωστο δείγμα εκπαίδευσης να μπορεί να ταξινομηθεί με όσο το δυνατό μεγαλύτερη ακρίβεια.
iii. Να μπορούν να ενημερώνονται (update), όταν διατεθούν περισσότερα δεδομένα.
iv. Να έχουν όσο πιο απλή δομή γίνεται.

Στη στατιστική η τεχνική της εξαγωγής δέντρων αποφάσεων ξεκίνησε με τη δημιουργία ιεραρχικής ταξινόμησης για διερεύνηση ερευνητικών δεδομένων [26]. Διάφορα στατιστικά προγράμματα, όπως το AID [27], το MAID [28], το THAID [29], και το CHAID [30] κατασκεύασαν δυαδικά διαχωριστικά δέντρα (binary segmentation trees), τα οποία αποσκοπούσαν στην ανακάλυψη των σχέσεων μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Στην αναγνώριση προτύπων τα δέντρα αποφάσεων χρησιμοποιήθηκαν στην επεξήγηση εικόνων από απομακρυσμένους δορυφόρους, όπως ο LANDSAT στη δεκαετία του 1970 [31]. Στην επιστήμη της μηχανικής μάθησης (machine learning) τα δέντρα αποφάσεων χρησιμοποιήθηκαν, προκειμένου να αποφευχθεί το «μποτιλιάρισμα» (bottleneck) της απόκτησης γνώσης για έμπειρα συστήματα [32]. Τέλος, στη διαδοχική διάγνωση σφαλμάτων

(sequential fault diagnosis) οι αλγόριθμοι που χρησιμοποιούνται παίρνουν συχνά τη μορφή δέντρων αποφάσεων [33] και [34].

Η λογιστική παλινδρόμηση (logistic regression) αποτελείται από ανεξάρτητες μεταβλητές και από την εξαρτημένη μεταβλητή. Τα μοντέλα αυτά χρησιμοποιούνται για την εκτίμηση των παραγόντων που επηρεάζουν την εξαρτημένη μεταβλητή και παράγουν μία λειτουργική μορφή συνάρτηση f, καθώς και το παραμετρικό διάνυσμα a, προκειμένου να εκφρασθεί η δεσμευμένη πιθανότητα $P_{(y/x)}$ (όπου y είναι η εξαρτημένη και x η ανεξάρτητη μεταβλητή). Η παράμετρος a καθορίζεται από τα δεδομένα, χρησιμοποιώντας συνήθως τη μέθοδο της εκτίμησης maximum-likelihood [35].

Επίσης, οι ταξινομητές Bayes θεωρούν ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους (δεδομένης της τάξης), και αποθηκεύουν μία πιθανολογική περίληψη για κάθε τάξη, προκειμένου να κάνουν την ταξινόμηση. Βασίζονται στη στατιστική θεωρία κατηγοριοποίησης του Bayes [21]. Στόχος είναι να κατηγοριοποιηθεί ένα δείγμα X σε μια από τις δεδομένες κατηγορίες C_1, C_2, ..,C_n χρησιμοποιώντας ένα μοντέλο πιθανότητας που ορίζεται σύμφωνα με τη θεωρία του Bayes. Πρόκειται για κατηγοριοποιητές που κάνουν αποτίμηση πιθανοτήτων και όχι πρόβλεψη. Αυτό πολλές φορές είναι πιο χρήσιμο και αποτελεσματικό. Εδώ οι προβλέψεις έχουν έναν βαθμό και σκοπός είναι το αναμενόμενο κόστος να ελαχιστοποιείται. Κάθε κατηγορία χαρακτηρίζεται από μια εκ των προτέρων πιθανότητα. Υποθέτουμε πως το δεδομένο δείγμα X ανήκει σε μια τάξη C_i. Βασισμένοι στους ορισμούς και τα παραπάνω καθορίζουμε την εκ των υστέρων πιθανότητα.

Ο πιο γνωστός Bayesian κατηγοριοποιητής είναι ο naïve Bayesian κατηγοριοποιητής που υποθέτει πως η επίδραση ενός γνωρίσματος σε μια δεδομένη κατηγορία είναι ανεξάρτητη από τις τιμές των άλλων γνωρισμάτων [21]. Ένας άλλος Bayesian κατηγοριοποιητής είναι τα Bayesian Belief Networks. Είναι γραφικά μοντέλα που επιτρέπουν την παρουσίαση των εξαρτήσεων μεταξύ των υποσυνόλων των γνωρισμάτων.

Τα νευρωνικά δίκτυα (neural networks, NN) χρησιμοποιούνται για πρόβλεψη και κατηγοριοποίηση και είναι εμπνευσμένα από τη νευροφυσιολογία του ανθρώπινου εγκεφάλου, αποτελούνται από στοιχεία (νευρώνες) τα οποία συμπεριφέρονται κατά τρόπο

ανάλογο των πιο στοιχειωδών λειτουργιών των φυσιολογικών κυττάρων. Τα βήματα που χρησιμοποιούν τα νευρωνικά δίκτυα για να κατασκευάζουν ένα μοντέλο κατηγοριοποίησης ή πρόβλεψης είναι [22]:

i. Αναγνώριση των χαρακτηριστικών εισόδου και τις τάξεις εξόδου.
ii. Κατασκευή ενός δικτύου με την κατάλληλη τοπολογία.
iii. Επιλογή του σωστού συνόλου εκπαίδευσης.
iv. Εκπαίδευση του δικτύου με βάση ένα αντιπροσωπευτικό σύνολο δεδομένων που θα απεικονίζονται οι υπό μελέτη τάξεις έτσι ώστε να μεγιστοποιηθεί η δυνατότητα του δικτύου να τις αναγνωρίζει σωστά.

Ένα σωστά εκπαιδευμένο νευρωνικό δίκτυο μπορεί να παράγει αποδεκτά, από πλευράς ακρίβειας, αποτελέσματα σε σύντομο υπολογιστικό χρόνο. Η ιδιότητα αυτή των NN αποτελεί και το βασικό τους πλεονέκτημα. Επίσης τα όρια απόφασης (δηλαδή το όριο που καθορίζει τα σημεία στο χώρο που διαχωρίζουν τις δύο ή περισσότερες τάξεις) μπορεί να είναι μη γραμμικά.

Το μοντέλο που παράγεται από το δίκτυο εφαρμόζεται για να προβλέψει τις κατηγορίες των μη κατηγοριοποιημένων δειγμάτων [36]. Αποτελούνται από νευρώνες με βάση τη νευρωνική δομή του εγκεφάλου, οι οποίοι επεξεργάζονται ένα στοιχείο κάθε φορά και μαθαίνουν συγκρίνοντας τη κατηγοριοποίηση τους για μια εγγραφή με τη γνωστή πραγματική κατηγοριοποίηση της εγγραφής.

Οι ταξινομητές SVM αποτελούν αλγοριθμικές εφαρμογές ιδεών από τη στατιστική θεωρία-μάθηση. Οι ταξινομητές αυτοί δημιουργούν όρια που διαχωρίζουν τα δεδομένα σε τάξεις, λύνοντας ένα πρόβλημα βελτιστοποίησης - μεγιστοποίησης συνήθως δευτεροβάθμιας εξίσωσης με περιορισμούς [37]. Χρησιμοποιώντας διαφορετικές συναρτήσεις, το μοντέλο μπορεί να περιλαμβάνει διάφορους βαθμούς μη γραμμικότητας και ευελιξίας. Τα μοντέλα SVM παράγουν διχοτομική ταξινόμηση, που σημαίνει ότι δεν δίνεται η πιθανότητα βαθμού τάξης (probability of class membership).

2.2 Αλγόριθμοι ταξινόμησης δέντρων

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 [38], 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 [39], 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) [40]. 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) [2].

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 ποιο ποσοστό των πλειάδων 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 [41]. 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 [3], όπου οι

τιμές a_1, a_2,…., a_m ανήκουν σε ένα πεδίο Α. Η δεσμευμένη πιθανότητα $P(c_i/a_j)$ αντιπροσωπεύει την πιθανότητα να συμβαίνει το c_i, δεδομένου ότι συμβαίνει το a_j. Έτσι, το πεδίο με τη μικρότερη εντροπία χωρίζει καλύτερα το σύνολο εκπαίδευσης. Αναλυτικά, τα βήματα του αλγορίθμου ID3 έχουν ως εξής:

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

ii. 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.

2.3 Κανόνες δημιουργίας δέντρων

Στην παραγωγή των δέντρων αποφάσεων μεγάλο ρόλο παίζει η κατάταξη των χαρακτηριστικών της βάσης δεδομένων και ο ορισμός των τάξεων. Ένας κανόνας ελέγχου διαιρεί τα δεδομένα σε υποομάδες. Τα δέντρα αποφάσεων που έχουν ένα χαρακτηριστικό σε κάθε ενδιάμεσο κόμβο χαρακτηρίζονται σαν μονομεταβλητά (univariate). Ο διαχωρισμός γίνεται με δύο τρόπους: α) με μετρικές αποστάσεις (distance measures) και β) με μετρικές εξάρτησης (dependence measures). Στην πρώτη περίπτωση ανήκει η εντροπία, και έχει σαν σκοπό τη μεγιστοποίηση της συνολικής αμοιβαίας πληροφορίας. Επίσης, η ίδια έχει

εφαρμοσθεί στην αναγνώριση προτύπων με σκοπό την τοπική μεγιστοποίηση του πληροφοριακού οφέλους (information gain), δηλαδή τη μείωση στην εντροπία που οφείλεται στο διαχωρισμό κάθε κόμβου [42] - [44].

Σαν κανόνας διαίρεσης έχει προταθεί ο δείκτης διαφοροποίησης Gini (Gini index of diversity), και περιλαμβάνεται στον αλγόριθμο CART, ο οποίος έχει χρησιμοποιηθεί για την κατασκευή δέντρων και την αναγνώριση προτύπων [45]. Έχουν προταθεί επιπλέον και άλλοι κανόνες, όπως ο κανόνας twoing [7], και ο κανόνας MPI (mean posterior improvement) [46], που αποτελούν βελτιώσεις του δείκτη Gini. Η απόσταση Bhattacharya [47] και η στατιστική χ2 [48] - [50] είναι κάποια άλλα μεγέθη που βασίζονται σε αποστάσεις και έχουν κατά καιρούς χρησιμοποιηθεί στην παραγωγή δέντρων αποφάσεων. Η απόσταση Kolmogorov-Smirnoff ενώ είχε αρχικά εφαρμοσθεί στην παραγωγή δέντρων αποφάσεων με δύο τάξεις [51], [52], επεκτάθηκε στην παραγωγή δέντρων αποφάσεων με πολλές τάξεις [53]. Τέλος, έχει προταθεί μία απλή μέθοδος για το διαχωρισμό τάξεων η οποία υποθέτει ότι οι τιμές των χαρακτηριστικών ακολουθούν κατανομή Gaussian [54].

Τα δέντρα απόφασης που χρησιμοποιούν κανόνες διακλάδωσης με πάνω από ένα χαρακτηριστικό σε κάθε κόμβο ονομάζονται πολυμεταβλητά (multivariate). Στους κόμβους γίνεται ένας γραμμικός συνδυασμός των χαρακτηριστικών και γι' αυτό ονομάζονται γραμμικά δέντρα. Σε αυτήν την περίπτωση υπάρχει μεγάλη δυσκολία εξεύρεσης διακλαδώσεων σε αντίθεση με τα μονομεταβλητά δέντρα απόφασης. Χρησιμοποιούνται κάποιες ευριστικές μεθόδοι (heuristics), για να διευκολυνθεί η εξεύρεση γραμμικών διακλαδώσεων [55] που βασίζεται στην ανάλυση γραμμικής διακριτής ταξινόμησης (linear discriminant analysis). Ο διαφοροποιητής υπολογίζει το πιο γρήγορο χαρακτηριστικό και το αντίστοιχο κατώφλι. Η μέθοδος προϋποθέτει ότι τα καλύτερα χαρακτηριστικά σε κάθε κόμβο έχουν καθορισθεί από το χρήστη [24]. Σε αντίθεση, μία άλλη μέθοδος των Loh *et al*. επιλέγει τις μεταβλητές σε κάθε στάδιο, ανάλογα με τα επιθυμητά δεδομένα και τον τύπο των διακλαδώσεων [56].

Ο αλγόριθμος CART [7] χρησιμοποιεί τη μέθοδο hill climbing, όπου εφαρμόζεται γραμμικός συνδυασμός των χαρακτηριστικών, καθώς επίσης και ευριστικές hill climbing τεχνικές και

εξάλειψη των χαρακτηριστικών, προκειμένου να βρεθούν οι βέλτιστοι γραμμικοί συνδυασμοί σε κάθε κόμβο.

Τα νευρωνικά δίκτυα έχουν επίσης χρησιμοποιηθεί σε συνδυασμό με τα δέντρα αποφάσεων. Το αποτέλεσμα που προκύπτει μπορεί να περιγραφεί ως δέντρα αποφάσεων με μη-γραμμικές διακλαδώσεις. Μία μεθοδολογία μετατρέπει ένα μονομεταβλητό δέντρο σε νευρωνικό δίκτυο, στη συνέχεια εκπαιδεύεται και καταλήγει σε δίκτυα εντροπίας με δομή δέντρων [57]. Επίσης, έχουν χρησιμοποιηθεί δέντρα αποφάσεων με μικρά πολλαπλά δίκτυα σε κάθε κόμβο, τα οποία παράγουν μη γραμμικές, πολλαπλές διακλαδώσεις [58]. Τέλος, αξίζει να σημειωθεί ότι οι αρχικοί αλγόριθμοι ταξινόμησης έχουν επεκταθεί (οι οποίοι είχαν αρχικά σχεδιασθεί για συνεχείς αριθμητικές μεταβλητές), ώστε να συμπεριλάβουν και διακριτές τιμές αριθμητικών μεταβλητών [59], καθώς και κατηγορικές μεταβλητές με πολλαπλές τιμές [60]. Μία επέκταση του βασικού αλγορίθμου ID3, ο αλγόριθμος CID3, κάνει διάκριση μεταξύ αριθμητικών και μη-αριθμητικών γραμμικών μεταβλητών [61].

Ο αλγόριθμος C4.5 αναπτύχθηκε από τον Quinlan [6] και αποτελεί εξέλιξη του αλγορίθμου ID3. Ο καινούργιος αλγόριθμος σε σχέση με τον προκάτοχό του έχει τα εξής βασικά πλεονεκτήματα:

i. Δυνατότητα επεξεργασίας και διαχείρισης ποσοτικών κριτηρίων
ii. Δυνατότητα διαχείρισης δεδομένων με ελλιπή στοιχεία
iii. Αποφυγή της μεγάλης προσαρμογής στα δεδομένα του δείγματος εκμάθησης

Οι διαφορές μεταξύ των αλγορίθμων C4.5 και CART είναι:

* Τα χαρακτηριστικά στον CART είναι δυαδικά κωδικοποιημένα, ενώ ο C4.5 επιτρέπει τα χαρακτηριστικά να έχουνε περισσότερες κωδικοποιήσεις. (Ο C4.5 έχει εκδόσεις που παίρνει και συνεχόμενα χαρακτηριστικά και όχι κωδικοποιημένα). Ο CART χρησιμοποιεί το δείκτη Gini, ενώ ο C4.5 μπορεί να χρησιμοποιήσει διάφορα κριτήρια διαχωρισμού.
* Ο CART κλαδεύει το δέντρο χρησιμοποιώντας ένα μοντέλο κόστους της πολυπλοκότητας του οποίου οι παράμετροι υπολογίζονται από διασταυρωμένη επικύρωση, ενώ ο C4.5 χρησιμοποιεί ένα μόνο πέρασμα στον αλγόριθμο που

προέρχεται από διωνυμικό όριο εμπιστοσύνης.

Ο αλγόριθμος C5.0 είναι η εμπορική έκδοση του C4.5. Στους αλγόριθμους των δέντρων απόφασης το πιο σημαντικό στοιχείο είναι η μέθοδος που χρησιμοποιείται για να διαχωρίσει κάθε κόμβο στο δέντρο. Ο C5.0 χρησιμοποιεί την αναλογία του κέρδους πληροφορίας (information gain) για να εκτιμήσει το διαχωρισμό σε κάθε εσωτερικό κόμβο του δέντρου. Το κέρδος πληροφορίας μετρά την μείωση της εντροπίας στα δεδομένα που παράγονται από τη διάσπαση.

2.4 Μέθοδοι περιορισμού επέκτασης δέντρων

Είναι σημαντικό κατά τη διάρκεια της δημιουργίας ενός δέντρου να γνωρίζουμε σε ποιο σημείο ολοκληρώνεται το μέγεθος του δέντρου. Έχει αναφερθεί ότι η ποιότητα ενός δέντρου εξαρτάται περισσότερο από κανόνες που ορίζουν ορθά το σημείο στο οποίο το δέντρο σταματά να αναπτύσσεται, παρά από τους κανόνες διακλάδωσης [7]. Ο Breiman [7] έχει εισηγηθεί μια μέθοδο για την παραγωγή κατάλληλου μεγέθους ενός δέντρου απόφασης που είναι η μέθοδος του κλαδέματος (pruning method). Η διαδικασία αυτή εφαρμόζεται όταν δημιουργηθεί ένα δέντρο απόφασης και έπειτα καταργούνται τα υποδέντρα που δεν συνεισφέρουν στην ακρίβεια (generalization accuracy). Υπάρχουν και άλλες εισηγήσεις, όπως το κλάδεμα με το κόστος πολυπλοκότητας και το κλάδεμα με μειωμένο λάθος. Στην πρώτη περίπτωση [7] δημιουργούνται μικρά δέντρα και επιλέγεται ένα δέντρο σαν δέντρο κλαδέματος με κριτήριο την ακρίβεια ταξινόμησης. Στη δεύτερη περίπτωση δεν παράγονται πολλά δέντρα και επομένως είναι μια γρηγορότερη μέθοδος [62].

Επιπρόσθετες τεχνικές που εφαρμόζονται εκτός από το κλάδεμα στα δέντρα απόφασης, είναι οι ακόλουθες:

i. Μη διακλάδωση των κόμβων όταν δεν έχουν ένα συγκεκριμένο αριθμό πλειάδων που προκαθορίζεται [52].

ii. Δημιουργία ενός δέντρου με καλή δομή και ακολούθως διακλάδωση σε όλους τους κόμβους. Αυτό γίνεται κατορθωτό χρησιμοποιώντας τον αλγόριθμο ομαδοποίησης k-means [63].

iii. Ορισμός ενός κατωφλιού το οποίο σταματά την ανάπτυξη του δέντρου αν το κριτήριο διακλάδωσης γίνεται μικρότερο από αυτό το κατώφλι. Μπορεί να οριστεί για ολόκληρο το δέντρο ή για μεμονωμένους κόμβους.

iv. Δέντρα που μετατρέπονται σε κανόνες: αυτή η μεθοδολογία έχει προταθεί από τον Quinlan [6] και [62].

v. Μείωση δέντρων: έχουν χρησιμοποιηθεί διακριτές αρχές της θεωρίας αποφάσεων, προκειμένου να μειωθεί ένα δέντρο. Σύμφωνα με κάποιο κριτήριο προσδοκώμενου κόστους κάθε δέντρο το οποίο δεν μπορεί να μειωθεί περαιτέρω είναι βέλτιστο [64].

2.5 Σύγκριση με άλλες μεθόδους

Παράλληλα με τα δέντρα αποφάσεων υπάρχουν και άλλες μέθοδοι ανάλυσης δεδομένων όπως η στατιστική και η μηχανική μάθηση (machine learning). Για το λόγο αυτό έχουν γίνει μετρήσεις για να συγκριθεί η αποτελεσματικότητα της κάθε τεχνικής-μεθοδολογίας. Η σύγκριση αυτή δείχνει ότι οι περισσότερες μέθοδοι παράγουν ακριβείς ταξινομητές, αλλά δεν παρέχουν αρκετή πληροφόρηση σχετικά με τη δομή του προβλήματος. Έχει γίνει σύγκριση και με νευρωνικά δίκτυα και τα αποτελέσματα των δύο μεθόδων έχουν μόνο ελάχιστη απόκλιση [65]. Ο J. R. Quinlan [66] παρουσιάζει τα δέντρα αποφάσεων να είναι πολύ γρηγορότερα από τα νευρωνικά δίκτυα. Αντίθετα, τα νευρωνικά δίκτυα λαμβάνουν υπόψη τους κάποιες στατιστικές πληροφορίες, κάτι που δεν συμβαίνει με τα δέντρα αποφάσεων [67]. Παρ' όλα αυτά έχει επίσης διατυπωθεί ότι τα δέντρα αποφάσεων δεν μπορούν να λάβουν υπόψη τους κάποιες στατιστικές πληροφορίες, οι οποίες είναι διαθέσιμες στα νευρωνικά δίκτυα [67]. Επίσης, οι perceptrons πολλαπλών στρώσεων (multi-layer) έχουν συγκριθεί με

τον αλγόριθμο CART με ή χωρίς γραμμικούς συνδυασμούς και έχει αναφερθεί ότι δεν υπάρχει μεγάλη διαφορά στην ακρίβεια (accuracy) [68]. Ο αλγόριθμος CART έχει επίσης συγκριθεί με πολλαπλή γραμμική παλινδρόμηση και ανάλυση διακριτής ταξινόμησης και προαναφέρθηκε ότι είναι πιο αποτελεσματικός σε περιπτώσεις δεδομένων με πολύ «θόρυβο», καθώς και σε περιπτώσεις δεδομένων με πολλές ελλιπείς τιμές [69]. Ο αλγόριθμος C4.5 έχει συγκριθεί με τη λογιστική παλινδρόμηση [70].

Γενικότερα, έχει αναφερθεί ότι τα βασικά πλεονεκτήματα των ταξινομητών δέντρων αποφάσεων είναι τα ακόλουθα [24]:

i. Η γνώση που αποκτάται από προταξινομημένα παραδείγματα ξεπερνάει το «μποτιλιάρισμα» της απόκτησης γνώσης από έναν domain expert.
ii. Οι μέθοδοι εξαγωγής δέντρων αποφάσεων είναι διερευνητικές και όχι επαγωγικές. Επίσης είναι μη-παραμετρικές. Αρκούν μερικές παραδοχές σχετικά με το μοντέλο και την κατανομή των δεδομένων, ώστε τα δέντρα να μοντελοποιήσουν ένα μεγάλο εύρος κατανομής δεδομένων.
iii. Η ιεραρχική διάκριση που παράγουν τα δέντρα αποφάσεων προσφέρει καλύτερη και αποτελεσματικότερη χρήση των διαθέσιμων χαρακτηριστικών (features).
iv. Τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν τόσο σε προβλήματα με διακριτές τιμές, όσο και σε ημιτελή προβλήματα (σε διακριτά προβλήματα η εξαρτημένη μεταβλητή μπορεί να προβλεφθεί απόλυτα από τις ανεξάρτητες μεταβλητές, κάτι το οποίο δεν ισχύει σε ημιτελή προβλήματα).
v. Τα δέντρα αποφάσεων ταξινομούν τα δεδομένα με βάση μία σειρά εύκολων και κατανοητών κριτηρίων.
vi. Σε σχέση με κάποιες στατιστικές μεθόδους, οι ταξινομητές δέντρων αποφάσεων χειρίζονται τα unimodal και multimodal δεδομένα με τον ίδιο τρόπο.

Οι ταξινομητές δέντρων αποφάσεων έχουν επίσης διάφορα μειονεκτήματα, τα οποία συνοψίζονται ως εξής [25]:

i. Κάποια λάθη μπορεί να συγκεντρώνονται από επίπεδο σε επίπεδο σε ένα μεγάλο δέντρο. Είναι λοιπόν εξαιρετικά δύσκολο να μεγιστοποιηθεί τόσο η ακρίβεια (accuracy), όσο και η αποτελεσματικότητα (efficiency) [69]. Έτσι, για κάποια δεδομένη ακρίβεια, πρέπει να ικανοποιείται ένα ποσοστό αποτελεσματικότητας.

ii. Οι επικαλύψεις (overlap) (όταν δηλαδή δύο ενδιάμεσοι κόμβοι έχουν τουλάχιστον μία κοινή κλάση) μπορεί να προκαλέσουν μεγάλο αριθμό τερματικών σε σχέση με τον πραγματικό αριθμό των κλάσεων, κάτι το οποίο αυξάνει το χρόνο κατασκευής και το χώρο μνήμης που χρειάζεται για την επεξεργασία.

iii. Επίσης, μπορεί να υπάρχουν δυσκολίες σχετικά με το σχεδιασμό ενός ταξινομητή δέντρου αποφάσεων.

Κεφάλαιο 3: Αλγόριθμοι συσχέτισης

3.1 Γενικά

Οι αλγόριθμοι εξαγωγής κανόνων συσχέτισης (association rules) είναι μία από τις σημαντικότερες τεχνικές εξόρυξης από δεδομένα. Είναι μια σύγχρονη μέθοδος καθότι εμφανίστηκε μόλις το 1993 [72] και αναφερόταν στην εξαγωγή συσχετίσεων στα πεδία βάσεων δεδομένων. Από τότε έχει διεξαχθεί ενδελεχής έρευνα και έχει αποδειχθεί η εφαρμογή της σε πολλούς τομείς. Οι πληροφορίες που μπορούν να περιγραφούν και συγκεντρωθούν από τους κανόνες συσχέτισης είναι ιδιαίτερα σημαντικές και αφορούν πολλαπλές εφαρμογές. Το πιο χαρακτηριστικό παράδειγμα εφαρμογής των κανόνων συσχέτισης είναι η ανάλυση του «καλαθιού της νοικοκυράς» (market-basket analysis), όπου μια συναλλαγή, δηλαδή η αγορά των προϊόντων (για παράδειγμα το περιεχόμενο ενός καλαθιού υπεραγοράς) αντιμετωπίζεται σαν μία μεμονωμένη συναλλαγή. Αναλύεται ένας αριθμός τέτοιων συναλλαγών, ώστε να εξαχθούν πρότυπα τα οποία θα αναδείξουν τις αγοραστικές τάσεις των πελατών. Το κατάστημα μπορεί να χρησιμοποιεί τέτοιες πηγές πληροφοριών για διάφορους σκοπούς όπως την προώθηση των προϊόντων, την τοποθέτηση των προϊόντων στα ράφια ενός καταστήματος και τη διαχείριση των αποθεμάτων.

Οι κανόνες συσχέτισης παρουσιάζονται με τη μορφή $X \rightarrow Y$, όπου τα X και Y είναι οι τιμές των πεδίων που παρουσιάζονται μέσα στους κανόνες. Ο κανόνας $X \rightarrow Y$ δείχνει ότι οι τιμές αυτές παρουσιάζονται μαζί μέσα στις εγγραφές. Με βάση τους κανόνες αυτούς μετά από διαλογή γίνεται η συσχέτιση των πεδίων και ο ορισμός των προτύπων. Κάθε κανόνας έχει δύο μετρικές, την υποστήριξη (support) και την εμπιστοσύνη (confidence). Αυτές οι μετρικές καθορίζουν το ποσοστό εφαρμογής του κανόνα στο σύνολο των εγγραφών. Η υποστήριξη μετράει ουσιαστικά την ισχύ του κανόνα, δηλαδή είναι το ποσοστό των συναλλαγών που περιέχουν το Y επί του αριθμού των συναλλαγών που περιέχουν το X. Η εμπιστοσύνη είναι το ποσοστό εμφάνισης του X και του Y μαζί στο σύνολο της βάσης δεδομένων, δηλαδή πόσο

συχνά συμβαίνει το πρότυπο αυτό στη βάση δεδομένων. Όσο πιο μεγάλοι είναι αυτοί οι αριθμοί, τόσο πιο «δυνατός» είναι ο κανόνας. Ο χρήστης πρέπει να καθορίσει την ελάχιστη εμπιστοσύνη και υποστήριξη που επιθυμεί να έχει ο κανόνας. Εδώ πρέπει να σημειωθεί ότι δεν υπάρχει κάποιος προκαθορισμένος αριθμός που πρέπει να χρησιμοποιηθεί από τον χρήστη. Ο ίδιος, ανάλογα με το πρόβλημα που μελετά, τα δεδομένα που διαθέτει και το τι επιθυμεί να αναδείξει, θέτει την ελάχιστη εμπιστοσύνη και υποστήριξη που κρίνει ορθή. Σίγουρα θεωρείται πολύ σημαντική η εμπειρία που έχει ο χρήστης, ώστε να γίνει η σωστή επιλογή του κατώτατου ορίου [73]. Η πολυπλοκότητα των αλγορίθμων και η δυσκολία επιλογής των χρήσιμων κανόνων, από το σύνολο των κανόνων που προκύπτουν, είναι βασικά προβλήματα που αφορούν την εύρεση κανόνων συσχέτισης. Το πρώτο πρόβλημα αφορά τον αριθμό των κανόνων, ο οποίος αυξάνεται εκθετικά με τον αριθμό των πεδίων. Οι πιο πρόσφατοι αλγόριθμοι που εξάγουν κανόνες συσχέτισης, μπορούν να μειώσουν αποτελεσματικά τον αριθμό αυτό, με τον καθορισμό ενός κατώτατου ορίου στην εμπιστοσύνη και την υποστήριξη, που αφορά τη μέτρηση της ποιότητας των κανόνων. Το δεύτερο πρόβλημα αφορά τους χρήσιμους κανόνες που συνήθως προκύπτουν και αποτελούν μόνο ένα μικρό ποσοστό του συνόλου των κανόνων. Το πρόβλημα αυτό ερευνάται σε σχέση με την υποστήριξη προς το χρήστη, όταν αναζητά κανόνες μέσα στους κανόνες που έχουν εξαχθεί, καθώς και με την ανάπτυξη επιπλέον μέτρων ποιότητας στους κανόνες.

3.2 Παρουσίαση προβλήματος συσχέτισης

Έστω $X= \{x_1, \ldots\ldots, x_n\}$ ένα σύνολο από προϊόντα. Έστω D ένα σύνολο από αγορές, και κάθε αγορά T είναι μία λίστα από προϊόντα, όπου T είναι ένα υποσύνολο του X. Κάθε υποσύνολο από προϊόντα $X \cup X$ ονομάζεται λίστα προϊόντων (itemset). Ο κανόνας συσχέτισης είναι της μορφής $X \rightarrow Y$, όπου $X \subset X$, $Y \subset X$ και $X \cap Y = 0$. Ο κανόνας $X \rightarrow Y$, σε ένα σύνολο από αγορές D, ισχύει με υποστήριξη c, αν το c% των αγορών που ανήκουν στο D και περιέχουν το X, περιέχουν επίσης και το Y. Ο κανόνας $X \rightarrow Y$ έχει εμπιστοσύνη s, αν το s% των αγορών

που ανήκουν στο D περιέχουν το XUY [8]. Η υποστήριξη προκύπτει από την εμπιστοσύνη: $c(X \rightarrow Y) = s(XUY)/ s(X)$. Ένα βασικό πρόβλημα στην εξαγωγή κανόνων συσχέτισης είναι ο μεγάλος αριθμός κανόνων που εξάγονται. Για την ακρίβεια, ο αριθμός των κανόνων αυξάνεται εκθετικά, καθώς το |X| αυξάνεται. Προκειμένου να μειωθεί ο αριθμός των κανόνων που εξάγονται, εισάγεται ο περιορισμός της ελάχιστης υποστήριξης και της ελάχιστης εμπιστοσύνης (minimum support και minimum confidence). Έτσι, το πρόβλημα εύρεσης κανόνων συσχέτισης μπορεί να διαιρεθεί στα δύο υπό-προβλήματα [8]:

i. Εύρεση όλων των συνδυασμών των προϊόντων που έχουν εμπιστοσύνη πάνω από την ελάχιστη εμπιστοσύνη (η οποία καθορίζεται από το χρήστη). Όλοι αυτοί οι συνδυασμοί ονομάζονται μεγάλες λίστες από προϊόντα (large itemsets) και όλοι οι υπόλοιποι συνδυασμοί μικρές λίστες από προϊόντα (small itemsets).
ii. Εξόρυξη κανόνων συσχέτισης, βασισμένη στις μεγάλες λίστες από προϊόντα, οι οποίες έχουν μεγαλύτερη υποστήριξη από αυτή που καθορίζει ο χρήστης. Η γνώση για τις τιμές της εμπιστοσύνης όλων των υπολιστών του X διασφαλίζεται μέσω της ιδιότητας της εμπιστοσύνης των προϊόντων: όλα τα υποσύνολα μίας μεγάλης λίστας από προϊόντα πρέπει να είναι και αυτά μεγάλες λίστες από προϊόντα [8]. Έτσι, το πρόβλημα της εξόρυξης κανόνων συσχέτισης εντοπίζεται στην εύρεση των μεγάλων λιστών από προϊόντα, δεδομένης μίας ελάχιστης εμπιστοσύνης.

3.3 Αλγόριθμοι συσχέτισης

Οι αλγόριθμοι συσχέτισης ανάλογα με τους κανόνες που εξάγουν μπορούν να χωριστούν σε κανονικούς-Boolean, χωρικούς, γενικευμένους, ποιοτικούς κ.α. Επίσης χωρίζονται σε διαδοχικούς και παράλληλους αλγόριθμους ανάλογα με την αρχιτεκτονική του επεξεργαστή που χρησιμοποιείται. Λαμβάνομε ακόμη υπ' όψη τον τρόπο δημιουργίας υποψήφιων λιστών, δηλαδή: i) μία δυναμική στρατηγική δημιουργεί τις λίστες κατά τη διάρκεια της σάρωσης των

δεδομένων ενώ ii) μία υβριδική τεχνική δημιουργεί κάποιες υποψήφιες λίστες πριν και κάποιες κατά τη διάρκεια. Πιο κάτω περιγράφονται οι βασικότεροι τύποι αλγορίθμων.

3.3.1 Διαδοχικοί αλγόριθμοι

Οι διαδοχικοί αλγόριθμοι έχουν ταξινομημένες λίστες κατά όνομα προϊόντος. Σχεδιάζονται να λειτουργούν με διαδοχικούς μηχανισμούς σε έναν κεντρικό επεξεργαστή. Υπάρχουν πολλοί αλγόριθμοι με αυτή τη φιλοσοφία, με αρχικό τον AIS [72]. Ο αλγόριθμος αυτός στόχευε στην εξαγωγή ποιοτικών κανόνων, δημιουργώντας και μετρώντας λίστες κατά τη διάρκεια της σάρωσης της βάσης δεδομένων. Στο διάβασμα της εγγραφής καθορίζεται ποια από τα προϊόντα που είχαν βρεθεί ως μεγάλες λίστες στην προηγούμενη επανάληψη περιλαμβάνονται στη εγγραφή. Έτσι δημιουργούνται νέες υποψήφιες λίστες από προϊόντα με την επέκταση αυτών των μεγάλων λιστών με άλλα προϊόντα της εγγραφής. Για υπολογισμούς μεγάλων λιστών υπάρχει ο αλγόριθμος SETM [74], ο οποίος έχει υλοποιηθεί άμεσα στην SQL. Όπως και ο AIS, έτσι και ο αλγόριθμος αυτός δημιουργεί υποψήφιες λίστες βασισμένος σε εγγραφές που διαβάζει από τη βάση δεδομένων. Για να χρησιμοποιήσει την SQL για δημιουργία υποψήφιων λιστών, ο αλγόριθμος διαχωρίζει τις υποψήφιες λίστες από το μέτρημα (counting). Το πρόβλημα των πολλών λιστών που δημιουργούνται έρχεται να λύσει ο αλγόριθμος Apriori [8]. Η εύρεση των μεγάλων λιστών, όπως αναφέρθηκε και στην προηγούμενη ενότητα, βασίζεται στο ότι μία λίστα από προϊόντα είναι μεγάλη, αν κάθε υποσύνολό της είναι μεγάλη λίστα από δεδομένα. Η εύρεση αυτή γίνεται κατόπιν πολλών επαναλήψεων στη βάση δεδομένων. Κατά την πρώτη επανάληψη, υπολογίζεται η εμπιστοσύνη κάθε προϊόντος και επίσης ποια από αυτά είναι μεγάλες λίστες. Σε κάθε επόμενη επανάληψη, λαμβάνονται υπόψη μόνο οι μεγάλες λίστες από προϊόντα που είχαν βρεθεί στην προηγούμενη επανάληψη, χωρίς να λαμβάνονται υπόψη οι εγγραφές. Από τις νέες λίστες δημιουργούνται νέες υποψήφιες μεγάλες λίστες. Κατόπιν μετράται η εμπιστοσύνη των λιστών αυτών και καθορίζεται ποιες από αυτές είναι τελικά μεγάλες λίστες. Ο αλγόριθμος ξεκινάει

πάλι, λαμβάνοντας υπόψη τις μεγάλες λίστες που καθορίστηκαν στην προηγούμενη επανάληψη. Αναλυτικά, τα βήματα του αλγορίθμου Apriori έχουν ως εξής [8]:

i. Εύρεση των προϊόντων που έχουν εμπιστοσύνη μεγαλύτερη από την ελάχιστη εμπιστοσύνη, δηλαδή το σύνολο L_1= μεγάλες λίστες από ένα προϊόν.
ii. Από k=2 και όσο το L_{k-1} δεν είναι κενό:
iii. α) εύρεση του συνόλου C_k των υποψηφίων μεγάλων λιστών από k προϊόντα με βάση το L_{k-1}

    β) εύρεση της εμπιστοσύνης των υποψήφιων μεγάλων λιστών και δημιουργία συνόλου L_k = μεγάλες λίστες από k προϊόντα.
iv. Για κάθε στοιχείο των $L_1 \ldots L_n$ εύρεση εκείνων που έχουν υποστήριξη μεγαλύτερη από την ελάχιστη υποστήριξη. Στο πρώτο βήμα ο αλγόριθμος μετράει την εμπιστοσύνη του κάθε προϊόντος ξεχωριστά, ώστε να σχηματιστούν οι μεγάλες λίστες μεγέθους ενός προϊόντος. Στο δεύτερο βήμα (που αποτελείται από δύο υποβήματα) μεγάλες λίστες από k-1 προϊόντα που βρέθηκαν στην προηγούμενη επανάληψη, χρησιμοποιούνται για να δημιουργηθούν οι υποψήφιες μεγάλες λίστες από k προϊόντα (C_k). Κατόπιν, υπολογίζεται η εμπιστοσύνη των υποψηφίων μεγάλων λιστών από k προϊόντα. Το βήμα αυτό τερματίζεται, όταν δεν υπάρχουν υποψήφιες μεγάλες λίστες. Τέλος, στο τρίτο βήμα, υπολογίζεται η υποστήριξη κάθε μεγάλης λίστας προϊόντων και εξάγονται κανόνες, από τους οποίους γίνονται αποδεκτοί εκείνοι που έχουν υποστήριξη μεγαλύτερη από την ελάχιστη υποστήριξη.

Ο αλγόριθμος Apriori δημιουργεί λίγες υποψήφιες λίστες προϊόντων και οι περισσότεροι κανόνες συσχέτισης βασίζονται σε αυτό τον αλγόριθμο. Ο αλγόριθμος AprioriTID είναι μία παραλλαγή του βασικού αλγορίθμου Apriori [8]. Σε αυτόν τον αλγόριθμο η βάση δεδομένων χρησιμοποιείται στην αρχή. Μετά την πρώτη επανάληψη δεν χρησιμοποιείται για υπολογισμό της εμπιστοσύνης των υποψήφιων μεγάλων λιστών, αλλά γίνεται χρήση μίας κωδικοποίησης των υποψηφίων μεγάλων λιστών που είχε χρησιμοποιηθεί στην προηγούμενη επανάληψη. Σε

επόμενες επαναλήψεις το μέγεθος της κωδικοποίησης αυτής μπορεί να γίνει πολύ μικρότερο από τον αριθμό των συναλλαγών στη βάση δεδομένων.

Μία επέκταση των Apriori και AprioriTID είναι ο αλγόριθμος AprioriHybrid [8], που συνδυάζει τα καλύτερα χαρακτηριστικά και των δύο αλγορίθμων. Συγκεκριμένα, ενώ ο Apriori στις αρχικές επαναλήψεις δίνει γρηγορότερα αποτελέσματα, ο AprioriTID δίνει γρηγορότερα αποτελέσματα σε μεταγενέστερες επαναλήψεις. Έτσι, ο AprioriHybrid χρησιμοποιεί τον Apriori σε αρχικά στάδια και στη συνέχεια χρησιμοποιεί τον AprioriTID. Η αλλαγή αυτή από τον έναν αλγόριθμο στον άλλον, φυσικά, συμπεριλαμβάνει ένα κόστος.

Ο αλγόριθμος PARTITION [75] χωρίζει τη βάση δεδομένων σε μικρά κομμάτια, έτσι ώστε να μπορούν να αναλύονται ξεχωριστά και αποτελεσματικά, προκειμένου να βρεθούν οι μεγάλες λίστες. Οι μεγάλες αυτές λίστες στη συνέχεια συνδυάζονται, ώστε να δημιουργηθούν οι υποψήφιες μεγάλες λίστες. Όμως χρειάζεται ακόμη ένα επιπλέον σάρωμα της βάσης για να επιβεβαιωθεί ότι οι τοπικές μεγάλες λίστες από προϊόντα είναι επίσης και ολικές. Ο αλγόριθμος OCD (Off-line Candidate Determination) [76] βασίζεται στην ιδέα ότι μικρά δείγματα συνήθως επιτυγχάνουν να βρουν μεγάλες λίστες με μεγαλύτερη ευκολία. Ακολουθεί διαφορετική τεχνική από τον Apriori για τον καθορισμό των υποψήφιων λιστών. Χρησιμοποιεί όλη τη διαθέσιμη πληροφορία από προηγούμενες επαναλήψεις, προκειμένου να χωρίσει τις υποψήφιες λίστες μεταξύ των επαναλήψεων, κρατώντας την επανάληψη όσο πιο απλή γίνεται. Με αυτό τον τρόπο απορρίπτει και υποψήφιες μικρές λίστες. Ο αλγόριθμος Sampling [77] μειώνει τις επαναλήψεις σε μία ή δύο. Ένα δείγμα επιλέγεται αρχικά από τη βάση δεδομένων, το οποίο μπορεί να χωρέσει στη βασική μνήμη. Στη συνέχεια, οι μεγάλες λίστες βρίσκονται από το δείγμα. Ο αλγόριθμος DHP (Direct Hashing and Pruning) [78] είναι ιδιαίτερα αποτελεσματικός για την εύρεση υποψήφιων λιστών 2 προϊόντων. Τέλος, ο αλγόριθμος CARMA (Continuous Association Rule Mining Algorithm) [79] εμφανίζει online τους κανόνες συσχέτισης και επιτρέπει στο χρήστη να αλλάξει την ελάχιστη εμπιστοσύνη και υποστήριξη σε κάθε συναλλαγή κατά τη διάρκεια της πρώτης επανάληψης.

Ένας από τα πιο γρήγορους και πιο δημοφιλής αλγορίθμους για εξόρυξη συχνών αντικειμένων είναι ο αλγόριθμος FP-growth [14]. Είναι βασισμένος σε ένα πρόθεμα

αναπαράστασης δέντρου της βάσης δεδομένων (που ονομάζεται FP-tree), το οποίο μπορεί να σώσει μεγάλα ποσά μνήμης για την αποθήκευση των συναλλαγών. Ο βασική ιδέα του αλγορίθμου FP-growth μπορεί να περιγραφεί ως ένα επαναληπτικό σύστημα εξουδετέρωσης: σε ένα στάδιο προεπεξεργασίας διαγράφει όλα τα στοιχεία από τις συναλλαγές που δεν είναι συχνά σε ατομική βάση, δηλαδή, δεν εμφανίζονται με βάση ενός ελαχίστου αριθμού συναλλαγών που καθορίζει ο χρήστης. Στη συνέχεια, επιλέγει όλες εκείνες τις συναλλαγές που περιέχουν το λιγότερο συχνά αντικείμενο (λιγότερο συχνά μεταξύ των συχνών) και διαγράφει αυτό το στοιχείο από αυτές. Επιστρέφει για να επεξεργαστεί την αποκτηθείσα μειωμένη βάση δεδομένων . Στην επιστροφή, αφαιρεί το επεξεργασμένο στοιχείο επίσης από τη βάση δεδομένων όλων των συναλλαγών και αρχίζει πέρα από την αρχή, δηλαδή, επεξεργάζεται το δεύτερο συχνό στοιχείο κ.λ.π. Σε αυτά τα βήματα επεξεργασίας το δέντρο προθέματος, που ενισχύεται από τις συνδέσεις μεταξύ των κλάδων, αξιοποιείται για να βρει γρήγορα στις συναλλαγές ένα δεδομένο στοιχείο και επίσης για να αφαιρέσει αυτό το στοιχείο από τις συναλλαγές αφότου έχει υποβληθεί σε επεξεργασία.

3.3.2 Παράλληλοι αλγόριθμοι

Αλγόριθμοι που επικεντρώνονται στο πώς θα γίνει παράλληλη η λειτουργία της εύρεσης μεγάλων λιστών από προϊόντα λέγονται παράλληλοι. Σχεδιάζονται για να λειτουργούν με παράλληλο τρόπο σε πολυεπεξεργαστές. Για την παράλληλη εύρεση κανόνων, θα πρέπει να γίνουν κάποιες αλλαγές-αντικαταστάσεις, σχετικά με την χρησιμοποίηση της διαθέσιμης μνήμης, τους υπολογισμούς, την επικοινωνία, την πληροφορία που παρέχει το συγκεκριμένο πρόβλημα κ.α. Οι παράλληλοι αλγόριθμοι είναι βασισμένοι στο σειριακό αλγόριθμο Apriori. Τρεις γνωστοί αλγόριθμοι που ανήκουν σε αυτή την κατηγορία έχουν προταθεί από τους Agrawal *et al*. [80]. Ο αλγόριθμος CM (Count Distribution algorithm) επικεντρώνεται στην ελαχιστοποίηση της επικοινωνίας και χρησιμοποιεί μία βασική αρχή επιτρέποντας «μειωμένους παράλληλους υπολογισμούς σε μη χρησιμοποιούμενους επεξεργαστές,

προκειμένου να αποφευχθεί η επικοινωνία» [80]. Με τον τρόπο αυτό, κάθε επεξεργαστής υπολογίζει ανεξάρτητα μεγάλες λίστες από προϊόντα και στη συνέχεια ανταλλάσσει την πληροφορία αυτή με τους υπόλοιπους επεξεργαστές, προκειμένου να εξαχθούν κανόνες σφαιρικά.

Ο αλγόριθμος DD (Data Distribution) επιχειρεί να αξιοποιήσει τη συνολική μνήμη του συστήματος πιο αποτελεσματικά [68]. Κάθε επεξεργαστής υπολογίζει αμοιβαίως αποκλειόμενες υποψήφιες λίστες. Καθώς ο αριθμός των επεξεργαστών αυξάνεται, ένας μεγάλος αριθμός υποψήφιων λιστών μπορεί να μετρηθεί σε κάθε επανάληψη. Τέλος, ο αλγόριθμος Candidate Distribution χωρίζει τόσο τα δεδομένα, όσο και τις υποψήφιες λίστες, με τέτοιο τρόπο, ώστε κάθε επεξεργαστής να μπορεί να δουλεύει ανεξάρτητα. Ένας άλλος αλγόριθμος που ανήκει σε αυτή την κατηγορία είναι ο PDM (Parallel Data Mining) [81], ο οποίος προσπαθεί να παραλληλίσει τον DHP [78]. Ο αλγόριθμος DMA (Distributed Mining Algorithm) [82] εξάγει κανόνες από κατανεμημένες βάσεις δεδομένων. Ο αλγόριθμος δημιουργεί έναν μικρό αριθμό υποψήφιων λιστών, βασισμένος σε τεχνικές περικοπής (pruning) και τεχνικές μείωσης επικοινωνίας μηνυμάτων (communication message reduction), και τελικά χρειάζεται μόνο O(n) μέγεθος μηνυμάτων για κάθε υποψήφια λίστα. Ο αλγόριθμος IDD (Intelligent Data Distribution) [83] αποτελεί μία βελτίωση του αλγορίθμου DD. Ο αλγόριθμος χρησιμοποιεί αποτελεσματικά τη μνήμη των παράλληλων υπολογιστών, υιοθετώντας ένα σχήμα διαχωρισμού υποψήφιων λιστών και χρησιμοποιεί αποτελεσματικούς μηχανισμούς επικοινωνίας για τη μεταφορά των δεδομένων μεταξύ των επεξεργαστών. Επίσης, ο αλγόριθμος HD (Hybrid Distribution) [82] χωρίζει δυναμικά τις υποψήφιες λίστες, προκειμένου να διατηρήσει καλή ισορροπία κατά τη μεταφορά των δεδομένων. Τέλος, ο ίδιος μπορεί να μετατραπεί στον αλγόριθμο CD σε μεταγενέστερες επαναλήψεις. Οι Shintani *et al*. [84] προτείνουν έναν αλγόριθμο για εξαγωγή κανόνων με ιεραρχική ταξινόμηση. Η διαθέσιμη μνήμη χρησιμοποιείται πλήρως, ανακαλύπτοντας υποψήφιες λίστες, οι οποίες μπορούν να αναλύονται τοπικά χωρίς επικοινωνία μεταξύ των επεξεργαστών. Τέλος, ο αλγόριθμος SH [85] δημιουργεί υποψήφιες λίστες ανεξάρτητα σε κάθε επεξεργαστή, καθώς σαρώνει τη βάση

δεδομένων και όχι εκ των προτέρων από τις προηγούμενες μεγάλες λίστες, όπως ο σειριακός αλγόριθμος Apriori.

3.3.3 Ποσοτικοί και γενικευμένοι αλγόριθμοι

Οι αρχικοί αλγόριθμοι για εξαγωγή κανόνων συσχέτισης χρησιμοποιούσαν κατηγορικά δεδομένα. Αν τα δεδομένα ήταν μεικτά, δηλαδή κατηγορικά και ποσοτικά χρησιμοποιούνταν ποσοτικοί αλγόριθμοι συσχέτισης που χώριζαν τις ποσότητες σε διαστήματα. Αν ένα ποσοτικό πρόβλημα μπορεί να μετατραπεί σε πρόβλημα με Boolean κανόνες, τότε κάθε αλγόριθμος εύρεσης κανονικών (Boolean) κανόνων συσχέτισης, μπορεί να χρησιμοποιηθεί για την εύρεση ποσοτικών κανόνων συσχέτισης [80], [86]. Οι τιμές των κατηγορικών μεταβλητών καταγράφονται από μία σειρά διαδοχικών ακέραιων αριθμών. Οι ποσοτικές μεταβλητές αν δεν χωριστούν σε διαστήματα, παραμένουν σαν διαδοχικοί ακέραιοι. Αν τα διαστήματα είναι πολύ μικρά, μερικοί κανόνες μπορεί να μην έχουν ελάχιστη εμπιστοσύνη. Από την άλλη μεριά, εάν τα διαστήματα είναι πολύ μεγάλα, τότε μερικοί κανόνες μπορεί να μην έχουν ελάχιστη υποστήριξη. Το πρόβλημα αυτό μπορεί να λυθεί με δύο τρόπους: α. Συνδυασμός κοντινών διαστημάτων / τιμών, και β. Εισαγωγή της «μέγιστης εμπιστοσύνης». Τα διαστήματα σταματούν να συνδυάζονται αν η συνδυασμένη επιβεβαίωσή τους ξεπερνάει την τιμή αυτή [80].

Οι γενικευμένοι αλγόριθμοι χρησιμοποιούν δεδομένα με ιεραρχική δομή για τη δημιουργία κανόνων συσχέτισης σε διαφορετικά επίπεδα της ιεραρχίας [87]. Οι κανόνες που δημιουργούνται για προϊόντα σε υψηλότερο επίπεδο στην ιεραρχία, παρέχουν αύξηση στην εμπιστοσύνη και την υποστήριξη. Ένας γενικευμένος κανόνας συσχέτισης είναι όπως ο κανονικός κανόνας συσχέτισης και έχει τη μορφή $X \rightarrow Y$, με τη διαφορά ότι κανένα προϊόν του Y μπορεί να είναι σε ψηλότερο επίπεδο της ιεραρχίας από ένα προϊόν του X [87]. Σχετικά με την εύρεση γενικευμένων κανόνων συσχέτισης υπάρχει το πρόβλημα δημιουργίας κανόνων για όλα τα επίπεδα. Έχουν προταθεί διάφοροι παράλληλοι αλγόριθμοι για τη

δημιουργία γενικευμένων κανόνων συσχέτισης [88]. Οι κανόνες συσχέτισης σε διαφορετικά επίπεδα στην ιεραρχία ονομάζονται κανόνες συσχέτισης πολλαπλών επιπέδων (multiple-level association rules) [89]. Δημιουργούνται από πάνω προς τα κάτω με ένα αλγόριθμο όπως τον Apriori.

3.3.4 Χωρικοί και χρονολογικοί αλγόριθμοι

Οι χωρικοί αλγόριθμοι περιλαμβάνουν τοπικές πληροφορίες για την αποθήκευση των δεδομένων. Οι πληροφορίες αυτές είναι γεωγραφικά δεδομένα όπως γεωγραφικό μήκος-γεωγραφικό πλάτος, ταχυδρομικοί κώδικες, διευθύνσεις και άλλα. Σε αυτή τη μέθοδο εξάγονται κανόνες συσχέτισης με χωρικές λειτουργίες (για παράδειγμα «κοντά», «δίπλα σε» κλπ.). Ένας χωρικός κανόνας είναι της μορφής $X \rightarrow Y$, όπου ή το X ή το Y ή και τα δύο μπορεί να είναι χωρική μεταβλητή [90]. Οι χωρικοί αλγόριθμοι συσχέτισης μπορούν να φανούν ιδιαίτερα χρήσιμοι, όταν συνδυάζονται με Γεωγραφικά Συστήματα Πληροφοριών (ΓΣΠ).

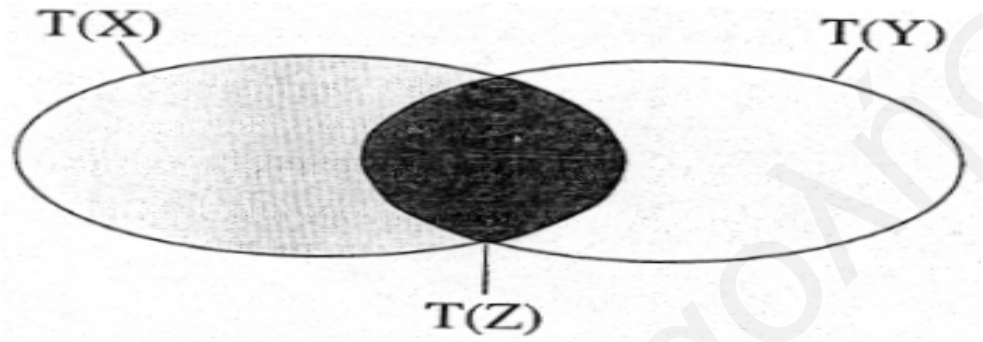
Οι χρονολογικοί κανόνες συσχέτισης είναι παρόμοιοι με τους χωρικούς, με τη διαφορά ότι συμπεριλαμβάνουν και τον παράγοντα χρόνο. Οι χωρικοί-χρονικοί κανόνες συμπεριλαμβάνουν τόσο χρόνο (time), όσο και χώρο (space).

3.4 Ιδιότητες κανόνων συσχέτισης

Η απαίτηση τα X, Y να είναι χωρίς κοινά μέλη δεν είναι απολύτως αναγκαία, καθότι δεν οδηγούμαστε σε κανόνες χωρίς νόημα, αλλά σε περιττούς και ασήμαντους κανόνες. Το $X \rightarrow X$ για παράδειγμα είναι τετριμμένα αληθές και το $X \rightarrow X \cup Y$ είναι ισοδύναμο με το $X \rightarrow Y$ και επομένως όχι ενδιαφέρον. Το ηγούμενο μέρος ενός κανόνα μπορεί να είναι κενό. Έτσι κάθε διεργασία (transaction) θεωρείται ότι υποστηρίζει το κενό σύνολο (itemset) και κατ' επέκταση ολόκληρη η βάση ικανοποιεί τον ηγούμενο όρο. Η υποστήριξη ενός τέτοιου κανόνα

είναι ίση με τη σχετική συχνότητα εμφάνισης του ακόλουθου μέρους. Απαιτούμε το ακόλουθο μέρος Y να μην είναι κενός, για τον ίδιο λόγο που απαιτούμε το ακόλουθο και το ηγούμενο μέρος να μην έχουν κενά στοιχεία.

Στον κανόνα $X \cup Y \rightarrow Z$, T(X): σύνολο διεργασιών του X, T(Y): σύνολο διεργασιών του Y, T(Z): σύνολο διεργασιών του Z



Σχήμα 3.1: παράδειγμα σύνθεσης κανόνων

Ακολουθούν οι εφτά βασικές ιδιότητες για τη σύνθεση των κανόνων συσχέτισης [71]:

Ιδιότητα 1 - Συνθήκη Υποστήριξη (Support) για Υποσύνολα.

Εάν για τα σύνολα A, B ισχύει $A \subseteq B$, τότε supp(A) ≥ supp(B), επειδή όλες οι διεργασίες στο D που περιέχουν το B, αναγκαστικά περιέχουν και το A.

Ιδιότητα 2 - Τα Υπερσύνολα όχι αποδεδειγμένων μεγάλων / συχνών (σπάνια) συνόλων, είναι επίσης σπάνια.

Στην περίπτωση όπου για ένα σύνολο A ισχύει supp(A) < Smin, τότε κανένα υπερσύνολο B του A δεν θα είναι συχνό / μεγάλο, λόγω του ότι ισχύει supp(B) ≤ supp (A) < Smin από την προηγούμενη ιδιότητα.

Ιδιότητα 3 - Τα Υποσύνολα συχνά σύνολα είναι επίσης συχνά.

Εάν ένα σύνολο B είναι συχνό / μεγάλο στο D (δηλαδή supp(B) ≥ Smin), τότε και κάθε υποσύνολο A του B είναι επίσης συχνό / μεγάλο στο D, καθώς supp(A) ≥ supp(B) ≥ Smin σύμφωνα με την 1. Στην πράξη εάν το σύνολο $A=\{\iota_1, \iota_2, \ldots \iota_\kappa\}$

είναι μεγάλο, τότε και όλα τα κ (κ-1) υποσύνολα του θα είναι επίσης μεγάλα. Δεν ισχύει το αντίθετο!

Ιδιότητα 4 - Δεν Επιτρέπεται η Σύνθεση των Κανόνων.

Εάν οι κανόνες $X \rightarrow Z$ και $Y \rightarrow Z$ λαμβάνουν χώρα στο D, αυτό δεν σημαίνει απαραίτητα ότι και ο κανόνας $X \cup Y \rightarrow Z$ είναι αληθής στο D. Ας θεωρήσουμε την περίπτωση όπου $X \cap Y = \varnothing$ και το Z συνεπάγεται (υποστηρίζεται) στο D, εάν και μόνο αν είτε το X, είτε το Y υποστηρίζεται από τις διεργασίες της βάσης. Στην περίπτωση αυτή το σύνολο $X \cup Y$ έχει υποστήριξη 0 και κατ' επέκταση ο κανόνας $X \cup Y \rightarrow Z$ έχει 0% εμπιστοσύνη. Το ίδιο ισχύει και για την σύνθεση κανόνων με το ίδιο ηγούμενο μέρος: $X \rightarrow Y \wedge X \rightarrow Z \not\Rightarrow X \rightarrow Y \cup Z$.

Ιδιότητα 5 - Διαχωρισμός των Κανόνων.

Εάν ισχύει ο κανόνας $X \cup Y \rightarrow Z$, δεν είναι σίγουρο ότι ισχύουν και οι κανόνες $X \rightarrow Z$ και $Y \rightarrow Z$. Το γεγονός αυτό εμφανίζεται στην περίπτωση όπου για παράδειγμα το Z εμφανίζεται σε μια διεργασία, εάν και μόνο αν εμφανίζονται σε αυτό τόσο το X όσο και το Y, δηλαδή εάν $supp(X \cup Y) = supp(Z)$. Εάν οι υποστηρίξεις για το X και το Y είναι πολύ μεγαλύτερες από το $supp(X \cup Y)$, οι δυο κανόνες ($X \rightarrow Z$ και $Y \rightarrow Z$) δεν έχουν την απαιτούμενη εμπιστοσύνη. Η περίπτωση αυτή σχηματικά αποδίδεται στο Σχήμα 3.1. Οι κύκλοι αντιστοιχούν στο σύνολο των διεργασιών που υποστηρίζουν τα αντίστοιχα σύνολα. Εντούτοις το αντίθετο ισχύει, δηλ. $X \rightarrow Y \cup Z \Rightarrow X \rightarrow Y \wedge X \rightarrow Z$, λόγω του ότι $supp(XY) \geq supp(XYZ)$ και $sup(XZ) \geq supp(XYZ)$. Έτσι οι τιμές τόσο της υποστήριξης, όσο και της εμπιστοσύνης μικρότερων κανόνων, αυξάνονται συγκριτικά με τις αντίστοιχες τιμές του αυθεντικού κανόνα. Δυστυχώς αυτό δεν βοηθάει ιδιαίτερα κατά τη διάρκεια κατασκευής των εξαγόμενων κανόνων, διότι εμείς επιθυμούμε τη κατασκευή μεγαλύτερων κανόνων από άλλους μικρότερους και όχι το αντίστροφο.

Ιδιότητα 6 - Δεν ισχύει η Μεταβατικότητα.

Εάν ισχύουν οι κανόνες $X \rightarrow Y$ και $Y \rightarrow Z$ δεν μπορούμε να ισχυριστούμε ότι ισχύει και ο $X \rightarrow Z$. Υποθέτουμε για παράδειγμα ότι ισχύει $T(X) \subset T(Y) \subset T(Z)$ και ότι η ελάχιστη αποδεκτή εμπιστοσύνη είναι Cmin. Ας θεωρήσουμε ότι $conf(X \rightarrow Y)=conf(Y \rightarrow Z)=Cmin$. Βασισμένοι τώρα στις σχετικές τιμές των υποστηρίξεων έχουμε ότι: $conf(X \rightarrow Z) = C2min < Cmin < 1$, το οποίο σημαίνει ότι η εμπιστοσύνη δεν είναι αρκετή για να ισχύσει ο κανόνας.

Ιδιότητα 7 - Συμπέρασμα για το εάν ισχύει ένας Κανόνας.

Η δεύτερη ιδιότητα υποδεικνύει ότι εάν ένας κανόνας της μορφής $A \rightarrow (L-A)$ δεν έχει την ελάχιστη εμπιστοσύνη, τότε ούτε και ο κανόνας $B \rightarrow (L-B)$ πρόκειται να την έχει, για τα σύνολα L, A, B και $B \subseteq A$. Χρησιμοποιώντας τη σχέση $supp(B) \geq supp(A)$ (ιδιότητα 1) και τον ορισμό της εμπιστοσύνης διαπιστώνουμε ότι $conf(B \rightarrow (L-B)) = supp(L) / supp(B) \leq supp(L) / supp(A) < c$. Ανάλογα εάν ισχύει ο κανόνας $(L-C) \rightarrow C$, τότε θα ισχύουν και όλοι οι κανόνες της μορφής $(L-D) \rightarrow D$, όπου $D \subseteq C$ και $D \neq 0$, διότι το ακόλουθο μέρος απαιτείται να είναι μη κενό. Η παρούσα ιδιότητα χρησιμεύει στην επιτάχυνση της διαδικασίας ανακάλυψης των κανόνων, όταν όλα τα συχνά σύνολα και οι υποστηρίξεις τους έχουν καθοριστεί.

Κεφάλαιο 4: Κριτήρια διαχωρισμού δέντρων αποφάσεων, μέτρα κανόνων και αξιολόγηση μοντέλων

4.1 Κριτήρια διαχωρισμού

Υπάρχουν πολλά κριτήρια διαχωρισμού τα οποία μπορούν να χαρακτηριστούν με διάφορους τρόπους ανάλογα με: α) την προέλευση του μέτρου: θεωρία της πληροφορίας, εξάρτηση και απόσταση και β) τη δομή του μέτρου: κριτήριο πρόσμιξης (impurity criterion), κριτήριο ομαλής πρόσμιξης, δυαδικό κριτήριο. Πιο κάτω περιγράφονται τα πιο γνωστά κριτήρια διαχωρισμού που αναφέρονται στη βιβλιογραφία και τα οποία υλοποιήθηκαν σε αυτή την διατριβή:

Information Gain [62]

Το κέρδος πληροφορίας (Information Gain) είναι ένα κριτήριο βασισμένο στην έρευνα του Claude Shannon στη θεωρία της πληροφορίας [91].

$$InfoGain(A) = Info(D) - Info_A(D) \qquad \text{(Εξίσωση 4.1)}$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$
$$p_i = \pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha(\tau\acute{\alpha}\xi\eta\_i\_\sigma\tau o\_\sigma\acute{\upsilon}\nu o\lambda o\_\delta\epsilon\delta o\mu\acute{\epsilon}\nu\omega\nu\_D)$$
$$m => \acute{o}\lambda\epsilon\varsigma\_o\iota\_\tau\iota\mu\acute{\epsilon}\varsigma\_\tau\eta\varsigma\_\tau\acute{\alpha}\xi\eta\varsigma \qquad \text{(Εξίσωση 4.2)}$$

Το κέρδος πληροφορίας είναι πολύ σχετικό με τον υπολογισμό του μέγιστου πιθανού (Maximum Likelihood Estimation, MLE), τη δημοφιλή στατιστική μέθοδο. Χρησιμοποιείται για να κάνει συναγωγές στις παραμέτρους πιθανότητας διασποράς από μια βάση δεδομένων. Είναι ένα κριτήριο βασισμένο στην πρόσμιξη που χρησιμοποιεί την εντροπία σαν μέτρο. Τα περισσότερα μέτρα επιλογής του κατάλληλου χαρακτηριστικού λαμβάνουν υπόψη τους την εντροπία του συνόλου εκπαίδευσης, αλλά αναλόγως με την εφαρμογή για την οποία

κατασκευάστηκαν, επηρεάζονται και από άλλα χαρακτηριστικά του προβλήματος. Πολύ διαδεδομένο είναι το μέτρο επιλογής κέρδος (Gain), το οποίο αφορά τη μείωση της εντροπίας που αποκτάμε εκχωρώντας το συγκεκριμένο χαρακτηριστικό στον κόμβο του δέντρου που αναπτύσσουμε.

Εντροπία για τον κόμβο t :

$$\boldsymbol{Entropy}(\boldsymbol{t}) = -\sum_{j=1}^{c} \boldsymbol{p}(\boldsymbol{j} \mid \boldsymbol{t}) \log \boldsymbol{p}(\boldsymbol{j} \mid \boldsymbol{t}) \qquad \text{(Εξίσωση 4.3)}$$

$\boldsymbol{p}(\boldsymbol{j} \mid \boldsymbol{t})$ σχετική συχνότητα της τάξης j στον κόμβο t,

c αριθμός τάξεων.

Μετράει την ομοιογένεια ενός κόμβου.

Μέγιστη τιμή log(c) όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις τάξεις (που σημαίνει τη λιγότερο ενδιαφέρουσα πληροφορία).

Ελάχιστη τιμή (0.0) όταν όλες οι εγγραφές ανήκουν σε μία τάξη (που σημαίνει την πιο ενδιαφέρουσα πληροφορία)

$$\boldsymbol{Info}_{\boldsymbol{A}}(\boldsymbol{D}) = \sum_{j=1}^{v} \frac{|\boldsymbol{D}_j|}{|\boldsymbol{D}|} \times \boldsymbol{Info}(\boldsymbol{D}_j) \qquad \text{(Εξίσωση 4.4)}$$

$|\boldsymbol{D}_j| = αριθμός\_των\_περιπτώσεων\_με\_τιμή\_χαρακτηριστικού\_\boldsymbol{j}\_στο\_σύνολο\_\boldsymbol{D}$
$|\boldsymbol{D}| = ολικός\_αριθμός\_των\_περιπτώσεων\_στο\_σύνολο\_\boldsymbol{D}$
$\boldsymbol{D}_j = υποσύνολο\_του\_\boldsymbol{D}\_που\_περιέχει\_την\_τιμή\_χαρακτηριστικού\_\boldsymbol{j}$
$\boldsymbol{v}$ => όλες_οι_τιμές_των_χαρακτηριστικών

Παρόλο που το κέρδος πληροφορίας είναι συνήθως ένα καλό κριτήριο για να αποφασιστεί αν ένα χαρακτηριστικό είναι ενδιαφέρον, δεν είναι τέλειο. Το πρόβλημα εμφανίζεται όταν το κριτήριο εφαρμόζεται σε χαρακτηριστικά που έχουν πολλές διαφορετικές τιμές. Όταν έχουμε αυτή την περίπτωση, τότε χρησιμοποιείται το κριτήριο Gain ratio.

Gain Ratio [19]

Τροποποίηση του κέρδους πληροφορίας, αντιμετωπίζει το πρόβλημα μεροληψίας. Υπολογίζει τον αριθμό και το μέγεθος του υποσυνόλου των υποδειγμάτων κάθε κλάδου για την επιλογή χαρακτηριστικού.

$$GainRatio(A) = \frac{InfoGain(A)}{SplitInfo_A(D)} \qquad \text{(Εξίσωση 4.5)}$$

$$\boldsymbol{SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)}$$

$$\boldsymbol{v} => \boldsymbol{όλες\_οι\_τιμές\_των\_χαρακτηριστικών}$$

(Εξίσωση 4.6)

Εναλλακτικά, μπορούμε να λάβουμε υπόψη μας τον αριθμό των κόμβων

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO} \qquad \text{(Εξίσωση 4.7)}$$

Όπου

$$\boldsymbol{SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}} \qquad \text{(Εξίσωση 4.8)}$$

SplitINFO: εντροπία της διάσπασης

- Μεγάλος αριθμός μικρών διασπάσεων (υψηλή εντροπία) τιμωρείται
- Χρησιμοποιείται στο C4.5

Gini Index [14]

Το Gini index είναι κριτήριο βασισμένο στην πρόσμιξη που μετρά τις αποκλίσεις μεταξύ της πιθανότητας διασποράς και τις στοχευόμενες τιμές των χαρακτηριστικών. Χρησιμοποιείται στο αλγόριθμο CART.

$$GiniGain(D) = Gini(D) - \sum_{j=1}^{v} p_j \times Gini(D_j) \qquad \text{(Εξίσωση 4.9)}$$

$$v => όλες\_οι\_τιμές\_των\_χαρακτηριστικών$$

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$
$$m => όλες\_οι\_τιμές\_της\_τάξης \qquad \text{(Εξίσωση 4.10)}$$

Ευρετήριο Gini για τον κόμβο t :

$$GINI(t) = 1 - \sum_{j=1}^{c} [p(j \mid t)]^2 \qquad \text{(Εξίσωση 4.11)}$$

$\boldsymbol{p(j \mid t)}$ σχετική συχνότητα της τάξης j στον κόμβο t (ποσοστό εγγραφών της τάξης j στον κόμβο t), c αριθμός τάξεων.

Ελάχιστη τιμή (0.0) όταν όλες οι εγγραφές ανήκουν σε μία τάξη (που σημαίνει την πιο ενδιαφέρουσα πληροφορία).

Μέγιστη τιμή (1 - 1/c) όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις τάξεις (που σημαίνει τη λιγότερο ενδιαφέρουσα πληροφορία) και εξαρτάται από τον αριθμό των τάξεων.

Το Gini index έχει χρησιμοποιηθεί από πολλούς όπως για παράδειγμα από τους Gelfand *et al*. [45].

Distance Measure [92]

Τα χαρακτηριστικά είναι δυνατόν να είναι σε κοντινή απόσταση με κάποια χαρακτηριστικά και σε μακρινή απόσταση από άλλα χαρακτηριστικά, σύμφωνα με μία απόσταση.

$$DM(A) = \frac{Gini(D)}{-\sum_{j=1}^{v}\sum_{i=1}^{m} p_{ij} \times \log_2(p_{ij})}$$
$$v => όλες\_οι\_τιμές\_των\_χαρακτηριστικών\ v \qquad \text{(Εξίσωση 4.12)}$$
$$m => όλες\_οι\_τιμές\_της\_τάξης$$

Likelihood Ratio Chi-squared statistics [93]

Αυτό το κριτήριο είναι χρήσιμο για τη μέτρηση της στατιστικής σημασίας του κριτηρίου Information Gain.

$$G^2(A, D) = 2 \times \ln(2) \times |D| \times InfoGain(A) \qquad \text{(Εξίσωση 4.13)}$$

DKM Criterion

Το DKM είναι κριτήριο διαχωρισμού βασισμένο στην πρόσμιξη και σχεδιασμένο για χαρακτηριστικά με δυαδικές τάξεις [66] και [92]. Ορίζεται ως

$$\text{DKM(y,S)} = 2.\sqrt{\left(\frac{|\sigma y = c1S|}{|S|}\right).\left(\frac{|\sigma y = c2S|}{|S|}\right)} \qquad \text{(Εξίσωση 4.14)}$$

Θεωρητικά έχει αποδειχθεί ότι αυτό το κριτήριο χρειάζεται μικρότερα δέντρα για την πρόκληση σοβαρού λάθους σε σχέση με άλλα κριτήρια.

4.2 Μέτρα κανόνων

Τα μέτρα ενδιαφέροντος των κανόνων αναφέρονται στους κανόνες που είναι της μορφής Α → Β. 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.

Υποστήριξη (Support) [13]

Η υποστήριξη είναι ένα μέτρο που εφαρμόζεται κυρίως σε κανόνες συσχέτισης. Λαμβάνοντας υπόψη τόσο το αριστερό όσο και το δεξιό μέρος του κανόνα, παίρνουμε το ποσοστό των περιπτώσεων που ικανοποιούν τη σχέση

$$\textbf{Support} = \frac{p(A \cap B)}{N} \qquad \text{(Εξίσωση 4.15)}$$

Ν είναι ο συνολικός αριθμός των περιπτώσεων

Κατά συνέπεια η υποστήριξη δείχνει τη σημαντικότητα του κανόνα αφού αυτό το μέτρο μας λέει πόσο συχνά εμφανίζεται αυτός ο κανόνας μέσα στα δεδομένα.

Εμπιστοσύνη (Confidence) [13]

Η εμπιστοσύνη είναι το μέτρο που δείχνει το ποσοστό των περιπτώσεων που καλύπτονται τόσο από το αριστερό όσο και από το δεξί μέρος του κανόνα.

$$\textbf{Confidence} = \frac{\boldsymbol{p(A \cap B)}}{\boldsymbol{A}} \qquad \text{(Εξίσωση 4.16)}$$

Το διάστημα των πιθανών τιμών της εμπιστοσύνης είναι μεταξύ 0 και 1, όπου όσο πιο κοντά στο 1 είναι το αποτέλεσμα, τόσο πιο σημαντικός δείχνει να είναι ο κανόνας.

Κάλυψη (Coverage) [95]

Η κάλυψη ενός κανόνα μας δείχνει το ποσοστό των περιπτώσεων που έχει ο κανόνας στο αριστερό μέρος.

$$\textbf{Coverage} = \boldsymbol{A} / \boldsymbol{N} \qquad \text{(Εξίσωση 4.17)}$$

Η κάλυψη παίρνει τιμές από το 0 μέχρι το 1 και όσο πλησιάζει η τιμή στο 1 τόσο πιο ενδιαφέροντας αποδεικνύεται να είναι ο κανόνας.

Επικράτηση (Prevalence) [95]

Είναι η πιθανότητα να ισχύει το Β και παίρνει τιμές μεταξύ 0 και 1.

$$\textbf{Prevalence} = \boldsymbol{P(B)} \qquad \text{(Εξίσωση 4.18)}$$

Δύναμη (Leverage) [95]:

Το μέτρο αυτό, το οποίο αποτελεί ένα μέτρο σπουδαιότητας του κανόνα, περιλαμβάνει την εμπιστοσύνη και την κάλυψη του. Είναι το ποσοστό των πρόσθετων περιπτώσεων που

καλύπτονται και από το αριστερό και το δεξιό μέρος του κανόνα, σε σχέση με το αναμενόμενο αν τα δύο μέρη ήταν ανεξάρτητα. Ορίζεται από

$$\textbf{Leverage}= p(A \mid B) - [p(B).p(A)] \qquad \text{(Εξίσωση 4.19)}$$

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.

Lift [95]

Το μέτρο Lift, ορίζει την εμπιστοσύνη διαιρούμενη με το ποσοστό όλων των περιπτώσεων. Είναι ένα μέτρο σπουδαιότητας του κανόνα, αλλά είναι ανεξάρτητη από την κάλυψη.

$$\textbf{Lift}= \frac{p(A \cap B)}{p(A).p(B)} \qquad \text{(Εξίσωση 4.20)}$$

Παίρνει πραγματικές θετικές τιμές.

- Όταν η τιμή τείνει στο 1, τα δύο μέρη είναι ανεξάρτητα και έτσι ο κανόνας δεν παρουσιάζει κάποιο ενδιαφέρον.
- Όταν η τιμή τείνει στο $+\infty$ τότε: α) Αν $B \subseteq A$ ή $A \subseteq B$ τότε ο κανόνας δεν παρουσιάζει κάποιο ενδιαφέρον, β) Το P(B) τείνει στο 0 δείχνει ότι ο κανόνας δεν είναι σημαντικός ή το $P(B \mid A)$ τείνει στο 1 τότε δείχνει ότι ο κανόνας είναι ενδιαφέροντας.
- Όταν το lift=0 σημαίνει ότι $P(B \mid A)=0 \Leftrightarrow P(B \cap A)=0$ και ο κανόνας δεν είναι σημαντικός.

Πεποίθηση (conviction) [96]

$$\textbf{Conviction}= \frac{n - p(b)}{(1 - Confidence\ )} \qquad \text{(Εξίσωση 4.21)}$$

Προτάθηκε από τον Brin [96] και το ν συμβολίζει τον αριθμό των συναλλαγών στη βάση δεδομένων.

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

Ιδιαιτερότητα (Peculiarity) [95]

Είναι ένα μέτρο που βασίζεται στην απόσταση των κανόνων. Χρησιμοποιείται για να καθορίσει το βαθμό που ένα αντικείμενο δεδομένων διαφέρει από άλλα παρόμοια αντικείμενα δεδομένων. Ο ορισμός της συνάρτησης είναι:

$$\mathbf{PF(x_i)}= \sum_{j=1}^{n} \sqrt{N(x_i, x_j)} \qquad \text{(Εξίσωση 4.22)}$$

Τα x_i και x_j είναι τιμές ιδιοτήτων, το n είναι ο αριθμός διαφορετικών τιμών ιδιοτήτων και $N(x_i, x_j)$ είναι η εννοιολογική απόσταση μεταξύ x_i και x_j. Η εννοιολογική διαφορά δίνεται από:

$\mathbf{N(x_i,x_j) = \left| x_i,x_j \right|}$

Προστιθέμενη Αξία (Added Value) [95]

Ορίζει την διαφορά μεταξύ της τελικής απάντησης με την άμεση και έμμεση απάντηση. Είναι η διαφορά του confidence με την πιθανότητα να ισχύει το B.

$$\boldsymbol{AddedValue} = \boldsymbol{P(B / A) - P(B)} \qquad \text{(Εξίσωση 4.23)}$$

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

4.3 Αξιολόγηση μοντέλου και μέτρα αξιοπιστίας

Σημαντικό κριτήριο σε ένα μοντέλο δεδομένων εκπαίδευσης (training data), είναι η ακρίβεια. Αυτό το κριτήριο είναι μια ένδειξη για τα μελλοντικά δεδομένα όπου το μοντέλο δεν έχει εκπαιδευτεί. Όσο μεγαλύτερη είναι η ακρίβεια, τόσο πιο επιτυχημένο είναι το μοντέλο. Γνωστές τεχνικές αξιολόγησης των μοντέλων είναι:

i. **Μέθοδος Hold-out [13]:** Διαχωρισμός του συνόλου των δεδομένων σε δύο τυχαία υποσύνολα, το σύνολο εκπαίδευσης και το σύνολο ελέγχου (training and

testing dataset). Ανάλογα με τη φύση των δεδομένων, αυτά χωρίζονται σε δύο σύνολα, στις περισσότερες όμως περιπτώσεις ο διαχωρισμός γίνεται από 50-70% με 50-30% αντίστοιχα. Τα δεδομένα εκπαίδευσης είναι αυτά που ορίζουν το μοντέλο. Κατόπι χρησιμοποιείται το μοντέλο αυτό στα δεδομένα ελέγχου. Μπορούμε αυτή τη διαδικασία να την κάνουμε περισσότερες φορές, οπόταν η ακρίβεια του μοντέλου θα είναι ο μέσος όρος της ακρίβειας που παίρνουμε κάθε φορά.

ii. **k-fold cross validation.[13]:** Τα δεδομένα χωρίζονται σε k περίπου ίσα μέρη (folds). Γίνονται k επαναλήψεις, όπου σε κάθε επανάληψη το ένα υποσύνολο είναι το υποσύνολο ελέγχου και τα υπόλοιπα υποσύνολα χρησιμοποιούνται για εκπαίδευση. Η ακρίβεια υπολογίζεται από τον συνολικό αριθμό των σωστών κατηγοριοποιήσεων σε όλες τις επαναλήψεις δια τον συνολικό αριθμό του συνόλου των δεδομένων.

iii. **Bootstraping [13]:** Βασίζεται στη μέθοδο k-fold cross validation. Το k ορίζεται σύμφωνα με τον αριθμό των αρχικών δειγμάτων. Τα σύνολα εκπαίδευσης επιλέγονται με δειγματοληψία με τη μέθοδο αντικατάστασης και παράλειψης (replacement and leave-one-out). Σε κάθε επανάληψη, ο ταξινομητής εκπαιδεύεται από το σύνολο των k-1 δειγμάτων που επιλέγονται τυχαία από το σύνολο των αρχικών δειγμάτων. Ο έλεγχος του ταξινομητή εκτελείται χρησιμοποιώντας το υπόλοιπο υποσύνολο.

iv. **Σύγχυση μήτρας (confusion matrix):** Μια σύγχυση μήτρας (confusion matrix) [96] περιέχει πληροφορίες σχετικά με την πραγματική και την προβλεπόμενη ταξινόμηση που έγινε σε ένα σύστημα ταξινόμησης. Οι επιδόσεις των συστημάτων αυτών αξιολογούνται συχνά με τη χρήση των δεδομένων στη μήτρα. Ο παρακάτω πίνακας δείχνει τη σύγχυση μήτρας για τον ταξινομητή με δύο κατηγορίες.

Οι εγγραφές στη μήτρα σύγχυσης έχουν την εξής έννοια:

• α είναι ο αριθμός των σωστών προβλέψεων όταν ένα δείγμα είναι αρνητικό,

• b είναι ο αριθμός των εσφαλμένων προβλέψεων ότι ένα δείγμα είναι θετικό,

• c είναι ο αριθμός των εσφαλμένων προβλέψεων όταν ένα δείγμα είναι αρνητικό, και

• d είναι ο αριθμός των σωστών προβλέψεων όταν ένα δείγμα είναι θετικό.

Πίνακας 4.1: Σύγχυση μήτρας

		Προβλεπόμενο	
		Αρνητικό	Θετικό
Πραγματικό	Αρνητικό	a	b
	Θετικό	c	d

Έχουν οριστεί πολλοί συνήθεις όροι για τη μήτρα σύγχυσης με δύο κατηγορίες.

• Το ποσοστό αληθινά θετικών (TP) είναι το ποσοστό των θετικών περιπτώσεων που έχουν αναγνωριστεί σωστά, όπως υπολογίζεται με την εξίσωση:

$$TP = \frac{d}{c+d}$$

• Το ποσοστό ψευδώς θετικών (FP) είναι το ποσοστό των αρνητικών περιπτώσεων που είχε εσφαλμένα χαρακτηρισθεί ως θετική, όπως υπολογίζεται με την εξίσωση:

$$FP = \frac{b}{a+b}$$

• Το ποσοστό αληθινά αρνητικών (TN) ορίζεται ως το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν σωστά, όπως υπολογίζεται με την εξίσωση:

$$TN = \frac{a}{a+b}$$

• Το ποσοστό ψευδώς αρνητικών (FN) είναι το ποσοστό των θετικών περιπτώσεων που είχαν εσφαλμένα χαρακτηριστεί σαν αρνητικά, όπως υπολογίζεται με την εξίσωση:

$$FN = \frac{c}{c+d}$$

Ευαισθησία (sensitivity) [95]

Η ευαισθησία (sensitivity) δείχνει πόσο καλά μπορεί ο κατηγοριοποιητής να αναγνωρίσει τα θετικά δείγματα. Ορίζεται σαν

$$\text{Ευαισθησία} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \qquad \text{(Εξίσωση 4.24)}$$

όπου το αληθινό θετικό αντιστοιχεί στον αριθμό των πραγματικά θετικών δειγμάτων και το θετικό είναι ο αριθμός θετικών δειγμάτων.

Ειδικότητα (specificity) [95]

Η ειδικότητα (specificity) δείχνει πόσο καλά μπορεί ο κατηγοριοποιητής να αναγνωρίσει τα αρνητικά δείγματα. Ορίζεται σαν

$$\text{Ειδικότητα} = = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FN}} \qquad \text{(Εξίσωση 4.25)}$$

όπου το αληθινό αρνητικό αντιστοιχεί στον αριθμό των πραγματικά αρνητικών δειγμάτων και το αρνητικό είναι ο αριθμός αρνητικών δειγμάτων.

Ακρίβεια (precision) [95]

Η ακρίβεια (precision) δείχνει πόσα από τα παραδείγματα που ο ταξινομητής έχει ταξινομήσει ως θετικά είναι πραγματικά θετικά. Όσο πιο μεγάλη η ακρίβεια, τόσο μικρότερος ο αριθμός των ψευδώς θετικών.

$$\text{Ακρίβεια} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \qquad \text{(Εξίσωση 4.26)}$$

Ανάκληση (Recall) [95]

Η ανάκληση είναι το αληθινά θετικό ποσοστό που κατάφερε ο ταξινομητής να βρει .

$$\text{Ανάκληση} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \qquad \text{(Εξίσωση 4.27)}$$

Ορθότητα (accuracy) [99]

Η ορθότητα (AC) είναι το ποσοστό του συνολικού αριθμού των προβλέψεων που ήταν σωστές. Αυτό καθορίζεται με την εξίσωση:

$$\text{Accuracy} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} + \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FN}} \qquad \text{(Εξίσωση 4.28)}$$

Η ορθότητα που προσδιορίζεται με την πιο πάνω εξίσωση ενδέχεται να μην είναι το πιο κατάλληλο μέτρο όταν ο αριθμός των αρνητικών περιπτώσεων είναι πολύ μεγαλύτερος από τον αριθμό των θετικών [99]. Ας υποθέσουμε ότι υπάρχουν 1.000 περιπτώσεις, 995 από τις οποίες είναι αρνητικές και 5 θετικές. Αν το σύστημα τις κατατάσσει όλες ως αρνητικές, η ορθότητα θα είναι 99.5%, παρόλο που ο ταξινομητής έχασε όλες τις θετικές περιπτώσεις.

Διαγώνιο πηλίκο (Odds-Ratio (OR)) [95]

Είναι ένα μέτρο σύγκρισης παραγόντων κινδύνου και εφαρμόζεται περισσότερο σε μελέτες ασθενών και φυσιολογικών μαρτύρων. Σε πολλές μελέτες έχει την έννοια του Σχετικού κινδύνου. Πολλές φορές οι ερευνητές συγχύζουν το Relative Risk με το Odds Ratio.

Πίνακας 4.2: Υπολογισμός διαγώνιου πηλίκου

	Περιπτώσεις: X = 1	Όχι περιπτώσεις: X = 0
Έκθεση: Y =1	A	B
Μη έκθεση: Y = 0	C	D

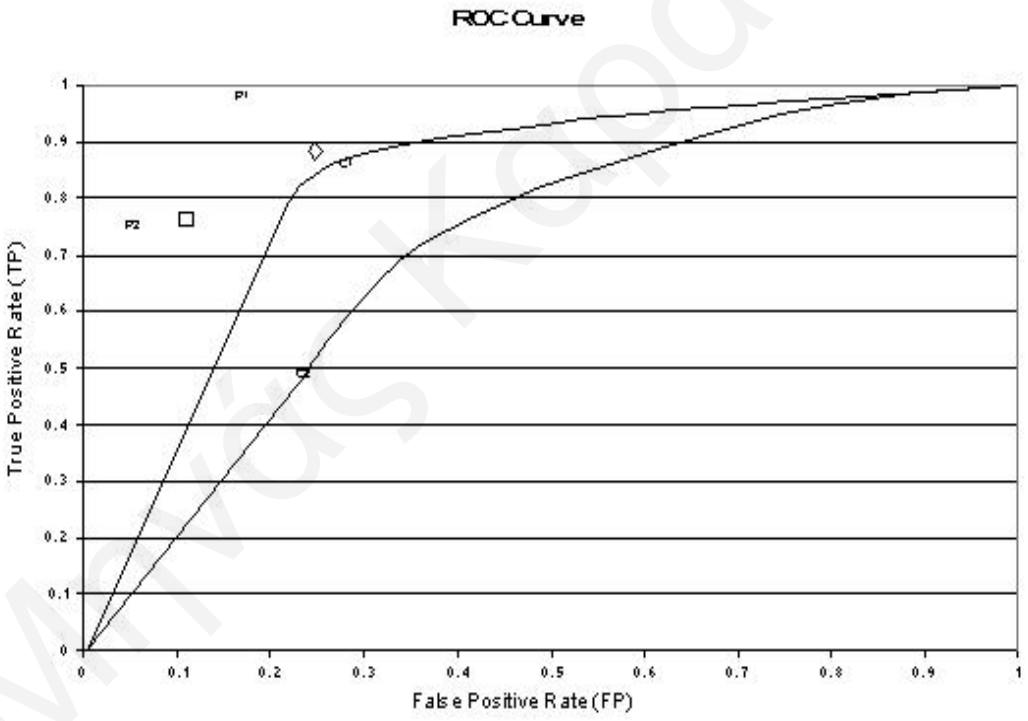
$$\text{Odds Ratio} = \frac{[A/(A+C) \div C/(A+C)]}{[B/(B+D) \div D/(B+D)]} = \frac{[A/C]}{[B/D]} = \frac{[AD]}{[BC]} \qquad \text{(Εξίσωση 4.29)}$$

ROC γραφήματα [100]

Εκτός από τη σύγχυση μήτρας, τα γραφήματα ROC αποτελούν ένα άλλο τρόπο για την εξέταση της απόδοσης των ταξινομητών. Ένα γράφημα ROC είναι μια περιοχή με το ποσοστό ψευδώς θετικών στον άξονα X και το ποσοστό αληθινά θετικών στον άξονα Y. Το σημείο (0,1) είναι ο τέλειος ταξινομητής: κατατάσσει όλες τις θετικές και αρνητικές περιπτώσεις ορθά. Είναι (0,1), επειδή το ψευδώς θετικό ποσοστό είναι 0 και το αληθινά θετικό ποσοστό 1.

Το σημείο (0,0) αντιπροσωπεύει τον ταξινομητή που προβλέπει όλες τις περιπτώσεις να είναι αρνητικές, ενώ το σημείο (1,1) αντιστοιχεί στον ταξινομητή που προβλέπει ότι σε κάθε περίπτωση είναι θετικές. Το σημείο (1,0) είναι ο ταξινομητής που εκτιμά ως λανθασμένες όλες τις ταξινομήσεις.

Σε πολλές περιπτώσεις, ο ταξινομητής έχει μια παράμετρο που μπορεί να προσαρμοστεί ώστε να αυξηθεί το αληθινά θετικό (TP) όταν έχει αυξηθεί το ψευδώς θετικό (FP) ή να μειώσει το FP όταν έχει μειωθεί το TP. Κάθε ρύθμιση παραμέτρων παρέχει ένα (FP, TP) ζεύγος και μια σειρά από τέτοια ζεύγη μπορούν να χρησιμοποιηθούν για τη δημιουργία μιας ROC καμπύλης. Ένας μη παραμετρικός ταξινομητής εκπροσωπείται από ένα μόνο σημείο ROC, το οποίο αντιστοιχεί σε ένα (FP, TP) ζεύγος.



Σχήμα 4.1: Παράδειγμα ενός ROC γραφήματος

Το παραπάνω σχήμα δείχνει ένα παράδειγμα ενός ROC γραφήματος με δύο καμπύλες ROC τη C1 και τη C2 και δύο ROC σημεία τα P1 και P2. Μη παραμετρικοί αλγόριθμοι παράγουν ενιαίο ROC σημείο για ένα συγκεκριμένο σύνολο δεδομένων.

Χαρακτηριστικά του ROC γραφήματος:

• Η καμπύλη ROC ή το σημείο είναι ανεξάρτητα από τη διανομή της τάξης ή το κόστος λάθους [101].

• Ένα γράφημα ROC ενσωματώνει όλες τις πληροφορίες που περιέχονται στη μήτρα σύγχυσης, δεδομένου ότι το FN αποτελεί το συμπλήρωμα του TP και το TN αποτελεί το συμπλήρωμα του FP [100].

• Οι ROC καμπύλες παρέχουν ένα οπτικό εργαλείο για την εξέταση της διαχωριστικής γραμμής μεταξύ της ικανότητας του ταξινομητή να αναγνωρίσει σωστά θετικές περιπτώσεις και του αριθμού των αρνητικών υποθέσεων που έχουν λανθασμένα ταξινομηθεί.

Κεφάλαιο 5: Εφαρμογές εξόρυξης γνώσης στην καρδιολογία

5.1 Μελέτες αξιολόγησης κινδύνου σε καρδιακά επεισόδια

Ο Πίνακας 5.1 παρουσιάζει κλινικές μελέτες που έγιναν από ερευνητές ή ομάδες ερευνητών, τη χρονολογία που έκαναν τη μελέτη, που επικεντρώθηκαν στη μελέτη τους και ποιους παράγοντες μελέτησαν, τη μεθοδολογία που χρησιμοποίησαν, ως επίσης και τα αποτελέσματα που εξήγαγαν. Πιο κάτω περιγράφονται αυτές οι μελέτες:

Euroaspire study group, 1997 [102]

Στη πρώτη μελέτη που έκανε η ομάδα της EUROASPIRE με τίτλο: 'A European Society of Cardiology survey of secondary prevention of coronary heart disease: principal results. EUROASPIRE Study Group. European Action on Secondary Prevention through Intervention to Reduce Events', περιγράφονται τα ακόλουθα ευρήματα:

Οι τρεις μεγάλες ευρωπαϊκές επιστημονικές εταιρείες στην καρδιαγγειακή ιατρική, η Ευρωπαϊκή Εταιρεία Καρδιολογίας (ESC), η Ευρωπαϊκή Εταιρεία Αθηροσκλήρωσης και η Ευρωπαϊκή Εταιρεία Υπέρτασης, δημοσίευσαν τον Οκτώβριο του 1994 κοινές συστάσεις για την πρόληψη της στεφανιαίας νόσου στην κλινική πράξη. Ασθενείς με στεφανιαία νόσο, ή άλλη σημαντική αθηροσκληρωτική νόσο, τέθηκαν σε προτεραιότητα για την πρόληψη. Έγινε μια ευρωπαϊκή έρευνα (EUROASPIRE) σε αυτόν τον τομέα υπό την αιγίδα της Ευρωπαϊκής Εταιρείας Καρδιολογίας για να περιγράψει την τρέχουσα κλινική πρακτική σε σχέση με την δευτερογενή πρόληψη της στεφανιαίας νόσου. Οι στόχοι της EUROASPIRE ήταν:

(i) να καθορίσει κατά πόσον οι κύριοι παράγοντες κινδύνου για στεφανιαία νόσο έχουν καταγραφεί στα ιατρικά αρχεία,

(ii) να καταμετρήσει τους μεταβαλλόμενους παράγοντες κινδύνου και να περιγράψει την τρέχουσα διαχείριση τους μετά από τη νοσηλεία, και

(iii) να διαπιστωθεί αν έχουν ελεγχθεί πρώτου βαθμού συγγενείς.

Η έρευνα πραγματοποιήθηκε σε επιλεγμένες γεωγραφικές περιοχές και νοσοκομεία σε εννέα ευρωπαϊκές χώρες. Είχαν εντοπιστεί (ασθενείς μέχρι 70 ετών) με τις εξής διαγνώσεις: αορτοστεφανιαία παράκαμψη, αγγειοπλαστική, οξύ έμφραγμα του μυοκαρδίου και ισχαιμία του μυοκαρδίου χωρίς έμφραγμα. Η συλλογή δεδομένων βασίστηκε στην αναδρομική ανασκόπηση των ιατρικών φακέλων των νοσοκομείων και μια συνέντευξη και εξέταση των ασθενών.

Τα αποτελέσματα των μελετών της ομάδας της EUROASPIRE είναι: 4863 ιατρικοί φάκελοι εξετάστηκαν από τους οποίους 25% ήταν γυναίκες. Σε 3569 ασθενείς έγιναν συνεντεύξεις (προσαρμοσμένο ποσοστό ανταπόκρισης 85%) με μέση ηλικία τα 61 έτη. Δεκαεννέα τοις εκατό των ασθενών κάπνιζαν, 25% ήταν υπέρβαροι (ΔΜΣ > ή = 30 $kg.m^{-2}$), 53% είχε αυξημένη αρτηριακή πίεση (συστολική BP >= 140 και/ή διαστολική BP >= 90 mmHg), 44 % είχε προβάλει ολική χοληστερόλη του πλάσματος (ολική χοληστερόλη > = 5.5 mmol/l) και 18% ήταν διαβητικοί. Φάρμακο που αναφέρθηκε στη συνέντευξη ήταν: αντιαιμοπεταλιακό 81%, β-αναστολείς, 54% (58% στις μετά-μυοκαρδίου ασθενείς). Από τους ασθενείς που λαμβάνουν φάρμακα για μείωση της αρτηριακής πίεσης (δεν είναι πάντα που προβλέπεται για τη θεραπεία της υπέρτασης) 50% είχαν συστολική πίεση BP > 140 mmHg και 21% > 160 mmHg, και όσων λαμβάνουν φάρμακα μείωσης των λιπιδίων, 49% είχε ολική χοληστερόλη > 5.5 mmol/l και 13% > 6.5 mmol/l. Τριάντα επτά τοις εκατό των ασθενών είχαν οικογενειακό ιστορικό πρόωρης στεφανιαίας νόσου σε πρώτου βαθμού συγγενή, αλλά μόνο το 21% των ασθενών ανέφερε ότι συμβούλευσε τους συγγενείς τους να ελέγχονται για στεφανιαία νόσο.

Euroaspire II study group, 2001 [103]

Ο κύριος στόχος της δεύτερης έρευνας EUROASPIRE II [103] ήταν να διαπιστωθεί σε ασθενείς με στεφανιαία νόσο κατά πόσον τηρούνται οι συστάσεις των 'Joint European

Societies' για την πρόληψη της στεφανιαίας νόσου και αν ακολουθούνται στην κλινική πρακτική.

Η έρευνα πραγματοποιήθηκε κατά το διάστημα 1999-2000 σε 15 ευρωπαϊκές χώρες: Βέλγιο, Τσεχία, Φινλανδία, Γαλλία, Γερμανία, Ελλάδα, Ουγγαρία, Ιρλανδία, Ιταλία, Κάτω Χώρες, Πολωνία, Σλοβενία, Σουηδία, Ισπανία και το Ηνωμένο Βασίλειο, σε επιλεγμένες γεωγραφικές περιοχές και 47 κέντρα. Ασθενείς, άνδρες και γυναίκες μικρότεροι των 70 χρόνων είχαν εντοπιστεί με τις εξής διαγνώσεις: επέμβαση αορτοστεφανιαίας παράκαμψης, αγγειοπλαστική, οξύ έμφραγμα του μυοκαρδίου και ισχαιμία του μυοκαρδίου. Η συλλογή δεδομένων βασίστηκε σε ανασκόπηση των ιατρικών φακέλων και συνέντευξη και αξιολόγηση των κινδύνων τουλάχιστον 6 μήνες μετά την εισαγωγή σε νοσοκομείο.

Υπήρχαν 8181 ιατρικοί φάκελοι (25% γυναίκες) και εξετάστηκαν 5556 ασθενείς (προσαρμοσμένο ποσοστό συμμετοχής 76%). Τα αποτελέσματα αυτής της μελέτης έδειξαν ότι το 21% των ασθενών κάπνιζαν, 31% ήταν παχύσαρκοι, το 50% είχε αυξημένη αρτηριακή πίεση (συστολική αρτηριακή πίεση 140 mmHg και/ή διαστολική αρτηριακή πίεση 90 mmHg), το 58% είχε αυξημένη ολική χοληστερόλη (σύνολο 5 mmol/l χοληστερόλη-1) και 20% ανέφεραν ιατρικό ιστορικό διαβήτη. Όσον αφορά τη φαρμακευτική αγωγή των ασθενών κατά την εισαγωγή τους, την εξαγωγή τους και έπειτα στο χρόνο της συνέντευξης τους, η χρήση της ασπιρίνης ήταν 47%, 90% και 86%, των Βήτα-αναστολέων 44%, 66% και 63% και των αναστολέων ΜΕΑ 24%, 38% και 38% αντίστοιχα. Με την εξαίρεση των αντιαιμοπεταλιακών φαρμάκων, υπήρχαν μεγάλες διαφορές ως προς τη χρήση των φαρμακευτικών αγωγών μεταξύ των χωρών.

Rea *et al.*, 2002 [104]

Οι Rea *et al.* [104] ασχολήθηκαν με τον παράγοντα κάπνισμα σε ασθενείς με στεφανιαία νόσο, όπου παραμένουν ερωτήματα σχετικά με τη σημασία του καπνίσματος και τη διακοπή του καπνίσματος μετά από ένα περιστατικό εμφράγματος του μυοκαρδίου.

Ο στόχος ήταν να εκτιμηθεί η σχέση μεταξύ του καπνίσματος και των κινδύνων για την επανάληψη ενός στεφανιαίου επεισοδίου. Μελετήθηκαν 2619 ασθενείς, άτομα που επιβίωσαν μετά από ένα (πρώτο) έμφραγμα του μυοκαρδίου. Χρησιμοποιήθηκε το μέτρο 'σχετικός κίνδυνος' (RR), που αξιολογείται με τη χρήση Cox αναλογικών κινδύνων ανάλυσης παλινδρόμησης, για επαναλήψεις στεφανιαίων επεισοδίων σε μη καπνιστές (άτομα που δεν είχαν ιστορικό καπνίσματος), πρώην καπνιστές (άτομα που είχαν σταματήσει το κάπνισμα πριν το επεισόδιο), άτομα που σταμάτησαν το κάπνισμα μετά από έμφραγμα και τους ενεργούς καπνιστές (τα άτομα που συνέχισαν το κάπνισμα μετά από έμφραγμα).

Τα αποτελέσματα της έρευνας έδειξαν ότι κατά τον χρόνο των περιστατικών του μυοκαρδίου, 33.6% των ασθενών δεν ήταν καπνιστές, 35.5% ήταν πρώην καπνιστές και 30.9% ήταν ενεργοί καπνιστές. Από τα 808 άτομα που ήταν ενεργοί καπνιστές κατά τη στιγμή του συμβάντος του μυοκαρδίου, 449 σταμάτησαν το κάπνισμα κατά τη διάρκεια της νοσηλείας ή μετά τη νοσηλεία. Με τους μη καπνιστές ως ομάδα αναφοράς, το μέτρο RR για επαναλαμβανόμενα στεφανιαία επεισόδια (n = 433) ήταν 1.17 (95% CI, 0.93 με 1.43) για τους πρώην καπνιστές και 1.51 (CI, 1.10 με 2.07) για τους ενεργούς καπνιστές. Με τους μη καπνιστές ως ομάδα αναφοράς, η RR για τα άτομα που σταμάτησαν το κάπνισμα μετά από έμφραγμα ήταν 1.62 (CI, 1.02 με 2.61), εφόσον η διάρκεια της διακοπής ήταν μεταξύ 0 και 6 μηνών, 1.60 (CI, 0.97 σε 2.60), αν η διάρκεια ήταν μεταξύ 6 και 18 μηνών, 1.48 (CI, 0.76 με 2.51), αν η διάρκεια ήταν μεταξύ 18 και 36 μηνών και 1.02 (CI, 0.54 με 1.86), αν η διάρκεια ήταν περισσότερο από 36 μήνες (p = 0.01).

Wuensche, 2003 [105]

Ο Wuensche [105] στην εργασία του "the visualization and measurement of left ventricular deformation" αναφέρει τα ακόλουθα: «Ενώ έχει σημειωθεί ιατρική πρόοδος στη διάγνωση και τη θεραπεία των καρδιακών παθήσεων, παραμένει όμως ο μεγαλύτερος δολοφόνος στο δυτικό κόσμο. Καρδιαγγειακές παθήσεις προκαλούν σημαντική νοσηρότητα και η πρόγνωση μετά την καρδιακή ανεπάρκεια είναι κακή. Μια βελτιωμένη κατανόηση του μηχανισμού της καρδίας βοηθά τη διάγνωση και τη θεραπεία των καρδιακών παθήσεων».

Έχουν αναπτύξει ένα εργαλείο σχεδιασμένο για την απεικόνιση των βιοϊατρικών προτύπων. Αυτή η εργασία εξηγεί τις τεχνικές για τα βιοϊατρικά πεπερασμένα και τα πρότυπα στοιχείων και καταδεικνύει την αίτησή τους στα βιοϊατρικά σύνολα στοιχείων με το να χρησιμοποιήσει για παράδειγμα δύο πρότυπα μιας υγιούς και ασθενούς ανθρώπινης αριστερής κοιλίας. Οι συνεισφορά αυτού του εγγράφου είναι τριπλή: πρώτα εφαρμόζουν τις τεχνικές που χρησιμοποιούνται παραδοσιακά - μηχανική και υπολογιστική ρευστή δυναμική στα βιοϊατρικά στοιχεία.

Προτείνουν επίσης μερικές βελτιώσεις και τροποποιήσεις. Αφετέρου σχηματίζουν μια νέα άποψη στους μηχανισμούς της υγιούς και ασθενούς αριστερής κοιλίας και διευκολύνουν την κατανόηση της σύνθετης παραμόρφωσης του μυός της καρδίας από τις νέες απεικονίσεις. Τέλος εισάγουν επίσης σε αυτήν την διαδικασία ένα εργαλείο που σχεδιάστηκε για την απεικόνιση των βιοϊατρικών στοιχείων.

Wang *et al.*, 2005 [106]

Οι Wang *et al.* [106] είχαν κάνει μια έρευνα με την εξίσωση του Framingham. Ο στόχος αυτής της δουλειάς ήταν να προσδιοριστεί ο βαθμός στον οποίο η εξίσωση του Framingham προβλέπει τον κίνδυνο για στεφανιαία νόσο (ΣΝ) σε ιθαγενείς στην Αυστραλία. Συμμετείχαν 687 άτομα ηλικίας 20-74 χρονών και βρίσκονταν υπό παρακολούθηση από μια αρχική εξέταση στα έτη 1992-1995 έως 31 Δεκεμβρίου 2003. Μια πρωτότυπη εξίσωση Framingham χρησιμοποιήθηκε για την πρόβλεψη του κινδύνου στεφανιαίας νόσου κατά τη διάρκεια της παρακολούθησης των τιμών των παραδοσιακών παραγόντων κινδύνου, η οποία περιελάμβανε την ηλικία, το φύλο, το συνολικό επίπεδο χοληστερόλης, τη λιποπρωτεΐνη υψηλής πυκνότητας (HDL) επιπέδου χοληστερόλης, την αρτηριακή πίεση , τη παρουσία του διαβήτη, και το κάπνισμα.

Η προβλεπόμενη συχνότητα εμφάνισης στεφανιαίας νόσου χρησιμοποιώντας την εξίσωση του Framingham ήταν 4.4 ανά 1000 άτομα-έτη, ενώ η συχνότητα εμφάνισης ήταν 11 (95% CI, 8.7-13.9) ανά 1000 άτομα-έτη. Ο παρατηρούμενος αριθμός των εκδηλώσεων στεφανιαίας νόσου (68) ήταν 2.5 φορές του προβλεπόμενου αριθμού (27) χρησιμοποιώντας την εξίσωση

του Framingham. Η συχνότητα εμφάνισης ήταν περίπου τέσσερις και τρεις φορές της προβλεπόμενης επίπτωσης για τις ομάδες ηλικίας < 35 και 35-44 ετών, αντίστοιχα, και περίπου διπλάσια από το προβλεπόμενο ποσοστό για όσους είναι άνω των 45 ετών. Η εξίσωση του Framingham ήταν μια ιδιαίτερα αξιόπιστη προάγγελος για τις γυναίκες, ιδιαίτερα τις νεότερες γυναίκες, στις οποίες το παρατηρούμενο ποσοστό στεφανιαίας νόσου ήταν 30 φορές της προβλεπόμενης τιμής.

Bambrick, 2005 [107]

Η Bambrick [107] έθεσε σαν στόχο να διαπιστώσει αν ο δείκτης μάζας σώματος (BMI), όριο που ορίζεται για παχυσαρκία (30 kg/m^2) αντικατοπτρίζει επαρκώς τον κίνδυνο σε μια κοινότητα με υψηλό ποσοστό του διαβήτη τύπου 2. Είχε στοιχεία για πέντε παράγοντες κινδύνου διαβήτη (ηλικία, ΔΜΣ, περιφέρεια μέσης (WC), υπέρταση και οικογενειακό ιστορικό) και γλυκόζη νηστείας (FG) που είχαν ληφθεί από ένα τυχαίο δείγμα των 117 ενηλίκων (62 γυναίκες και 55 άνδρες) που δεν είχαν διαγνωσθεί με διαβήτη. Έγινε ανάλυση γραμμικής παλινδρόμησης μεταξύ ΔΜΣ, WC και FG, και διεξήχθησαν αναλύσεις με την ευαισθησία και την ιδιαιτερότητα στην πρόβλεψη αυξημένων FG και υπέρτασης. Οι παράγοντες δείκτης μάζας σώματος ΔΜΣ ≥ 30 kg/m^2 και κεντρική παχυσαρκία που εκτιμάται από την περιφέρεια μέσης (γυναίκες ≥ 88 εκατοστά, άνδρες ≥ 102 εκατοστά) ήταν σημαντικά και θετικά συνδεδεμένοι. Μεταξύ των γυναικών, η κεντρική παχυσαρκία ήταν σχεδόν καθολική, κάτι που συμβαίνει σε ΔΜΣ κάτω του ορίου των 20-25. Η περιφέρεια μέσης συσχετίστηκε γραμμικά με άλλους παράγοντες κινδύνου διαβήτη. Η περιφέρεια μέσης ≥ 88 εκατοστών ήταν πιο ευαίσθητη αλλά λιγότερο συγκεκριμένη από το ΔΜΣ ≥ 30 στην πρόβλεψη αυξημένης γλυκόζης νηστείας και της υπέρτασης μεταξύ των γυναικών, ενώ ο ΔΜΣ ≥ 25 στους άνδρες τείνει να είναι τόσο πιο ευαίσθητος και πιο συγκεκριμένος από τους δύο παράγοντες, ΔΜΣ ≥ 30 και περιφέρεια μέσης ≥ 102 εκατοστά.

Marshall, 2008 [108]

Ο Marshall [108], σε αυτήν την έρευνά του, αναφέρει ότι για τον εντοπισμό ασυμπτωματικών ασθενών υψηλού κινδύνου καρδιαγγειακών νοσημάτων απαιτείται η αξιολόγηση των παραγόντων κινδύνου. Η παροχή πρωτοβάθμιας περίθαλψης πρέπει συνεπώς να προσδιορίζει τους ασθενείς που δεν είχαν καρδιαγγειακή νόσο και είναι ύψιστης προτεραιότητας για την αξιολόγηση του καρδιαγγειακού κινδύνου. Μία προσέγγιση είναι να δώσει προτεραιότητα σε ασθενείς κάνοντας μια αρχική αξιολόγηση για τον κίνδυνο καρδιακού επεισοδίου. Αυτή η εκ των προτέρων εκτίμηση του καρδιαγγειακού κινδύνου προέρχεται από στοιχεία για τους παράγοντες κινδύνου, τα οποία ανήκουν συνήθως σε ηλεκτρονικά ιατρικά αρχεία. Σε ασθενείς με άγνωστη αρτηριακή πίεση και επίπεδα χοληστερόλης οι τιμές αυτές αντικαθίστανται από προεπιλεγμένες τιμές που προέρχονται από τα εθνικά στοιχεία της έρευνας. Σε αυτή την εργασία αναλύονται τα χαρακτηριστικά χρησιμοποιώντας στρατηγική για την αναγνώριση των ασθενών υψηλού κινδύνου.

Έχουν δημιουργηθεί καμπύλες με τα χαρακτηριστικά για τη χρησιμοποίηση μιας εκ των προτέρων εκτίμησης του καρδιαγγειακού κινδύνου για τον εντοπισμό των ασθενών σε ποσοστό (δεκαετή καρδιαγγειακού) κινδύνου μεγαλύτερο του 20%. Αυτό συγκρίθηκε με τις στρατηγικές που χρησιμοποιούν την ηλικία, ή διαβητική κατάσταση και αντιυπερτασική θεραπεία για τον εντοπισμό ασθενών υψηλού κινδύνου.

Η περιοχή κάτω από την καμπύλη για την προηγούμενη εκτίμηση του καρδιαγγειακού κινδύνου, που υπολογίζεται χρησιμοποιώντας ελάχιστα στοιχεία (0.933, 95% CI: 0.925 με 0.941) είναι σημαντικά μεγαλύτερη από ότι για μια στρατηγική επιλογή με βάση την ηλικία (0.892, 95% CI: 0.882 με 0.902), ή διαβήτη και υπέρταση (0.608, 95% CI: 0.584 με 0.632).

Χρησιμοποιώντας στοιχεία των βάσεων δεδομένων που συλλέχθηκαν κατά την πρωτοβάθμια φροντίδα είναι δυνατόν να προσδιοριστούν τα άτομα που διατρέχουν υψηλό κίνδυνο καρδιαγγειακής νόσου. Η τεχνολογία των πληροφοριών, για να βοηθήσει τους ασθενείς πρωτοβάθμιας φροντίδας που δίνει προτεραιότητα στη πρόληψη των καρδιαγγειακών παθήσεων, μπορεί να βελτιώσει την αποτελεσματικότητα της αξιολόγησης του κινδύνου των καρδιαγγειακών.

Euroaspire III Kotseva *et al.*, 2009 [109]

Ο στόχος της έρευνας (EUROASPIRE III) [109] 'Ευρωπαϊκή Δράση για την Πρωτοβάθμια και Δευτεροβάθμια πρόληψη με την παρέμβαση για τη μείωση επεισοδίων' ήταν να προσδιοριστεί, εάν οι κατευθυντήριες γραμμές από τις 'Joint European Societies' για την πρόληψη των καρδιαγγειακών νοσημάτων ακολουθούνται στην καθημερινή κλινική πράξη και να περιγράψει τον τρόπο ζωής, τους παράγοντες κινδύνου και την θεραπευτική αντιμετώπιση σε ασθενείς με στεφανιαία νόσο στην Ευρώπη.

Η έρευνα EUROASPIRE III πραγματοποιήθηκε κατά την περίοδο 2006-2007 σε 76 κέντρα από επιλεγμένες γεωγραφικές περιοχές σε 22 χώρες στην Ευρώπη. Ασθενείς, με την κλινική διάγνωση της στεφανιαίας νόσου, που εντοπίστηκαν, παρακολουθήθηκαν, πέρασαν από συνεντεύξεις και εξετάζονταν τουλάχιστον 6 μήνες μετά την εκδήλωση της στεφανιαίας νόσου.

Σε αυτή την έρευνα υπήρχαν δεκατρείς χιλιάδες εννιακόσιοι τριάντα πέντε ιατρικά φακέλοι ασθενών (27% γυναίκες) και εξετάστηκαν 8.966 ασθενείς. Το 17% των ασθενών κάπνιζαν, το 35% ήταν παχύσαρκοι, το 56% είχαν αρτηριακή πίεση ≥ 140/90 mmHg (≥ 130/80 σε άτομα με σακχαρώδη διαβήτη), το 51% είχαν επίπεδα της ολικής χοληστερόλης ορού ≥ 4.5 mmol / l και το 25% ανέφεραν ιστορικό διαβήτη εκ των οποίων 10% είχε στη γλυκόζη πλάσματος νηστείας λιγότερο από 6.1 mmol / l και το 35% μία γλυκοζυλιωμένη αιμοσφαιρίνη A1c λιγότερο από το 6.5%. Η χρήση των φαρμάκων ήταν: αντιπλατελέτες 91%, β-αποκλειστές 80%, αναστολείς του μετατρεπτικού ενζύμου της αγγειοτενσίνης / αγγειοτασίνης-αναστολείς των υποδοχέων 71%, αναστολείς των διαύλων ασβεστίου 25% και 78% στατίνες.

Πίνακας 5.1: Επιλεγμένες Κλινικές μελέτες αξιολόγησης των παραγόντων κινδύνου σε καρδιακά επεισόδια

Ερευνητής	Χρόνος	Αναφορά	Ιατρικό πρόβλημα	Αλγόριθμοι/Μέθοδοι	Αποτελέσματα
Ομάδα Euroaspire	1997	[102]	Καταγραφή των παραγόντων κινδύνου από ιατρικά αρχεία ασθενών, μέτρηση των παραγόντων κινδύνου και του ιστορικού στην οικογένεια.	Συλλογή δεδομένων, επανεξέταση των νοσοκομειακών ιατρικών αρχείων, μέγεθος του δείγματος και την ισχύ των υπολογισμών	Ελάττωση νοσηρότητας και θνησιμότητας σε καρδιακές παθήσεις και βελτίωση της πιθανότητας επιβίωσης των ασθενών
Euroaspire study group	II2001	[103]	Οι συστάσεις για την πρόληψη της στεφανιαίας νόσου που ακολουθείται στην κλινική πράξη	Επανεξέταση των ιατρικών φακέλων και τη συνέντευξη και αξιολόγηση των κινδύνων, τουλάχιστον 6 μήνες μετά την εισαγωγή σε νοσοκομείο	Ανθυγιεινό τρόπο ζωής των ασθενών που δημιουργούν υψηλές τιμές στους μεταβολλόμενους παράγοντες κινδύνου
Rea *et al.*	2002	[104]	Σημασία του καπνίσματος μετά από περιστατικό του εμφράγματος του μυοκαρδίου	Cox ανάλυση παλινδρόμησης αναλογικού κινδύνου	Συσχέτιση καπνίσματος ως παράγοντας κινδύνου
Wuensche	2003	[105]	Η πληρέστερη κατανόηση της καρδιακής μηχανικής για τη διάγνωση και τη θεραπεία των καρδιακών παθήσεων	Χρησιμοποιώντας ετικέτες μαγνητικής τομογραφίας (MRI)	Προτείνουν κάποιες μικρές βελτιώσεις και τροποποιήσεις, απόκτηση νέας άποψης για τη μηχανική των υγιών και νοσούντων στην αριστερή κοιλία
Wang *et al*.	2005	[106]	Framingham CHD για αυτόχθονες	Διάρκεια της παρακολούθησης και τις τιμές των παραδοσιακών παραγόντων κινδύνου	Η εξίσωση Framingham παράγει σημαντικές υποτιμήσεις των κινδύνων στεφανιαίας νόσου σε αυτόχθονες
Bambrick	2005	[107]	Για να προσδιοριστεί αν ο Δείκτης Μάζας Σώματος αντανακλά ικανοποιητικά τον κίνδυνο σε μια κοινότητα με αυτόχθονες	Γραμμική παλινδρόμηση	Ποιοι παράγοντες κινδύνου είναι γραμμικά συνδεδεμένοι
Marshall	2008	[108]	Για τον εντοπισμό των ασθενών ψηλού κινδύνου χωρίς καρδιαγγειακή νόσο	Framingham	ROC καμπύλη για την προ-εκτίμηση του καρδιαγγειακού κινδύνου
Euroaspire III Kotseva *et al.*	2009	[109]	αν οι κατευθυντήριες γραμμές του Joint European Societies στην καρδιαγγειακή πρόληψη ακολουθείται στην καθημερινή κλινική πράξη και να περιγράψει τον τρόπο ζωής, παράγοντα κινδύνου και θεραπευτική αντιμετώπιση	ασθενείς, με την κλινική διάγνωση της στεφανιαίας νόσου, εντοπίστηκαν εκ των υστέρων και, στη συνέχεια ακολούθησε, συνέντευξη και εξετάστηκαν τουλάχιστον 6 μήνες μετά το στεφανιαίο επεισόδιο τους	17% των ασθενών κάπνιζαν τσιγάρα, το 35% ήταν παχύσαρκοι και το 53% παχύσαρκοι, το 56% είχαν αρτηριακή πίεση ≥ 140/90 mmHg, το 51% είχαν ολική χοληστερόλη ≥ 4,5 mmol / l και 25% διαβήτη εκ των οποίων 10% είχαν γλυκόζη κάτω των 6,1 mmol / l και 35%% το γλυκοζυλιωμένη αιμοσφαιρίνη A1c λιγότερο από 6,5

5.2 Μελέτες με δέντρα απόφασης για αξιολόγηση των παραγόντων κινδύνου για καρδιακά επεισόδια

Σε αυτό το υποκεφάλαιο περιγράφονται εν συντομία μελέτες που έχουν γίνει για τους παράγοντες καρδιακών επεισοδίων με εξόρυξη δεδομένων, με την χρήση των δέντρων απόφασης. Ο Πίνακας 5.2 περιγράφει επιλεγμένες μελέτες και τη χρονολογία που έχουν γίνει. Παρατίθεται το ιατρικό πρόβλημα, ο αλγόριθμος / μέθοδος που χρησιμοποιήθηκε και τα αποτελέσματα που έχουν εξαχθεί.

Tsien *et al.*, 1998 [110]

Οι Tsien *et al.* [110] αναφέρουν ότι η έγκαιρη και ακριβής διάγνωση του εμφράγματος του μυοκαρδίου (MI) σε ασθενείς οι οποίοι παρουσιάζονται στις πρώτες βοήθειες με πόνο στο στήθος είναι ένα σημαντικό πρόβλημα στην ιατρική των εκτάκτων περιστατικών. Έχει αναπτυχθεί ένας αριθμός από βοηθήματα με στόχο να βοηθήσει το πρόβλημα, αλλά στη γενική χρήση δεν υπάρχει πρόοδος. Τεχνικές μηχανικής μάθησης, συμπεριλαμβανομένων δέντρων απόφασης και λογιστικής αναγωγής (LR), έχουν τη δυνατότητα να δημιουργήσουν μια απλή αλλά ακριβή απόφαση. Και οι δύο, ένα δέντρο απόφασης (Tree FT) και ένα μοντέλο LR (FT LR) έχουν αναπτυχθεί για να προβλέψουν την πιθανότητα ενός ασθενή με πόνο στο στήθος να έχει ένα επεισόδιο MI και βασίζεται μόνο σε στοιχεία που διατίθενται κατά τη στιγμή της παρουσίασης του ER. Προηγούμενα μοντέλα που έχουν ήδη δημοσιευθεί, το δέντρο απόφασης Goldman [111] και η εξίσωση Kennedy LR [112], αξιολογήθηκαν με τα ίδια σύνολα δεδομένων. Τα αποτελέσματα έδειξαν ότι τα δέντρα ταξινόμησης , τα οποία έχουν ορισμένα πλεονεκτήματα έναντι των LR μοντέλων, μπορούν εξίσου να χρησιμοποιηθούν για τη διάγνωση ασθενών με έμφραγμα μυοκαρδίου.

Colombet *et al.*, 2000 [113]

Οι Colombet *et al.* [113] επισημαίνουν ότι η εκτίμηση του κινδύνου πολλαπλών παραλλαγών είναι πλέον απαραίτητη για τον καθορισμό κατευθυντήριων γραμμών για την πρόληψη των καρδιαγγειακών επεισοδίων. Οι περιορισμοί των υφιστάμενων στατιστικών μοντέλων οδήγησε στη διερεύνηση μεθόδων μηχανικής μάθησης. Η μελέτη αυτή αξιολογεί την εφαρμογή και την εκτέλεση ενός δέντρου απόφασης (CART) και ενός MLP (Multi Layer Perceptron) για να προβλέψουν τον κίνδυνο σε καρδιαγγειακά επεισόδια από τα πραγματικά δεδομένα. Τα δεδομένα της μελέτης χωρίστηκαν τυχαία στο σύνολο εκπαίδευσης (n = 10.296) και στο σύνολο ελέγχου (n = 5.148). Έγινε ανάλυση της καμπύλης ROC για τα κριτήρια της επίδοσης. Περιοχές κάτω από την καμπύλη ROC όπου το 95% του διαστήματος εμπιστοσύνης είναι 0.78 (0.75 – 0.81). 0.78 (0.75 – 0.80) και 0.76 (0.73 – 0.79), εξετάστηκαν αντίστοιχα για λογιστική αναγωγή, MLP και CART. Αυτές οι μέθοδοι μπορούν να συμπληρώσουν τα υπάρχοντα στατιστικά μοντέλα και να συμβάλουν στην ερμηνεία του καρδιαγγειακού κινδύνου.

Gamberger *et al.*, 2000[114]

Οι Gamberger *et al.* [114] στην εργασία τους για τον εντοπισμό των ομάδων ψηλού κινδύνου για ασθένειες καρδιακών νοσημάτων αναφέρουν τα ακόλουθα:

Το ίδρυμα για την καρδιαγγειακή πρόληψη και την αποκατάσταση Ζάγκρεμπ, είναι ένα σημαντικό κροατικό κέντρο για τις καρδιακές παθήσεις. Η ομάδα SolEuNet, που συνεργάζεται με τον καρδιολόγο, έχει χρησιμοποιήσει επιτυχώς τα στοιχεία που συλλέχθηκαν υπό μορφή αρχείων ασθενών στα έτη 1999 και 2000 για να ανακαλύψει τις ομάδες πληθυσμών υψηλού κινδύνου για τις στεφανιαίες καρδιακές παθήσεις. Η ενεργός ομάδα εξόρυξης δεδομένων μαζί με τον ειδικό καρδιολόγο, έχει οδηγήσει στη γνώση που θα χρησιμοποιηθεί στη διαλογή στοχευόμενων πληθυσμών για την έγκαιρη ανίχνευση ασθενειών.

Οι στεφανιαίες καρδιακές παθήσεις (CHD) προκαλούνται από τη στένωση (αθηρωματική πλάκα) μιας ή και περισσοτέρων από τις στεφανιαίες αρτηρίες. Ο μειωμένος ανεφοδιασμός

του αίματος που προκαλεί το μειωμένο ανεφοδιασμό οξυγόνου της εξαρτώμενης περιοχής του μυοκαρδίου, είναι γνωστό ως στηθάγχη. Οι πιο ακραίες συνέπειες είναι το μυοκαρδιακό έμφραγμα και ο (καρδιακός) θάνατος. Οι στεφανιαίες καρδιακές παθήσεις αποτελούν μια από τις συχνότερες αιτίες θανάτου παγκόσμια και ένα σημαντικό πρόβλημα στην ιατρική πρακτική. Η σημερινή πρόληψη CHD στηρίζεται ουσιαστικά σε τρεις σημαντικά διαφορετικούς άξονες:

i. Γενική εκπαίδευση ολόκληρου του πληθυσμού για τους γνωστούς παράγοντες κινδύνου, ειδικά για τους παράγοντες τρόπου ζωής.
ii. Πρακτική διαλογής του παράγοντα κινδύνου από τη συλλογή δεδομένων η οποία εκτελείται σε τρία διαφορετικά στάδια:
    a. πληροφορίες και φυσικά αποτελέσματα εξέτασης,
    b. αποτελέσματα των εργαστηριακών εξετάσεων,
    c. αποτελέσματα του ηλεκτροκαρδιογραφήματος (ECG).
iii. Οι εξετάσεις πρόληψης στα εξειδικευμένα όργανα που περιλαμβάνουν ηλεκτροκαρδιογράφημα κατά την άσκηση ECG και το υπερηχοκαρδιογράφημα.

Pordorelec *et al.*, 2002 [115]

Οι Pordorelec *et al.* [115] επισημαίνουν ότι στην λήψη απόφασης στην ιατρική (ταξινόμηση, διάγνωση, κ.λ.π.), υπάρχουν πολλές περιπτώσεις όπου η απόφαση πρέπει να γίνει αποτελεσματικά και αξιόπιστα. Εννοιολογικά απλά μοντέλα λήψης απόφασης με τη δυνατότητα αυτόματης μάθησης είναι τα πλέον κατάλληλα για την εκτέλεση αυτών των καθηκόντων. Τα δέντρα απόφασης είναι μια αξιόπιστη τεχνική για αποτελεσματική λήψη αποφάσεων, που παρέχει υψηλή ακρίβεια ταξινόμησης με μια απλή αναπαράσταση των γνώσεων που συγκεντρώνονται και έχουν χρησιμοποιηθεί σε διάφορους τομείς της ιατρικής για λήψεις αποφάσεων. Στο άρθρο παρουσιάζουν τα βασικά χαρακτηριστικά των διαγραμμάτων αποφάσεων και τις επιτυχείς εναλλακτικές λύσεις για την παραδοσιακή προσέγγιση επαγωγής με έμφαση στις υπάρχουσες και μελλοντικές πιθανές εφαρμογές στην ιατρική.

Voss *et al.*, 2002 [116]

Οι Voss *et al.* [116] εκτιμούν ότι η λογιστική παλινδρόμηση (LR) χρησιμοποιείται συνήθως για την εκτίμηση του κινδύνου της στεφανιαίας νόσου. Έχει γίνει μια επιδημιολογική μελέτη, PROCAM χρησιμοποιώντας νευρωνικά δίκτυα για τη βελτίωση της εκτίμησης των κινδύνων της LR με την ανάλυση των παραγόντων κινδύνου για στεφανιαία νόσο μεταξύ των ανδρών και των γυναικών κατά την εργασία στη Βόρεια Γερμανία. Χρησιμοποίησαν MLP (Multi Layer Perceptron) και PNN (Probabilistic Neural Networks) για την εκτίμηση του κινδύνου εμφράγματος του μυοκαρδίου ή οξύ στεφανιαίο θάνατο (στεφανιαίων επεισοδίων) κατά τη διάρκεια 10 ετών παρακολούθησης μεταξύ 5159 ανδρών ηλικίας 35 - 65 ετών κατά την πρόσληψη στο PROCAM. Σε αυτήν την ομάδα καταγράφηκαν 325 περιστατικά με στεφανιαία νόσο. Αξιολογήθηκε η απόδοση της κάθε διαδικασίας με τη μέτρηση της περιοχής στο πλαίσιο του δέκτη-καμπύλης των χαρακτηριστικών λειτουργίας (AUROC).

Το AUROC του MLP ήταν μεγαλύτερο από εκείνο του PNN (0.897 έναντι 0.872), και τα δύο υπερέβησαν τα AUROC για LR των 0.840. Αν 'ψηλός κίνδυνος' ορίζεται ως ένας κίνδυνος εκδήλωσης μεγαλύτερος του 20% σε 10 χρόνια, η LR κατατάσσεται (8.4% των ανδρών) ως υψηλού κινδύνου, 36.7% εκ των οποίων υπέστησαν μια εκδήλωση σε 10 χρόνια (45.8% του συνόλου των εκδηλώσεων). Το MLP 7.9% που ταξινομείται ως υψηλού κινδύνου, 64% εκ των οποίων υπέστησαν ένα γεγονός (74.5% του συνόλου των εκδηλώσεων), ενώ με την PNN, μόνο το 3.9% ήταν σε υψηλό κίνδυνο, 58.6% εκ των οποίων υπέστησαν ένα επεισόδιο (33.5% του συνόλου των εκδηλώσεων). Οι μελέτες δείχνουν ότι περίπου ένας στους τρεις με στεφανιαία επεισόδια μπορεί να προληφθεί στα 5 χρόνια με θεραπεία. Η ανάλυση δείχνει ότι η χρήση του MLP για να εντοπίζονται τα άτομα υψηλού κινδύνου, όπως οι υποψήφιοι για θεραπεία, θα επιτρέψει την πρόληψη του 25% των στεφανιαίων επεισοδίων σε μεσήλικες άνδρες, σε σύγκριση με το 15% και 11% με LR και PNN, αντίστοιχα.

Madigan *et al.*, 2006 [117]

Οι Madigan *et al.* [117] είχαν σκοπό σε αυτήν την έρευνα να κατανοήσουν την απόδοση των πρακτικών υγειονομικής περίθαλψης που γίνεται στα σπίτια στις ΗΠΑ. Οι σχέσεις μεταξύ

των παραγόντων των ασθενών της υγειονομικής περίθαλψης στο σπίτι και των χαρακτηριστικών του οργανισμού δεν είναι καλά κατανοητοί. Ειδικότερα, η απαλλαγή προορισμού και η διάρκεια παραμονής δεν έχουν μελετηθεί χρησιμοποιώντας μια προσέγγιση εξόρυξης δεδομένων που μπορούν να παρέχουν γνώσεις που δεν έχουν παραχθεί με παραδοσιακές στατιστικές αναλύσεις.

Τα στοιχεία ελήφθησαν από 2000 ασθενείς σε μια έρευνα για τρεις συγκεκριμένες συνθήκες (χρόνια αποφρακτική πνευμονική νόσος, καρδιακή ανεπάρκεια και η αντικατάσταση ισχίου), και αντιπροσωπεύουν περίπου 580 ασθενείς από ολόκληρη την Αμερική. Η προσέγγιση εξόρυξης δεδομένων που χρησιμοποίησαν ήταν ο αλγόριθμος CART (Classification and Regression Trees). Ο στόχος τους ήταν διπλός: i. ο καθορισμός των οδηγών των αποτελεσμάτων των υπηρεσιών υγειονομικής περίθαλψης στο σπίτι (απαλλαγή προορισμού και τη διάρκεια παραμονής) και ii. την εξέταση της εφαρμογής της επαγωγής με την εξόρυξη δεδομένων με τα στοιχεία της υγειονομικής περίθαλψης στο σπίτι.

Αποτελέσματα: ασθενείς ηλικίας 85 και άνω ήταν η κινητήρια δύναμη για την απαλλαγή προορισμού και τη διάρκεια παραμονής και για τις τρεις προϋποθέσεις.

Bayat *et al.*, 2009 [118]

Η μελέτη των Bayat *et al.* [118] συγκρίνει την αποτελεσματικότητα των δικτύων Bayesian και των δέντρων απόφασης για την πρόβλεψη πρόσβασης σε λίστα αναμονής για μεταμόσχευση νεφρού, σε ένα γαλλικό δίκτυο υγειονομικής περίθαλψης. Τα στοιχεία ήταν από 809 ασθενείς που ξεκίνησαν θεραπεία υποκατάστασης των νεφρών. Τα δεδομένα χωρίστηκαν τυχαία σε ένα σύνολο εκπαίδευσης (90%) και ένα σύνολο ελέγχου (10%). Το δίκτυο Bayesian και το δέντρο απόφασης CART χτίστηκαν για το σύνολο της εκπαίδευσης. Οι επιδόσεις τους συγκρίθηκαν με τα αποτελέσματα του συνόλου ελέγχου.

Η ηλικία βρέθηκε να είναι ο σημαντικότερος παράγοντας και στα δύο μοντέλα. Και τα δύο μοντέλα ήταν εξαιρετικά ευαίσθητα και συγκεκριμένα: ευαισθησία 90% (95% CI: 76.8 - 100), ειδικότητα 96.7% (95% CI: 92.2 - 100). Η παρουσίαση των αποτελεσμάτων των δύο μεθόδων έδειξε ότι τα δέντρα απόφασης ήταν πιο κατανοητά από τους γιατρούς. Οι προσεγγίσεις αυτές

παρέχουν γνώσεις σχετικά με την τρέχουσα διαδικασία φροντίδας. Αυτή η γνώση θα μπορούσε να χρησιμοποιηθεί για τη βελτιστοποίηση της διαδικασίας της υγειονομικής περίθαλψης.

Πίνακας 5.2: Μελέτες με δέντρα απόφασης για την αξιολόγηση των παραγόντων κινδύνου σε καρδιακά επεισόδια

Ερευνητής	Χρόνος	Αναφορά	Ιατρικό πρόβλημα	Αλγόριθμοι / Μέθοδοι	Αποτελέσματα
Tsien *et al.*	1998	[110]	Διάγνωση του εμφράγματος του μυοκαρδίου	C4.5, παλινδρόμηση	Απόδοση των μοντέλων
Colombet *et al.*	2000	[113]	Μοντέλα πρόβλεψης του καρδιαγγειακού κινδύνου	MLP CART	Αυτοί οι μέθοδοι μπορούν να συμπληρώσουν τα υπάρχοντα στατιστικά μοντέλα και να συμβάλουν στην ερμηνεία του κινδύνου
Gamberger *et al.*	2000	[114]	Είναι δυνατόν να αποφασίσει να επισκεφθεί κανείς ένα καρδιολόγο πριν ακόμα αντιληφθεί πραγματικά προβλήματα υγείας	μη συγκλίσεις σε ιατρικές ταξινομήσεις, και ακατάλληλο ορισμό του παράγοντα κινδύνου	περιγραφή των πέντε επιμέρους ομάδων του πληθυσμού με δυσανάλογα υψηλό κίνδυνο στεφανιαίας νόσου
Podgorelec *et al.*	2002	[115]	Η πραγμάτωση μιας σωστής απόφασης στην ιατρική	ID3, C4.5, CART	Επιτυχείς εναλλακτικές λύσεις για την παραδοσιακή προσέγγιση της επαγωγής
Voss *et al.*	2002	[116]	Εκτίμηση του κινδύνου της στεφανιαίας νόσου	MLP, PNN	Επιδόσεις των μοντέλων
Madigan *et al.*	2006	[117]	Κατ' οίκον περίθαλψη	CART	Οι ασθενείς άνω των 85 ετών έχουν διαφορετικά αποτελέσματα ανάλογα με τις συνθήκες που έχει ο καθένας
Bayat *et al.*	2009	[118]	Σύγκριση των Bayesian Networks και των δέντρων απόφασης	Bayesian Networks CART	Τα δέντρα απόφασης είναι πιο εύκολο να κατανοηθούν

MLP: Multi Layer Perception, PNN: Propabilistic Neural Network

5.3 Μελέτες με κανόνες συσχέτισης για αξιολόγηση των παραγόντων κινδύνου για καρδιακά επεισόδια

Σε αυτό το υποκεφάλαιο περιγράφονται εν συντομία μελέτες που έχουν γίνει για τους παράγοντες καρδιακών επεισοδίων με εξόρυξη δεδομένων, με την χρήση των κανόνων συσχέτισης. Ο Πίνακας 5.3 περιγράφει επιλεγμένες μελέτες και τη χρονολογία που έχουν γίνει. Παρατίθεται το ιατρικό πρόβλημα, ο αλγόριθμος / μέθοδος που χρησιμοποιήθηκε και τα αποτελέσματα που έχουν εξαχθεί.

Ordonez *et al*, 2001 [119]

Οι Ordonez *et al.* [119] στην εργασία αυτή περιγράφουν τις εμπειρίες τους για την ανακάλυψη των κανόνων συσχέτισης σε ιατρικά δεδομένα με στόχο την πρόβλεψη των καρδιακών παθήσεων. Έχουν επικεντρωθεί σε δύο θέματα σε αυτό το πρόγραμμα: τη χαρτογράφηση των ιατρικών δεδομένων σε μορφή πλειάδων κατάλληλα για τους κανόνες συσχέτισης και τον εντοπισμό χρήσιμων περιορισμών. Με βάση αυτές τις πτυχές εισάγουν ένα βελτιωμένο αλγόριθμο για να ανακαλύψουν περιορισμένους κανόνες συσχέτισης. Παρουσιάζουν ένα πειραματικό μέρος εξηγώντας αρκετούς ενδιαφέροντες κανόνες που ανακάλυψαν.

Karban *et al.*, 2004 [120]

Οι Karban *et al.* [120] αναφέρουν ότι ο κανόνας συσχέτισης εκφράζει τη σχέση μεταξύ παραδοχής και συνέπειας. Η μέθοδος αυτή είναι ευρύτερη σε σχέση με τους "κλασικούς" κανόνες συσχέτισης που εξάγονται από τον αλγόριθμο Apriori. Μπορούν να εκφραστούν διάφορα είδη επιπτώσεων ή ισοδυναμίας, καθώς και οι σχέσεις που αντιστοιχούν σε στατιστική έννοια υποθέσεων δοκιμών. Αυτή η έννοια των κανόνων συσχέτισης μπορεί να τροποποιηθεί για να περιγράψει ενδιαφέρουσες σχέσεις σε ζευγάρια των συνδεδεμένων συνόλων. Ο κανόνας SDS περιγράφει τη σχέση μεταξύ δύο συνόλων A και B και το δεδομένο

ιδιοκτησίας. Η συνήθης ερμηνεία, είναι να βρεθούν ζευγάρια στα σύνολα που διαφέρουν σημαντικά με το συγκεκριμένο χαρακτηριστικό ή να βρεθούν μεγάλες διαφορές μεταξύ των συνόλων A και B. Αυτό το άρθρο περιγράφει τα κίνητρα των κανόνων SDS και την ομοιότητα με τους κανόνες συσχέτισης. Ακολούθως περιγράφει τον τρόπο ορισμού των συνόλων A και B με τα χαρακτηριστικά που προέρχονται από την ανάλυση των δεδομένων μήτρας και δίνει κάποια αποτελέσματα αυτής της μεθόδου που εφαρμόζεται για ιατρικά δεδομένα.

Exarchos *et al.*, 2005 [121]

Οι Exarchos *et al.* [121] παρουσιάζουν μια αυτοματοποιημένη μέθοδο, η οποία ανιχνεύει παροδικά γεγονότα στο ηλεκτροεγκεφαλογράφημα και τα κατατάσσει ως επιληπτικά καρφιά, μυϊκή δραστηριότητα, τα μάτια να ανοιγοκλείνουν και την έντονη δραστηριότητα άλφα. Βασίζεται στους αλγόριθμους εξόρυξης δεδομένων και περιλαμβάνει τέσσερα στάδια: (I) προεπεξεργασία EEG και παροδική ανίχνευση εκδηλώσεων, (II) ομαδοποίηση των παροδικών εκδηλώσεων και εξαγωγή χαρακτηριστικών, (III) διακριτοποίηση των χαρακτηριστικών και (IV) τη συσχέτιση κανόνων εξόρυξης και την ταξινόμηση. Η μεθοδολογία έχει αξιολογηθεί με χρήση ενός συνόλου δεδομένων από 25 EEG και η παραγόμενη συνολική ακρίβεια είναι 84.35%. Το σημαντικότερο πλεονέκτημα αυτής της προσέγγισής είναι το γεγονός ότι είναι σε θέση να παρέχει ερμηνεία για τις αποφάσεις που λαμβάνονται, δεδομένου ότι βασίζεται σε ένα σύνολο κανόνων συσχέτισης.

Ordonez, 2006 [122]

Ο Ordonez σε μια άλλη εργασία του [122] επισημαίνει ότι οι κανόνες συσχέτισης αντιπροσωπεύουν μια ελπιδοφόρο τεχνική για να βρεθούν τα κρυμμένα πρότυπα σε ένα ιατρικό σύνολο δεδομένων. Το κύριο ζήτημα για τους κανόνες συσχέτισης σε μια ιατρική βάση δεδομένων, είναι ο μεγάλος αριθμός κανόνων που ανακαλύπτονται, το μεγαλύτερο μέρος των οποίων είναι άσχετο. Τέτοιος αριθμός κανόνων καθιστά την αναζήτηση αργή και την ερμηνεία από τον εμπειρογνώμονα δύσκολη. Σε αυτήν την εργασία, οι περιορισμοί

αναζήτησης εισάγονται για να βρουν μόνο τους ιατρικά σημαντικούς κανόνες συσχέτισης και να καταστήσουν την αναζήτηση αποδοτικότερη. Στους ιατρικούς όρους, οι κανόνες συσχέτισης αφορούν τις μετρήσεις αιματώματος καρδίας και τους παράγοντες κινδύνου σε ένα ασθενή, και το βαθμό στένωσης σε τέσσερις συγκεκριμένες αρτηρίες. Η ιατρική σημασία του κανόνα συσχέτισης αξιολογείται με τις συνηθισμένες μετρικές υποστήριξης και εμπιστοσύνης, αλλά και το lift. Οι κανόνες συσχέτισης συγκρίνονται με τους κανόνες που εξάγονται από τα δέντρα απόφασης, μια γνωστή τεχνική μηχανικής μάθησης. Τα δέντρα απόφασης, συγκρινόμενα με τους κανόνες συσχέτισης, αποδεικνύονται να είναι λιγότερο επαρκή για την πρόβλεψη αρτηριακών παθήσεων.

Τα πειράματα παρουσιάζουν ότι τα δέντρα απόφασης τείνουν να βρουν λίγους απλούς κανόνες, οι περισσότεροι κανόνες έχουν κάπως χαμηλή υποστήριξη, οι περισσότερες διασπάσεις ιδιοτήτων είναι διαφορετικές από τις ιατρικά κοινές διασπάσεις και οι περισσότεροι κανόνες αναφέρονται στα πολύ μικρά σύνολα ασθενών. Αντίθετα, οι κανόνες συσχέτισης περιλαμβάνουν γενικά τους απλούστερους κανόνες, λειτουργούν καλά, η υποστήριξη του κανόνα είναι ψηλότερη και οι κανόνες αναφέρονται γενικά στα μεγαλύτερα σύνολα ασθενών.

Kuo *et al.*, 2007 [123]

Οι Kuo *et al.* [123] αναφέρονται στον τομέα οντολογίας με γνώμονα την προσέγγιση για την εξόρυξη δεδομένων σε μια ιατρική βάση δεδομένων που περιέχει κλινικά δεδομένα για ασθενείς που υποβάλλονται σε θεραπεία με χρόνια νεφρική νόσο. Κάθε εγγραφή στο σύνολο δεδομένων αποτελείται από ένα μεγάλο αριθμό (έως 96) των ποσοτικών και ποιοτικών μετρήσεων που αντιπροσωπεύουν την φυσιολογική κατάσταση του συγκεκριμένου ασθενή σε μια συγκεκριμένη ημέρα της θεραπείας. Μια από τις προκλήσεις της εξόρυξης σε ένα τέτοιο σύνολο δεδομένων είναι το γεγονός ότι η σημασία πολλών χαρακτηριστικών δεν είναι εύκολα κατανοητή από κάποιον που δεν είναι εξοικειωμένος με τον τομέα της νεφρικής νόσου και της θεραπείας και δεν είναι σαφές ποια από τα χαρακτηριστικά είναι χρήσιμα στον τομέα της εξόρυξης δεδομένων.

Γίνεται διερεύνηση της δυνατότητας να χρησιμοποιηθεί στον τομέα της ιατρικής οντολογίας ως πηγή εξόρυξης γνώσης και έκφρασης της γνώσης σε μια χρήσιμη μορφή. Έχουν περιγράψει μια προσέγγιση στην οποία χρησιμοποιείται η οντολογία για να κατηγοριοποιήσει τα δεδομένα έτσι ώστε να τα προετοιμάσει για εξόρυξη κανόνων συσχέτισης στα δεδομένα. Οι κανόνες που εξάχθηκαν, εξετάστηκαν στη συνέχεια από έναν ειδικό, προκειμένου να αξιολογηθεί η χρησιμότητά τους. Καταλήγουν στο συμπέρασμα ότι η οντολογία με γνώμονα την εξόρυξη δεδομένων μπορεί να επιτύχει περισσότερα αποτελέσματα από ότι η απλή εξόρυξη δεδομένων.

Concaro *et al.*, 2009 [124]

Οι Concaro *et al.* [124] εστιάζουν την προσοχή τους στην παροχή ιατροφαρμακευτικής βοήθειας του Σακχαρώδη Διαβήτη. Έχουν δείξει την εφαρμογή ενός αλγορίθμου για την εξαγωγή χρονικών κανόνων συσχέτισης σε ακολουθίες με υβριδικά επεισόδια. Αυτή η μέθοδος επιτρέπει τη σωστή εκμετάλλευση της ενσωμάτωσης των διαφόρων πηγών πληροφόρησης υγειονομικής περίθαλψης, και μπορεί να χρησιμοποιηθεί για την αξιολόγηση της καταλληλότητας της ροής παροχής φροντίδας για συγκεκριμένες παθήσεις, προκειμένου να αξιολογήσει εκ νέου ή να βελτιώσει τις ακατάλληλες πρακτικές που οδηγούν σε μη ικανοποιητικά αποτελέσματα.

Πίνακας 5.3: Μελέτες με κανόνες συσχέτισης για αξιολόγηση των παραγόντων κινδύνου για καρδιακά επεισόδια

Ερευνητής	Χρόνος	Αναφορά	Ιατρικό πρόβλημα	Αλγόριθμοι/Μέθοδοι	Αποτελέσματα
Ordonez *et al.*	2001	[119]	Κανόνες συσχέτισης για πρόβλεψη καρδιακών παθήσεων	Χαρτογράφηση αλγόριθμου για κανόνες συσχέτισης	Οι κανόνες είναι υποχρεωμένοι να περιλαμβάνουν ένα μέγιστο αριθμό στοιχείων ώστε να καταστούν απλούστεροι και πιο γενικοί
Karban *et al.*	2004	[120]	SDS-κανόνες και κανόνες συσχέτισης	SDS	Ο εμπειρογνώμονας μπορεί ημι-αυτόματα να γενικεύσει ή να βελτιώσει τις γνώσεις που αποκτήθηκαν
Exarchos *et al.*	2005	[121]	EEG παροδική ανίχνευση επεισοδίου και ταξινόμηση	CAR, CBA	Αυτή η προσέγγιση δείχνει συγκρίσιμη ή καλύτερη απόδοση
Ordonez	2006	[122]	Σύγκριση κανόνων συσχέτισης και δέντρων απόφασης για την πρόβλεψη ενός επεισοδίου	κανόνες συσχέτισης και δέντρα απόφασης	Οι κανόνες συσχέτισης περιλαμβάνουν απλούστερους κανόνες, λειτουργούν καλά με τα ομαδοποιημένα χαρακτηριστικά του χρήστη, υπάρχει αξιοπιστία στους κανόνες και οι κανόνες αναφέρονται σε μεγαλύτερα σύνολα των ασθενών
Kuo *et al.*	2007	[123]	προσέγγιση μιας κύριας οντολογίας με γνώμονα την εξόρυξη δεδομένων σε μια ιατρική βάση δεδομένων που περιέχει κλινικά δεδομένα σχετικά με τους ασθενείς	κανόνες συσχέτισης	Βρέθηκαν 4 παράγοντες κινδύνου για θανάτους καρδιαγγειακών επεισοδίων
Concaro *et al.*	2009	[124]	Εξόρυξη κλινικών δεδομένων με διαβήτη με χρονικούς κανόνες συσχέτισης	TAR αλγόριθμος συσχέτισης	Η υποβληθείσα ανάλυση υπογραμμίζει τις κύριες δυνατότητες της εφαρμογής των χρονικών κανόνων συσχέτισης για την εξόρυξη των βάσεων δεδομένων

CAR: Class Association Rules CBA: Classification Based on Association

Κεφάλαιο 6: Μεθοδολογία

6.1 Γενικά

Έχοντας σαν δεδομένο τους παράγοντες που είναι σημαντικοί για την πρόκληση ενός καρδιακού επεισοδίου, όπως επίσης μια βάση δεδομένων με ασθενείς που έχουν υποστεί ένα καρδιακό επεισόδιο, θα πρέπει να δημιουργήσουμε ένα σύστημα απλό, εύκολο και γρήγορο, έτσι ώστε να το έχει κάθε ειδικός στη διάθεσή του και να το χρησιμοποιεί σε κάθε περίπτωση για να παίρνει γρήγορες και σωστές αποφάσεις.

Το πρόβλημα μπορούμε να το χωρίσουμε σε δύο μέρη:

i. την εφαρμογή των σταδίων της εξόρυξης δεδομένων και
ii. τη δημιουργία των μοντέλων για την εξαγωγή των κανόνων.

Μοντέλα που μελετήθηκαν:

Μετά από υποδείξεις των γιατρών δημιουργήθηκαν μοντέλα με βάσει τους παράγοντες και τα επεισόδια:

- 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)

Εξάχθηκαν κανόνες με δέντρα απόφασης, με πέντε διαφορετικές μετρικές διαχωρισμού και κανόνες συσχέτισης με τους αλγόριθμους Akamas και Apriori, υπολογίζοντας τη στατιστική σημαντικότητα και το ποσοστό πάθησης ενός επεισοδίου του κάθε κανόνα. Ο αλγόριθμος AKAMAS έχει σχεδιαστεί και υλοποιηθεί από εμάς για να αποφευχθούν τα προβλήματα που παρουσιάζει ο αλγόριθμος Apriori και να δίνει τους καλύτερους κανόνες, που αυτοί δεν θα είναι πολλές δεκάδες ή ακόμη και εκατοντάδες. Για την εξαγωγή αυτών των κανόνων χρησιμοποιήθηκε η μέθοδος holdout. Με τα αποτελέσματα της κάθε μεθόδου έγινε ο υπολογισμός των σημαντικότερων παραγόντων κινδύνου σε κάθε μοντέλο.

Για κάθε ασθενή έχει υπολογιστεί το ποσοστό του κινδύνου να πάθει ένα επεισόδιο με την εξίσωση του Framingham [125]. Σε κάθε κανόνα επιλέγονται όλοι οι ασθενείς που αντιστοιχούν σε αυτό τον κανόνα. Υπολογίζεται ο μέσος όρος του ποσοστού του κινδύνου για ένα επεισόδιο από τους ασθενείς που ανήκουν σε αυτό τον κανόνα. Το ποσοστό αυτό χωρίζεται σε τρεις κατηγορίες: χαμηλό (Low) από 0 μέχρι 5%, μεσαίο από 5% μέχρι 15% και ψηλό πάνω από 15%. Επιπλέον έχει υπολογιστεί το μέτρο χ^2 (chi-square) [126], το οποίο είναι καθοριστικό για να υπολογιστεί το μέτρο p-value [127], που δείχνει αν κάποιος κανόνας είναι στατιστικά σημαντικός ή όχι.

6.2 Βάση 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. Έτσι καταλήξαμε στα πεδία που αναφέρονται στο κεφάλαιο 6.3, τα οποία κωδικοποιήσαμε με βάση τις οδηγίες των γιατρών και των διεθνών και ευρωπαϊκών προδιαγραφών. Η αρχική βάση δεδομένων περιείχε 1200 ασθενείς με τριών ειδών καρδιαγγειακά νοσήματα: i. Έμφραγμα Μυοκαρδίου (MI), ii. Αγγειοπλαστική (PCI) και iii. 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 υπ' όψη οι ακόλουθες προϋποθέσεις:

i. 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.
ii. Παράγοντες που περικλείονταν σε άλλους παράγοντες δεν λήφθηκαν υπ' όψη, για παράδειγμα το ύψος και το βάρος του ασθενή που περιέχονται στον παράγοντα Δείκτη Μάζας Σώματος (ΔΜΣ).
iii. Παράγοντες που είχαν πολλές ελλιπείς τιμές και δεν υπήρχε η ευχέρεια ανάκτησης αυτών των τιμών, έχουν αφαιρεθεί.

Στον Πίνακα 6.1 που ακολουθεί, παρουσιάζεται μια γενική αναφορά και περιγραφή των πεδίων της βάσης δεδομένων.

Πίνακας 6.1: 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 (myocardial infarction, MI), 213 με αγγειοπλαστική (percutaneous coronary intervention, PCI) και 215 με στεφανιαία παράκαμψη (Coronary Artery Bypass Graft surgery, CABG). Ο Πίνακας 6.2 παρουσιάζει τη βάση δεδομένων κατά μοντέλο.

Πίνακας 6.2: Κατανομή περιπτώσεων ανά τάξη

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

6.3 Κωδικοποίηση των χαρακτηριστικών

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: i. κλινικοί και ii. βιοχημικοί. Επίσης οι παράγοντες μπορούν να διαχωριστούν σε δύο κατηγορίες: i. στους μεταβαλλόμενους και ii. 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 τρία μοντέλα:

i. μοντέλο με παράγοντες πριν από το επεισόδιο,

ii. μοντέλο με παράγοντες μετά το επεισόδιο και

iii. μοντέλο με παράγοντες πριν και μετά το επεισόδιο.

Ο Πίνακας 6.3 παρουσιάζει την κωδικοποίηση των χαρακτηριστικών που επιλέγηκαν για να μελετηθούν.

Πίνακας 6.3: 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

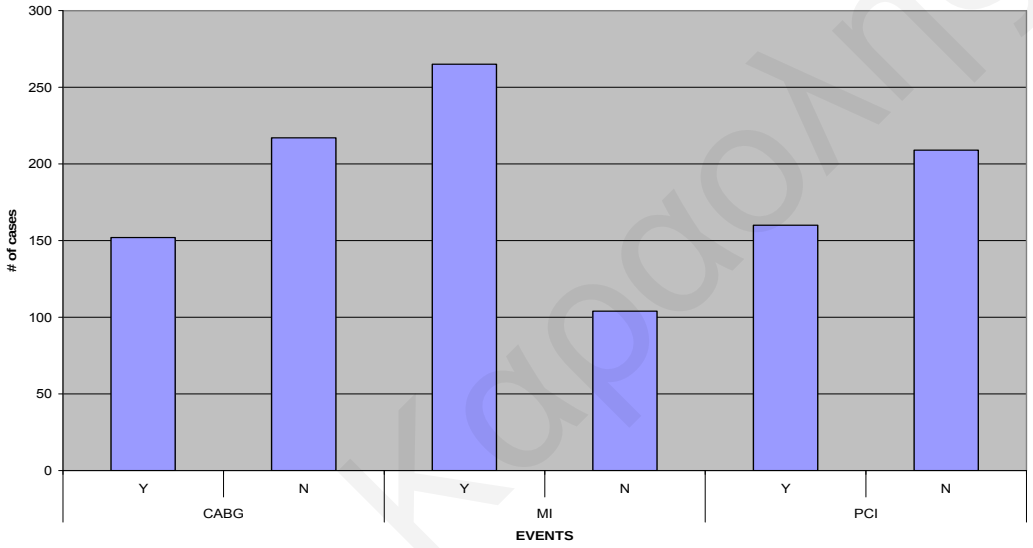
Στον Πίνακα 6.4 παρουσιάζεται η αριθμητική κατανομή των ασθενών που παρουσιάζουν το κάθε χαρακτηριστικό ανά κωδικοποίηση.

Πίνακας 6.4: Κατανομή περιστατικών βάσει των κωδικοποιημένων χαρακτηριστικών

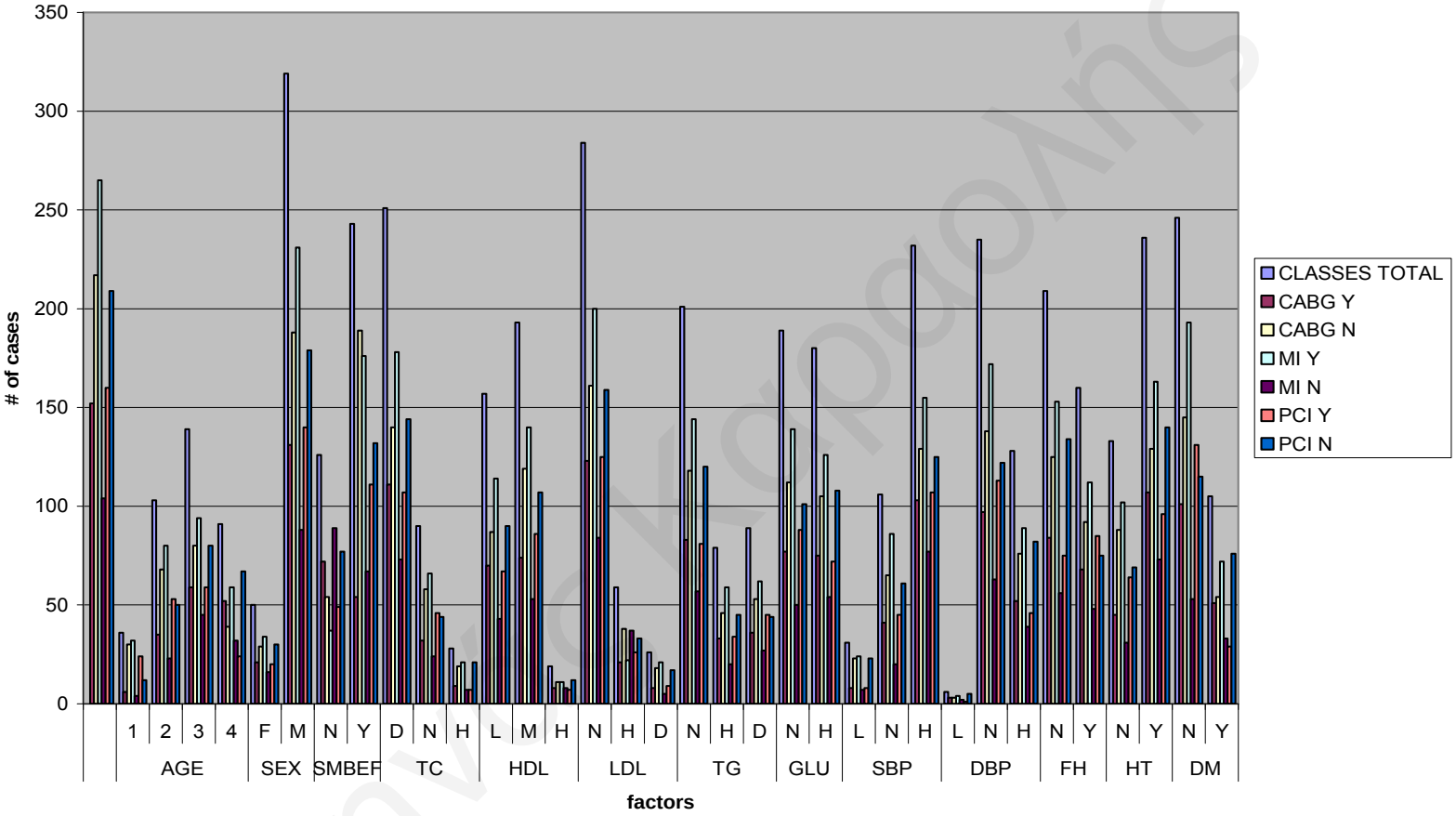
Τάξεις			MI		PCI		CABG	
	Τιμή	Αριθμός Περιστατικών	Y 265	N 104	Y 160	N 209	Y 152	N 217
AGE	1	36	32	4	24	12	6	30
	2	103	80	23	53	50	35	68
	3	139	94	45	59	80	59	80
	4	91	59	32	24	67	52	39
SEX	F	50	34	16	20	30	21	29
	M	319	231	88	140	179	131	188
SMBEF	N	126	37	89	49	77	72	54
	Y	243	176	67	111	132	54	189
TC	D	251	178	73	107	144	111	140
	N	90	66	24	46	44	32	58
	H	28	21	7	7	21	9	19
HDL	L	157	114	43	67	90	70	87
	M	193	140	53	86	107	74	119
	H	19	11	8	7	12	8	11
LDL	N	284	200	84	125	159	123	161
	H	59	22	37	26	33	21	38
	D	26	21	5	9	17	8	18
TG	N	201	144	57	81	120	83	118
	H	79	59	20	34	45	33	46
	D	89	62	27	45	44	36	53
GLU	N	189	139	50	88	101	77	112
	H	180	126	54	72	108	75	105
SBP	L	31	24	7	8	23	8	23
	N	106	86	20	45	61	41	65
	H	232	155	77	107	125	103	129
DBP	L	6	4	2	1	5	3	3
	N	235	172	63	113	122	97	138
	H	128	89	39	46	82	52	76
FH	N	209	153	56	75	134	84	125
	Y	160	112	48	85	75	68	92
HT	N	133	102	31	64	69	45	88
	Y	236	163	73	96	140	107	129
DM	N	246	193	53	131	115	101	145
	Y	105	72	33	29	76	51	54

Το Σχήμα 6.1 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

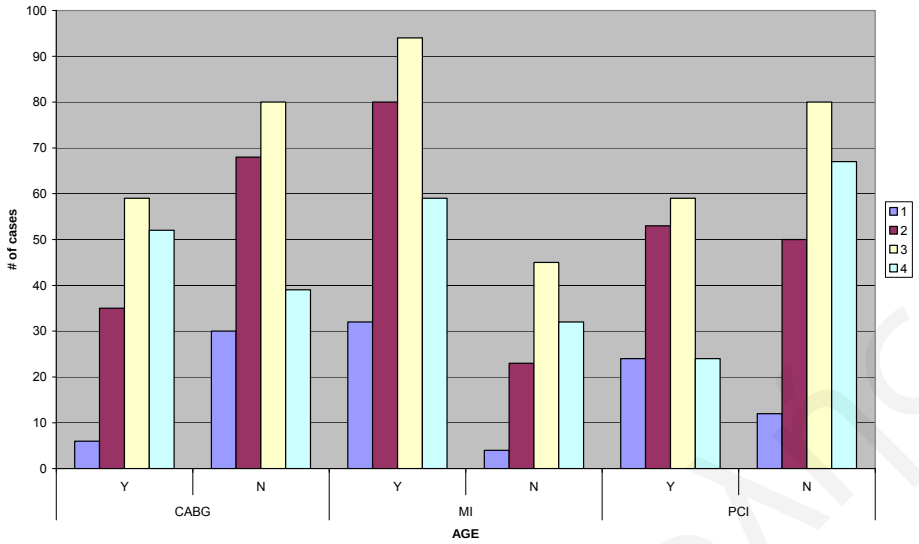
παρουσιάζονται περισσότερες δοσοληψίες με τιμή ‘Y’ παρά με ‘N’. Ενώ το Σχήμα 6.2 παρουσιάζει τον αριθμό των ασθενών για κάθε χαρακτηριστικό. Παρατηρούμε ότι παρουσιάζονται περισσότερες περιπτώσεις να είναι άνδρας (SEX = M), και πολλές περιπτώσεις με κανονική ποσότητα των λιποπρωτεϊνών χαμηλής πυκνότητας στο αίμα (LDL = N). Στα Σχήματα 6.3 μέχρι 6.15 παρουσιάζονται οι κατανομές των παραγόντων για τα τρία επεισόδια, CABG, MI και PCI.



Σχήμα 6.1: Κατανομή περιστατικών για τις τάξεις CABG, MI και PCI

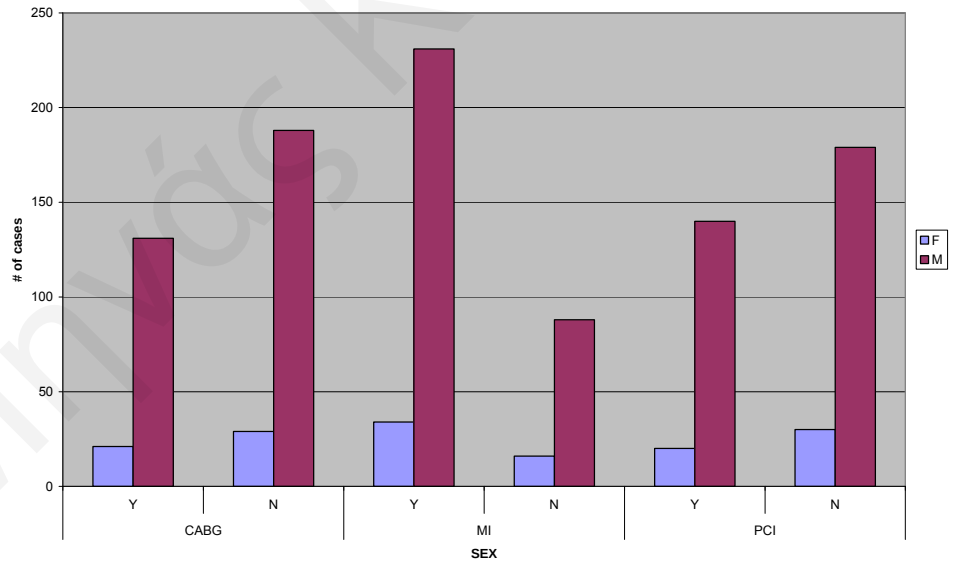


Σχήμα 6.2: Αριθμός περιστατικών έναντι κωδικοποιημένων χαρακτηριστικών για τις τάξεις CABG, MI και PCI



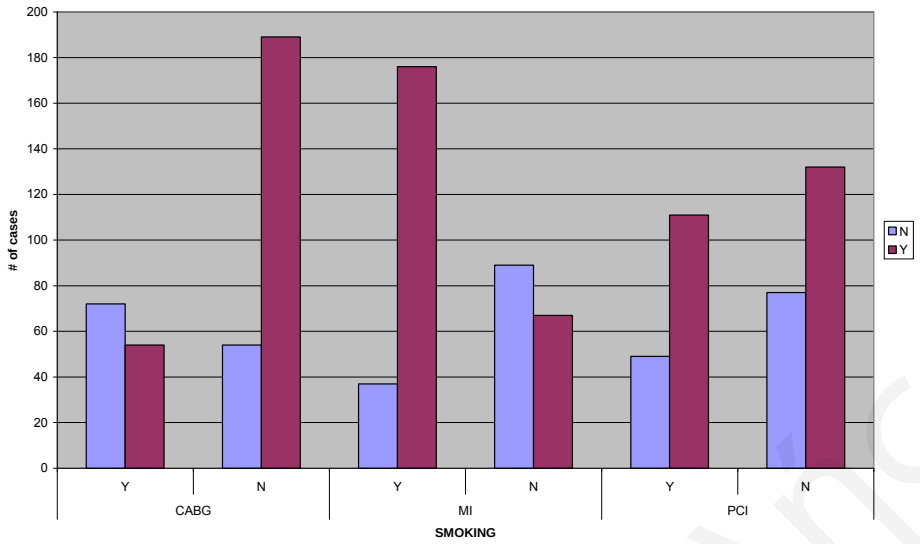
Ηλικίες: 1: 34-50, 2: 51-60, 3: 61-70 και 4: 71-85

Σχήμα 6.3: Κατανομή περιστατικών ανά κωδικοποίηση ηλικίας στις τάξεις CABG, MI και PCI

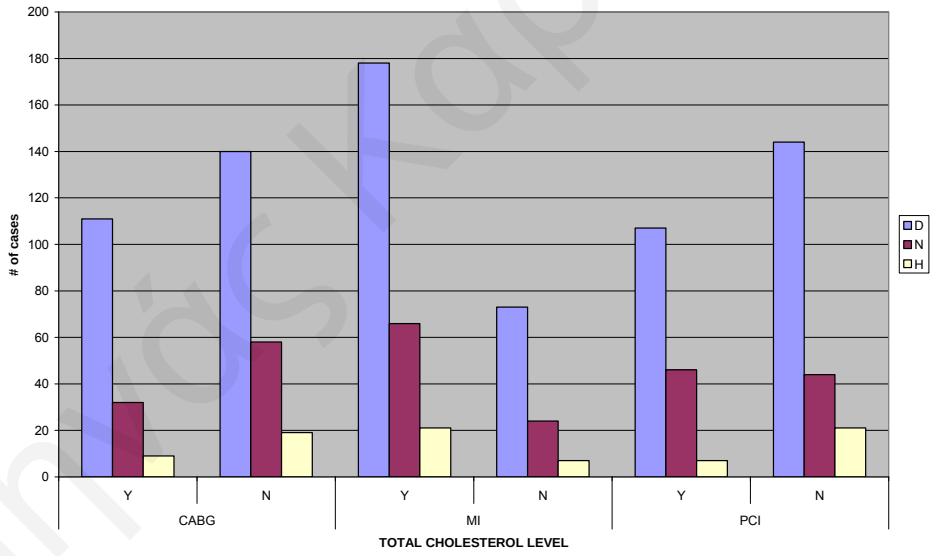


Φύλο: M:άντρες και F: γυναίκες

Σχήμα 6.4: Κατανομή περιστατικών ανά φύλο στις τάξεις CABG, MI και PCI

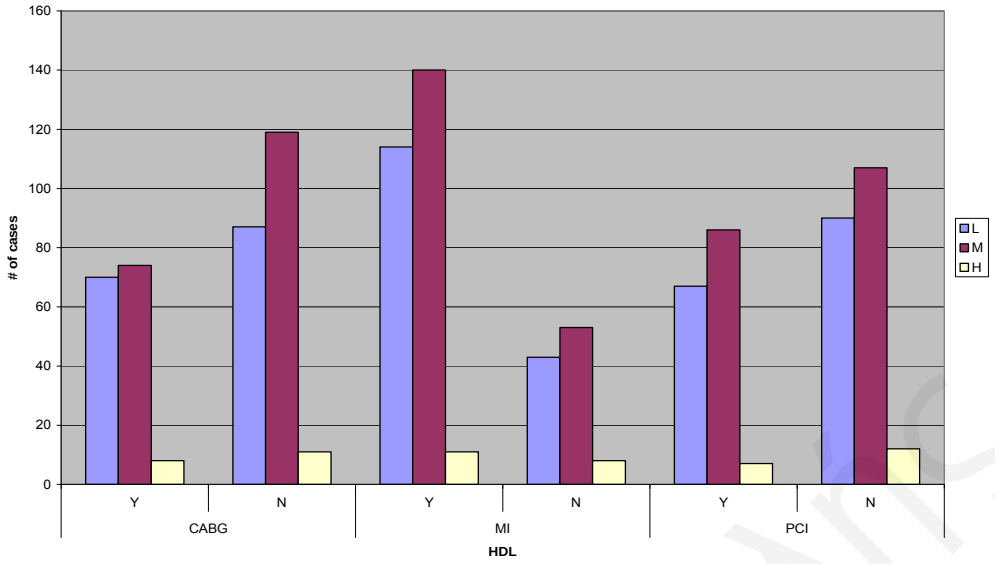


Σχήμα 6.5: Κατανομή περιστατικών με το χαρακτηριστικό κάπνισμα στις τάξεις CABG, MI και PCI



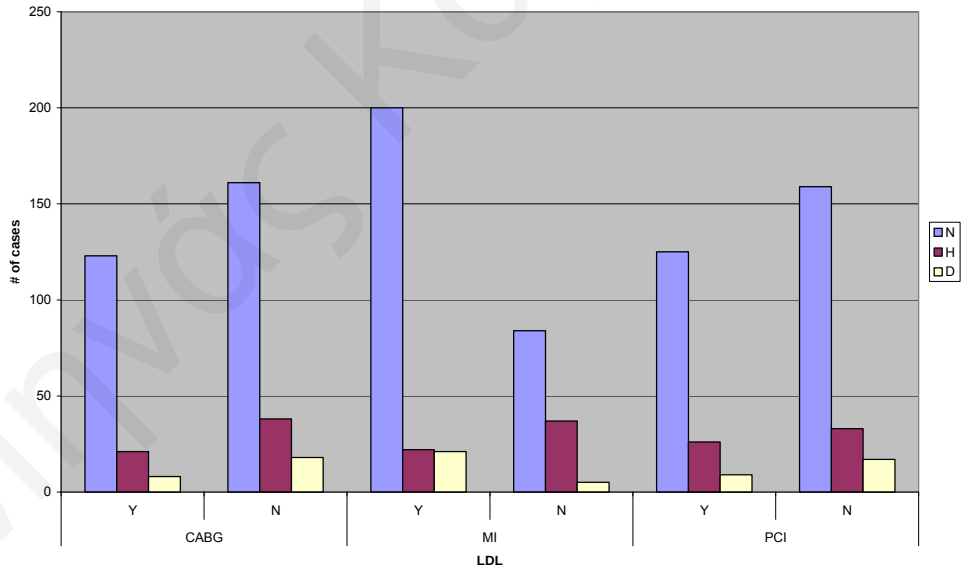
Ολική χοληστερόλη: D: χαμηλή, N: κανονική και H: επικίνδυνη

Σχήμα 6.6: Κατανομή περιστατικών με το χαρακτηριστικό ολικής χοληστερόλης (TC) στις τάξεις CABG, MI και PCI



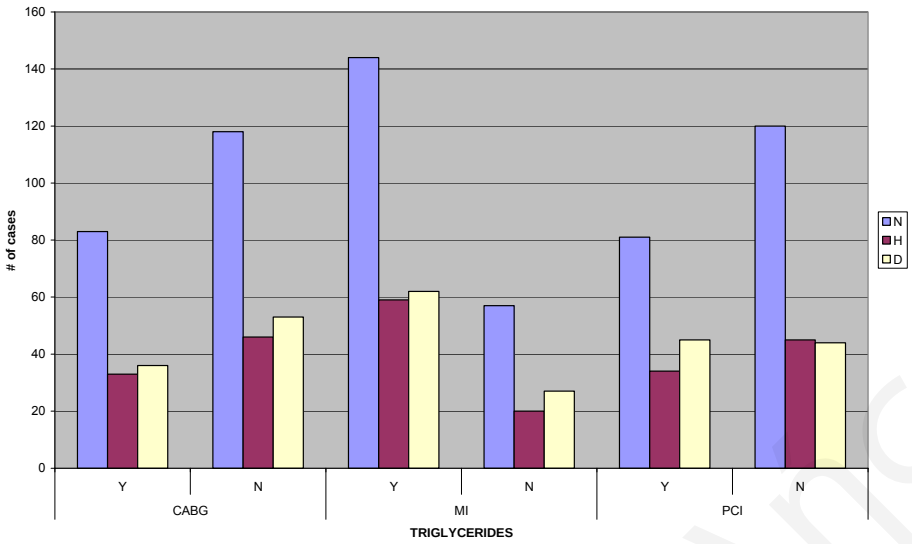
HDL: L: χαμηλή, M: κανονική και H: επικίνδυνη

Σχήμα 6.7: Κατανομή περιστατικών με το χαρακτηριστικό HDL στις τάξεις CABG, MI και PCI



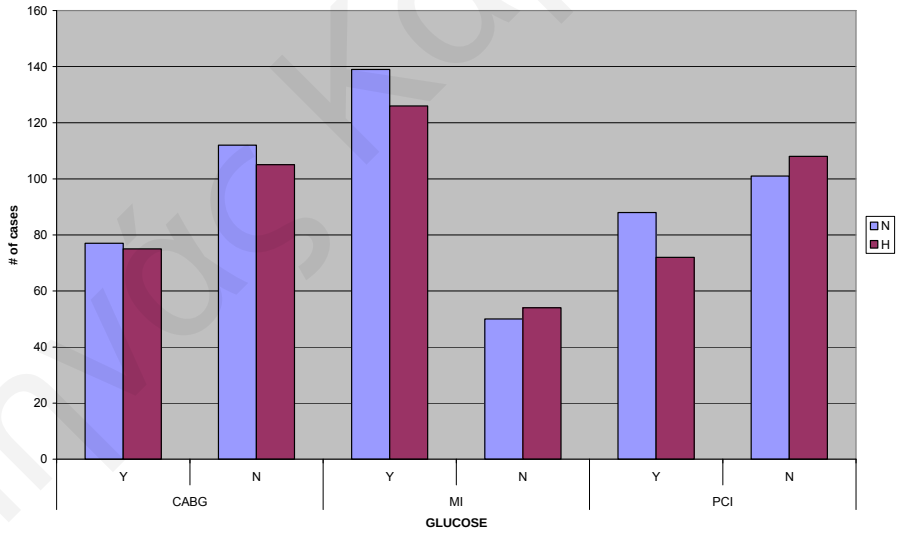
LDL: D: χαμηλή, N: κανονική και H: επικίνδυνη

Σχήμα 6.8: Κατανομή περιστατικών με το χαρακτηριστικό LDL στις τάξεις CABG, MI και PCI



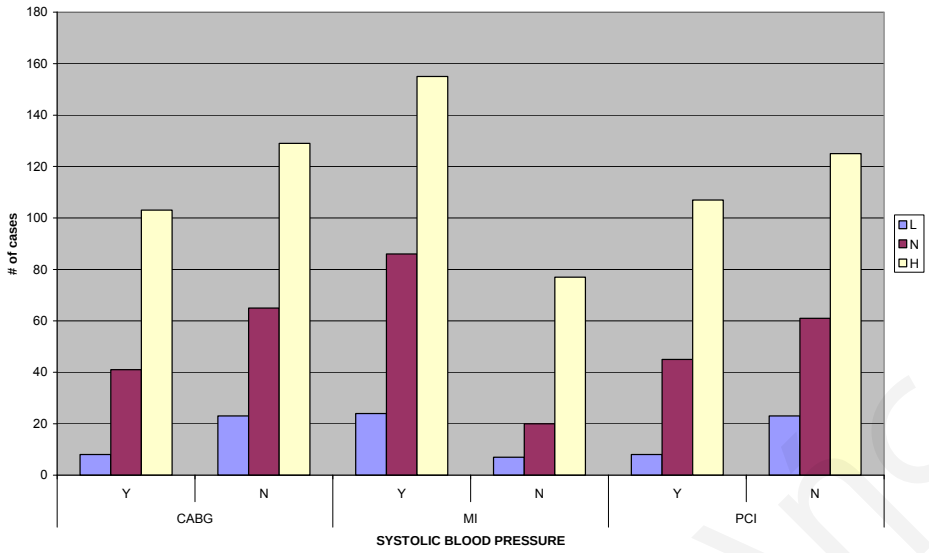
TG: D: χαμηλή, N: κανονική και H: επικίνδυνη

Σχήμα 6.9: Κατανομή περιστατικών με το χαρακτηριστικό TG στις τάξεις CABG, MI και PCI



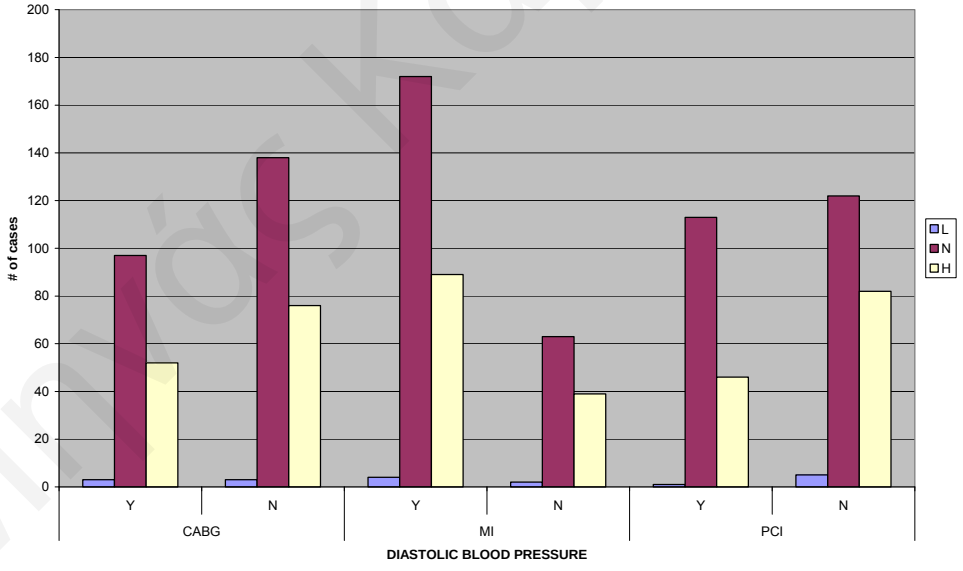
GLU: N: κανονική και H: επικίνδυνη

Σχήμα 6.10: Κατανομή περιστατικών με το χαρακτηριστικό GLU στις τάξεις CABG, MI και PCI



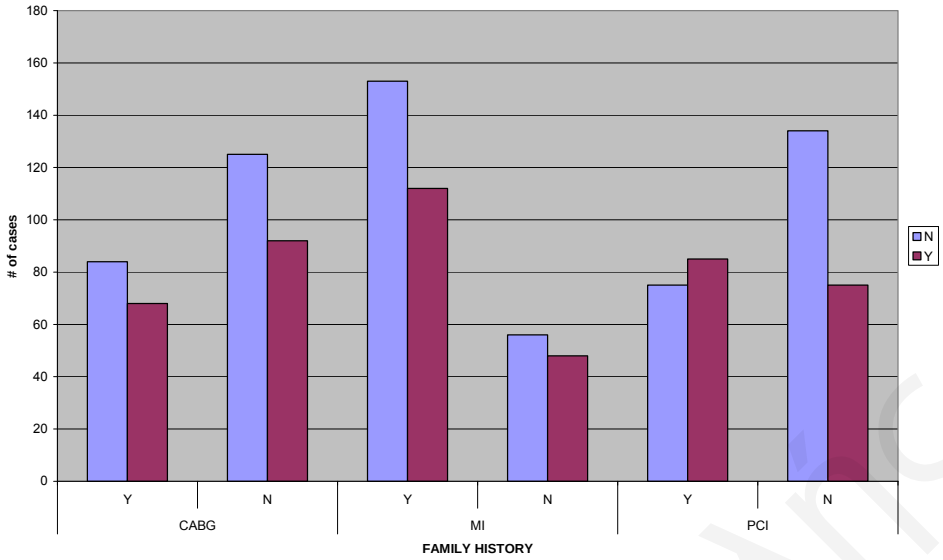
SBP: L: χαμηλή, M: κανονική και H: ψηλή

Σχήμα 6.11: Κατανομή περιστατικών με το χαρακτηριστικό SBP στις τάξεις CABG, MI και PCI

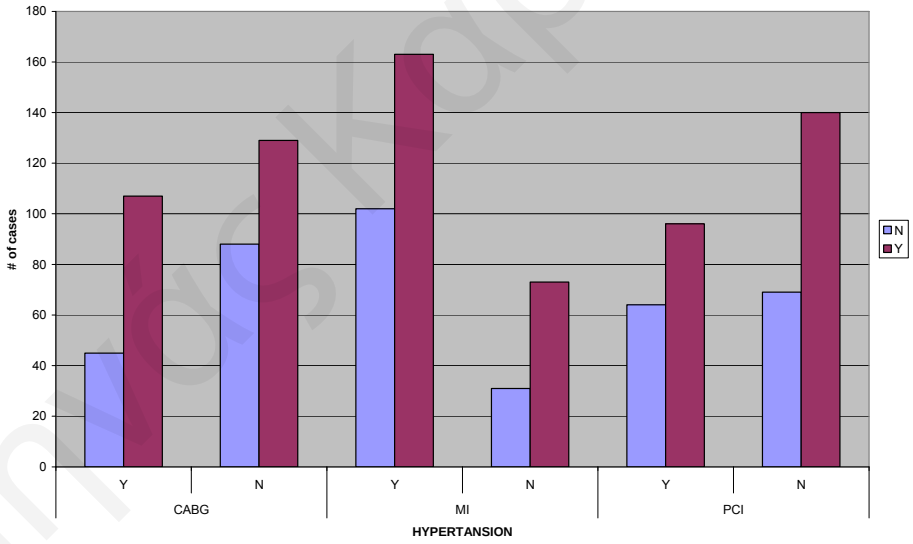


DBP: L: χαμηλή, M: κανονική και H: ψηλή

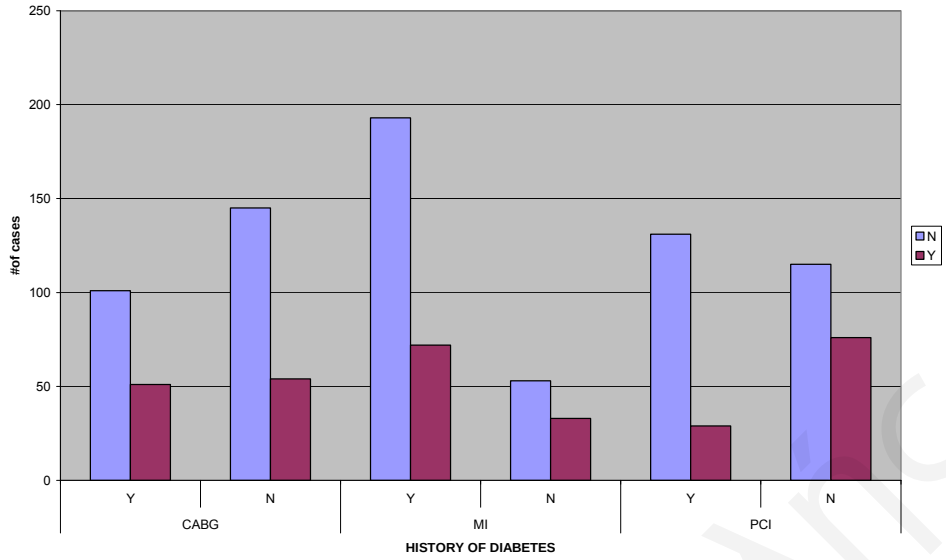
Σχήμα 6.12: Κατανομή περιστατικών με το χαρακτηριστικό DBP στις τάξεις CABG, MI και PCI



Σχήμα 6.13: Κατανομή περιστατικών με το χαρακτηριστικό FH στις τάξεις CABG, MI και PCI



Σχήμα 6.14: Κατανομή περιστατικών με το χαρακτηριστικό HT στις τάξεις CABG, MI και PCI



Σχήμα 6.15: Κατανομή περιστατικών με το χαρακτηριστικό DM στις τάξεις CABG, MI και PCI

6.4 Δέντρα απόφασης

Τα δέντρα απόφασης κατασκευάζονται χρησιμοποιώντας μόνο εκείνα τα γνωρίσματα που είναι σε θέση να διακρίνουν τις έννοιες προς εκμάθηση. Για να χτίσουμε ένα δέντρο απόφασης, πρέπει αρχικά να επιλέξουμε ένα υποσύνολο περιπτώσεων από το σύνολο των δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση (υποσύνολο δεδομένων εκπαίδευσης-trainingset). Αυτό το υποσύνολο (δεδομένα ελέγχου-testset) χρησιμοποιείται έπειτα από τον αλγόριθμο για να κατασκευάσει το δέντρο απόφασης. Τα υπόλοιπα δεδομένα, τα δεδομένα trainingset, χρησιμοποιούνται στην εξέταση της ακρίβειας του κατασκευασμένου δέντρου. Εάν το δέντρο απόφασης ταξινομεί τις περιπτώσεις σωστά, η διαδικασία ολοκληρώνεται. Εάν μια περίπτωση είναι ανακριβώς ταξινομημένη, η περίπτωση προστίθεται στο επιλεγμένο υποσύνολο των trainingset και ένα νέο δέντρο κατασκευάζεται.

6.4.1 Αλγόριθμος C4.5

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 [13]. Στο Σχήμα 6.16 περιγράφονται τα βήματα εκτέλεσης του αλγορίθμου, ο οποίος είναι βασισμένος στον αλγόριθμο C4.5. Η διαφορά του αλγόριθμου που περιγράφεται από τον αλγόριθμο C4.5 είναι ότι ο χρήστης έχει την επιλογή του κριτήριου διαχωρισμού, από τα πέντε κριτήρια που έχουν υλοποιηθεί [128].

Αλγόριθμος δημιουργίας δέντρου απόφασης: C4.5 βασισμένος στο [6]

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 6.16: Ψευδοκώδικας αλγόριθμου δημιουργίας δέντρου απόφασης

6.4.2 Κλάδεμα (Pruning)

Στον αλγόριθμο δέντρων απόφασης μπορεί να αναπτυχθεί το κάθε κλαδί αρκετά έτσι ώστε να ταξινομηθούν σωστά τα παραδείγματα εκπαίδευσης. Παρόλο που αυτή είναι μια καλή στρατηγική στην ουσία μπορεί να οδηγήσει σε δυσκολίες όταν υπάρχει θόρυβος στα δεδομένα ή όταν το δείγμα του συνόλου εκπαίδευσης είναι πολύ μικρό. Και στις δύο αυτές περιπτώσεις ο αλγόριθμος μπορεί να δημιουργήσει ένα δέντρο το οποίο να παρουσιάζει υπερεκπαίδευση (over fitting). Η υπερεκπαίδευση είναι μια σημαντική δυσκολία για την κατασκευή δέντρων απόφασης όπως και σε άλλες μεθόδους. Γι' αυτό υλοποιήθηκε το κλάδεμα για την αποφυγή αυτής της κατάστασης [129].

Υπάρχουν δύο τρόποι για την πρόληψη της υπερεκπαίδευσης [130]:

i. Ο αλγόριθμος κλαδέματος από κάτω προς τα πάνω [131] έχει δύο φάσεις: Στην πρώτη φάση, τη φάση της ανάπτυξης, δημιουργείται ένα βαθύ δέντρο. Στη δεύτερη φάση, τη φάση του κλαδέματος, αυτό το δέντρο κλαδεύεται για αποφυγή της υπερεκπαίδευσης.
ii. Ο αλγόριθμος κλαδέματος από πάνω προς τα κάτω [132] χρησιμοποιεί ένα κριτήριο το οποίο είναι αρμόδιο για τη διακοπή της ανάπτυξης του δέντρου όταν αυτό θεωρηθεί αναγκαίο, βάσει του υπολογισμού αυτού του ιδίου.

Επιλέχθηκε ο αλγόριθμος κλαδέματος από κάτω προς τα πάνω χρησιμοποιώντας τον υπολογισμό του λάθους με τη μέθοδο του Laplace. Όταν δημιουργείται το δέντρο και από τον κόμβο παράγεται το φύλλο, γίνεται ο υπολογισμός του λάθους:

$$E(D) = \frac{N - n + k - 1}{N + k}$$

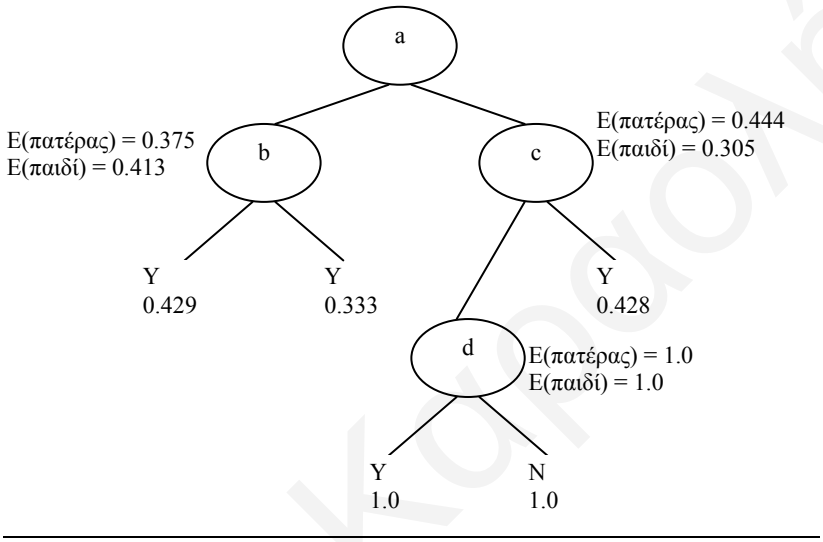
$$D => σύνολο\_δεδομένων$$
$$C => μεγαλύτερη\_τιμή\_τάξης\_στο\_D$$
$$k => αριθμός\_των\_τιμών\_της\_τάξης$$
$$N => αριθμός\_των\_περιπτώσεων\_στο\_D$$
$$n => αριθμός\_των\_περιπτώσεων\_που\_έχουν\_τιμή\_τάξης\_C$$

(Εξίσωση 6.3)

Κατά τη διάρκεια που ο αλγόριθμος γυρίζει πίσω στη ρίζα του δέντρου, το λάθος του κλαδιού περνά στον κόμβο του πατέρα. Ο κόμβος του πατέρα υπολογίζει το συνολικό λάθος όλων των

παιδιών του όπως επίσης και το δικό του. Αν το λάθος του πατέρα είναι μικρότερο από το συνολικό λάθος των παιδιών του, τότε ο κόμβος πατέρας κλαδεύεται και αντικαθίσταται από το παιδί που έχει τη μεγαλύτερη τιμή της τάξης. Αν το λάθος του πατέρα είναι μεγαλύτερο από το συνολικό λάθος των παιδιών του, τότε δεν χρειάζεται περισσότερο κλάδεμα σε αυτό το μονοπάτι και η τιμή του λάθους που επιστρέφει είναι μηδέν.

Παράδειγμα:



Σχήμα 6.17: Παράδειγμα κλαδέματος

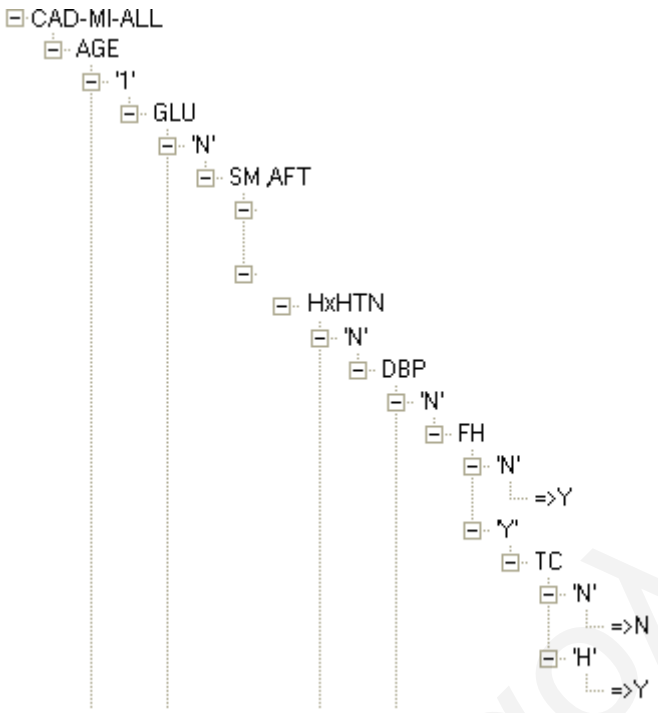
Ο κόμβος b έχει 4 περιπτώσεις με την τιμή της τάξης να είναι Y και 2 περιπτώσεις με την τιμή της τάξης να είναι N. Γι' αυτό το λάθος του είναι $E(b) = \frac{N - n + k - 1}{N + k} = \frac{6 - 4 + 2 - 1}{6 + 2} = \frac{3}{8} = 0.375$. Τα παιδιά του έχουν τιμές λάθους 0.429 και 0.333. Ο συνολικός αριθμός λάθους των παιδιών είναι $E(b's\_\text{παιδιά}) = (\frac{5}{6}) \times 0.429 + (\frac{1}{6}) \times 0.333 = 0.413$. Το E(b) είναι μικρότερο από το E(b's παιδιά), γι' αυτό ο κόμβος b κλαδεύεται και αντικαθίσταται με το παιδί που έχει τη μεγαλύτερη τιμή της τάξης.

Ο κόμβος d έχει 1 περίπτωση με τιμή Y και 1 περίπτωση με τιμή N. Γι' αυτό το λάθος του είναι $E(d) = \frac{N - n + k - 1}{N + k} = \frac{2 - 1 + 2 - 1}{2 + 2} = \frac{4}{4} = 1.0$. Τα παιδιά του έχουν το ίδιο λάθος 1.0. Το ολικό λάθος των παιδιών είναι $E(d's\_παιδιά) = (\frac{1}{2}) \times 1.0 + (\frac{1}{2}) \times 1.0 = 1.0$. Επειδή τα E(d) και E(d's παιδιά) είναι ίσα, δεν υπάρχει κλάδεμα. Γι' αυτό στο ίδιο μονοπάτι δεν γίνεται άλλο κλάδεμα και ο κόμβος d δίνει μηδενικό λάθος στον κόμβο c.

Ο κόμβος c έχει 4 περιπτώσεις με την τιμή της τάξης να είναι Y και 3 περιπτώσεις με την τιμή της τάξης να είναι N. Το λάθος του κόμβου c είναι $E(c) = \frac{N - n + k - 1}{N + k} = \frac{7 - 4 + 2 - 1}{7 + 2} = \frac{3}{9} = 0.444$. Τα παιδιά του έχουν το λάθος 0.0 και 0.428. Το συνολικό λάθος των παιδιών είναι $E(c's\_παιδιά) = (\frac{2}{7}) \times 0.0 + (\frac{5}{7}) \times 0.428 = 0.305$. Το E(c) έχει μεγαλύτερο λάθος από τα παιδιά του και έτσι δεν γίνεται κλάδεμα.

6.4.3 Εξαγωγή κανόνων

Όταν τελειώσει το κτίσιμο του δέντρου, τότε εξάγονται οι κανόνες. Για κάθε μονοπάτι του δέντρου δημιουργείται ένας κανόνας όπως δίδεται στο Σχήμα 6.18.



Σχήμα 6.18: Τμήμα δέντρου απόφασης εξαγόμενο από το μοντέλο MI

(για την κωδικοποίηση των χαρακτηριστικών βλέπε Πίνακα 6.3).

Κανόνες που εξάγονται βάσει του Σχήματος 6.18:

AGE=’1’ AND GLU = ‘N’ AND SMAFT=’N’ → MI =’YES’

AGE=’1’ AND GLU = ‘N’ AND SMAFT=’Y’

AND HxHTN=’N’ AND DBP=’N’ AND FH=’N’ → MI =’YES’

AGE=’1’ AND GLU = ‘N’ AND SMAFT=’Y’

AND HxHTN=’N’ AND DBP=’N’ AND FH=’Y’

AND TC=’N’ → MI =’NO’

AGE=’1’ AND GLU = ‘N’ AND SMAFT=’Y’

AND HxHTN=’N’ AND DBP=’N’ AND FH=’Y’

AND TC=’Y’ → MI =’YES’

6.4.4 Αξιολόγηση

Όταν δημιουργηθεί το δέντρο και εξαχθούν οι κανόνες, τότε διεξάγεται η αξιολόγηση. Στην αρχή οι περιπτώσεις χωρίζονται σε δύο σύνολα, το σύνολο της εκπαίδευσης και το σύνολο της ελέγχου. Το σύνολο της εκπαίδευσης χρησιμοποιείται για την κατασκευή του δέντρου. Το σύνολο της ελέγχου χρησιμοποιείται για την αξιολόγηση των αποτελεσμάτων. Το ποσοστό των δύο συνόλων που χρησιμοποιήθηκε σε αυτή την εργασία είναι 50% και 50% αντίστοιχα (βλέπε Πίνακα 6.21). Έχουν χρησιμοποιηθεί οι μέθοδοι holdout και 10-fold cross validation.
Για κάθε περίπτωση στο σύνολο της ελέγχου λαμβάνεται υπόψη η τιμή της τάξης. Αν ο κανόνας που έχει εξαχθεί από το δέντρο απόφασης έχει προβλέψει σωστά την τιμή της τάξης, τότε αυτή η περίπτωση είναι σωστή.
Χρησιμοποιώντας την ίδια διαδικασία, δημιουργείται η σύγχυση μήτρας (confusion matrix). Η σύγχυση μήτρας είναι ένας πίνακας που συγκρίνει την αναμενόμενη τιμή με αυτή που βρέθηκε βάσει του μοντέλου (Βλέπε Πίνακα 4.1) [95].
Τέλος, χρησιμοποιώντας την σύγχυση μήτρας υπολογίζονται τα μέτρα ακρίβειας για κάθε τιμή της τάξης [126].

6.5 Αλγόριθμοι συσχέτισης

6.5.1 Αλγόριθμος Apriori

Ο αλγόριθμος Apriori έχει προταθεί από τους R. Agrawal και R. Srikant το 1994 [8]. 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 αναζήτηση, level-wise, η οποία είναι μια επαναλαμβανόμενη τεχνική που χρησιμοποιεί τα k-itemsets 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 χαρακτηριστικό στη βάση δεδομένων, και μετά συλλέγει τα χαρακτηριστικά που ικανοποιούν την ελάχιστη υποστήριξη, 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, μέχρι που να μην μπορεί να βρεθεί άλλο σύνολο με k-itemsets, 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 υποστήριξης (min_sup), τότε το I δεν είναι συχνό, P(I) < min_sup. Εάν το χαρακτηριστικό A προστίθεται στο σύνολο χαρακτηριστικών 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 $k \geq 2$, 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 χαρακτηριστικά (k-itemsets) από την ένωση του συνόλου L_{k-1} με τον εαυτό του. Το σύνολο με τα υποψήφια σύνολα χαρακτηριστικών καλείται C_k. Εάν το l_i είναι μέλος του L_{k-1}, τότε το $l_i[j]$ αναφέρεται στο χαρακτηριστικό j του συνόλου χαρακτηριστικών l_i. Ο Apriori θεωρεί ότι τα χαρακτηριστικά στα σύνολα είναι ταξινομημένα σε αλφαβητική σειρά. Για κάποιο σύνολο χαρακτηριστικών l_i με (k-1) χαρακτηριστικά, τα χαρακτηριστικά 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) χαρακτηριστικά 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 χαρακτηριστικών που ανήκουν στο Ck, μπορεί να είναι συχνά εμφανιζόμενα κι άλλα όχι, όμως όλα τα συχνά εμφανιζόμενα σύνολα k χαρακτηριστικών (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 την ελάχιστη υποστήριξη, τότε αυτό το σύνολο χαρακτηριστικών προστίθεται στο Lk. Αυτό γίνεται, όπως αναφέρει η Apriori ιδιότητα, οποιοδήποτε (k-1)-itemset σύνολο χαρακτηριστικών δεν είναι συχνό τότε δεν μπορεί να είναι υποσύνολο κάποιου k-itemset σύνολο χαρακτηριστικών. Έτσι επειδή το σύνολο Ck, μπορεί να γίνει αρκετά μεγάλο, τα σύνολα χαρακτηριστικών που δεν είναι συχνά αφαιρούνται.

6.5.1.1 Περιγραφή ψευδοκώδικα αλγόριθμου Apriori

Στο Σχήμα 6.19 παρουσιάζεται ο ψευδοκώδικας του αλγόριθμου Apriori και οι σχετικές διαδικασίες [133]:

i. Καταρχάς ο αλγόριθμος δέχεται μια βάση δεδομένων με δοσοληψίες. Η βάση δεδομένων δοσοληψιών 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 που συνθέτουν την δοσοληψία.

ii. Στο πρώτο βήμα βρίσκονται όλα τα συχνά σύνολα αντικειμένων με 1 χαρακτηριστικό και φυλάγονται στο σύνολο L1

iii. Στο βήμα 3 γίνεται 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 χαρακτηριστικών και μετά χρησιμοποιεί την ιδιότητα Apriori για να αφαιρέσει αυτά που δεν είναι συχνά.

iv. Στα βήματα 4 – 10 γίνεται αναζήτηση στη βάση δεδομένων, για να υπολογιστούν τα σύνολα χαρακτηριστικών που εμφανίζονται στην βάση. Στο βήμα 9 ο αλγόριθμος βρίσκει τα υποψήφια σύνολα χαρακτηριστικών που έχουν μεγαλύτερο από την ελάχιστη υποστήριξη και τα προσθέτει στο σύνολο Lk.

v. Στο τελικό βήμα 11 γίνεται η ένωση όλων των συχνών συνόλων χαρακτηριστικών των Lk στο L. Έτσι μετά τη διαδικασία για εξαγωγή κανόνων συσχέτισης ο αλγόριθμος μπορεί να χρησιμοποιήσει το σύνολο L.

Αλγόριθμος: Apriori. Εύρεση των συχνών συνόλων χαρακτηριστικών (itemsets) χρησιμοποιώντας την επαναλαμβανόμενη τεχνική level-wise βασισμένη στα παραγωγή υποψηφίων βασισμένο στο [8].

Είσοδος:

- D, βάση δεδομένων με δοσοληψίες
- min-sup, ο ελάχιστος αριθμός υποστήριξης.

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

Μέθοδος:

(1) L_1 =find_frequent_l-itemsets(D);
(2) for(k = 2;$L_{k-1} \neq 0$;k++){
(3) C_k = apriori_gen(L_{k-1});
(4) for each transaction $t \in D$ { //scan D for counts
(5) C_t = subset(C_k,t); //get the subsets of t that are candidates
(6) for each candidate $c \in C_t$
(7) c.count++;
(8) }
(9) $L_k = \{c \in C_k | c.count \geq min\text{-}sup\}$
(10) }
(11) return $L = U_k L_k$;

procedure apriori_gen(L_{k-1} :frequent (k-1)-itemsets)
(1) for each itemset $l_1 \in L_{k-1}$
(2) for each itemset $l_2 \in L_{k-1}$
(3) if(($l_1[1]=l_2[1]$) ∧ ($l_1[2 =l_2[2]$) ∧ … ∧ (l_1 [k-2]= l_2 [k-2]) ∧ (l_1 [k-1]]< l_2[k-1])) then {
(4) $c = l_1 \bowtie l_2$; //join step: generate candidates
(5) if has_infrequent_subset(c, L_{k-1}) then
(6) delete c; //prune step: remove unfruitful candidate
(7) else add c to C_k;
(8) }
(9) return C_k;

procedure has_infrequent_subset(c: candidate *k-itemset;*
L_{k-1}*:* frequent *(k* - 1)-itemsets); //use prior knowledge
(1) for each *(k* - 1)-subset s of c
(2) if $s \in L_{k-1}$ then
(3) return TRUE;
(4) return FALSE;

Σχήμα 6.19: Ψευδοκώδικας Αλγόριθμου Apriori

6.5.1.2 Παράδειγμα εκτέλεσης αλγόριθμου Apriori

Έχοντας μία μικρή βάση δεδομένων με ασθενείς με έμφραγμα του μυοκαρδίου, που παρουσιάζονται στον Πίνακα 6.5, θα εφαρμόσουμε τον αλγόριθμο Apriori για να βρούμε τα συχνά σύνολα χαρακτηριστικών. Θεωρούμε ως ελάχιστη υποστήριξη το 0.4 (40%), που σημαίνει ότι θα πρέπει ένα σύνολο χαρακτηριστικών να εμφανίζεται τουλάχιστο 11 φορές, δεδομένου ότι στο παράδειγμα μας έχουμε 28 περιστατικά.

Πίνακας 6.5: Βάση δεδομένων δοσοληψιών για ασθενείς με έμφραγμα του μυοκαρδίου (MI)

No	SEX	SMBEF	HDL	GLU	HT	MI
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	H	N	Y
20	M	Y	M	N	Y	Y
21	M	N	L	N	N	Y
22	M	N	M	N	N	Y
23	F	Y	M	N	N	Y
24	M	Y	M	N	N	Y
25	M	Y	L	N	Y	N
26	M	Y	H	N	N	N
27	F	N	H	N	Y	N
28	M	Y	M	N	Y	N

Βλέπε Πίνακα 6.3 για τις κωδικοποιήσεις των χαρακτηριστικών

Βήμα 1: Καταρχάς γίνεται μια αναζήτηση στη βάση δεδομένων για να βρεθούν όλα τα σύνολα χαρακτηριστικών με 1 χαρακτηριστικό και οι φορές που αυτά εμφανίζονται στην βάση δεδομένων. Όλα τα σύνολα χαρακτηριστικών αποθηκεύονται στο σύνολο C1. Σημειώνουμε ότι, το χαρακτηριστικό τάξης δεν εισάγεται στα υποψήφια σύνολα χαρακτηριστικών. Αυτό θα προστεθεί στο τέλος της διαδικασίας ώστε να διασφαλιστεί ότι δεν θα εξαχθούν κανόνες συσχέτισης που να μην συμπεριλαμβάνουν το χαρακτηριστικό για την τάξη. Τα αποτελέσματα του συνόλου C1 παρουσιάζονται στο πιο κάτω Πίνακα 6.6.

Πίνακας 6.6: Παραγόμενο σύνολο C1 (παράδειγμα αλγόριθμου Apriori)

C1			
Χαρακτηριστικό	Συχνότητα	Support MI = N	Support MI = Y
SEX = M	24	6	18
SEX = F	4	2	2
SMBEF = N	5	3	2
SMBEF = Y	23	6	17
HDL = H	3	2	1
HDL = L	8	3	5
HDL = M	17	3	14
GLU = N	20	6	14
GLU = H	8	2	6
HT = N	19	3	16
HT = Y	9	5	4

Βήμα 2: Από το C1 επιλέγουμε τα σύνολα χαρακτηριστικών, που όταν ενωθούν με κάποιο αντικείμενο της τάξης να έχουν υποστήριξη μεγαλύτερη από την ελάχιστη υποστήριξη. Αυτά αποθηκεύονται στο σύνολο L1. Δηλαδή θα αφαιρεθούν τα σύνολα χαρακτηριστικών που εμφανίζονται λιγότερο από 12 φορές. Τα αποτελέσματα του συνόλου L1 παρουσιάζονται στο πιο κάτω Πίνακα 6.7.

Πίνακας 6.7 Παραγόμενο σύνολο L1 (παράδειγμα αλγόριθμου Apriori)

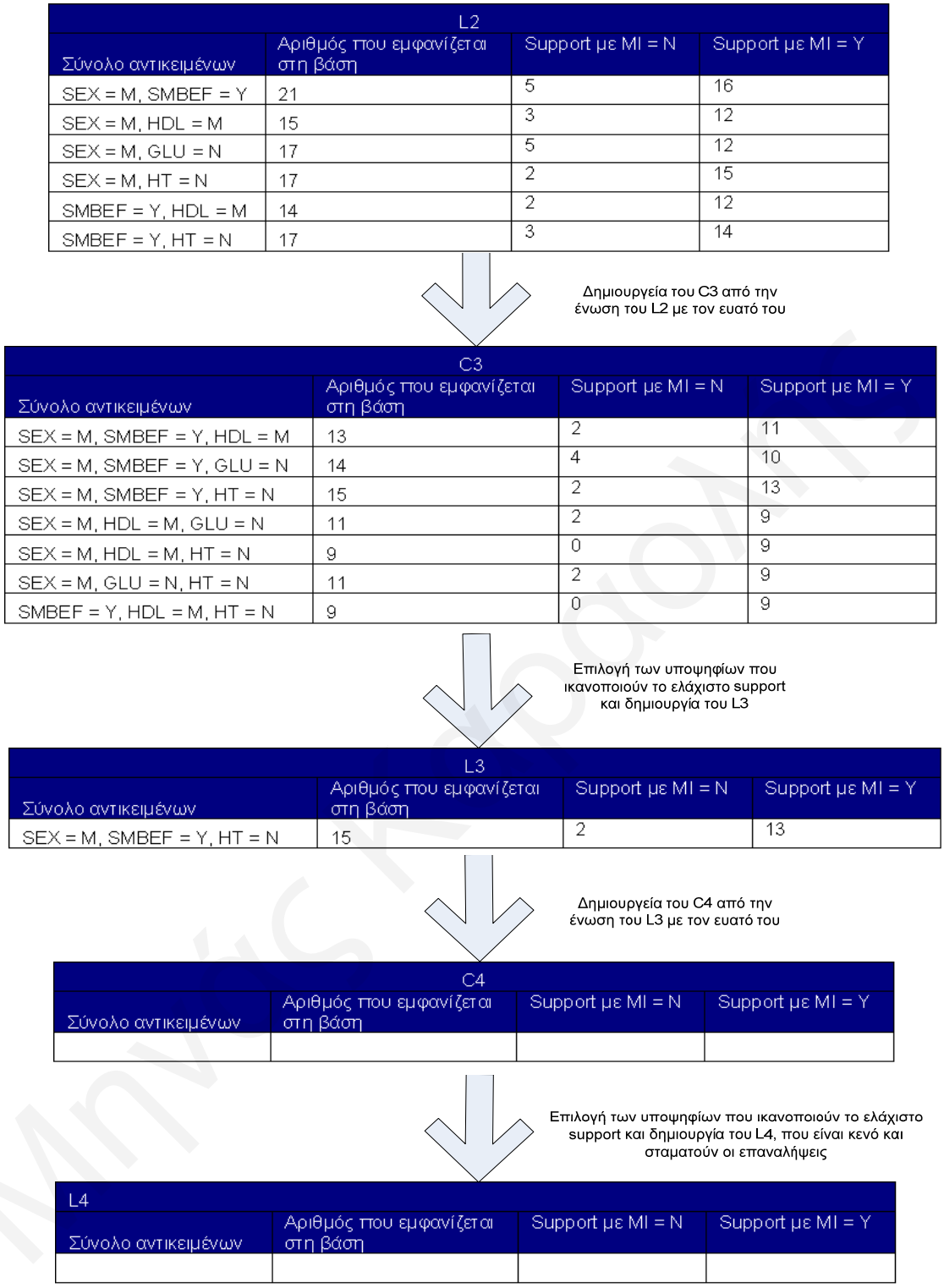
L1			
Χαρακτηριστικό	Συχνότητα	Support MI = N	Support MI = Y
SEX = M	24	6	18
SMBEF = Y	23	6	17
HDL = M	17	3	14
GLU = N	20	6	14
HT = N	19	3	16

Βήμα 3: Σε αυτό το βήμα θα γίνει η ένωση του L1 με τον εαυτό του για κτιστεί το σύνολο C2. Για κάθε σύνολο χαρακτηριστικών του L1 γίνεται συνένωση με τα υπόλοιπα. Τα αποτελέσματα του συνόλου C2 παρουσιάζονται στο πιο κάτω Πίνακα 6.8.

Πίνακας 6.8: Παραγόμενο σύνολο C2 (Παράδειγμα αλγόριθμου Apriori)

C2			
Χαρακτηριστικό	Συχνότητα	Support MI = N	Support MI = Y
SEX = M, SMBEF = Y	21	5	16
SEX = M, HDL = M	15	3	12
SEX = M, GLU = N	17	5	12
SEX = M, HT = N	17	2	15
SMBEF = Y, HDL = M	14	2	12
SMBEF = Y, GLU = N	15	4	11
SMBEF = Y, HT = N	17	3	14
HDL = M, GLU = N	13	2	11
HDL = M, HT = N	10	0	10
GLU = N ,HT = N	12	2	10

Βήμα 4: Επανάληψη των Βημάτων 2 και 3 μέχρι το Lk να είναι κενό, σε αυτή τη περίπτωση, οι επαναλήψεις γίνονται μέχρι που γίνεται το L4 κενό. Στο Σχήμα 6.20, παρουσιάζονται τα υπόλοιπα βήματα.



Σχήμα 6.20: Παραγόμενα σύνολα χαρακτηριστικών (παράδειγμα αλγόριθμου Apriori)

6.5.1.3 Διαδικασία εξόρυξης κανόνων συσχέτισης από τα εξαγόμενα συχνά σύνολα χαρακτηριστικών

Αφού ολοκληρωθεί ο αλγόριθμος Apriori, το σύνολο L, με όλα τα συχνά σύνολα χαρακτηριστικών, προωθούνται 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 τάξης, και ελέγχεται εάν το μέτρο ικανοποιείται. Εάν λάβουμε όλα τα μη κενά σύνολα L από το προηγούμενο παράδειγμα τότε μπορούμε να εξάξουμε τους ακόλουθους κανόνες συσχέτισης του Πίνακα 6.9. Για την αξιολόγηση των κανόνων θα χρησιμοποιήσουμε το μέτρο αξιολόγησης κανόνων, confidence (εμπιστοσύνη). Σε ένα κανόνα $A \rightarrow B$, η εμπιστοσύνη, είναι η πιθανότητα να ισχύει το B νοουμένου ότι ισχύει το A, και υπολογίζεται από την Εξίσωση 4.16 (Κεφάλαιο 4.2).

Οι κανόνες που ικανοποιούν το ελάχιστο όριο υποστήριξης και εμπιστοσύνης, εξάγονται και παρουσιάζονται στο Πίνακα 6.10. Ο Πίνακας 6.11 παρουσιάζει τους κανόνες όπως εξάγονται από το σύστημα.

Πίνακας 6.9: Παραγόμενοι κανόνες συσχέτισης (παράδειγμα αλγόριθμου Apriori)

Κανόνες Συσχέτισης		Support	Confidence
SEX = M	→ MI = 'Y'	0.64	0.75
SMBEF = Y	→ MI = 'Y'	0.61	0.74
HDL = M	→ MI = 'Y'	0.50	0.82
GLU = N	→ MI = 'Y'	0.50	0.70
HT = N	→ MI = 'Y'	0.57	0.84
SEX = M, SMBEF = Y	→ MI = 'Y'	0.57	0.76
SEX = M, HDL = M	→ MI = 'Y'	0.43	0.80
SEX = M, GLU = N	→ MI = 'Y'	0.43	0.71
SEX = M, HT = N	→ MI = 'Y'	0.54	0.88
SMBEF = Y, HDL = M	→ MI = 'Y'	0.43	0.86
SMBEF = Y, HT = N	→ MI = 'Y'	0.50	0.82
MBEF = Y, HT = N	→ MI = 'Y'	0.46	0.87
SEX = M	→ MI = 'N'	0.21	0.25
SMBEF = Y	→ MI = 'N'	0.21	0.26
HDL = M	→ MI = 'N'	0.11	0.18
GLU = N	→ MI = 'N'	0.21	0.30
HT = N	→ MI = 'N'	0.11	0.16
SEX = M, SMBEF = Y	→ MI = 'N'	0.18	0.24
SEX = M, HDL = M	→ MI = 'N'	0.11	0.20
SEX = M, GLU = N	→ MI = 'N'	0.18	0.29
SEX = M, HT = N	→ MI = 'N'	0.07	0.12
SMBEF = Y, HDL = M	→ MI = 'N'	0.07	0.12
SMBEF = Y, HT = N	→ MI = 'N'	0.11	0.18
HDL = M, GLU = N	→ MI = 'N'	0.07	0.15
SEX = M, SMBEF = Y, HT = N	→ MI = 'N'	0.07	0.13

Πίνακας 6.10: Κανόνες που εξάγονται από τον αλγόριθμο Apriori

Κανόνες Συσχέτισης		Support	Confidence
HDL = M	→ MI = 'Y'	0.50	0.82
HT = N	→ MI = 'Y'	0.57	0.84
SEX = M, HDL = M	→ MI = 'Y'	0.43	0.80
SEX = M, HT = N	→ MI = 'Y'	0.54	0.88
SMBEF = Y, HDL = M	→ MI = 'Y'	0.43	0.86
SMBEF = Y, HT = N	→ MI = 'Y'	0.50	0.82
SEX = M, SMBEF = Y, HT = N	→ MI = 'Y'	0.46	0.87

Πίνακας 6.11: Παρουσίαση αποτελεσμάτων για τάξη (class) MI

SEX	SMBEF	HDL	GLU	HT	class	Support	Confidence
		M			Y	0.5	0.82
				N	Y	0.57	0.84
	Y	M			Y	0.43	0.86
M		M			Y	0.43	0.8
	Y			N	Y	0.5	0.82
M				N	Y	0.54	0.88
M	Y			N	Y	0.46	0.87

6.5.2 Αλγόριθμος AKAMAS

Ο αλγόριθμος AKAMAS που έχει σχεδιαστεί και υλοποιηθεί από εμάς είναι μια παραλλαγή του αλγόριθμου Apriori, με την διαφορά ότι δεν χρησιμοποιεί την επαναλαμβανόμενη τεχνική που χρησιμοποιεί τα k-itemset για να κτίσει τα (k+1)-itemsets.

Στην αρχή, ο αλγόριθμος βρίσκει τα συχνά εμφανιζόμενα 1-itemsets (το σύνολο χαρακτηριστικών με 1 μόνο χαρακτηριστικό). Ο αλγόριθμος αναζητά και συναθροίζει τις φορές που εμφανίζεται κάθε χαρακτηριστικό στη βάση δεδομένων, και μετά συλλέγει τα χαρακτηριστικά που ικανοποιούν την ελάχιστη υποστήριξη, στο σύνολο L1.

Αφού συλλέξει όλα τα σύνολα χαρακτηριστικών με 1 χαρακτηριστικό τα οποία ικανοποιούν το ελάχιστο όριο υποστήριξης, τότε κάνει όλους τους δυνατούς συνδυασμούς μεταξύ των

συνόλων χαρακτηριστικών του συνόλου L1 και χτίζει κανόνες συσχέτισης. Πρώτα για κάθε σύνολο χαρακτηριστικών του L1, γίνεται ένωση του με όλα τα υπόλοιπα σύνολα χαρακτηριστικών, και δημιουργούνται έτσι σύνολα χαρακτηριστικών με 2 χαρακτηριστικά (2-itemsets). Στην επόμενη επανάληψη για κάθε σύνολο χαρακτηριστικών του L1 θα γίνει ένωση του με δύο άλλα σύνολα χαρακτηριστικών και θα κτιστούν έτσι όλοι οι δυνατοί συνδυασμοί με 3 χαρακτηριστικά, 3-itemsets. Οι επαναλήψεις συνεχίζονται μέχρι να γίνουν όλοι οι δυνατοί συνδυασμοί μεταξύ των συχνών συνόλων χαρακτηριστικών με 1 χαρακτηριστικό.

Παράλληλα, ενώ οι δυνατοί συνδυασμοί των συνόλων χαρακτηριστικών δημιουργούνται, κτίζονται παράλληλα και οι κανόνες συσχέτισης. Για κάθε καινούργιο σύνολο χαρακτηριστικών γίνεται η ένωσή του και με ένα από τα χαρακτηριστικά τάξης, κι έτσι δημιουργείται ένας κανόνας συσχέτισης. Για να γίνει κάποιο ξεκαθάρισμα των κανόνων, γίνεται αξιολόγηση του εξαγόμενου κανόνα με κάποιο μέτρο (ή με περισσότερα από ένα μέτρα) αξιολόγησης κανόνων, που ο χρήστης εισάγει στο σύστημα. Παραδείγματος χάριν ελέγχετε εάν ο κανόνας ικανοποιεί το ελάχιστο όριο της υποστήριξης και της εμπιστοσύνης. Επομένως, εάν σε κάποια επανάληψη, κανένας κανόνας δεν ικανοποιεί το ελάχιστο όριο των μέτρων αξιολόγησης, τότε ο αλγόριθμος σταματά. Για παράδειγμα στην επανάληψη όπου κτίζονται κανόνες με 4-itemsets, αλλά κανένας από τους κανόνες δεν ικανοποιεί το ελάχιστο όριο των μέτρων, τότε ο αλγόριθμος σταματά και επιστρέφει όλους τους κανόνες με λιγότερα από 4-itemsets και ικανοποιούν το όριο των μέτρων.

Η υλοποίηση του αλγόριθμου AKAMAS επήλθε από την ανάγκη να επιλέγονται τα σημαντικότερα πρότυπα. Με τη χρήση των μέτρων που έχουν υλοποιηθεί γίνεται καλύτερη επιλογή των κανόνων. Τα μέτρα που χρησιμοποιούνται στον αλγόριθμο Apriori φάνηκαν σε κάποιες περιπτώσεις να μην είναι ικανοποιητικά για να μπορούμε να επισημάνουμε τα καλύτερα πρότυπα, αφού έστω και αν θέσουμε ψηλά κατώφλια στα δύο αυτά μέτρα, υπάρχουν φορές που εξάγονται εκατοντάδες κανόνες και μάλιστα με το ίδιο ποσοστό υποστήριξης και εμπιστοσύνης.

Αλγόριθμος: AKAMAS. Εύρεση των συνόλων χαρακτηριστικών (itemsets)

Είσοδος:

- D, βάση δεδομένων με δοσοληψίες
- min-sup, η ελάχιστη υποστήριξη
- min_conf, η ελάχιστη εμπιστοσύνη

Έξοδος: Association_rules, σύνολο με όλους τους δυνατούς κανόνες που ικανοποιούν την ελάχιστη υποστήριξη και εμπιστοσύνη.

Μέθοδος:

(1) L_1 =find_frequent_1-itemsets(D);

(2) for(k = 2;Association_rules$_{k-1} \neq 0$;k++){

(3) for each itemset $l_1 \in L_1$ {

(4) new_rule = l1

(5) for each itemset $l_2 \in L_1$ && $l_2 < l_1$ {

(6) new_rule = new_rule U l_2

(7) if (new_rule.length = k){

(8) if(support(new_rule)≥ min_supp &&

confidence(new_rule)≥ min_supp)

(9) Association_rule$_k$ = Association_rule$_k$ U new_rule

(10) new_rule = new_rule – l_2

(11) }

(12) }

(13) }

(14) return Association_rules = U_k Association_rules$_k$;

Σχήμα 6.21: Ψευδοκώδικας αλγόριθμου AKAMAS

6.5.2.1 Περιγραφή Ψευδοκώδικα Αλγόριθμου AKAMAS

Στο Σχήμα 6.21 παρουσιάζεται ο ψευδοκώδικας του αλγόριθμου AKAMAS:

1. Καταρχάς ο αλγόριθμος δέχεται μια βάση δεδομένων με δοσοληψίες. Όπως επίσης δέχεται από τον χρήστη ένα ελάχιστο όριο για υποστήριξη και εμπιστοσύνη.
2. Στο πρώτο βήμα αναγνωρίζονται όλα τα συχνά σύνολα χαρακτηριστικών με 1 χαρακτηριστικό και φυλάγονται στο σύνολο L1.
3. Μετά για κάθε επανάληψη k (Βήμα 2), όπου k είναι ο αριθμός των χαρακτηριστικών που θα έχει ο κανόνας, θα γίνεται επανάληψη μέχρι να μην υπάρχουν κανόνες με (k-1)-itemsets που να ικανοποιούν το ελάχιστο όριο υποστήριξης και εμπιστοσύνης.
4. Κάθε σύνολο χαρακτηριστικών που ανήκει στο L1 (Βήμα 3) θα ενώνεται με άλλα k-1 σύνολα χαρακτηριστικών που ανήκουν επίσης στο L1, για να παραχθεί κανόνας με k-itemsets. Στην αρχή εισάγεται το l1 (Βήμα 4) στον κανόνα. Μετά κάθε άλλο σύνολο χαρακτηριστικών l2 που ανήκει στο L1, και l2 είναι μικρότερου το l1 προστίθεται στον κανόνα, για να βεβαιωθούμε ότι δεν θα παραχθούν ίδιοι κανόνες (Βήμα 5 και 6). Ελέγχεται εάν ο κανόνας έχει k-itemsets (Βήμα 6) εάν όχι τότε θα προχωρήσει με άλλο l2, αλλιώς θα γίνει έλεγχος εάν ο κανόνας ικανοποιεί την ελάχιστη υποστήριξη και εμπιστοσύνη (Βήμα 8) και θα προστεθεί στο σύνολο των κανόνων. Τελικά θα αφαιρεθεί το l2 από τον κανόνα για να γίνουν άλλοι δυνατοί συνδυασμοί (Βήμα 10) .
5. Στο τέλος ο αλγόριθμος θα επιστρέψει τους κανόνες συσχέτισης στο σύστημα παρουσίασης των κανόνων (Βήμα 14).

6.5.2.2 Παράδειγμα εκτέλεσης αλγόριθμου AKAMAS

Έχοντας την βάση δεδομένων του παραδείγματος εκτέλεσης του αλγόριθμου Apriori, θα εφαρμόσουμε τον αλγόριθμο AKAMAS. Θεωρούμε την ελάχιστη υποστήριξη 0.4 (40%) και ελάχιστη εμπιστοσύνη 0.8 (γίνεται εισαγωγή δεδομένων από το χρήστη, το αρχείο με τη βάση δεδομένων και τα μέτρα αξιόλογης όπως υποστήριξη και εμπιστοσύνη).

Βήμα 1: Καταρχάς γίνεται αναζήτηση στη βάση δεδομένων για να βρεθούν όλα τα σύνολα πλειάδων με 1 χαρακτηριστικό και οι φορές που εμφανίζονται στην βάση δεδομένων. Όλα τα σύνολα χαρακτηριστικών αποθηκεύονται στο σύνολο C1 (Πίνακας 6.12).

Πίνακας 6.12: Σύνολο χαρακτηριστικών C1 (παράδειγμα εφαρμογής αλγόριθμου AKAMAS)

C1	
Σύνολο χαρακτηριστικών	Συχνότητα
SEX = M	24
SEX = F	4
SMBEF = N	5
SMBEF = Y	23
HDL = H	3
HDL = L	8
HDL = M	17
GLU = N	20
GLU = H	8
HT = N	19
HT = Y	9

Βήμα 2: Από το C1 επιλέγουμε τα σύνολα χαρακτηριστικών, που όταν ενωθούν με κάποιο χαρακτηριστικό της τάξης, έχουν υποστήριξη μεγαλύτερη από την ελάχιστη

υποστήριξη, και τα αποθηκεύουμε στο σύνολο L1 (Πίνακας 6.13). Δηλαδή θα αφαιρεθούν τα σύνολα χαρακτηριστικών που εμφανίζονται λιγότερο από 12 φορές.

Πίνακας 6.13: Σύνολο συχνών πλειάδων L1 (παράδειγμα εφαρμογής αλγόριθμου AKAMAS)

L1	
Σύνολο χαρακτηριστικών	Συχνότητα
SEX = M	24
SMBEF = Y	23
HDL = M	17
GLU = N	20
HT = N	19

Βήμα 3: Από το L1 δημιουργούνται όλοι οι κανόνες συσχέτισης. Επιλέγονται μόνο αυτοί που ικανοποιούν το ελάχιστο όριο υποστήριξης και εμπιστοσύνης. Στον πιο κάτω πίνακα (Πίνακας 6.14) παρουσιάζονται οι κανόνες με 1-χαρακτηριστικό που επιλέγονται.

Πίνακας 6.14: Επιλογή κανόνων με 1-χαρακτηριστικό

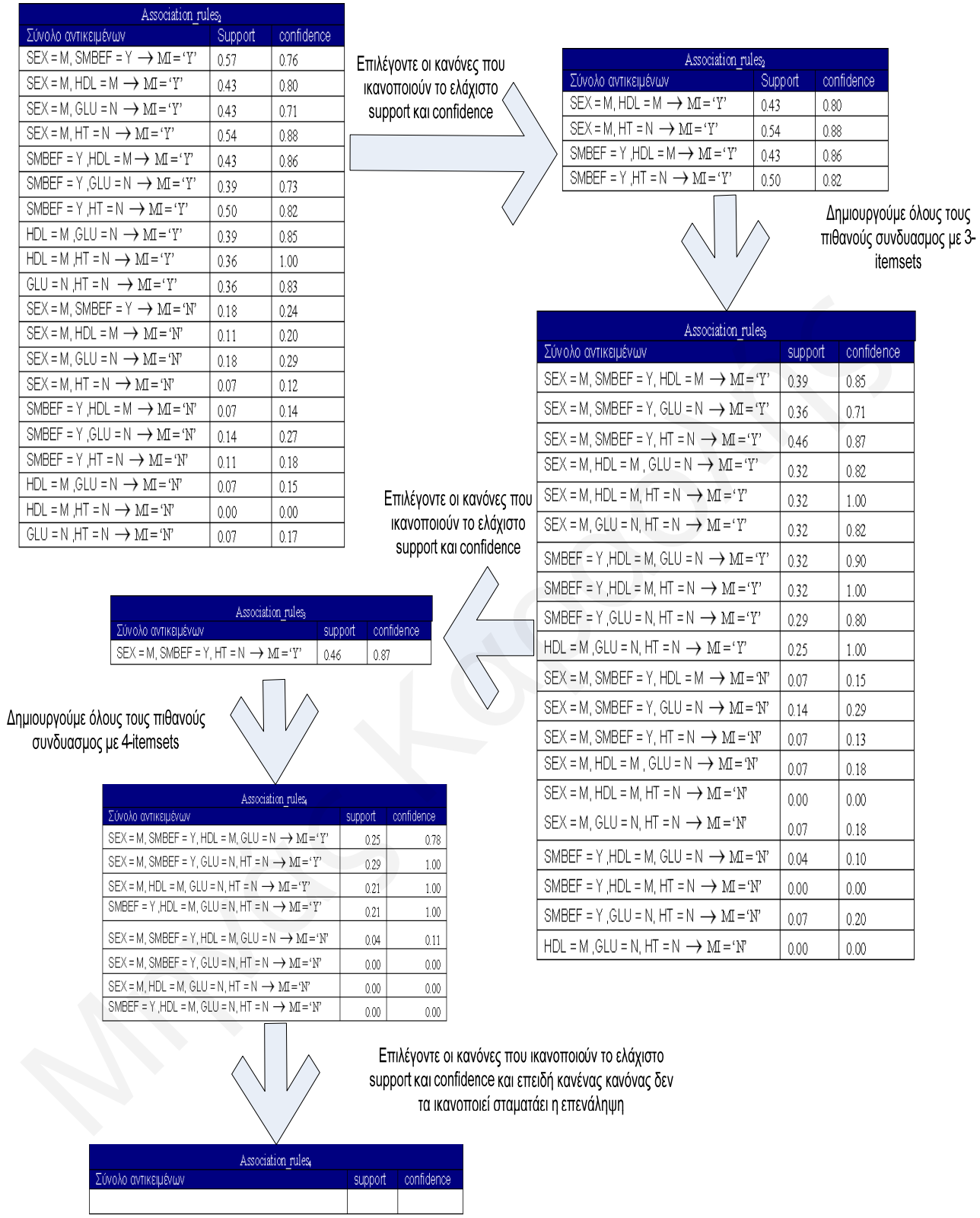
$Association\_rules_1$		
Κανόνες Συσχέτισης	Support	Confidence
SEX = M → MI = 'Y'	0.64	0.75
SMBEF = Y → MI = 'Y'	0.61	0.74
HDL = M → MI = 'Y'	0.50	0.82
GLU = N → MI = 'Y'	0.50	0.70
HT = N → MI = 'Y'	0.57	0.84
SEX = M → MI = 'N'	0.21	0.25
SMBEF = Y → MI = 'N'	0.21	0.26
HDL = M → MI = 'N'	0.11	0.18
GLU = N → MI = 'N'	0.21	0.30
HT = N → MI = 'N'	0.11	0.16

⇓

$Association\_rules_1$		
Κανόνες Συσχέτισης	Support	Confidence
HDL = M → MI = 'Y'	0.50	0.82
HT = N → MI = 'Y'	0.57	0.84

Βήμα 4: Σε αυτό το βήμα θα γίνει η ένωση του κάθε συνόλου χαρακτηριστικών του L1 με όλα τα υπόλοιπα και δημιουργία των κανόνων συσχέτισης με 2-itemsets. Από τους κανόνες συσχέτισης επιλέγονται μόνο αυτοί που ικανοποιούν την ελάχιστη υποστήριξη και εμπιστοσύνη. Αυτό το βήμα επαναλαμβάνεται μέχρι να μην μπορούν να δημιουργηθούν άλλοι κανόνες που να ικανοποιούν τις συνθήκες. (Στο Σχήμα 6.22 παρουσιάζεται η συνέχεια της εκτέλεσης)

Βήμα 5: Τέλος γίνεται παρουσίαση όλων των παραγόμενων κανόνων οι οποίοι ικανοποιούν τα ελάχιστα όρια υποστήριξης και εμπιστοσύνης. Στον Πίνακα 6.15 παρουσιάζονται οι κανόνες που έχουν εξαχθεί από το σύστημα.



Σχήμα 6.22: Παραγόμενοι κανόνες συσχέτισης (παράδειγμα εφαρμογής αλγόριθμου AKAMAS)

Πίνακας 6.15: Παρουσίαση αποτελεσμάτων από το εργαλείο (παράδειγμα εφαρμογής αλγόριθμου AKAMAS)

SEX	SMBEF	HDL	GLU	HT	Class	Support	Confidence
		M			Y	0.5	0.82
				N	Y	0.57	0.84
	Y	M			Y	0.43	0.86
M		M			Y	0.43	0.8
	Y			N	Y	0.5	0.82
M				N	Y	0.54	0.88
M	Y			N	Y	0.46	0.87

6.5.3 Αξιολόγηση Κανόνων

Χρησιμοποιώντας τους αλγόριθμους χωρίς να εφαρμόζεται κάποιο μέτρο αξιολόγησης ή επιλογής των υποψήφιων κανόνων, τότε το σύστημα θα εξαγάγει πάρα πολλούς κανόνες, που δεν θα βοηθήσουν το χρήστη να εξάγει κάποια σημαντική γνώση. 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 ένα 2×2 πίνακα πιθανοτήτων.

Όπως έχει αναφερθεί προηγουμένως, κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων, τα μέτρα αξιολόγησης μπορούν να χρησιμοποιηθούν με τρεις τρόπους:

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

Η πρώτη προσέγγιση, χρησιμοποιείται από τους αλγόριθμους Apriori και AKAMAS για να αφαιρεθούν σύνολα πλειάδων που δεν είναι συχνά εμφανιζόμενα. Το μέτρο που χρησιμοποιείται είναι η υποστήριξη, η οποία, για ένα κανόνα συσχέτισης $A \to B$, δείχνει την πιθανότητα εμφάνισης του Α και του Β μαζί στα δεδομένα (Εξίσωση 4.15).

Το σύστημα που έχει υλοποιηθεί, χρησιμοποιεί διάφορα άλλα αντικειμενικά μέτρα αξιολόγησης κανόνων για ξεκαθάρισμα των εξαγόμενων κανόνων ή της ταξινόμησης τους.

Ο χρήστης μπορεί να επιλέξει ποία από τα μέτρα θέλει να παρουσιάζονται μαζί με τους εξαγόμενους κανόνες συσχέτισης. Για παράδειγμα ο χρήστης επιλέγει να παρουσιάζονται τα μέτρα υποστήριξη, εμπιστοσύνη, ακρίβεια και σχετικός κίνδυνος και τα αποτέλεσμα που παρουσιάζονται στον Πίνακα 6.16. Επομένως ο χρήστης μπορεί να ταξινομήσει τους κανόνες με βάσει την υπολογισμένη τιμή των μέτρων αξιολόγησης:

* **Support**
* **Confidence**
* **Coverage**
* **Prevalence**
* **Recall**
* **Specificity**
* **Accuracy**
* **Lift/Interest**
* **Leverage**
* **Added Value/Change of Support**
* **Relative Risk**
* **Odds Ratio**
* **Conviction**

Πίνακας 6.16: Παρουσίαση κανόνων και μέτρων αξιολόγησης

SEX	SMBEF	HDL	GLU	HT	Class	Support	Confidence	Accuracy	Relative Risk
		M			Y	0.5	0.82	0.68	1.51
				N	Y	0.57	0.84	0.75	1.89
	Y	M			Y	0.43	0.86	0.64	1.5
M		M			Y	0.43	0.8	0.61	1.3
	Y			N	Y	0.5	0.82	0.68	1.51
M				N	Y	0.54	0.88	0.75	1.94
M	Y			N	Y	0.46	0.87	0.68	1.61

Η τρίτη προσέγγιση εφαρμόζεται επίσης για αυτά τα μέτρα που έχουν αναφερθεί πιο πάνω. Ο χρήστης μπορεί να εισάγει κάποιο ελάχιστο όριο για οποιοδήποτε μέτρο αξιολόγησης κανόνων. Τότε το σύστημα θα παρουσιάσει μόνο τους κανόνες που τα ικανοποιούν. Για παράδειγμα ο χρήστης βάζει κάποιο ελάχιστο όριο την υποστήριξη (0.5), εμπιστοσύνη (0.8) και ακρίβεια (0.7), και τα αποτελέσματα παρουσιάζονται όπως δίδεται στον Πίνακα 6.17.

Πίνακας 6.17: Παρουσίαση κανόνων που ικανοποιούν τα όρια των μέτρων

SEX	SMBEF	HDL	GLU	HT	class	Support	Confidence	Accuracy	Relative Risk
				N	Y	0.57	0.84	0.75	1.89
M				N	Y	0.54	0.88	0.75	1.94

Το πρόβλημα που παρουσιάζεται όσον αφορά τα μέτρα αξιολόγησης είναι ποιο ή ποια μέτρα πρέπει να επιλεγούν έτσι ώστε να απομονωθούν οι καλύτεροι κανόνες.

Για να επιλεγούν τα καλύτερα μέτρα για την αξιολόγηση των κανόνων έχει δημιουργηθεί ένας αλγόριθμος όπου με την εξαγωγή όλων των κανόνων σε κάθε μέθοδο και κάθε μοντέλο, γίνονται τα πιο κάτω βήματα:

---

1. Εφαρμόζεται ο αλγόριθμος συσχέτισης
2. Γίνεται κωδικοποίηση των μέτρων που έχουν εξαχθεί από τους κανόνες
3. Χρησιμοποιείται ο αλγόριθμος των δέντρων απόφασης για να κατηγοριοποιήσει τα μέτρα βάσει της τάξης του κανόνα
4. Δημιουργία δέντρου με χαρακτηριστικά τα μέτρα αξιολόγησης
5. Επιλογή των σημαντικότερων κανόνων του βήματος 1 εφαρμόζοντας τους κανόνες για τα μέτρα που παράχθηκαν στο βήμα 4

Σχήμα 6.23: Αλγόριθμος Μετα-Επιλογής Κανόνων Συσχέτισης βάσει πολλαπλών μέτρων

Η κωδικοποίηση των μέτρων αξιολόγησης έχει γίνει με βάση των τιμών που είχαν μετά την εξαγωγή των κανόνων λαμβάνοντας υπόψη τη μέγιστη τιμή, την ελάχιστη τιμή και την κατανομή των τιμών, όπως δίνεται στον Πίνακα 6.18.

Πίνακας 6.18: Κωδικοποίηση των μέτρων

	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

Με τη χρήση του αλγόριθμου για δέντρα απόφασης έχουν δημιουργηθεί δέντρα απόφασης για όλα τα μοντέλα και για όλες τις τάξεις. Ο πιο κάτω πίνακας (Πίνακας 6.19) δείχνει τα σημαντικότερα μέτρα για το έμφραγμα του μυοκαρδίου πριν το επεισόδιο εφαρμόζοντας τον αλγόριθμο 10 φορές.

Πίνακας 6.19: Παρουσίαση του δέντρου απόφασης με τα μέτρα για τα εμφράγματα μυοκαρδίου πριν από το επεισόδιο

# of Run	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Σωστά ταξινομημένα (%)
1	Accuracy 3	Support 3			29	98.60%
2	Accuracy 3	Support 2	Recall 3		22	98.19%
3	Accuracy 3	Support 2	Recall 3	Coverage 3	10	98.40%
4	Accuracy 3	Support 3	Specificity 2		2	98.34%
5	Accuracy 2	Coverage 3	Support 2		25	97.98%
6	Accuracy 3	Support 3	Specificity 3		1	97.98%
7	Accuracy 3	Support 2	Recall 3	Coverage 3	10	98.45%
8	Accuracy 3	Support 2	Coverage 3		10	99.30%
9	Accuracy 3	Support 2	Recall 3		22	97.98%
10	Accuracy 3	Support 3	Specificity 3		1	98.45%

6.6 Υπολογισμός κινδύνου βάσει της εξίσωσης Framingham

Η εξίσωση του Framingham [125] αποτελείται από σημαντικούς παράγοντες πάθησης ενός καρδιακού επεισοδίου. Ο κάθε παράγοντας έχει τη δική του βαρύτητα μέσα στην εξίσωση, γι' αυτό και συνοδεύεται από μια σταθερά. Το Σχήμα 6.24 παρουσιάζει την υλοποίηση του αλγόριθμου υπολογισμού του κινδύνου με την εξίσωση του Framingham [134]. Με αυτή την εξίσωση έχει υπολογιστεί ο κίνδυνος πάθησης ενός επεισοδίου για κάθε ασθενή. Κάνοντας χρήση το ποσοστό του κινδύνου πάθησης ενός επεισοδίου των ασθενών υπολογίζεται το ποσοστό πάθησης ενός επεισοδίου στους κανόνες που εξάχθηκαν από τα μοντέλα που μελετήθηκαν.

Αλγόριθμος Framingham Event Risk

Είσοδος:
Rules dataset
Events Dataset
Non Coded Events Dataset

Έξοδος:
EventRiskValue

For Each Rule
1. EventRiskValue=0
2. Find the Number of Attributes in the Rule
3. Find the value of each attribute
4. For Each Row in Events Dataset
    a. Check if the attribute's value(s) is/are satisfied
    b. Add the event number in Event table
5. Read Non Coded Events Dataset
6. For Each Event number in Event table
    a. Find appropriate event in Non Coded Events Dataset
    b. 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.

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

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

7. End Loop
8. Event Risk = Average(ER for each rule)
9. Return ***Event Risk***

End Loop

Σχήμα 6.24: Ψευδοκώδικας Αλγόριθμου υπολογισμού της εξίσωσης Framingham

6.7 Στατιστική ανάλυση κανόνων

Ένα ενδιαφέρον μέτρο που δείχνει τη στατιστική σημαντικότητα ενός εξαγόμενου κανόνα είναι η τιμή p (p-value). Για να υπολογιστεί αυτό το μέτρο χρειάζεται πρώτα να υπολογιστεί το μέτρο χ^2 (chi-square) όπως επίσης και ο βαθμός ελευθερίας (degree of freedom).

Η χ^2 δοκιμή (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 είδη σύγκρισης:

i. τον έλεγχο καλής προσαρμογής (test of Goodness of fit) και
ii. τον έλεγχο της ανεξάρτητης προσαρμογής (test of independence)

Κατά τον έλεγχο καλής προσαρμογής ελέγχεται το 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.

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).

$$\chi^2 = \sum_{i=1}^{n} \frac{(\boldsymbol{Oi} - \boldsymbol{Ei})^2}{\boldsymbol{Ei}} \qquad \text{(Εξίσωση 6.1)}$$

όπου

χ^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.

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 6.20: Παρουσίαση αποτελεσμάτων που παρατηρήθηκαν (observed)

		SEX		
		M	***F***	Σύνολο
Class	***Y***	50	20	70
	N	20	10	30
	Σύνολο	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

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

Το chi-square statistic υπολογίζεται από την εξίσωση 6.1 και για αυτό τον κανόνα ισούται με

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

Στο Σχήμα 6.25 παρουσιάζεται ο αλγόριθμος για τον υπολογισμό του χ^2.

Αλγόριθμος chi square test

Είσοδος:
Rules dataset
Events Dataset

Έξοδος:
ChiSquareValue

Calculate Chi Square Test Value
For Each Rule

1. ChiSquareValue = 0
2. Find the number of attributes in the rule
3. Find the value of each attribute
4. Create the Chi Square ***Observed*** Table (Class[2] x NumberOfAttributes^2)
5. 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
6. Create the Chi Square ***Expected*** Table
7. 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. $$ChiSquareValue = \sum_{i=0,j=0}^{i=2,j=NumberOfAttributes^2} \frac{(Observed_{(i,j)} - Expected_{(i,j)})^2}{Expected_{(i,j)}}$$

8. Return ***ChiSquareValue***

End Loop

Σχήμα 6.25: Ψευδοκώδικας Αλγόριθμου υπολογισμού χ^2 (chi-square test)

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 σχέση:

$$P-Value = \frac{(1/2)^{k/2}}{\Gamma(k/2)} \chi^{k/2-1} e^{-\chi/2} \qquad \text{(Εξίσωση 6.2)}$$

όπου

χ = η τιμή του chi-square,

k = ο βαθμός ελευθερίας (degree of freedom) και

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

Παράδειγμα υπολογισμού του p-value

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 6.5

c = το πλήθος των στηλών του Πίνακα 6.5

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).

Στο Σχήμα 6.26 παρουσιάζεται ο αλγόριθμος για τον υπολογισμό της τιμή του p. Αν η τιμή του p είναι μικρότερη από 0.05 τότε ο κανόνας θεωρείται στατιστικά σημαντικός

Αλγόριθμος P – Value

Είσοδος:
Rules dataset
Chi Square Test values

Έξοδος:
P-Value

Calculate Chi Square Test Value

For Each Rule

1. P-Value = 0
2. Find the Number of Attributes in the Rule
3. Calculate Degrees of Freedom
   1. Degrees of Freedom = (c – 1) x (r-1)
      = [(#of Attributes)^2 – 1] x [(ClassValues) – 1]
4. Calculate P-Value

   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 6.26: Ψευδοκώδικας Αλγόριθμου υπολογισμού p-value

6.8 Αξιολόγηση μοντέλων

Έκτος από την αξιολόγηση των κανόνων, θα πρέπει επίσης να αξιολογήσουμε την αξιοπιστία των παραγόμενων κανόνων, δηλαδή κατά πόσον αυτοί οι κανόνες συμπεριφέρονται το ίδιο σε μια άλλη βάση δεδομένων που χρησιμοποιείται για έλεγχο (testing). Για την αξιολόγηση της αξιοπιστίας των κανόνων, θα πρέπει η βάση μας να μοιραστεί σε εκπαίδευση (training), και έλεγχο (testing), στην οποία θα εφαρμοστούν οι αλγόριθμοι και θα εξαχθούν οι ενδιαφέρον κανόνες. Για τον έλεγχο (testing) θα μελετηθεί η αξιοπιστία όλων των εξαγόμενων κανόνων. Σε αυτή την βάση δεδομένων ελέγχεται το ποσοστό επιτυχίας του κάθε κανόνα που εξάχθηκε από τη βάση εκπαίδευσης. Ο διαχωρισμός των περιπτώσεων σε εκπαίδευση και έλεγχο παρουσιάζεται στον Πίνακα 6.21.

Πίνακας 6.21: Παρουσίαση του διαχωρισμού των περιπτώσεων σε εκπαίδευση και έλεγχο στις τάξεις MI, PCI και CABG

	Μοντέλο	MI	PCI	CABG
		N/Tr/Ev	N/Tr/Ev	N/Tr/Ev
Επεισόδιο	Yes	378/75/75	72/36/36	86/43/43
	No	150/75/75	274/36/36	307/43/43
	Ολικό	528/150/150	346/72/72	392/86/86

N: Συνολικός αριθμός περιπτώσεων, Tr και Ev δίνουν τον αριθμό των περιπτώσεων εκπαίδευσης και ελέγχου αντίστοιχα.

Έχουν εφαρμοστεί διάφορες τεχνικές για να γίνει έλεγχος και αξιολόγηση των μοντέλων που χρησιμοποιήσαμε. Έχει χρησιμοποιηθεί η μέθοδος holdout [13] και η μέθοδος random subsampling [13], όπου εφαρμόστηκε η μέθοδος holdout δέκα φορές για τους κανόνες που εξάχθηκαν από τα δέντρα απόφασης και πέντε φορές για τη μέθοδο της συσχέτισης. Εκτός από τη μέθοδο holdout, έχει χρησιμοποιηθεί και η μέθοδος 10-fold cross validation, τόσο για την εξαγωγή κανόνων, όσο και για να γίνει έλεγχος για τη σωστή ταξινόμηση χρησιμοποιώντας τα πέντε κριτήρια διαχωρισμού που έχουν αναφερθεί [13]. Η εκτέλεση των μεθόδων έγινε με δεδομένα εκπαίδευσης και δεδομένα ελέγχου, κάνοντας χρήση το 70-30%

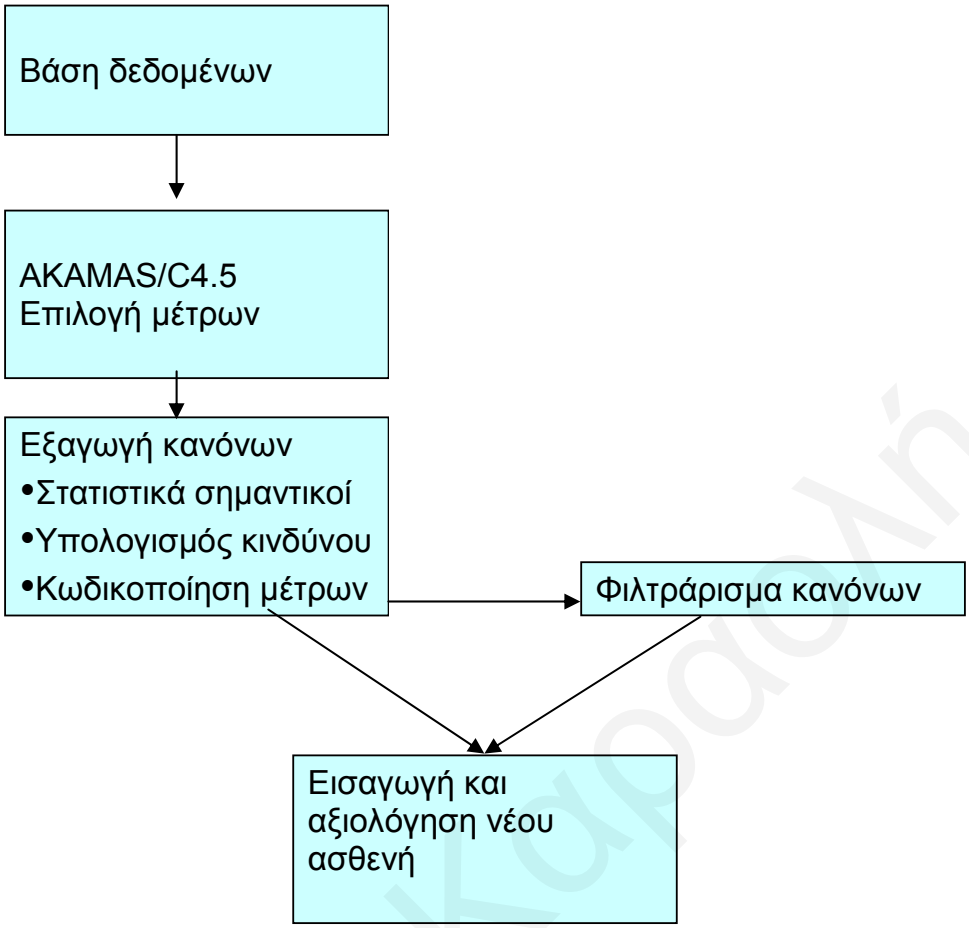
και το 50-50% στο διαχωρισμό. Τα αποτελέσματα που παρουσιάζονται έχουν δεδομένα εκπαίδευσης 50% και δεδομένα ελέγχου 50% (βλέπε Πίνακα 6.21). Σε όλες τις περιπτώσεις οι κανόνες που εξάχθηκαν ήταν οι ίδιοι ή σχεδόν οι ίδιοι.

Από τους εξαγόμενους κανόνες από όλα τα μοντέλα στη μέθοδο της συσχέτισης έχουν επιλεγεί οι κανόνες που είναι στατιστικά σημαντικοί, για να διερευνηθεί η αξιολόγηση τους. Στα δέντρα απόφασης αυτό δεν ήταν εφικτό γιατί πολλά μοντέλα δεν είχαν εξάγει κανόνες που να είναι στατιστικά σημαντικοί.

6.9 Εφαρμογή συστήματος εξαγωγής κανόνων από τον καρδιολόγο

Από τη βάση δεδομένων εκτελείται με τα δέντρα απόφασης και τη μέθοδο της συσχέτισης το μοντέλο για το επεισόδιο που θέλει ο καρδιολόγος να μελετήσει. Οι εξαγόμενοι κανόνες έχουν ταυτόχρονα και την πληροφορία κατά πόσο είναι στατιστικά σημαντικοί, όπως επίσης και το ποσοστό κινδύνου πάθησης επεισοδίου. Με τα μέτρα των εξαγόμενων κανόνων εκτελείται ο αλγόριθμος των δέντρων απόφασης και εξάγονται τα σημαντικότερα μέτρα. Τα μέτρα αυτά εφαρμόζονται στους εξαγόμενους κανόνες και έτσι γίνεται ένα φιλτράρισμα των κανόνων (Σχήμα 6.27).

Έχει δημιουργηθεί ένα εργαλείο που έχοντας όλους τους εξαγόμενους κανόνες όπως επίσης και τους φιλτραρισμένους κανόνες, μπορεί ο ειδικός γιατρός να εισάγει τα δεδομένα ενός νέου ασθενή και να του απομονώσει όλους εκείνους τους κανόνες που έχουν τις τιμές στους παράγοντες που έχει και ο νέος ασθενής. Δίνεται προτεραιότητα στους φιλτραρισμένους κανόνες. Με αυτό τον τρόπο μπορεί ο γιατρός να δει το ποσοστό κινδύνου που έχει ο νέος ασθενής να πάθει ένα επεισόδιο. Πέραν τούτου μπορεί να αλλάξει στον ασθενή ένα μεταβαλλόμενο παράγοντα για να δει κατά πόσο το ποσοστό κινδύνου μειώνεται και έτσι να μπορέσει να λάβει απόφαση για τη φαρμακευτική θεραπεία που θα δώσει στον νέο ασθενή.



Σχήμα 6.27: Προτεινόμενο σύστημα

Κεφάλαιο 7: Αποτελέσματα

7.1 Γενικά

Εφαρμόζοντας όλα τα στάδια της εξόρυξης δεδομένων και τη μεθοδολογία που προτείναμε για να αναπτύξουμε ένα ολοκληρωμένο σύστημα όσον αφορά την εξαγωγή κανόνων και της επικινδυνότητας πάθησης ενός επεισοδίου σε ιατρικά περιστατικά και ιδιαίτερα σε καρδιαγγειακές παθήσεις, έχουμε υλοποιήσει τα πιο κάτω:

1. Αφαιρέσαμε από τη βάση δεδομένων τους παράγοντες κινδύνου που δεν είχαν καμιά σημασία με βάσει τις υποδείξεις των ειδικών.
2. Καθορίσαμε τη βάση δεδομένων και συμπληρώσαμε ελλιπείς τιμές, όπου αυτό ήταν εφικτό. Ασθενείς που εξακολουθούσαν να είχαν ελλιπείς τιμές τους αφαιρέσαμε από τη βάση δεδομένων.
3. Κωδικοποιήσαμε τους παράγοντες κινδύνου που ήταν στην βάση δεδομένων μετά από υποδείξεις των ειδικών ως επίσης βασισμένοι στις διεθνείς προδιαγραφές.
4. Υλοποιήσαμε τα μέτρα που θα χρησιμοποιούσαμε στο σύστημά μας για την εξαγωγή των κανόνων.
5. Υλοποιήσαμε τα κριτήρια διαχωρισμού που θα χρησιμοποιούσαμε στα δέντρα απόφασης.
6. Επιλέξαμε και υλοποιήσαμε τη μέθοδο κλαδέματος που θα χρησιμοποιούσαμε στον αλγόριθμο για την κατασκευή δέντρων απόφασης.
7. Υλοποιήσαμε αλγόριθμους συσχέτισης.
8. Δημιουργήσαμε νέο αλγόριθμο συσχέτισης.
9. Δημιουργήσαμε σύστημα επιλογής των σημαντικότερων κανόνων πολλαπλών μέτρων για τα δέντρα απόφασης και αλγόριθμους κανόνων συσχέτισης.

10. Δημιουργήσαμε ένα γραφικό περιβάλλον πολύ φιλικό προς τον χρήστη.

11. Δημιουργήσαμε τα μοντέλα που θα μελετούσαμε και των οποίων τα αποτελέσματα παρουσιάζομε σε αυτό το κεφάλαιο.

- Η βάση δεδομένων χωρίστηκε σε εκπαίδευση και έλεγχο με ποσοστό 50% και 50% αντίστοιχα, μέθοδος holdout
- Έγινε το holdout validation
- Δημιουργήθηκαν τα ακόλουθα μοντέλα με βάσει τους παράγοντες και τα επεισόδια:
  - 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)

7.2 Εξαγωγή κανόνων ταξινόμησης με δέντρα απόφασης

Χρησιμοποιήθηκαν πέντε κριτήρια διαχωρισμού, IG: Information Gain, GI: Gini Index, X2: Likelihood Ratio Chi-squared statistics, GR: Gain Ratio, DM: Distance Measure. Με τη χρήση του αλγόριθμου C4.5 έχουν χρησιμοποιηθεί τα μοντέλα πριν, μετά και πριν και μετά το επεισόδιο για το έμφραγμα μυοκαρδίου, αγγειοπλαστική και στεφανιαία παράκαμψη.
Οι εξαγόμενοι κανόνες από κάθε μοντέλο μελετήθηκαν ως προς τους παράγοντες που είχε ο κάθε κανόνας. Με τη μέθοδο της συχνότητας έχουν υπολογιστεί όλοι οι παράγοντες όλων των κανόνων σε κάθε μοντέλο και επιλέγηκαν οι τρεις σημαντικότεροι.

7.2.1 Κανόνες με δέντρα απόφασης για Έμφραγμα μυοκαρδίου, MI vs PCI ή CABG

Στον Πίνακα 7.1 παρουσιάζονται τα αποτελέσματα της ταξινόμησης από το μοντέλο MI vs PCI ή CABG για τα πέντε κριτήρια διαχωρισμού, χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά (B+A). Δίνεται η μέση τιμή για τα %CC, %TP και %FP.

Πίνακας 7.1: Μοντέλα για Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG για τα πέντε κριτήρια διαχωρισμού με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A). Δίνονται ο μέσος όρος (Me) (ελάχιστο (m) και μέγιστο (M)) για τις δέκα εκτελέσεις για το %CC, %TP και %FP. Για την ευαισθησία και την ειδικότητα δίνεται ο μέσος όρος

	%CC			%TP			%FP			Sensitivity			Specificity		
	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A
	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me	Me	Me	Me	Me	Me
DM	60(58,62)	59(58,62)	63(61,65)	71(57,67)	61(57,69)	65(57,71)	47(39,54)	43(40,45)	40(27,45)	59	59	65	63	59	64
GI	61(59,63)	61(59,63)	63(61,66)	67(55,71)	59(55,71)	63(57,76)	47(41,59)	36(33,48)	39(25,51)	59	60	62	61	62	64
GR	60(58,61)	59(59,59)	62(61,64)	65(53,72)	59(55,67)	65(53,67)	45(37,53)	41(36,49)	41(38,45)	59	59	62	61	59	62
IG	58(57,64)	61(60,63)	62(61,65)	64(60,76)	68(61,73)	67(53,68)	48(44,55)	45(41,49)	37(25,47)	58	60	63	60	64	63
X2	58(57,60)	61(59,63)	63(62,65)	65(63,73)	63(59,76)	64(59,72)	49(47,53)	39(35,59)	36(35,47)	57	62	64	59	61	64

Από τα μέτρα που εξάχθηκαν από τους κανόνες έγινε μια εκτέλεση με τον αλγόριθμο για δέντρα απόφασης έτσι ώστε να επιλεγούν τα σημαντικότερα για το φιλτράρισμα των κανόνων. Στον Πίνακα 7.2 παρουσιάζονται τα μέτρα που λήφθηκαν υπόψη για το φιλτράρισμα των κανόνων.

Πίνακας 7.2: Μέτρα των μοντέλων για Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A)

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

Στον Πίνακα 7.3α παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με έμφραγμα μυοκαρδίου με χαρακτηριστικά πριν το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η τέταρτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ειδικότητα (specificity) και η εμπιστοσύνη (confidence). Στο μοντέλο αυτό είχαμε 58% σωστή ταξινόμηση. Στον Πίνακα 7.3β παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με έμφραγμα μυοκαρδίου με χαρακτηριστικά μετά το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η τρίτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η προστιθέμενη αξία (added_value) και το διαγώνιο πηλίκο (odds ratio). Στο μοντέλο αυτό είχαμε 61% σωστή ταξινόμηση. Στον Πίνακα 7.3γ παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με έμφραγμα μυοκαρδίου με χαρακτηριστικά πριν και μετά το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η πέμπτη εκτέλεση (run). Τα

μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ειδικότητα (specificity) και η προστιθέμενη αξία (added_value). Στο μοντέλο αυτό είχαμε 62% σωστή ταξινόμηση. Με τη βοήθεια της εξίσωσης του Framingham έχει υπολογιστεί ο κίνδυνος πάθησης ή μη πάθησης ενός επεισοδίου για κάθε κανόνα. Επιπλέον έχει υπολογιστεί το μέτρο χ^2 (chi-square), το οποίο είναι καθοριστικό για να υπολογιστεί το μέτρο p-value, που δείχνει αν κάποιος κανόνας είναι στατιστικά σημαντικός.

Ο κανόνας 1.1 στον Πίνακα 7.3α δείχνει ότι άντρες ασθενείς στις ηλικίες 51-60 χρονών που έχουν διαβήτη, έχουν πάθει έμφραγμα μυοκαρδίου. Στον ίδιο πίνακα ο κανόνας 1.3 δείχνει ότι στις ίδιες ηλικίες με τον κανόνα 1.1 αλλά γυναίκες, παρόλο που έχουν και ιστορικό στην οικογένεια δεν παθαίνουν έμφραγμα μυοκαρδίου. Επίσης δείχνουν αυτοί οι κανόνες ότι δεν είναι στατιστικά σημαντικοί. Οι κανόνες 2.1 και 2.2 στον Πίνακα 7.3β δείχνουν ότι ασθενείς έχουν ψηλή χοληστερόλη, ψηλές λιποπρωτεΐνες ψηλής πυκνότητας, ψηλές λιποπρωτεΐνες χαμηλής πυκνότητας και ψηλά τριγλυκερίδια παθαίνουν επεισόδιο, ανεξάρτητα αν έχουν ψηλή διαστολική πίεση ή αν δεν έχουν. Έχουν και οι δύο κανόνες ψηλό ποσοστό πάθησης επεισοδίου και δεν είναι στατιστικά σημαντικοί. Οι κανόνες 3.3 και 3.4 στον Πίνακα 7.3γ δείχνουν ότι άντρες ασθενείς στις ηλικίες 51-60 χρονών που έχουν ψηλή διαστολική πίεση, ψηλή χοληστερόλη και ψηλά τριγλυκερίδια, το κάπνισμα και η ψηλή συστολική πίεση δεν παίζουν ρόλο. Οι κανόνες αυτοί δεν είναι στατιστικά σημαντικοί, αλλά είναι στα ψηλά ποσοστά πάθησης ενός επεισοδίου.

Πίνακας 7.3α: Επιλεγμένοι κανόνες για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά πριν το επεισόδιο (Β)

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
1.1	M	2	N	N	N	Y	Y	1	3	1	0.68	1	3	1	3	3	1	3	1	58.15	NS	H
1.2	M	2	N	N	Y	Y	N	1	3	1	0.32	1	3	1	3	3	1	3	1	58.15	NS	H
1.3	F	2	Y	N			N	1	3	1	0.32	1	3	1	3	3	1	3	1	20.71	NS	H
1.4	F	3	N	Y	N	Y	N	1	3	1	0.32	1	3	1	3	3	1	3	1	62.51	NS	H
1.5		3	Y	Y	N	Y	N	1	3	1	0.32	1	3	1	3	3	1	3	1	46.30	S	M
1.6		4	Y	Y	N	N	N	1	3	1	0.32	1	3	1	3	3	1	3	1	43.87	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, class: MI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.3β: Επιλεγμένοι κανόνες για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά μετά από το επεισόδιο (A)

	SMAFT	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
2.1	N	N	N	N	H	N	H	H	Y	1	1	1	0.67	1	3	1	3	1	3	3	3	103.2	NS	H
2.2	N	N	H	N	H	H	H	H	Y	1	1	1	0.67	1	1	1	3	1	3	3	3	98.45	NS	H
2.3	N	N	N	Y	H	H	H	N	Y	1	1	1	0.67	1	1	1	3	1	3	3	3	80.51	NS	H
2.4	Y	N	N	Y	N	N	N	N	Y	1	1	1	0.67	1	3	1	3	1	3	3	3	99.43	NS	H
2.5	Y	H	N	N	N	N	N	N	Y	1	1	1	0.67	1	1	1	3	1	3	3	3	76.22	NS	H
2.6	Y	H	H	N	N	N	N	H	Y	1	1	1	0.67	1	1	1	3	1	3	3	3	71.59	NS	H
2.7	Y	H	N	N	H	H	H	N	Y	1	1	1	0.67	1	1	1	3	1	3	3	3	111.8	NS	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

SMAFT: Καπνιστής μετά το επεισόδιο, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη,
LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: MI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.3γ: Επιλεγμένοι κανόνες για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG, με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

	SEX	AGE	FH	SMAFT	SMBEF	HxHTN	HxDM	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
3.1	M	2	N	N	Y	N	N	N	N	N	N	N	N	H	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	98.4	NS	H
3.2	M	2	N	N	Y	N	N	N	H	N	N	H	N	N	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	112	NS	H
3.3	M	2	N	N	N	N	N	H	H	N	H	H	N	H	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	109	NS	H
3.4	M	2	N	N	Y	N	N	N	H	N	H	N	N	H	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	111	NS	H
3.5	M	2	Y	N	Y	Y	Y	H	N	Y	N	N	N	H	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	118	NS	H
3.6	M	3	N	N	Y	N	N	N	H	N	N	H	N	N	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	99.2	NS	H
3.7	M	3	N	N	Y	N	N	N	H	N	H	H	N	N	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	101	NS	H
3.8	M	3	N	N	Y	Y	N	N	H	N	H	H	N	H	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	110	NS	H
3.9	M	3	Y	N	Y	Y	N	H	N	Y	N	N	N	N	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	100	NS	H
3.10	M	3	Y	N	Y	Y	Y	H	H	Y	N	N	N	H	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	108	NS	H
3.11	M	4	N	N	N	N	N	N	N	N	N	N	N	N	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	97.4	NS	H
3.12	F	4	Y	N	N	Y	N	H	N	Y	N	N	N	N	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	104	NS	H
3.13	M	4	N	N	Y	Y	Y	H	N	Y	N	N	N	N	Y	1	1	1	0.68	1	3	1	3	1	3	3	1	119	NS	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, SMAFT: Καπνιστής μετά το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη, LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: MI

S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός

Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Στον Πίνακα 7.4 παρουσιάζονται οι εξαγόμενοι κανόνες. Η στήλη L δείχνει τους κανόνες που έχουν χαμηλό ποσοστό κινδύνου, η M δείχνει τους κανόνες που έχουν μέτριο ποσοστό κινδύνου και η στήλη H αυτούς που έχουν ψηλό ποσοστό κινδύνου για κάθε μοντέλο. Επίσης παρουσιάζεται ο αριθμός των κανόνων που είναι στατιστικά σημαντικοί και αυτών που δεν παρουσιάζουν στατιστική σημαντικότητα

Πίνακας 7.4: Εξαγόμενος αριθμός κανόνων δέντρων απόφασης για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG.

	Κίνδυνος επεισοδίου			Στατιστική ανάλυση Κανόνων		Σύνολο κανόνων
	L	M	H	S	NS	
B	0 (0%)	14 (22%)	50 (78%)	12 (19%)	52 (81%)	64 (100%)
A	0 (0%)	19 (48%)	21 (52%)	6 (15%)	34 (85%)	40 (100%)
B+A	0 (0%)	28 (42%)	38 (58%)	17 (26%)	49 (74%)	66 (100%)

Οι τρεις κυριότεροι παράγοντες για MI από τα πέντε κριτήρια διαχωρισμού χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά το επεισόδιο (B+A) φαίνονται στον Πίνακα 7.5. Τα δύο πρώτα μοντέλα, B και A, είχαν εντελώς διαφορετικούς παράγοντες κινδύνου και γι' αυτό το λόγο βλέπουμε στον πίνακα να μην υπάρχουν κοινοί παράγοντες. Στο μοντέλο B+A όπου είναι όλοι οι παράγοντες των δύο προηγούμενων μοντέλων, εμφανίζονται σαν σημαντικοί παράγοντες και από τα δύο προηγούμενα μοντέλα. Σε αυτό το σημείο φαίνεται ότι μεταξύ των δύο μοντέλων (B και A) δεν υπάρχει κάποιο που να είναι πιο σημαντικό από το άλλο.

Πίνακας 7.5: Κυριότεροι παράγοντες για MI vs PCI ή CABG

Κριτήριο Διαχωρ.	B			A			B+A		
IG	AGE	SMBEF	HxHTN	SBP	SMAFT	DBP	AGE	SMAFT	SBP
GI	AGE	HxHTN	SMBEF	SBP	SMAFT	DBP	AGE	SBP	SMBEF
X2	AGE	HxHTN	SMBEF	SMAFT	SBP	DBP	AGE	DBP	HxHTN
GR	AGE	HxHTN	SMBEF	SBP	SMAFT	DBP	SBP	SMAFT	HxHTN
DM	AGE	HxHTN	SMBEF	SBP	DBP	SMAFT	AGE	SBP	SMBEF

Χρησιμοποιώντας τη μέθοδο του στατιστικού ελέγχου του Wilcoxon [135], δεν βρέθηκαν σημαντικές διαφορές στα κριτήρια διαχωρισμού που μελετήθηκαν. Η απόδοση ήταν γύρω στο 60% για το %CC για τα μοντέλα B, A και B+A για όλα τα κριτήρια διαχωρισμού. Την καλύτερη απόδοση είχε το μοντέλο B+A, όπου η μέση τιμή του %CC ήταν μεταξύ 62 και 63%. Το καλύτερο μοντέλο παρουσιάστηκε με το κριτήριο διαχωρισμού GI για το B+A με %CC = 66%. Οι κυριότεροι παράγοντες για το μοντέλο B είναι ηλικία, υπέρταση και κάπνισμα πριν το επεισόδιο, για το μοντέλο A, συστολική πίεση, κάπνισμα μετά το επεισόδιο και διαστολική πίεση και για το μοντέλο B+A, ηλικία, κάπνισμα και υπέρταση.

Με τη μέθοδο του Wilcoxon [135] έγινε στατιστική ανάλυση των μοντέλων που μελετήθηκαν για τα κριτήρια διαχωρισμού και τις τρεις τάξεις MI, PCI και CABG. Ο Πίνακας 7.6 δείχνει τη στατιστική ανάλυση των μέτρων διαχωρισμού με βάση το μέτρο της σωστής ταξινόμησης (correct classification) για το μοντέλο· MI vs PCI ή CABG. Σε όλες τις περιπτώσεις φαίνεται ότι δεν υπάρχει στατιστική σημαντικότητα. Η ανάλυση των μοντέλων που μελετήθηκαν με τα κριτήρια διαχωρισμού που χρησιμοποιήθηκαν, εκτελώντας τον αλγόριθμο με το κριτήριο της πληροφορίας του κέρδους, έδειξε ότι δεν υπάρχει στατιστική σημαντικότητα (Πίνακας 7.7).

Πίνακας 7.6: Στατιστική ανάλυση για το μέτρο %CC των κριτηρίων διαχωρισμού για τα μοντέλο MI vs PCI ή CABG

	IG	GI	X2	GR	DM
IG		NS	NS	NS	NS
GI			NS	NS	NS
X2				NS	NS
GR					NS
DM					

S: statistically significant; NS: non statistically significant

Πίνακας 7.7: Στατιστική ανάλυση των παραγόντων κινδύνου πριν, μετά, πριν και μετά για τα κριτήρια διαχωρισμού για τα μοντέλο MI vs PCI ή CABG

	B vs A	B vs B+A	A vs B+A
IG	NS	NS	NS
GI	NS	NS	NS
X2	NS	NS	NS
GR	NS	NS	NS
DM	NS	NS	NS

S: statistically significant; NS: non statistically significant

7.2.2 Κανόνες με δέντρα απόφασης για Αγγειοπλαστική, PCI vs MI ή CABG

Στον Πίνακα 7.8 παρουσιάζονται τα αποτελέσματα της ταξινόμησης από το μοντέλο PCI vs MI ή CABG για τα πέντε κριτήρια διαχωρισμού, χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά (B+A). Δίνεται η μέση τιμή για τα %CC, %TP και %FP.

Πίνακας 7.8: Μοντέλα για Αγγειοπλαστική, PCI vs MI ή CABG για τα πέντε κριτήρια διαχωρισμού με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A). Δίνονται ο μέσος όρος (Me) (ελάχιστο (m) και μέγιστο (M)) για τις δέκα εκτελέσεις για το %CC, %TP και %FP. Για την ευαισθησία και την ειδικότητα δίνεται ο μέσος όρος

	%CC			%TP			%FP			Sensitivity			Specificity		
	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A
	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me	Me	Me	Me	Me	Me
DM	64(63,65)	65(61,71)	65(64,68)	69(64,78)	72(67,78)	69(64,75)	42(33,47)	42(36,56)	39(33,47)	63	62	64	66	67	67
GI	61(61,64)	67(65,68)	67(63,70)	67(50,86)	69(50,75)	67(56,69)	39(28,64)	42(14,50)	31(22,42)	63	64	69	65	64	64
GR	63(61,70)	64(64,65)	65(64,67)	67(56,82)	67(53,83)	72(53,72)	44(31,50)	39(25,56)	39(22,44)	65	63	65	63	65	67
IG	63(61,65)	67(64,75)	67(65,70)	64(53,72)	72(67,78)	58(56,64)	36(31,42)	39(28,50)	22(22,31)	63	65	71	63	69	65
X2	63(60,64)	65(63,72)	65(63,65)	69(56,69)	72(58,78)	72(58,78)	36(33,44)	36(33,42)	42(28,53)	61	64	63	65	65	68

Από τα μέτρα που εξάχθηκαν από τους κανόνες έγινε μια εκτέλεση με τον αλγόριθμο για δέντρα απόφασης έτσι ώστε να επιλεγούν τα σημαντικότερα για το φιλτράρισμα των κανόνων. Στον Πίνακα 7.9 παρουσιάζονται τα μέτρα που λήφθηκαν υπόψη για το φιλτράρισμα των κανόνων.

Πίνακας 7.9: Μέτρα των μοντέλων για Αγγειοπλαστική, PCI vs MI ή CABG χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A)

Μοντέλο	Μέτρο 1	Μέτρο 2	Μέτρο 3	Μέτρο 4	Αρ. Κανόνων	Αρ. Εκτέλεσης
PCI vs MI ή CABG (B)	Convicion	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

Στον Πίνακα 7.10α παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με αγγειοπλαστική πριν το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η πέμπτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η πεποίθηση (conviction) και η ειδικότητα (specificity). Στο μοντέλο αυτό είχαμε 66.67% σωστή ταξινόμηση. Στον Πίνακα 7.10β παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με αγγειοπλαστική μετά το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η τέταρτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η δύναμη (leverage), η εμπιστοσύνη (confidence) και η ειδικότητα (specificity). Στο μοντέλο αυτό είχαμε 54.55% σωστή ταξινόμηση. Στον Πίνακα 7.10γ παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με αγγειοπλαστική πριν και μετά το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η δέκατη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ειδικότητα (specificity) και η εμπιστοσύνη (confidence). Στο μοντέλο αυτό είχαμε 63.88% σωστή ταξινόμηση. Με τη βοήθεια της εξίσωσης του Framingham έχει υπολογιστεί ο κίνδυνος πάθησης ή μη ενός επεισοδίου για κάθε κανόνα. Επιπλέον έχει

υπολογιστεί το μέτρο chi-square, το οποίο είναι καθοριστικό για να υπολογιστεί το μέτρο p-value, που δείχνει αν κάποιος κανόνας είναι στατιστικά σημαντικός.

Οι κανόνες 1.1 και 1.2 στον Πίνακα 7.10α δείχνουν ότι στις γυναίκες ασθενείς σε μικρές ηλικίες (μέχρι 50 χρονών) το κάπνισμα δεν είναι καθοριστικός παράγοντας για την πάθηση ενός επεισοδίου Εμφράγματος μυοκαρδίου. Οι ασθενείς που αντιπροσωπεύουν αυτοί οι δύο κανόνες είναι με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου. Επίσης δείχνουν αυτοί οι κανόνες ότι είναι στατιστικά σημαντικοί. Οι κανόνες 2.9 και 2.10 στον Πίνακα 7.10β δείχνουν ότι στους ασθενείς με συστολική πίεση, ψηλές λιποπρωτεϊνες χαμηλής πυκνότητας και ψηλά τριλγυκερίδια, καθοριστικός παράγοντας ενός επεισοδίου είναι το κάπνισμα και η χοληστερόλη. Και οι δύο κανόνες έχουν ψηλό ποσοστό πάθησης ενός επεισοδίου και επίσης αυτοί οι κανόνες δεν είναι στατιστικά σημαντικοί. Ο κανόνας 3.8 στον Πίνακα 7.10γ δείχνει ότι στους άντρες μεγάλης ηλικίας (70 χρονών και άνω) το ιστορικό στην οικογένεια είναι καθοριστικός παράγοντας Δείχνει επίσης ότι είναι κανόνας με ψηλό ποσοστό κινδύνου επεισοδίου και είναι στατιστικά σημαντικός.

Πίνακας 7.10α: Επιλεγμένοι κανόνες για το Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά πριν από το επεισόδιο (Β)

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
1.1	F	1	N	N	N	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	79.69	S	H
1.2	F	1	N	Y	N	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	79.69	S	H
1.3	M	1	N	N	Y	Y	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	79.69	S	H
1.4	F	2	N	Y	N	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	88.68	S	H
1.5	F	3	Y	Y	Y	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	75.53	NS	H
1.6	M	3	Y	N	N	Y	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	75.53	NS	H
1.7	F	4	N	Y	Y	N	N	1	3	1	0.60	1	3	1	3	3	1	3	1	85.04	S	H
1.8	F	4	N	Y	Y	Y	N	1	3	1	0.60	1	3	1	3	3	1	3	1	85.04	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, class: PCI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.10β: Επιλεγμένοι κανόνες για το Μοντέλο Αγγειπλαστική, PCI vs MI ή CABG, με χαρακτηριστικά μετά από το επεισόδιο (A)

	SMAFT	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
2.1		N	N	N	H		N	N	N	1	3	1	0.60	1	3	1	3	3	1	3	1	65.83	NS	H
2.2	N	N	N	N	H		H	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	90.41	NS	M
2.3		H	N	N	N		H	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	65.83	NS	H
2.4	N	H	N	N	H		H	N	N	1	3	1	0.60	1	3	1	3	3	1	3	1	90.41	NS	H
2.5		N	N	Y	N		H	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	65.83	NS	H
2.6	N	H	N	Y	H		N	N	N	1	3	1	0.60	1	3	1	3	3	1	3	1	90.41	NS	H
2.7	Y	H	N	Y	H		N	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	90.41	NS	H
2.8	Y	H	N	N		N	N	H	N	1	3	1	0.60	1	3	1	3	3	1	3	1	88.95	NS	H
2.9		H	N	N	N	H	N	H	N	1	3	1	0.60	1	3	1	3	3	1	3	1	88.81	NS	H
2.10	Y	H	N	N	H	H	N	H	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	93.23	NS	H
2.11	Y		N	Y	N		N	H	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	51.52	NS	H
2.12	Y	N	H	N	N	H	N	H	N	1	3	1	0.60	1	3	1	3	3	1	3	1	80.81	NS	H
2.13	Y	H	H	N	N	H	N	H	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	89.76	NS	H
2.14	N	N	H	N	H	N	N		Y	1	3	1	0.40	1	3	1	3	3	1	3	1	84.84	NS	M
2.15	N	N	H	Y	H	H	H	N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	78.98	NS	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

SMAFT: Καπνιστής μετά το επεισόδιο, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη,
LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: PCI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.10γ: Επιλεγμένοι κανόνες για το Μοντέλο Αγγειπλαστική, PCI vs MI ή CABG με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

	SEX	AGE	FH	SMAFT	SMBEF	HxHTN	HxDM	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
3.1	M	2	N		N					Y			N		Y	1	3	1	0.40	1	3	1	3	3	1	3	1	87.11	S	H
3.2		2		Y	Y	Y	N							H	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	71.38	NS	M
3.3		3	N		N	N		N	N				N		N	1	3	1	0.60	1	3	1	3	3	1	3	1	111.9	NS	M
3.4		3	N	N	N	N		H	N	N					N	1	3	1	0.60	1	3	1	3	3	1	3	1	99.60	NS	H
3.5		3	Y			Y			N	Y		H		N	N	1	3	1	0.60	1	3	1	3	3	1	3	1	131.2	NS	H
3.6	M	3		N	Y			N	H	N				N	N	1	3	1	0.60	1	3	1	3	3	1	3	1	113.4	NS	H
3.7	M	4	N	N			Y	H	N				N		N	1	3	1	0.60	1	3	1	3	3	1	3	1	103.8	NS	H
3.8	M	4	Y				N		N					N	Y	1	3	1	0.40	1	3	1	3	3	1	3	1	104.3	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, SMAFT: Καπνιστής μετά το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη, LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: PCI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Στον Πίνακα 7.11 παρουσιάζονται οι εξαγόμενοι κανόνες. Η στήλη L δείχνει τους κανόνες που έχουν χαμηλό ποσοστό κινδύνου, η M δείχνει τους κανόνες που έχουν μέτριο ποσοστό κινδύνου και η στήλη H αυτούς που έχουν ψηλό ποσοστό κινδύνου για κάθε μοντέλο. Επίσης παρουσιάζεται ο αριθμός των κανόνων που είναι στατιστικά σημαντικοί και αυτών που δεν παρουσιάζουν στατιστική σημαντικότητα

Πίνακας 7.11: Εξαγόμενος αριθμός κανόνων δέντρων απόφασης για το Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG.

	Κίνδυνος επεισοδίου			Στατιστική ανάλυση Κανόνων		Σύνολο κανόνων
	L	M	H	S	NS	
B	0 (0%)	16 (29%)	39 (71%)	37 (67%)	18 (33%)	55 (100%)
A	0 (0%)	19 (39%)	30 (61%)	3 (6%)	46 (97%)	49 (100%)
B+A	0 (0%)	33 (37%)	57 (63%)	39 (43%)	51 (57%)	90 (100%)

Οι τρεις κυριότεροι παράγοντες για PCI από τα πέντε κριτήρια διαχωρισμού χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά το επεισόδιο (B+A) φαίνονται στον Πίνακα 7.12.

Πίνακας 7.12: Κυριότεροι παράγοντες για PCI vs MI ή CABG

Κριτήριο Διαχωρ.	B			A			B+A		
IG	FH	AGE	HxDM	DBP	LDL	SMAFT	HxDM	DBP	FH
GI	AGE	HxHTN	FH	DBP	LDL	SMAFT	DBP	FH	HxHTN
X2	FH	HxHTN	HxDM	DBP	LDL	SMAFT	DBP	HxHTN	AGE
GR	FH	HxHTN	HxDM	DBP	SMAFT	LDL	HxDM	FH	DBP
DM	FH	HxHTN	HxDM	DBP	LDL	SMAFT	FH	DBP	HxDM

Το μοντέλο PCI έχει ελαφρώς καλύτερη απόδοση από το μοντέλο MI. Καλύτερη απόδοση παρουσιάστηκε στα μοντέλα A και B+A, με μέση τιμή του %CC να κυμαίνεται μεταξύ 65 και 67%. Παρόμοια απόδοση φάνηκε σε όλα τα κριτήρια διαχωρισμού. Οι πιο σημαντικοί παράγοντες κινδύνου πριν το επεισόδιο είναι ηλικία, ιστορικό στην οικογένεια, υπέρταση και

ιστορικό διαβήτη, για τους παράγοντες μετά το επεισόδιο, διαστολική πίεση, χαμηλή πυκνότητα λιποπροτεϊνης και κάπνισμα μετά το επεισόδιο και για τους παράγοντες πριν και μετά το επεισόδιο, ιστορικό διαβήτη, διαστολική πίεση, ιστορικό στην οικογένεια, υπέρταση και ηλικία.

Με τη μέθοδο του Wilcoxon [135] έγινε στατιστική ανάλυση των μοντέλων που μελετήθηκαν για τα κριτήρια διαχωρισμού και τις τρεις τάξεις MI, PCI και CABG. Ο Πίνακας 7.13 δείχνει τη στατιστική ανάλυση των μέτρων διαχωρισμού με βάση το μέτρο της σωστής ταξινόμησης (correct classification) για το μοντέλο PCI vs MI ή CABG. Σε όλες τις περιπτώσεις φαίνεται ότι δεν υπάρχει στατιστική σημαντικότητα. Η ανάλυση των μοντέλων που μελετήθηκαν με τα κριτήρια διαχωρισμού που χρησιμοποιήθηκαν, εκτελώντας τον αλγόριθμο με το κριτήριο της πληροφορίας του κέρδους, έδειξε ότι δεν υπάρχει στατιστική σημαντικότητα (Πίνακας 7.14).

Πίνακας 7.13: Στατιστική ανάλυση για το μέτρο %CC των κριτηρίων διαχωρισμού για τα μοντέλο PCI vs MI ή CABG

	IG	GI	X2	GR	DM
IG		NS	NS	NS	NS
GI			NS	NS	NS
X2				NS	NS
GR					NS
DM					

S: statistically significant; NS: non statistically significant

Πίνακας 7.14: Στατιστική ανάλυση των παραγόντων κινδύνου πριν, μετά, πριν και μετά για τα κριτήρια διαχωρισμού για τα μοντέλο MI vs PCI ή CABG

	B vs A	B vs B+A	A vs B+A
IG	NS	NS	NS
GI	NS	NS	NS
X2	NS	NS	NS
GR	NS	NS	NS
DM	NS	NS	NS

S: statistically significant; NS: non statistically significant

7.2.3 Κανόνες με δέντρα απόφασης για Στεφανιαία Παράκαμψη, CABG vs MI ή PCI

Στον Πίνακα 7.15 παρουσιάζονται τα αποτελέσματα της ταξινόμησης από το μοντέλο CABG για τα πέντε κριτήρια διαχωρισμού, χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά (B+A). Δίνεται η μέση τιμή για τα %CC, %TP και %FP.

Πίνακας 7.15: Μοντέλα για Στεφανιαία παράκαμψη, CABG vs MI ή PCI για τα πέντε κριτήρια διαχωρισμού με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A). Δίνονται ο μέσος όρος (Me) (ελάχιστο (m) και μέγιστο (M)) για τις δέκα εκτελέσεις για το %CC, %TP και %FP. Για την ευαισθησία και την ειδικότητα δίνεται ο μέσος όρος

	%CC			%TP			%FP			Sensitivity			Specificity		
	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A
	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me	Me	Me	Me	Me	Me
DM	71(70,72)	61(59,67)	69(69,71)	67(63,72)	77(58,81)	70(58,74)	28(19,30)	49(40,58)	33(21,35)	73	59	70	71	67	70
GI	69(69,71)	63(61,65)	69(67,71)	67(58,74)	67(56,72)	74(72,74)	28(21,35)	42(30,42)	37(33,40)	70	63	67	68	64	70
GR	69(66,71)	63(61,66)	69(69,75)	67(65,74)	70(61,74)	74(65,77)	35(26,37)	44(28,49)	30(26,40)	67	62	69	68	65	71
IG	69(67,73)	66(63,69)	70(70,71)	70(63,77)	74(65,79)	65(63,65)	35(23,40)	42(33,47)	23(11,26)	67	67	73	70	68	68
X2	69(67,73)	63(61,65)	69(67,72)	72(63,81)	72(63,79)	74(72,77)	33(21,44)	47(42,58)	37(30,42)	67	61	67	69	66	71

Από τα μέτρα που εξάχθηκαν από τους κανόνες έγινε μια εκτέλεση με τον αλγόριθμο για δέντρα απόφασης έτσι ώστε να επιλεγούν τα σημαντικότερα για το φιλτράρισμα των κανόνων. Στον Πίνακα 7.16 παρουσιάζονται τα μέτρα που λήφθηκαν υπόψη για το φιλτράρισμα των κανόνων.

Πίνακας 7.16: Μέτρα των μοντέλων για Στεφανιαία παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A)

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

Στον Πίνακα 7.17α παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με στεφανιαία παράκαμψη πριν το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η τρίτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η δύναμη (leverage), η εμπιστοσύνη (confidence) και η ειδικότητα (specificity). Στο μοντέλο αυτό είχαμε 68% σωστή ταξινόμηση. Στον Πίνακα 7.17β παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με στεφανιαία παράκαμψη μετά το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η όγδοη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ειδικότητα (specificity) και η εμπιστοσύνη (confidence). Στο μοντέλο αυτό είχαμε 57.35% σωστή ταξινόμηση. Στον Πίνακα 7.17γ παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με στεφανιαία παράκαμψη πριν και μετά το επεισόδιο. Το κριτήριο διαχωρισμού που χρησιμοποιήθηκε ήταν το κέρδος πληροφορίας (information gain) και επιλέγηκε η τέταρτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ειδικότητα (specificity), η δύναμη (leverage) και

η εμπιστοσύνη (confidence). Παρόμοια με τα προηγούμενα δύο μοντέλα που μελετήσαμε, με τη βοήθεια της εξίσωσης του Framingham έχει υπολογιστεί ο κίνδυνος πάθησης ή μη πάθησης ενός επεισοδίου για κάθε κανόνα. Επιπλέον έχει υπολογιστεί το μέτρο chi-square, το οποίο είναι καθοριστικό για να υπολογιστεί το μέτρο p-value, που δείχνει αν κάποιος κανόνας είναι στατιστικά σημαντικός.

Οι κανόνες 1.5 και 1.6 στον Πίνακα 7.17α δείχνουν ότι άντρες ασθενείς στις ηλικίες 70 χρονών και άνω που έχουν διαβήτη δεν παίζει ρόλο αν έχουν ιστορικό στην οικογένεια και υπέρταση για να πάθουν επεισόδιο. Οι κανόνες αυτοί δεν είναι στατιστικά σημαντικοί, αλλά είναι σε ψηλά επίπεδα κινδύνου. Οι κανόνες 2.4 και 2.5 στον Πίνακα 7.17β δείχνουν ότι στους ασθενείς που έχουν ψηλά τριγλυκερίδια και διαστολική πίεση, η συστολική πίεση είναι καθοριστικός παράγοντας. Σε αυτούς τους κανόνες το ποσοστό κινδύνου είναι στα ψηλά επίπεδα. . Οι κανόνες αυτοί δεν είναι στατιστικά σημαντικοί. Οι κανόνες 3.4 και 3.5 στον Πίνακα 7.17γ δείχνουν ότι ασθενείς στις ηλικίες 60 μέχρι 70 χρονών με ψηλή γλυκόζη, καθοριστικός παράγοντας είναι ο διαβήτης. Ο κανόνας 3.4 κατατάσσεται σε αυτούς με μεσαίο κίνδυνο, ενώ ο 3.5 σε αυτούς με ψηλά επίπεδα πάθησης επεισοδίου.

Πίνακας 7.17α: Επιλεγμένοι κανόνες για το Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν από το επεισόδιο (Β)

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
1.1	F	2	N	Y	Y	N	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	49.56	NS	H
1.2	M	2	N	N	Y	Y	N	1	3	1	0.59	1	3	1	3	3	1	3	1	49.56	NS	H
1.3	M	3	Y	N	N	Y	N	1	3	1	0.59	1	3	1	3	3	1	3	1	64.53	NS	H
1.4		4	Y	Y	N	N	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	45.49	S	H
1.5	M	4	N	N	N	Y	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	65.20	NS	H
1.6	M	4	Y	N	Y	Y	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	65.20	NS	H
1.7	F	4		Y	Y	N	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	36.17	NS	H
1.8	F	4	N	Y	Y	Y	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	65.20	NS	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, class: CABG
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.17β: Επιλεγμένοι κανόνες για το Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά μετά από το επεισόδιο (A)

	SMAFT	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
2.1	N	N	N	N	H		H	N	N	1	3	1	0.59	1	3	1	3	3	1	3	1	96.17	NS	M
2.2	N	N	N	N		H	H	H	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	93.58	NS	H
2.3	N	H	N	N	H	N		N	N	1	3	1	0.59	1	3	1	3	3	1	3	1	102.3	NS	H
2.4	N	N	H	N	N	H		H	N	1	3	1	0.59	1	3	1	3	3	1	3	1	102.3	NS	H
2.5	N	H	H	N	N			H	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	76.81	NS	H
2.6	N	N	H	N	H	N		N	N	1	3	1	0.59	1	3	1	3	3	1	3	1	102.9	NS	M
2.7	N	N		Y	H	N	N	H	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	95.68	NS	M
2.8	N			Y		N	H	H	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	44.13	S	M
2.9	N	N	N	Y	N	H		H	N	1	3	1	0.59	1	3	1	3	3	1	3	1	102.9	NS	H
2.10	N	H	H	Y	H	H	H		Y	1	3	1	0.41	1	3	1	3	3	1	3	1	98.33	NS	M
2.11	Y	N	H		H	N	N	N	N	1	3	1	0.59	1	3	1	3	3	1	3	1	93.32	NS	H
2.12	Y	H	H		N		N	N	N	1	3	1	0.59	1	3	1	3	3	1	3	1	70.55	NS	H
2.13	Y	H	N	N	H		N		N	1	3	1	0.59	1	3	1	3	3	1	3	1	70.07	NS	H
2.14	Y	H	N	Y	H		N	N	N	1	3	1	0.59	1	3	1	3	3	1	3	1	96.17	NS	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

SMAFT: Καπνιστής μετά το επεισόδιο, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη,
LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: CABG
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.17γ: Επιλεγμένοι κανόνες για το Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

	SEX	AGE	FH	SMAFT	SMBEF	HxHTN	HxDM	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
3.1	M	2	N	N	Y	N	N			Y	N		N		N	1	3	1	0.59	1	3	1	3	3	1	3	1	95.24	NS	H
3.2		3		N	N	Y			H	N					Y	1	3	1	0.41	1	3	1	3	3	1	3	1	77.49	S	H
3.3		3	N	N	Y	Y		H		N	N				Y	1	3	1	0.41	1	3	1	3	3	1	3	1	87.64	NS	H
3.4		3		N		N	N	N		Y			N		N	1	3	1	0.59	1	3	1	3	3	1	3	1	96.43	NS	M
3.5		3		N			Y			Y	N	N	N	N	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	88.81	NS	H
3.6		3		N	Y		Y		N	Y	H				N	1	3	1	0.59	1	3	1	3	3	1	3	1	115.9	NS	H
3.7		4		N	Y	Y		H			N		N	N	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	75.89	NS	H
3.8		4		N		N			N			N		H	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	80.33	S	H
3.9	M	4		N					H					H	Y	1	3	1	0.41	1	3	1	3	3	1	3	1	51.60	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, SMAFT: Καπνιστής μετά το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη, LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: CABG
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Στον Πίνακα 7.18 παρουσιάζονται οι εξαγόμενοι κανόνες. Η στήλη L δείχνει τους κανόνες που έχουν χαμηλό ποσοστό κινδύνου, η M δείχνει τους κανόνες που έχουν μέτριο ποσοστό κινδύνου και η στήλη H αυτούς που έχουν ψηλό ποσοστό κινδύνου για κάθε μοντέλο. Επίσης παρουσιάζεται ο αριθμός των κανόνων που είναι στατιστικά σημαντικοί και αυτών που δεν παρουσιάζουν στατιστική σημαντικότητα

Πίνακας 7.18: Εξαγόμενος αριθμός κανόνων δέντρων απόφασης για το Μοντέλο Στεφανιαία παράκαμψη, CABG vs MI ή PCI.

	Κίνδυνος επεισοδίου			Στατιστική ανάλυση Κανόνων		Σύνολο κανόνων
	L	M	H	S	NS	
B	0 (0%)	18 (26%)	52 (74%)	11 (16%)	59 (84%)	70 (100%)
A	0 (0%)	18 (38%)	30 (62%)	9 (19%)	39 (81%)	48 (100%)
B+A	0(0%)	22 (35%)	40 (65%)	24 (39%)	38 (61%)	62 (100%)

Οι τρεις κυριότεροι παράγοντες για CABG από τα πέντε κριτήρια διαχωρισμού χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά (B+A) το επεισόδιο φαίνονται στον Πίνακα 7.19.

Πίνακας 7.19: Κυριότεροι παράγοντες για CABG vs MI ή PCI

Κριτήριο Διαχωρ.	B			A			B+A		
IG	AGE	HxHTN	SMBEF	SMAFT	SBP	DBP	AGE	SMBEF	HxDM
GI	AGE	HxDM	SMBEF	SMAFT	SBP	DBP	AGE	SMBEF	HxDM
X2	AGE	SMBEF	HxDM	SMAFT	SBP	DBP	AGE	SMBEF	SMAFT
GR	AGE	HxDM	SMBEF	SMAFT	SBP	DBP	AGE	SMAFT	HxDM
DM	AGE	HxDM	SMBEF	SMAFT	DBP	SBP	AGE	SMAFT	HxDM

Η πιο ψηλή απόδοση επιτεύχθηκε στο μοντέλο CABG, με τη μέση τιμή του %CC να είναι 70%. Όπως και στα προηγούμενα μοντέλα, δεν υπάρχει σημαντική διαφορά όσον αφορά τα διάφορα κριτήρια διαχωρισμού. Η ψηλότερη απόδοση είναι με το κριτήριο διαχωρισμού GR για το μοντέλο B+A, όπου η μέγιστη απόδοση είναι %CC = 75%. Οι πιο σημαντικοί παράγοντες για το μοντέλο πριν από ένα επεισόδιο είναι ηλικία, υπέρταση, διαβήτη και

κάπνισμα πριν το επεισόδιο, για τους παράγοντες μετά από ένα επεισόδιο κάπνισμα μετά το επεισόδιο, συστολική πίεση και διαστολική πίεση και για τους παράγοντες πριν και μετά, ηλικία, κάπνισμα πριν το επεισόδιο, κάπνισμα μετά το επεισόδιο και διαβήτη.

Με τη μέθοδο του Wilcoxon [135] έγινε στατιστική ανάλυση των μοντέλων που μελετήθηκαν για τα κριτήρια διαχωρισμού και τις τρεις τάξεις MI, PCI και CABG. Ο Πίνακας 7.20 δείχνει τη στατιστική ανάλυση των μέτρων διαχωρισμού με βάση το μέτρο της σωστής ταξινόμησης (correct classification) για το μοντέλο·CABG vs MI ή PCI. Σε όλες τις περιπτώσεις φαίνεται ότι δεν υπάρχει στατιστική σημαντικότητα. Η ανάλυση των μοντέλων που μελετήθηκαν με τα κριτήρια διαχωρισμού που χρησιμοποιήθηκαν, εκτελώντας τον αλγόριθμο με το κριτήριο της πληροφορίας του κέρδους, έδειξε ότι δεν υπάρχει στατιστική σημαντικότητα (Πίνακας 7.21).

Πίνακας 7.20: Στατιστική ανάλυση για το μέτρο %CC των κριτηρίων διαχωρισμού για τα μοντέλο CABG vs MI ή PCI

	IG	GI	X2	GR	DM
IG		NS	NS	NS	NS
GI			NS	NS	NS
X2				NS	NS
GR					NS
DM					

S: statistically significant; NS: non statistically significant

Πίνακας 7.21: Στατιστική ανάλυση των παραγόντων κινδύνου πριν, μετά, πριν και μετά για τα κριτήρια διαχωρισμού για τα μοντέλο MI vs PCI ή CABG

	B vs A	B vs B+A	A vs B+A
IG	NS	NS	NS
GI	NS	NS	NS
X2	NS	NS	NS
GR	NS	NS	NS
DM	NS	NS	NS

S: statistically significant; NS: non statistically significant

7.3 Εξαγωγή κανόνων συσχέτισης με τον αλγόριθμο AKAMAS

Με τη χρήση του αλγόριθμου AKAMAS έχουν χρησιμοποιηθεί τα μοντέλα πριν, μετά και πριν και μετά το επεισόδιο για το έμφραγμα μυοκαρδίου, αγγειοπλαστική και στεφανιαία παράκαμψη.

Οι εξαγόμενοι κανόνες από κάθε μοντέλο μελετήθηκαν ως προς τους παράγοντες που είχε ο κάθε κανόνας. Με τη μέθοδο της συχνότητας έχουν υπολογιστεί όλοι οι παράγοντες όλων των κανόνων σε κάθε μοντέλο και επιλέγηκαν οι τρεις σημαντικότεροι.

7.3.1 Κανόνες συσχέτισης για Έμφραγμα μυοκαρδίου, MI vs PCI ή CABG

Στον Πίνακα 7.22 παρουσιάζονται τα αποτελέσματα της ταξινόμησης από το μοντέλο MI vs PCI ή CABG, χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά (B+A). Έγιναν πέντε εκτελέσεις και δίνεται η μέση τιμή, η ελάχιστη και η μέγιστη για το %CC για όλους τους εξαγόμενους κανόνες, όπως επίσης και για τους στατιστικά σημαντικούς κανόνες.

Πίνακας 7.22: Ταξινόμηση για Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A). Δίνονται ο μέσος όρος (Me) (ελάχιστο (m) και μέγιστο (M)) για τις πέντε εκτελέσεις για το %CC, για όλους και για τους στατιστικά σημαντικούς κανόνες

%CC					
B		A		B+A	
Me(m,M)		Me(m,M)		Me(m,M)	
Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες	Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες	Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες
57(56,58)	58(57,60)	58(57,60)	59(57,61)	65(64,67)	88(87,89)

Από τα μέτρα που εξάχθηκαν από τους κανόνες έγινε μια εκτέλεση με τον αλγόριθμο για δέντρα απόφασης έτσι ώστε να επιλεγούν τα σημαντικότερα για το φιλτράρισμα των κανόνων. Στον Πίνακα 7.23 παρουσιάζονται τα μέτρα που λήφθηκαν υπόψη για το φιλτράρισμα των κανόνων.

Πίνακας 7.23: Μέτρα των μοντέλων για Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A)

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

Στον Πίνακα 7.24α παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με έμφραγμα μυοκαρδίου με χαρακτηριστικά πριν το επεισόδιο. Επιλέγηκε η πρώτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy) και η υποστήριξη (support). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 73.34% σωστή ταξινόμηση των μέτρων. Στον Πίνακα 7.24β παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με έμφραγμα μυοκαρδίου με χαρακτηριστικά μετά το επεισόδιο. Επιλέγηκε η πρώτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy) και η υποστήριξη (support). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 71.63% σωστή ταξινόμηση των μέτρων. Στον Πίνακα 7.24γ παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με έμφραγμα μυοκαρδίου με χαρακτηριστικά πριν και μετά το επεισόδιο. Επιλέγηκε η δεύτερη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων

είναι η ορθότητα (accuracy), η υποστήριξη (support) και η ανάκληση (recall). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 72.60% σωστή ταξινόμηση των μέτρων. Με τη βοήθεια της εξίσωσης του Framingham έχει υπολογιστεί ο κίνδυνος πάθησης ή όχι ενός επεισοδίου για κάθε κανόνα. Επιπλέον έχει υπολογιστεί το μέτρο chi-square, το οποίο είναι καθοριστικό για να υπολογιστεί το μέτρο p-value, που δείχνει αν κάποιος κανόνας είναι στατιστικά σημαντικός. Στους πίνακες μας για έμφραγμα μυοκαρδίου πριν, μετά και πριν και μετά το επεισόδιο, έχουμε κανόνες που είναι στατιστικά σημαντικοί. Επίσης το ποσοστό κινδύνου πάθησης επεισοδίου είναι στην ψηλή κατηγορία.

Οι κανόνες 1.8 και 1.10 στον Πίνακα 7.24α δείχνουν ότι ασθενείς καπνιστές παθαίνουν έμφραγμα μυοκαρδίου, με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου.. Επίσης αυτοί οι κανόνες είναι στατιστικά σημαντικοί. Οι κανόνες 2.5 και 2.13 στον Πίνακα 7.24β δείχνουν ότι ασθενείς με ψηλά ποσοστά λιποπρωτεϊνων χαμηλής πυκνότητας παθαίνουν έμφραγμα μυοκαρδίου. Αυτοί οι κανόνες δεν είναι στατιστικά σημαντικοί και οι ασθενείς που αντιπροσωπεύει είναι με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου. Οι κανόνες 3.2 και 3.5 στον Πίνακα 7.24γ δείχνουν ότι ασθενείς που καπνίζουν παθαίνουν έμφραγμα μυοκαρδίου. Επίσης αυτοί οι κανόνες είναι στατιστικά σημαντικοί και οι ασθενείς που αντιπροσωπεύουν είναι με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου.

Πίνακας 7.24α: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν από το επεισόδιο (Β)

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
1.1					Y		Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	1.04	NS	H
1.2				Y			Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	6.18	S	H
1.3			N				Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	0.44	NS	H
1.4						N	Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	0.03	NS	H
1.5	M						Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	1.68	NS	H
1.6				Y		N	Y	3	1	2	0.68	2	1	3	1	1	1	1	1	1	15.71	S	H
1.7	M		N				Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	2.46	NS	H
1.8	M			Y			Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	6.30	S	H
1.9	M					N	Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	2.45	NS	H
1.10	M			Y		N	Y	3	1	2	0.68	2	1	3	1	1	1	1	1	1	16.43	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, class: MI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.24β: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά μετά από το επεισόδιο (A)

	SMAFT	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
2.1		N							Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	6.71	S	M
2.2					N				Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	1.65	NS	H
2.3				N					Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	0.01	NS	H
2.4								N	Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	0.52	NS	H
2.5						H			Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	0.01	NS	H
2.6			N						Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	0.05	NS	H
2.7	N								Y	3	1	3	0.68	3	1	3	1	1	1	1	3	1	7.30	S	H
2.8							N		Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	0.85	NS	H
2.9	N				N				Y	3	1	2	0.68	2	1	3	1	1	1	1	1	1	15.24	S	H
2.10							N	N	Y	3	1	2	0.68	2	1	3	1	1	1	1	1	1	1.97	NS	H
2.11		N					N		Y	3	1	2	0.68	2	1	3	1	1	1	1	1	1	8.88	S	H
2.12					N		N		Y	3	1	2	0.68	3	1	3	1	1	1	1	1	1	2.38	NS	H
2.13						H	N		Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	1.22	NS	H
2.14			N				N		Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	1.09	NS	H
2.15	N						N		Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	8.31	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

SMAFT: Καπνιστής μετά το επεισόδιο, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη,
LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: MI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.24γ: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

	SEX	AGE	FH	SMAFT	SMBEF	HxHTN	HxDM	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
3.1								N							Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	6.71	S	M
3.2					Y										Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	6.18	S	H
3.3				N											Y	3	1	3	0.68	3	1	3	1	1	1	1	3	1	7.30	S	H
3.4	M											H			Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	10.11	S	H
3.5					Y								N		Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	13.56	S	H
3.6	M			N											Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	10.59	S	H
3.7				N									N		Y	3	1	3	0.68	3	1	3	1	1	1	1	1	1	8.31	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, SMAFT: Καπνιστής μετά το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη, LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: MI

S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός

Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Στον Πίνακα 7.25 παρουσιάζονται οι εξαγόμενοι κανόνες. Η στήλη L δείχνει τους κανόνες που έχουν χαμηλό ποσοστό κινδύνου, η M δείχνει τους κανόνες που έχουν μέτριο ποσοστό κινδύνου και η στήλη H αυτούς που έχουν ψηλό ποσοστό κινδύνου για κάθε μοντέλο. Επίσης παρουσιάζεται ο αριθμός των κανόνων που είναι στατιστικά σημαντικοί και αυτών που δεν παρουσιάζουν στατιστική σημαντικότητα

Πίνακας 7.25: Εξαγόμενος αριθμός κανόνων συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Έμφραγμα Μυοκαρδίου, MI vs PCI ή CABG.

	Κίνδυνος επεισοδίου			Στατιστική ανάλυση Κανόνων		Σύνολο κανόνων
	L	M	H	S	NS	
B	0 (0%)	441 (38%)	720 (62%)	323 (28%)	838 (72%)	1161 (100%)
A	23 (0.5%)	1744 (33.5%)	3421 (66%)	1530 (29%)	3658 (71%)	5188 (100%)
B+A	0 (0%)	796 (23%)	2610 (77%)	1125 (33%)	2281 (67%)	3406 (100%)

Ο Πίνακας 7.26 παρουσιάζει τους κυριότερους παράγοντες πάθησης ενός επεισοδίου εμφράγματος μυοκαρδίου στα μοντέλα πριν το επεισόδιο, μετά το επεισόδιο και πριν και μετά το επεισόδιο. Κύριοι παράγοντες είναι εκείνοι που εμφανίζονται πιο συχνά στους κανόνες. Οι παράγοντες φύλο, ιστορικό στην οικογένεια και κάπνισμα είναι οι πιο σημαντικοί πριν το επεισόδιο, όπως δείχνουν οι εξαγόμενοι κανόνες. Οι παράγοντες συστολική πίεση, διαστολική πίεση και χοληστερόλη είναι οι πιο σημαντικοί μετά το επεισόδιο και οι παράγοντες ηλικία, κάπνισμα πριν το επεισόδιο και υπέρταση, είναι οι πιο σημαντικοί για πριν και μετά το επεισόδιο.

Πίνακας 7.26: Κυριότεροι παράγοντες για MI vs PCI ή CABG

	B			A			B+A		
Παράγοντες	SEX	FH	SMBEF	SBP	DBP	TC	AGE	SMBEF	HxHTN

7.3.2 Κανόνες συσχέτισης για Αγγειοπλαστική, PCI vs MI ή CABG

Στον Πίνακα 7.27 παρουσιάζονται τα αποτελέσματα της ταξινόμησης από το μοντέλο MI vs PCI ή CABG, χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά (B+A).). Έγιναν πέντε εκτελέσεις και δίνεται η μέση τιμή, η ελάχιστη και η μέγιστη για το %CC για όλους τους εξαγόμενους κανόνες, όπως επίσης και για τους στατιστικά σημαντικούς κανόνες.

Πίνακας 7.27: Ταξινόμηση για Αγγειοπλαστική, PCI vs MI ή CABG με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A). Δίνονται ο μέσος όρος (Me) (ελάχιστο (m) και μέγιστο (M)) για τις πέντε εκτελέσεις για το %CC, για όλους και για τους στατιστικά σημαντικούς κανόνες

%CC					
B		A		B+A	
Me(m,M)		Me(m,M)		Me(m,M)	
Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες	Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες	Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες
56(55,59)	68(66,70)	54(53,56)	67(66,69)	58(58,59)	70(69,71)

Από τα μέτρα που εξάχθηκαν από τους κανόνες έγινε μια εκτέλεση με τον αλγόριθμο για δέντρα απόφασης έτσι ώστε να επιλεγούν τα σημαντικότερα για το φιλτράρισμα των κανόνων. Στον Πίνακα 7.28 παρουσιάζονται τα μέτρα που λήφθηκαν υπόψη για το φιλτράρισμα των κανόνων.

Πίνακας 7.28: Μέτρα των μοντέλων για Αγγειοπλαστική, PCI vs MI ή CABG χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A)

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

Στον Πίνακα 7.29α παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με αγγειοπλαστική πριν το επεισόδιο. Επιλέγηκε η τρίτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy) και η ανάκληση (recall). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 75.75% σωστή ταξινόμηση των μέτρων. Στον Πίνακα 7.29β παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με αγγειοπλαστική μετά το επεισόδιο. Επιλέγηκε η τέταρτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy), η ανάκληση (recall) και η κάλυψη (coverage). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 71.90% σωστή ταξινόμηση των μέτρων. Στον Πίνακα 7.29γ παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με αγγειοπλαστική πριν και μετά το επεισόδιο. Επιλέγηκε η πρώτη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy), η υποστήριξη (support), η ανάκληση (recall) και η κάλυψη (coverage). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 72.07% σωστή ταξινόμηση των μέτρων. Στους πίνακες μας για αγγειοπλαστική πριν, μετά και πριν και μετά το επεισόδιο, έχουμε κανόνες που είναι στατιστικά σημαντικοί. Επίσης το ποσοστό κινδύνου πάθησης επεισοδίου είναι στην ψηλή κατηγορία.

Οι κανόνες 1.1 1.2 στον Πίνακα 7.29α δείχνουν ότι η υπέρταση δεν είναι καθοριστικός παράγοντας. Επίσης αυτοί ο κανόνες είναι στατιστικά σημαντικοί και οι ασθενείς που αντιπροσωπεύουν είναι με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου. Οι κανόνες 2.3 και 2.7 στον Πίνακα 7.29β δείχνουν ότι ασθενείς με ψηλά ποσοστά λιποπρωτεϊνων ψηλής πυκνότητας παθαίνουν επεισόδιο. Αυτοί οι κανόνες δεν είναι στατιστικά σημαντικοί, οι ασθενείς που αντιπροσωπεύουν είναι με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου. Οι κανόνες 3.3 και 3.6 στον Πίνακα 7.29γ δείχνουν ότι ασθενείς καπνιστές κάνουν αγγειοπλαστική. Επίσης αυτοί οι κανόνες είναι στατιστικά σημαντικοί και οι ασθενείς που αντιπροσωπεύει είναι με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου.

Πίνακας 7.29α: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG με χαρακτηριστικά πριν από το επεισόδιο (Β)

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
1.1					Y		N	3	1	3	0.6	3	1	3	1	1	1	1	1	1	0.09	NS	H
1.2					Y		Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	0.09	NS	H
1.3				Y			Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	7.85	S	H
1.4			N				N	3	1	3	0.6	3	1	3	1	1	1	1	1	1	14.78	S	M
1.5						N	Y	3	1	3	0.4	3	2	3	1	1	1	1	1	1	10.48	S	H
1.6	M						N	3	1	3	0.6	3	1	3	1	1	1	1	3	1	4.40	S	H
1.7				Y		N	Y	2	1	2	0.4	3	1	3	1	1	1	1	1	1	21.20	S	H
1.8			N			N	N	3	1	3	0.6	3	1	3	1	1	1	1	1	1	32.89	S	M
1.9	M		N				N	3	1	3	0.6	3	1	3	1	1	1	1	1	1	18.68	S	M
1.10	M			Y			Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	9.24	S	H
1.11	M					N	Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	14.07	S	H
1.12	M			Y		N	Y	2	1	2	0.4	3	1	3	1	1	1	1	1	1	24.13	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, class: PCI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.29β: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG με χαρακτηριστικά μετά από το επεισόδιο (Α)

	SMAFT	SBP	DBP	GLU	TC	LDL	HDL	TG	Class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
2.1					N				N	3	1	3	0.6	3	1	2	1	1	1	1	1	1	0.26	NS	H
2.2				N					Y	2	1	3	0.4	3	1	2	1	1	1	1	1	1	0.03	NS	H
2.3						H			Y	2	1	3	0.4	3	1	2	1	1	1	1	1	1	0.95	NS	H
2.4			N						N	3	1	3	0.6	3	1	2	1	1	1	1	3	1	10.81	S	H
2.5	N								Y	3	1	3	0.4	3	1	2	1	1	1	1	1	1	1.81	NS	H
2.6							N		Y	3	1	3	0.4	3	1	2	1	1	1	1	1	1	0.43	NS	H
2.7						H	N		Y	2	1	3	0.4	3	1	2	1	1	1	1	1	1	1.36	NS	H
2.8			N				N		N	3	1	3	0.6	3	1	2	1	1	1	1	3	1	10.82	S	H
2.9	N						N		Y	2	1	3	0.4	3	1	2	1	1	1	1	1	1	2.38	NS	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

SMAFT: Καπνιστής μετά το επεισόδιο, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη,
LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: PCI
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.29γ: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

	SEX	AGE	FH	SMAFT	SMBEF	HxHTN	HxDM	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
3.1											N				Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	0.26	NS	H
3.2						Y									Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	0.09	NS	H
3.3					Y										Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	7.85	S	H
3.4									N						Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	10.18	S	H
3.5				N					N						Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	12.02	S	H
3.6					Y								N		Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	9.11	S	H
3.7	M								N						Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	14.15	S	H
3.8	M				Y										Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	9.24	S	H
3.9									N				N		Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	10.82	S	H
3.10				N			N								Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	12.27	S	H
3.11	M						N								Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	14.07	S	H
3.12	M			N			N								Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	15.76	S	H
3.13	M								N				N		Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	16.01	S	H
3.14	M				Y								N		Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	10.74	NS	H
3.15				N			N						N		Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	14.82	S	H
3.16	M						N						N		Y	2	1	3	0.4	3	1	3	1	1	1	1	1	1	16.64	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, SMAFT: Καπνιστής μετά το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη, LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: PCI

S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός

Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Στον Πίνακα 7.30 παρουσιάζονται οι εξαγόμενοι κανόνες. Η στήλη L δείχνει τους κανόνες που έχουν χαμηλό ποσοστό κινδύνου, η M δείχνει τους κανόνες που έχουν μέτριο ποσοστό κινδύνου και η στήλη H αυτούς που έχουν ψηλό ποσοστό κινδύνου για κάθε μοντέλο. Επίσης παρουσιάζεται ο αριθμός των κανόνων που είναι στατιστικά σημαντικοί και αυτών που δεν παρουσιάζουν στατιστική σημαντικότητα

Πίνακας 7.30: Εξαγόμενος αριθμός κανόνων συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Αγγειοπλαστική, PCI vs MI ή CABG

	Κίνδυνος επεισοδίου			Στατιστική ανάλυση Κανόνων		Σύνολο κανόνων
	L	M	H	S	NS	
B	0 (0%)	395 (34%)	772 (66%)	1058 (91%)	109 (9%)	1167 (100%)
A	31 (1%)	1796 (33%)	3461 (65%)	504 (10%)	4784 (90%)	5288 (100%)
B+A	0 (0%)	1034 (36%)	1868 (64%)	1466 (51%)	1436 (49%)	2902 (100%)

Ο Πίνακας 7.31 παρουσιάζει τους κυριότερους παράγοντες πάθησης ενός επεισοδίου εμφράγματος μυοκαρδίου στα μοντέλα πριν το επεισόδιο, μετά το επεισόδιο και πριν και μετά το επεισόδιο. Κύριοι παράγοντες είναι εκείνοι που εμφανίζονται πιο συχνά στους κανόνες. Οι παράγοντες ιστορικό στην οικογένεια, υπέρταση και διαβήτης είναι οι πιο σημαντικοί πριν το επεισόδιο, όπως δείχνουν οι εξαγόμενοι κανόνες. Οι παράγοντες διαστολική πίεση, λιποπρωτεΐνες χαμηλής πυκνότητας και κάπνισμα μετά το επεισόδιο είναι οι πιο σημαντικοί μετά το επεισόδιο και οι παράγοντες διαστολική πίεση, ιστορικό στην οικογένεια και υπέρταση είναι οι πιο σημαντικοί για πριν και μετά το επεισόδιο.

Πίνακας 7.31 Κυριότεροι παράγοντες για PCI vs MI ή CABG

	B			A			B+A		
Παράγοντες	FH	HxHTN	HxDM	DBP	LDL	SMAFT	DBP	FH	HxHTN

Για αυτό το μοντέλο χρησιμοποιήσαμε το μέτρο coverage. Λαμβάνοντας υπ' όψη μόνο ένα παράγοντα, υποθέτουμε ότι δεν υπάρχει διαφορά σε ασθενείς που έπαθαν επεισόδιο με εκείνους που δεν έχουν πάθει.

7.3.3 Κανόνες συσχέτισης για Στεφανιαία Παράκαμψη, CABG vs MI ή PCI

Στον Πίνακα 7.32 παρουσιάζονται τα αποτελέσματα της ταξινόμησης από το μοντέλο MI vs PCI ή CABG, χρησιμοποιώντας τους παράγοντες πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά (B+A). Έγιναν πέντε εκτελέσεις και δίνεται η μέση τιμή, η ελάχιστη και η μέγιστη για το %CC για όλους τους εξαγόμενους κανόνες, όπως επίσης και για τους στατιστικά σημαντικούς κανόνες.

Πίνακας 7.32: Ταξινόμηση για Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A). Δίνονται ο μέσος όρος (Me) (ελάχιστο (m) και μέγιστο (M)) για τις πέντε εκτελέσεις για το %CC, για όλους και για τους στατιστικά σημαντικούς κανόνες

%CC					
B		A		B+A	
Me(m,M)		Me(m,M)		Me(m,M)	
Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες	Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες	Όλοι οι κανόνες	Στατιστικά σημαντικοί κανόνες
61(61,62)	61(60,63)	59(56,61)	60(58,61)	59(56,60)	87(85,87)

Από τα μέτρα που εξάχθηκαν από τους κανόνες έγινε μια εκτέλεση με τον αλγόριθμο για δέντρα απόφασης έτσι ώστε να επιλεγούν τα σημαντικότερα για το φιλτράρισμα των κανόνων. Στον Πίνακα 7.33 παρουσιάζονται τα μέτρα που λήφθηκαν υπόψη για το φιλτράρισμα των κανόνων.

Πίνακας 7.33: Μέτρα των μοντέλων για Στεφανιαία παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν το επεισόδιο (B), μετά το επεισόδιο (A), και πριν και μετά (B+A)

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

Στον Πίνακα 7.34α παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με στεφανιαία παράκαμψη πριν το επεισόδιο. Επιλέγηκε η δεύτερη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy) και η ανάκληση (recall). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 71.92% σωστή ταξινόμηση των μέτρων. Στον Πίνακα 7.34β παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με στεφανιαία παράκαμψη μετά το επεισόδιο. Επιλέγηκε η δεύτερη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy), η υποστήριξη (support) και η ανάκληση (recall). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 70.37% σωστή ταξινόμηση των μέτρων. Στον Πίνακα 7.34γ παρουσιάζονται επιλεγμένοι κανόνες που έχουν εξαχθεί για τους ασθενείς με στεφανιαία παράκαμψη πριν και μετά το επεισόδιο. Επιλέγηκε η δέκατη εκτέλεση (run). Τα μέτρα που λήφθηκαν υπόψη για την επιλογή αυτών των κανόνων είναι η ορθότητα (accuracy) και η υποστήριξη (support). Η επιλογή αυτών των μέτρων έγινε κάνοντας χρήση του αλγόριθμου δέντρων απόφασης πάνω στους εξαγόμενους κανόνες από τη μέθοδο της συσχέτισης λαμβάνοντας υπόψη τα κωδικοποιημένα μέτρα. Στο μοντέλο αυτό είχαμε 72.35% σωστή ταξινόμηση των μέτρων.

Οι κανόνες 1.5 και 1.6 στον Πίνακα 7.34α δείχνουν ότι η υπέρταση είναι πιο σημαντικός παράγοντας από το κάπνισμα. Αυτοί ο κανόνες αντιπροσωπεύουν ασθενείς με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου. Οι κανόνες 2.1 και 2.4 στον Πίνακα 7.34β δείχνουν ότι ασθενείς με ψηλή συστολική πίεση παθαίνουν επεισόδιο. Αυτοί ο κανόνες αντιπροσωπεύουν ασθενείς με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου. Οι κανόνες 3.2 και 3.6 στον Πίνακα 7.34γ δείχνουν ότι ασθενείς με ψηλά ποσοστά λιποπρωτεϊνων χαμηλής πυκνότητας κάνουν αγγειοπλαστική. Επίσης αυτοί οι κανόνες αντιπροσωπεύουν ασθενείς με ψηλό ποσοστό κινδύνου πάθησης επεισοδίου.

Πίνακας 7.34α: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν από το επεισόδιο (Β)

	SEX	AGE	FH	SMBEF	HxHTN	HxDM	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
1.1		3					Y	1	1	1	0.41	2	1	3	1	1	1	1	1	1	0.61	NS	H
1.2				N			Y	1	1	2	0.41	2	1	3	1	1	1	1	1	1	0.43	NS	H
1.3	M				N		N	2	1	2	0.59	2	1	3	1	1	1	1	1	1	3.83	NS	H
1.4				Y		N	N	2	1	2	0.59	2	1	3	1	1	1	1	1	1	4.24	NS	H
1.5	M				Y		Y	1	1	2	0.41	2	1	3	1	1	1	1	1	1	3.83	NS	H
1.6	M			Y		N	N	2	1	2	0.59	2	1	3	1	1	1	1	1	1	4.46	NS	H
1.7	M		N			N	Y	1	1	2	0.41	2	1	3	1	1	1	1	1	1	8.55	NS	M

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, class: CABG
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.34β: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά μετά από το επεισόδιο (Α)

	SMAFT	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
2.1		H							Y	2	1	2	0.41	3	1	3	1	1	1	1	1	1	2.64	NS	H
2.2					N				Y	2	1	3	0.41	3	1	3	1	1	1	1	1	1	0.73	NS	H
2.3				N					Y	2	1	3	0.41	3	1	3	1	1	1	1	1	1	0.51	NS	H
2.4		H					N		Y	2	1	2	0.41	3	1	3	1	1	1	1	1	1	7.15	S	H
2.5	N					H			Y	2	1	2	0.41	3	1	3	1	1	1	1	1	1	18.99	S	H
2.6					N		N		Y	2	1	2	0.41	3	1	3	1	1	1	1	1	1	10.31	S	H
2.7	N		N						Y	2	1	3	0.41	3	1	3	1	1	1	1	1	1	19.87	S	H
2.8	N		N				N		Y	2	1	2	0.41	3	1	3	1	1	1	1	1	1	21.50	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

SMAFT: Καπνιστής μετά το επεισόδιο, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη,
LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας,, TG: Τριγλυκερίδια, class: CABG
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Πίνακας 7.34γ: Επιλεγμένοι κανόνες συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI με χαρακτηριστικά πριν και μετά από το επεισόδιο (B+A)

	SEX	AGE	FH	SMAFT	SMBEF	HxHTN	HxDM	SBP	DBP	GLU	TC	LDL	HDL	TG	class	Support	Confidence	Coverage	Prevalence	Recall	Specificity	Accuracy	Lift/Interest	Leverage	AddedValue	RelativeRisk	OddsRatio	Conviction	ChiSquare	P-Value	EventRisk
3.1								N							N	3	1	3	0.59	3	1	3	1	1	1	1	1	1	2.64	NS	M
3.2												H			N	3	1	3	0.59	3	1	3	1	1	1	1	1	1	1.91	NS	H
3.3							N								N	3	1	3	0.59	3	1	3	1	1	1	1	1	1	1.18	NS	H
3.4				N											Y	3	1	3	0.41	3	3	3	1	1	1	1	1	1	17.90	S	H
3.5					Y								N		N	3	1	3	0.59	3	1	3	1	1	1	1	1	1	4.57	NS	H
3.6												H	N		N	3	1	3	0.59	3	1	3	1	1	1	1	1	1	5.25	NS	H
3.7	M			N											Y	3	1	3	0.41	3	2	3	1	1	1	1	1	1	21.21	S	H
3.8				N									N		Y	3	1	3	0.41	3	2	3	1	1	1	1	1	1	19.30	S	H

(Βλέπε Πίνακα 6.3 για την κωδικοποίηση των χαρακτηριστικών και Πίνακα 6.18 για την κωδικοποίηση των μέτρων)

FH: Ιστορικό οικογένειας, SMBEF: Καπνιστής πριν το επεισόδιο, SMAFT: Καπνιστής μετά το επεισόδιο, HxHTN: Ιστορικό Υπέρτασης, HxDM: Ιστορικό Διαβήτη, SBP: Συστολική πίεση, DBP: Διαστολική πίεση, GLU: Γλυκόζη, TC: Χοληστερόλη, LDL: Λιποπρωτεΐνες χαμηλής πυκνότητας, HDL: Λιποπρωτεΐνες ψηλής πυκνότητας, TG: Τριγλυκερίδια, class: CABG
S: Στατιστικά σημαντικός κανόνας, όπου $p<0.05$, NS: όχι στατιστικά σημαντικός
Event_Risk: L: χαμηλό, M: μέτριο, H: ψηλό

Στον Πίνακα 7.35 παρουσιάζονται οι εξαγόμενοι κανόνες. Η στήλη L δείχνει τους κανόνες που έχουν χαμηλό ποσοστό κινδύνου, η M δείχνει τους κανόνες που έχουν μέτριο ποσοστό κινδύνου και η στήλη H αυτούς που έχουν ψηλό ποσοστό κινδύνου για κάθε μοντέλο. Επίσης παρουσιάζεται ο αριθμός των κανόνων που είναι στατιστικά σημαντικοί και αυτών που δεν παρουσιάζουν στατιστική σημαντικότητα

Πίνακας 7.35: Εξαγόμενος αριθμός κανόνων συσχέτισης με τον αλγόριθμο AKAMAS για το Μοντέλο Στεφανιαία Παράκαμψη, CABG vs MI ή PCI

	Κίνδυνος επεισοδίου			Στατιστική ανάλυση Κανόνων		Σύνολο κανόνων
	L	M	H	S	NS	
B	1 (0.1%)	450 37.9%)	741 (62%)	277 (23%)	915 (77%)	1192 (100%)
A	10 (0.3%)	1812 (49%)	1867 (50.7%)	1676 (45%)	2013 (55%)	3689 (100%)
B+A	0 (0%)	673 (37%)	1141 (63%)	944 (52%)	870 (48%)	1814 (100%)

Ο Πίνακας 7.36 παρουσιάζει τους κυριότερους παράγοντες πάθησης ενός επεισοδίου εμφράγματος μυοκαρδίου στα μοντέλα πριν το επεισόδιο, μετά το επεισόδιο και πριν και μετά το επεισόδιο. Κύριοι παράγοντες είναι εκείνοι που εμφανίζονται πιο συχνά στους κανόνες. Οι παράγοντες ηλικία, διαβήτης και κάπνισμα πριν το επεισόδιο είναι οι πιο σημαντικοί πριν το επεισόδιο, όπως δείχνουν οι εξαγόμενοι κανόνες. Οι παράγοντες κάπνισμα μετά το επεισόδιο, συστολική πίεση και διαστολική πίεση είναι οι πιο σημαντικοί μετά το επεισόδιο και οι παράγοντες ηλικία κάπνισμα πριν το επεισόδιο και διαβήτης είναι οι πιο σημαντικοί για πριν και μετά το επεισόδιο.

Πίνακας 7.36: Κυριότεροι παράγοντες για CABG vs MI ή PCI

	B			A			B+A		
Παράγοντες	AGE	HxDM	SMBEF	SMAFT	SBP	DBP	AGE	SMBEF	HxDM

Για αυτό το μοντέλο χρησιμοποιήσαμε το μέτρο relative risk. Υπάρχουν πολλοί κανόνες που παρουσιάζουν στατιστική σημασία και θεωρούνται οι πιο σημαντικοί.

7.3.4 Σύγκριση αλγορίθμων Apriori και AKAMAS

Όπως έχει αναφερθεί πιο πάνω κι οι δύο αλγόριθμοι χρησιμοποιούνται για εξαγωγή κανόνων συσχέτισης. Να σημειωθεί ότι εάν ο χρήστης επιλέξει να αξιολογούνται οι κανόνες συσχέτισης στον αλγόριθμο AKAMAS και με κάποιο όριο για υποστήριξη, όπως και στον Apriori, τότε οι αλγόριθμοι θα παράγουν τους ίδιους ακριβώς κανόνες.

Ο αλγόριθμός Apriori, χρησιμοποιεί τα συχνά σύνολα χαρακτηριστικών με κ-1 χαρακτηριστικά, που ικανοποιούν το ελάχιστο όριο υποστήριξης, για να παραχθούν τα συχνά σύνολα χαρακτηριστικών με κ χαρακτηριστικά. Επομένως, τα αποτελέσματα του αλγόριθμου Apriori εξαρτώνται από την υποστήριξη, και θα παραχθούν κανόνες με κ χαρακτηριστικά, μόνο εάν στην προηγούμενη επανάληψη, τα κ-1 χαρακτηριστικά ικανοποιούσαν την υποστήριξη. Ενώ αυτό δεν ισχύει στον αλγόριθμο AKAMAS.

Ο αλγόριθμος AKAMAS στηρίζεται στην υποστήριξη μόνο στην πρώτη επανάληψη, όπου βρίσκει τα συχνά σύνολα χαρακτηριστικών με 1 χαρακτηριστικό, τα οποία ικανοποιούν την ελάχιστη υποστήριξη. Μετά κτίζει όλους του δυνατούς συνδυασμούς συνόλων χαρακτηριστικών, για να βρει κανόνες συσχέτισης χωρίς να εξαρτάται από το πόσο συχνά αυτά τα σύνολα χαρακτηριστικών εμφανίζονται στη βάση στο σύστημα. Έτσι με αυτό το τρόπο μπορούν να βρεθούν όλοι οι δυνατοί κανόνες συσχέτισης και να αξιολογηθούν με οποιοδήποτε μέτρο αξιολόγησης κανόνων θέλει ο χρήστης, κι όχι μόνο με βάση την υποστήριξη.

Μια άλλη σημαντική διαφορά των δύο αλγόριθμων αφορά την ταχύτητα εξαγωγής κανόνων. Λόγω του ότι ο αλγόριθμος AKAMAS κτίζει όλους τους δυνατούς συνδυασμούς, χωρίς να απορρίπτει κάποιους κανόνες, για να ελαττώνει τον χώρο αναζήτησης, είναι πιο αργός από τον αλγόριθμο Apriori. Αυτό συμβαίνει σε μεγάλες βάσεις δεδομένων ή σε βάσεις δεδομένων όπου εξάγουν μεγάλο αριθμό κανόνων. Για παράδειγμα, αν χρησιμοποιήσω μια βάση δεδομένων για να εκτελέσω τους δύο αλγόριθμους με μεγάλο κατώφλι υποστήριξης, δεν θα υπάρχει διαφορά στον χρόνο εκτέλεσης γιατί ο αριθμός των εξαγόμενων κανόνων είναι πολύ μικρός. Αν χρησιμοποιήσω ένα μέτριο κατώφλι υποστήριξης, θα αντιληφθώ ότι ο Apriori

είναι ελαφρώς πιο γρήγορος. Αν όμως χρησιμοποιήσω κατώφλι υποστήριξης πολύ μικρό, τότε η διαφορά της εκτέλεσης των δύο αλγορίθμων θα αυξηθεί πολύ.

7.4 Αξιολόγηση αποτελεσμάτων

Λαμβάνοντας υπόψη το ποσοστό της σωστής ταξινόμησης στις 10 εκτελέσεις επιλέγεται η καταλληλότερη εκτέλεση και κατ' επέκταση οι κανόνες. Σε αυτή την εκτέλεση γίνεται επιλογή των καλύτερων μέτρων με τη βοήθεια της κατηγοριοποίησης των μέτρων. Εφαρμόζοντας τα καλύτερα μέτρα στους εξαγόμενους κανόνες μειώνονται οι κανόνες. Αυτοί είναι οι φιλτραρισμένοι κανόνες, που θεωρούνται και οι σημαντικότεροι στο μοντέλο. Μεγαλύτερη έμφαση δίνεται στους κανόνες που είναι στατιστικά σημαντικοί. Όπως φάνηκε και από τις εκτελέσεις που έγιναν, οι στατιστικά σημαντικοί κανόνες δίνουν μεγαλύτερα ποσοστά σωστής ταξινόμησης. Επίσης σε κάθε κανόνα λαμβάνεται υπόψη και το ποσοστό κινδύνου πάθησης ενός επεισοδίου.

Άρα σε κάθε μοντέλο δίνεται έμφαση στους φιλτραρισμένους κανόνες και μετά στους υπόλοιπους, όπως επίσης λαμβάνεται υπόψη αν κάποιος κανόνας είναι στατιστικά σημαντικός και αν το ποσοστό κινδύνου πάθησης ενός επεισοδίου βρίσκεται στα ψηλά επίπεδα.

Κεφάλαιο 8: Συζήτηση

Στόχος της διατριβής αυτής είναι η αξιολόγηση των παραγόντων κινδύνου σε καρδιαγγειακές βάσεις δεδομένων και την εξόρυξη κανόνων εκτίμησης κινδύνου βασισμένων σε αλγόριθμους δένδρων αποφάσεων και κανόνων συσχέτισης. Με τη βοήθεια των εξαγόμενων κανόνων θα μπορεί ο καρδιολόγος να αξιολογήσει ένα νέο ασθενή, συγκρίνοντας τις τιμές των παραγόντων του ασθενή με αυτές των κανόνων.

Σε αυτή τη διατριβή για την εξόρυξη γνώσης με την εξαγωγή κανόνων σε καρδιαγγειακές βάσεις δεδομένων είχαμε στη διάθεση μας μια βάση δεδομένων με πραγματικά δεδομένα. Η βάση δεδομένων περιείχε συμπτωματικούς ασθενείς με έμφραγμα μυοκαρδίου, αγγειοπλαστική και στεφανιαία παράκαμψη. Επίσης οι παράγοντες κινδύνου των ασθενών διακρίνονταν σε παράγοντες πριν από το επεισόδιο και παράγοντες μετά το επεισόδιο. Έτσι δημιουργήσαμε τα πιο κάτω 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)

Επιλέγηκαν οι μέθοδοι της συσχέτισης και της ταξινόμησης βασισμένη στα δέντρα απόφασης, για να εξαχθούν κανόνες. Χρησιμοποιήθηκαν οι αλγόριθμοι Apriori και AKAMAS για να εξαχθούν κανόνες συσχέτισης και ο αλγόριθμος C4.5 για να εξαχθούν κανόνες με τα δέντρα απόφασης. Για το λόγο ότι οι εξαγόμενοι κανόνες ήταν πολλοί και τα μέτρα της υποστήριξης και της εμπιστοσύνης δεν ήταν επαρκή για να μπορέσουμε να απομονώσουμε τους σημαντικότερους κανόνες, έχουμε υλοποιήσει και άλλα μέτρα, έτσι ώστε να καταστεί εύκολη η επιλογή των πιο σημαντικών κανόνων. Για κάθε μοντέλο χρησιμοποιήθηκαν διαφορετικά μέτρα για το φιλτράρισμα των κανόνων. Υλοποιήσαμε επίσης το μέτρο χ^2, που ήταν αναγκαίο για τον υπολογισμό του μέτρου p-value. Το p-value δείχνει κατά πόσο ένας κανόνας είναι στατιστικά σημαντικός. Επίσης υλοποιήσαμε το μέτρο Event_Risk, όπου υπολογίζει σε κάθε ασθενή και σε κάθε κανόνα τον κίνδυνο πάθησης ενός επεισοδίου.

Όσον αφορά τα δέντρα απόφασης, για τη δημιουργία του δέντρου χρειάζεται ένα κριτήριο διαχωρισμού. Υλοποιήσαμε πέντε κριτήρια διαχωρισμού και τα δοκιμάσαμε κάνοντας μετρήσεις του ποσοστού της σωστής ταξινόμησης. Όλα τα κριτήρια διαχωρισμού κυμαίνονταν στα ίδια ποσοστά σωστής ταξινόμησης. Επίσης επιλέξαμε και τον τρόπο κλαδέματος ενός δέντρου απόφασης για την αποφυγή της υπερεκπαίδευσης. Επιλέχθηκε ο αλγόριθμος κλαδέματος από κάτω προς τα πάνω χρησιμοποιώντας τον υπολογισμό του λάθους με τη μέθοδο του Laplace.

Στη συνέχεια για κάθε μοντέλο χρησιμοποιήσαμε τους αλγόριθμους AKAMAS και C4.5 και εξάγαμε κανόνες. Με βάση τη συχνότητα των παραγόντων που εμφανίζονται στους κανόνες,

απομονώσαμε τους κυριότερους από αυτούς. Στους εξαγόμενους κανόνες κωδικοποιήσαμε τα μέτρα και με τον αλγόριθμο C4.5 δημιουργήσαμε δέντρα απόφασης με τα κωδικοποιημένα μέτρα. Έτσι διακρίναμε για κάθε μοντέλο τα σημαντικότερα μέτρα, τα οποία χρησιμοποιήσαμε για να φιλτράρουμε τους εξαγόμενους κανόνες και να πάρουμε τους σημαντικότερους. Από τις χιλιάδες εξαγόμενους κανόνες καταλήξαμε να έχουμε για κάθε μοντέλο λιγότερους από είκοσι κανόνες.

Τέλος δημιουργήσαμε ένα εργαλείο όπου ο καρδιολόγος εισάγει παράγοντες κινδύνου σε ένα νέο ασθενή και του παρουσιάζονται οι κανόνες που αντιστοιχούν σε αυτούς τους παράγοντες. Με αυτούς τους κανόνες θα μπορεί ο καρδιολόγος να διαπιστώσει πιο ποσοστό κινδύνου έχει ο νέος ασθενής και επίσης να διαφοροποιήσει τους μεταβαλλόμενους παράγοντες για να δει αν χαμηλώνει το ποσοστό κινδύνου. Με αυτό τον τρόπο θα μπορεί να πάρει απόφαση για τη φαρμακευτική αγωγή που θα δώσει στον νέο ασθενή. Αν για παράδειγμα ο νέος ασθενής έχει ποσοστό κινδύνου 35% να πάθει επεισόδιο έχοντας στους μεταβαλλόμενους παράγοντες και τον παράγοντα τριγλυκερίδια, απομονώνει από τους κανόνες αυτό που έχει τους ίδιους παράγοντες και με τριγλυκερίδια κανονικά. Αν το ποσοστό είναι μικρότερο από αυτό που έχει ο νέος ασθενής, τότε ο καρδιολόγος θα δώσει στον ασθενή φαρμακευτική αγωγή, έτσι ώστε να φέρει τα ψηλά ποσοστά τριγλυκεριδίων του νέου ασθενή σε κανονική κατάσταση. Στα πρότυπα που θα δοθούν στον καρδιολόγο, θα δίνεται έμφαση στους φιλτραρισμένους κανόνες, τους στατιστικά σημαντικούς και στο ποσοστό κινδύνου που έχει ο κάθε κανόνας.

Τα παραπάνω αποτελέσματα και οι παράγοντες κινδύνου έχουν εξαχθεί και από άλλους ερευνητές [102]-[124]. Θα πρέπει να σημειωθεί ότι τα αποτελέσματα της μελέτης μας βασίζονται σε ασθενείς από την Πάφο και είναι συγκρίσιμα με άλλες μελέτες. Μια εξίσωση υπολογισμού του κινδύνου για ένα ειδικό πληθυσμό είναι αναγκαία όπως άλλωστε αναφέρεται και από άλλους ερευνητές [106].

Η επιλογή των δύο μεθόδων, της συσχέτισης και της κατηγοριοποίησης, δεν ήταν τυχαία για το λόγο του ότι οι δύο αυτές μέθοδοι εξάγουν κανόνες. Επιπλέον οι αλγόριθμοι που χρησιμοποιούνται κατατάσσονται στους καλύτερους αλγόριθμους [5].

Χρησιμοποιώντας τις βάσεις δεδομένων με τη μεθοδολογία που έχει αναπτυχθεί σε αυτή την εργασία, έχει διαπιστωθεί ότι η σωστή ταξινόμηση (correct classification) κυμαίνεται περίπου στο 65% για τους κανόνες με δέντρα απόφασης και περίπου στο 58% για τους κανόνες συσχέτισης. Αυτό σημαίνει ότι οι εξαγόμενοι κανόνες δέντρων απόφασης είναι πιο δυνατοί από εκείνους της συσχέτισης. Στην περίπτωση της μεθόδου της ταξινόμησης η επιλογή του κριτηρίου διαχωρισμού δεν είναι καθοριστική.

Στους εξαγόμενους κανόνες κάθε μοντέλου μελετήθηκαν οι παράγοντες που παρουσιάζονται μέσα στους κανόνες. Καταμετρήθηκαν οι παράγοντες που παρουσιάζονται στο μοντέλο. Οι παράγοντες εκείνοι που παρουσιάζονται τις περισσότερες φορές στο μοντέλο, θεωρούνται οι πιο σημαντικοί.

Στην εκτέλεση κάθε μοντέλου για την εξαγωγή κανόνων έχει εξαχθεί μαζί με τους κανόνες η πληροφορία κατά πόσο ένας κανόνας είναι στατιστικά σημαντικός. Τους κανόνες αυτούς τους χρησιμοποιήσαμε για να εκτελέσουμε ξανά το μοντέλο, όπου μας έδωσε καλύτερα αποτελέσματα σωστής ταξινόμησης. Αυτό είχε γίνει στους κανόνες συσχέτισης, όπου για όλους τους κανόνες είχαμε σωστή ταξινόμηση 58%, ενώ για τους στατιστικά σημαντικούς 69%. Στα δέντρα απόφασης δεν μπορούσε να γίνει η αξιολόγηση για το λόγο ότι οι στατιστικά σημαντικοί κανόνες ήταν πολύ λίγοι. Επιπλέον με τη βοήθεια των μέτρων έχουν φιλτραριστεί οι κανόνες για να καταλήξουμε στους πιο σημαντικούς. Οι φιλτραρισμένοι αυτοί κανόνες χρήζουν περαιτέρω διερεύνησης όσον αφορά την αξιολόγηση των κανόνων.

8.1 Κλινικές Μελέτες

Η ομάδα Euroaspire πραγματοποίησε μελέτες σχετικά με τους παράγοντες κινδύνου. Οι μελέτες της Euroaspire ενέπλεξαν διάφορους ευρωπαϊκούς πληθυσμούς και επιπλέον συμπεριέλαβαν και άλλους παράγοντες όπως η παχυσαρκία. Όλες οι μελέτες εξετάστηκαν από κοινού και έδωσαν συνδυασμένα αποτελέσματα όπως περιγράφονται στις μελέτες [102]-[104]. Τα αποτελέσματα των μελετών της ομάδας της EUROASPIRE είναι: 4863 ιατρικοί

φάκελοι εξετάστηκαν από τους οποίους 25% ήταν γυναίκες. Σε 3569 ασθενείς έγιναν συνεντεύξεις (προσαρμοσμένο ποσοστό ανταπόκρισης 85%) με μέση ηλικία τα 61 έτη. Δεκαεννέα τοις εκατό των ασθενών κάπνιζαν, 25% ήταν υπέρβαροι (Δείκτης Μάζας Σώματος (ΔΜΣ) > ή = 30 $kg.m^{-2}$), 53% είχε αυξημένη αρτηριακή πίεση (συστολική BP > ή = 140 και / ή διαστολική BP > ή = 90 mmHg), 44 % είχε προβάλει ολική χοληστερόλη του πλάσματος (ολική χοληστερόλη > = 5.5 mmol/l) και 18% ήταν διαβητικοί. Τα φάρμακο που αναφέρθηκαν στις συνεντεύξεις ήταν: αντιαιμοπεταλιακό 81%, β-αναστολείς, 54% (58% στις μετά-μυοκαρδίου ασθενείς). Από τους ασθενείς που λαμβάνουν φάρμακα για μείωση της αρτηριακής πίεσης (δεν είναι πάντα που προβλέπεται για τη θεραπεία της υπέρτασης) 50% είχαν συστολική πίεση BP > 140 mmHg και 21% > 160 mmHg, και όσων λαμβάνουν φάρμακα μείωσης των λιπιδίων, 49% είχε ολική χοληστερόλη > 5.5 mmol/l και 13%> 6.5 mmol/l. Τριάντα επτά τοις εκατό των ασθενών είχαν οικογενειακό ιστορικό πρόωρης στεφανιαίας νόσου σε πρώτου βαθμού συγγενή, αλλά μόνο το 21% των ασθενών ανέφερε ότι συμβούλευσε τους συγγενείς τους να ελέγχονται για στεφανιαία νόσο.

Μετά από αυτά η ομάδα κατέληξε στο συμπέρασμα ότι αυτή η έρευνα κατέδειξε υψηλό επιπολασμό στους μεταβαλλόμενους παράγοντες κινδύνου σε ασθενείς με στεφανιαία καρδιοπάθεια. Υπάρχουν σημαντικές δυνατότητες για τους καρδιολόγους και γιατρούς για την περαιτέρω μείωση της νοσηρότητας στεφανιαίας καρδιακής νόσου και θνησιμότητας και την αύξηση των πιθανοτήτων επιβίωσης των ασθενών.

Ένα γενικό αποτέλεσμα ήταν το γεγονός ότι οι ασθενείς δεν ακολουθούν τις συμβουλές και τις συστάσεις των γιατρών τους. Σε σύγκριση με τη μελέτη Euroaspire, τα συμπεράσματά μας σχετικά με τους μεταβαλλόμενους παράγοντες κινδύνου μετά την εκδήλωση ενός επεισοδίου είναι τα ακόλουθα [104]:

- 14% των ασθενών καπνίζουν μετά το επεισόδιο (16% στην Euroaspire)
- 22% των ασθενών έχουν ψηλή αρτηριακή πίεση (26% στην Euroaspire)
- 34% των ασθενών έχουν ψηλή χοληστερόλη (31% στην Euroaspire)
- 45% των ασθενών έχουν ψηλή low density lipoprotein (31% στην Euroaspire).

Στη μελέτη Euroaspire, το κάπνισμα, η αρτηριακή πίεση και η χοληστερόλη διαπιστώθηκε ότι είναι σημαντικοί παράγοντες κινδύνου [102] και [104]. Το συμπέρασμα ήταν ότι υπάρχουν μεγάλες αποκλίσεις μεταξύ των 15 χωρών στην επικράτηση των παραγόντων κινδύνου και τη χρήση των καρδιοπροστατευτικών φαρμακευτικών αγωγών [103]. Επίσης, εξακολουθεί να υπάρχει σημαντικό δυναμικό σε όλη την Ευρώπη για να βελτιώσει τα πρότυπα της προληπτικής φροντίδας, προκειμένου να μειωθεί ο κίνδυνος συνέχισης των ασθενειών και θανάτου σε ασθενείς με στεφανιαία νόσο.

Επιπλέον, πρόσθετες παρατηρήσεις που θα μπορούσαν να εξαχθούν από τη βάση δεδομένων που διερευνήθηκε σε αυτή τη μελέτη σχετικά με τους μη μεταβαλλόμενους παράγοντες κινδύνου σε σχέση με τη μελέτη Euroaspire [102] είναι και οι ακόλουθες:

- 14% των ασθενών ήταν γυναίκες (24.7% στην Euroaspire)
- 9% των ασθενών ήταν <= 50 χρονών (23.1% στην Euroaspire)
- 28% ήταν μεταξύ 51 και 60 χρονών (33.8% στην Euroaspire)
- 39% των ασθενών ήταν μεταξύ 61 και 70 χρονών (43.1% στην Euroaspire)
- 24% των ασθενών ήταν μεταξύ 71 και 84 χρονών.

Δεν βρέθηκαν γυναίκες ασθενείς κάτω των 50 χρονών, αλλά μόνο άντρες βρέθηκαν σε αυτή την ηλικία.

Οι Rea *et al.* [104] κατέληξαν στο συμπέρασμα ότι το κάπνισμα συσχετίστηκε με αυξημένο κίνδυνο για επαναλαμβανόμενα στεφανιαία επεισόδια. Συγκεκριμένα η μελέτη τους έδειξε ότι κατά τον χρόνο των περιστατικών του μυοκαρδίου, 33.6% των ασθενών δεν ήταν καπνιστές, 35.5% ήταν πρώην καπνιστές και 30.9% ήταν ενεργοί καπνιστές. Από τα 808 άτομα που ήταν ενεργοί καπνιστές κατά τη στιγμή του συμβάντος του μυοκαρδίου, 449 σταμάτησαν το κάπνισμα κατά τη διάρκεια της νοσηλείας ή μετά τη νοσηλεία. Με τους μη καπνιστές ως ομάδα αναφοράς, το μέτρο σχετικός κίνδυνος (RR), για επαναλαμβανόμενα στεφανιαία επεισόδια (n = 433) ήταν 1.17 (95% CI, 0.93 με 1.43) για τους πρώην καπνιστές και 1.51 (CI,

1.10 με 2.07) για τους ενεργούς καπνιστές. Με τους μη καπνιστές ως ομάδα αναφοράς, η RR για τα άτομα που σταμάτησαν το κάπνισμα μετά από έμφραγμα ήταν 1.62 (CI, 1.02 με 2.61), εφόσον η διάρκεια της διακοπής ήταν μεταξύ 0 και 6 μηνών, 1.60 (CI, 0.97 σε 2.60), αν η διάρκεια ήταν μεταξύ 6 και 18 μηνών, 1.48 (CI, 0.76 με 2.51), αν η διάρκεια ήταν μεταξύ 18 και 36 μηνών και 1.02 (CI, 0.54 με 1.86), αν η διάρκεια ήταν περισσότερο από 36 μήνες (p = 0.01).

Συμπέρασμα: μετά το περιστατικό εμφράγματος του μυοκαρδίου, το κάπνισμα συσχετίστηκε με αυξημένο κίνδυνο για επαναλαμβανόμενα επεισόδια. Σε άτομα που σταμάτησαν το κάπνισμα μετά από έμφραγμα, ο κίνδυνος μειώθηκε και είναι ίσος με εκείνο των μη καπνιστών που σταμάτησαν πριν από 3 χρόνια.

Τα αποτελέσματα της έρευνας αυτής ταυτίζονται με τα δικά μας, όπου λαμβάνοντας υπόψη όλα τα μοντέλα για τα διάφορα επεισόδια, το κάπνισμα φαίνεται να είναι ένας από τους κύριους παράγοντες πάθησης ενός επεισοδίου. Για το λόγο ότι έχουμε τον παράγοντα κάπνισμα πριν από ένα επεισόδιο και μετά από ένα επεισόδιο, έχουμε μελετήσει το κάπνισμα σαν δύο διαφορετικούς παράγοντες. Τα αποτελέσματά μας έδειξαν ότι το κάπνισμα τόσο πριν από ένα επεισόδιο, όσο και μετά από ένα επεισόδιο είναι ένας από τους κύριους παράγοντες πάθησης ενός επεισοδίου, ενόψει και του γεγονότος ότι 14% των ασθενών συνεχίζουν να καπνίζουν και μετά το επεισόδιο.

Οι Wang *et al.* [106] για να προβλέψουν τη στεφανιαία νόσο χρησιμοποίησαν τους εξής παράγοντες κινδύνου: ηλικία, φύλο, χοληστερόλη, HDL, αρτηριακή πίεση, διαβήτη και κάπνισμα. Η προβλεπόμενη συχνότητα εμφάνισης στεφανιαίας νόσου χρησιμοποιώντας την εξίσωση Framingham ήταν 4.4 ανά 1000 άτομα το χρόνο, ενώ η συχνότητα εμφάνισης ήταν 11.0 (95% CI, 8.7 με 13.9) ανά 1000 άτομα το χρόνο. Ο παρατηρούμενος αριθμός των εκδηλώσεων στεφανιαίας νόσου (68) ήταν 2.5 φορές του προβλεπόμενου αριθμού (27) χρησιμοποιώντας τη εξίσωση Framingham. Η συχνότητα εμφάνισης ήταν περίπου τέσσερις και τρεις φορές από την προβλεπόμενη επίπτωση για τις ομάδες ηλικίας < 35 και 35 - 44

ετών, αντίστοιχα, και περίπου διπλάσιο από το προβλεπόμενο ποσοστό για όσους είναι άνω των 45 ετών.

Συμπεράσματα: η εκτίμηση κινδύνου βάσει της εξίσωσης Framingham έδωσε σημαντικά χαμηλότερο αριθμό περιστατικών από τον πραγματικό αριθμό περιστατικών κινδύνου της στεφανιαίας νόσου που παρατηρήθηκε σε ιθαγενείς σε μια απομακρυσμένη κοινότητα, ιδίως για τις γυναίκες και τους νεότερους ενήλικες. Αυτό σημαίνει ότι οι παραδοσιακοί παράγοντες κινδύνου έχουν διαφορετικό αντίκτυπο ή / και ότι άλλοι παράγοντες συμβάλλουν στον κίνδυνο.

Στην εργασία αυτή γίνεται επίσης χρήση της εξίσωσης Framingham όπως περιγράφεται στην αναφορά [125], όπου ο κάθε παράγοντας κινδύνου διαδραματίζει διαφορετικό ρόλο στην εκτίμηση του κινδύνου πάθησης ενός επεισοδίου.

Ο Bambrick [107] μελέτησε πέντε παράγοντες κινδύνου διαβήτη (ηλικία, ΔΜΣ, περιφέρεια μέσης (WC), υπέρταση και οικογενειακό ιστορικό) και γλυκόζη νηστείας (FG).

Το συμπέρασμα του ήταν ότι στις γυναίκες, η κεντρική παχυσαρκία έχει καλύτερη δυνατότητα πρόβλεψης του διαβήτη και του κινδύνου καρδιαγγειακής νόσου από το Δείκτη Μάζας Σώματος ≥ 30, ο οποίος δεν αποτελεί αξιόπιστο δείκτη. Ο Δείκτης Μάζας Σώματος ≥ 25 ήταν μια καλή ένδειξη στους άνδρες.

Κατά συνέπεια ο Δείκτης Μάζας Σώματος είναι ένα χρήσιμο κλινικό εργαλείο για τον εντοπισμό ατόμων που διατρέχουν κίνδυνο για διαβήτη. Για τις γυναίκες, ο Δείκτης Μάζας Σώματος ≥ 25 θα μπορούσε περισσότερο να αντικατοπτρίζει τον κίνδυνο, ενώ η τρέχουσα τιμή της περιφέρειας της μέσης των 88 εκατοστών εξακολουθεί να είναι κατάλληλη. Για τους άνδρες, η μείωση τόσο στο Δείκτη Μάζας Σώματος ≥ 25 όσο και στην περιφέρεια της μέσης (90 εκατοστών), μπορούν να αντανακλούν καλύτερα το διαβήτη και τον κίνδυνο καρδιαγγειακής νόσου. Εκτός από τον Bambrick [107] αυτόν τον παράγοντα τον χρησιμοποίησε και η ομάδα της Euroaspire [109].

Στη δική μας μελέτη, παρόλο που ο Δείκτης Μάζας Σώματος υπήρχε στο πρωτόκολλο και αρχικά θέλαμε να τον συμπεριλάβουμε στην ανάλυση των παραγόντων κινδύνου, αυτό δεν

μπόρεσε να γίνει για τον λόγο ότι στους περισσότερους ασθενείς έλειπε η τιμή του παράγοντα αυτού.

8.2 Δέντρα απόφασης

Τα ακόλουθα πέντε διαφορετικά κριτήρια αποτέλεσαν αντικείμενο της έρευνας, information gain, gini index, likelihood ratio chi-squared statistics, gain ratio και distance measure. Τα κριτήρια αυτά είχαν στα μοντέλα παρόμοια απόδοση και δεν είχαν σημαντική διαφορά μεταξύ τους. Έτσι, κάθε ένα από τα κριτήρια διαχωρισμού που διερευνήθηκαν θα μπορούσε να χρησιμοποιηθεί για τα σύνολα δεδομένων σε αυτή τη μελέτη. Αυτό το αποτέλεσμα είναι σύμφωνο με τη μελέτη για την ανάπτυξη των μοντέλων με δέντρα απόφασης, που τεκμηριώνει ότι η επιλογή των κριτηρίων διαχωρισμού δεν κάνει μεγάλη διαφορά στην απόδοση του δέντρου απόφασης [136] - [138]. Επίσης, τα διάφορα κριτήρια διαχωρισμού, συμφωνούν στο ποιοί είναι οι σημαντικότεροι παράγοντες κινδύνου. Από ότι γνωρίζουμε δεν υπάρχει παρόμοια μελέτη στη βιβλιογραφία που να συγκρίνει τα πέντε κριτήρια διαχωρισμού που μελετήθηκαν σε αυτή τη μελέτη για το πρόβλημα της στεφανιαίας νόσου.

Σχετικά με το σύστημα ελέγχου του Εδιμβούργου, στα αποτελέσματα τους οι Tsien *et al.* [110] έδειξαν ότι τα Δέντρα FT, FT LR, και Kennedy LR εκτελούνται εξίσου καλά, με την καμπύλη ROC να είναι στις περιοχές του 94.04%, 94.28% και 94.30% για τη σωστή ταξινόμηση. Σε αντίθεση με προηγούμενες εργασίες [135], η παρούσα μελέτη δείχνει ότι τα δέντρα απόφασης έχουν ορισμένα πλεονεκτήματα και μπορούν να χρησιμοποιηθούν στη διάγνωση των ασθενών με ΜΙ.

Στη μελέτη μας χρησιμοποιήσαμε και εμείς δέντρα απόφασης γιατί είχαμε κι εμείς ασθενείς με ΜΙ και γιατί θέλαμε σαν αποτελέσματα να εξάγουμε κανόνες έτσι ώστε να τους συγκρίνουμε και αντιπαραθέσουμε με τους κανόνες συσχέτισης. Η διαφορά του ποσοστού της απόδοσης της σωστής ταξινόμησης στην πιο πάνω μελέτη με τη δική μας μελέτη είναι ότι

οι μετρήσεις οι δικές μας έγιναν κατά τη διάρκεια της χρήσης των δεδομένων ελέγχου, ενώ στην πιο πάνω μελέτη έγιναν κατά τη διάρκεια της δημιουργίας του δέντρου.

Στα αποτελέσματα τους οι Voss *et al.* [116] δείχνουν ότι περίπου ένας στους τρεις με στεφανιαίο επεισόδιο μπορεί να προληφθεί στα 5 χρόνια με θεραπεία. Η ανάλυση τους δείχνει ότι η χρήση του νευρωνικού δικτύου τύπου Multi Layer Perceptron MLP για να εντοπίζονται τα άτομα υψηλού κινδύνου, όπως και οι υποψήφιοι για θεραπεία, θα επιτρέψει την πρόληψη του 25% των στεφανιαίων επεισοδίων σε μεσήλικες άνδρες, σε σύγκριση με το 15% και 11% με LR και PNN, αντίστοιχα.

Οι ασθενείς στη δική μας μελέτη με βάση την εξίσωση του Framingham ήταν στην πλειοψηφία τους ασθενείς ψηλού κινδύνου (52%).

Οι Bayat *et al.* [118] στα αποτελέσματα τους έδειξαν ότι η ηλικία βρέθηκε να είναι ο σημαντικότερος παράγοντας και στα δύο μοντέλα, δηλαδή των Bayesian δικτύων και των δέντρων απόφασης. Και τα δύο μοντέλα ήταν εξαιρετικά ευαίσθητα και συγκεκριμένα: ευαισθησία 90.0%, ειδικότητα 96.7%. Η παρουσίαση των αποτελεσμάτων των δύο μεθόδων έδειξε ότι τα δέντρα απόφασης ήταν πιο κατανοητά από τους γιατρούς. Οι προσεγγίσεις αυτές παρέχουν γνώσεις σχετικά με την τρέχουσα διαδικασία φροντίδας. Αυτή η γνώση θα μπορούσε να χρησιμοποιηθεί για τη βελτιστοποίηση της διαδικασίας της υγειονομικής περίθαλψης.

Στα μοντέλα που μελετήσαμε βλέπουμε ότι συνολικά η ηλικία είναι ένας σημαντικός παράγοντας στην πάθηση ενός επεισοδίου. Επίσης στη μελέτη μας είχαμε κι εμείς κοντινά αποτελέσματα όσον αφορά τα μέτρα ευαισθησία και ειδικότητα. Συγκεκριμένα στο μέτρο ευαισθησία είχαμε το ποσοστό 91% και στο μέτρο ειδικότητα 94%.

8.2.1 Μοντέλα MI

Από τους εξαγόμενους κανόνες για τα μοντέλα του εμφράγματος μυοκαρδίου στο μοντέλο πριν από ένα επεισόδιο έχουμε κανόνες με τα ψηλότερα ποσοστά κινδύνου πάθησης ενός επεισοδίου. Ακολουθεί το μοντέλο πριν και μετά το επεισόδιο και τα πιο χαμηλά ποσοστά τα έχει το μοντέλο μετά το επεισόδιο. Σε κανένα από τα τρία μοντέλα δεν είχαμε ασθενείς με χαμηλό κίνδυνο. Αυτό οφείλεται κατά ένα βαθμό στο γεγονός ότι η βάση δεδομένων που μελετήσαμε έχει μόνο συμπτωματικούς ασθενείς με καρδιαγγειακά νοσήματα.

Οι πιο δυνατοί κανόνες όσον αφορά το ποσοστό κινδύνου είναι:

πριν το επεισόδιο (B)

SEX	AGE	SMBEF	HxDM	CLASS	Event_Risk
M	2	Y	Y	Y	33.36%

μετά το επεισόδιο (A)

SMAFT	SBP	GLU	CLASS	Event_Risk
Y	H	Y	Y	26.15%

πριν και μετά το επεισόδιο (B+A)

SEX	AGE	FH	SMBEF	HxHTN	SBP	GLU	LDL	CLASS	Event_Risk
M	2	Y	Y	Y	H	Y	H	Y	32.26%

8.2.2 Μοντέλα PCI

Στους κανόνες στην αγγειοπλαστική το μοντέλο με τα ψηλότερα ποσοστά κινδύνου πάθησης επεισοδίου είναι το πριν και μετά το επεισόδιο και ακολουθεί το μοντέλο πριν το επεισόδιο. Και εδώ πάλι το μοντέλο μετά το επεισόδιο έχει τα πιο χαμηλά ποσοστά κινδύνου. Όπως και στα εμφράγματα μυοκαρδίου, έτσι και εδώ δεν έχουμε ασθενείς με χαμηλό ποσοστό κινδύνου.

Οι πιο δυνατοί κανόνες όσον αφορά το ποσοστό κινδύνου είναι:

πριν το επεισόδιο (B)

SEX	AGE	FH	CLASS	Event_Risk
F	2	Y	Y	26.04%

μετά το επεισόδιο (A)

SMAFT	GLU	TG	CLASS	Event_Risk
Y	Y	H	Y	27.17%

πριν και μετά το επεισόδιο (B+A)

AGE	FH	SMBEF	HxHTN	GLU	LDL	CLASS	Event_Risk
2	Y	Y	Y	Y	H	Y	32.27%

8.2.3 Μοντέλα CABG

Τα μοντέλα της στεφανιαίας παράκαμψης είναι παρόμοια με αυτά της αγγειοπλαστικής. Και πάλι εδώ δεν έχουμε ασθενείς με χαμηλό ποσοστό κινδύνου. Στο μοντέλο πριν το επεισόδιο οι ασθενείς με ψηλό ποσοστό κινδύνου είναι περίπου τριπλάσιο από τους ασθενείς με μεσαίο ποσοστό κινδύνου. Στα άλλα δύο μοντέλα είναι περίπου οι μισοί ασθενείς με ψηλό ποσοστό κινδύνου και οι άλλοι μισοί με μέτριο ποσοστό κινδύνου.

Οι πιο δυνατοί κανόνες όσον αφορά το ποσοστό κινδύνου είναι:

πριν το επεισόδιο (B)

AGE	FH	SMBEF	CLASS	Event_Risk
4	Y	Y	Y	30.14%

μετά το επεισόδιο (A)

SBP	LDL	CLASS	Event_Risk
H	H	Y	27.90%

πριν και μετά το επεισόδιο (B+A)

AGE	FH	SMBEF	LDL	TG	CLASS	Event_Risk
4	Y	Y	H	H	Y	34.59%

8.3 Κανόνες συσχέτισης

Ο Ordonez [119] χρησιμοποίησε τον αλγόριθμο δέντρων απόφασης C4.5 και τους κανόνες συσχέτισης, για την πρόγνωση της καρδιακής νόσου βασισμένος σε 25 παράγοντες κινδύνου και τεκμηρίωσε ότι οι κανόνες συσχέτισης περιλαμβάνουν γενικά απλούστερους έξυπνους κανόνες από τους κανόνες δέντρων απόφασης. Η χρησιμότητα των κανόνων συσχέτισης στην ανάλυση των παραγόντων κινδύνου στεφανιαίας νόσου ερευνήθηκε επίσης από την ομάδα μας σε μια παρόμοια βάση δεδομένων με αυτή τη μελέτη [135]. Τα αποτελέσματα όσον αφορά τους σημαντικότερους παράγοντες κινδύνου ήταν παρόμοια. Τα αποτελέσματα της δικής μας μελέτης δείχνουν ότι και οι δύο μέθοδοι είναι σημαντικές. Απλοί και έξυπνοι κανόνες στη μέθοδο της συσχέτισης, πολλοί παράγοντες στους κανόνες με δέντρα απόφασης.

Οι Concaro *et al.* στην ανάλυση που παρουσιάζουν στην εργασία τους [124] υπογραμμίζουν τις κύριες δυνατότητες της εφαρμογής των κανόνων συσχέτισης για την εξόρυξη γνώσης από βάσεις δεδομένων στην υγεία. Ο αλγόριθμος που εφαρμόζεται επιτρέπει να εκμεταλλευτούν σωστά την ενσωμάτωση των διαφόρων πηγών πληροφοριών της ιατροφαρμακευτικής περίθαλψης, όπως τα διοικητικά και κλινικά δεδομένα. Η μέθοδος έχει αποδειχθεί ότι μπορεί να χαρακτηρίσει τις υποομάδες των θεμάτων, τονίζοντας ενδιαφέρουσες συχνές συσχετίσεις μεταξύ διαγνωστικών ή θεραπευτικών προτύπων και των προτύπων που σχετίζονται με την κλινική κατάσταση του ασθενούς. Λαμβάνοντας υπόψη την προοπτική ενός περιφερειακού οργανισμού ιατροφαρμακευτικής περίθαλψης, η μέθοδος αυτή θα μπορούσε να αξιοποιηθεί κατάλληλα για την αξιολόγηση της καταλληλότητας της ροής παροχής φροντίδας για συγκεκριμένες παθήσεις, προκειμένου να αξιολογήσει εκ νέου ή να βελτιώσει τις ακατάλληλες πρακτικές που οδηγούν σε μη ικανοποιητικά αποτελέσματα.

Οι Kuo *et al.* [123] έχουν διερευνήσει τη δυνατότητα να χρησιμοποιηθεί ο τομέας της ιατρικής οντολογίας ως πηγή εξόρυξης γνώσης και έκφρασης της γνώσης σε μια χρήσιμη

μορφή. Χρησιμοποίησαν μια βάση δεδομένων με ασθενείς που έχουν νεφρικά προβλήματα. Έχουν περιγράψει μια προσέγγιση στην οποία χρησιμοποιείται η οντολογία για να κατηγοριοποιήσει τα δεδομένα έτσι ώστε να τα προετοιμάσει για εξόρυξη κανόνων συσχέτισης στα δεδομένα. Οι κανόνες που εξάχθηκαν, εξετάστηκαν στη συνέχεια από έναν ειδικό, προκειμένου να αξιολογηθεί η χρησιμότητά τους. Κατάληξαν στο συμπέρασμα ότι η οντολογία με γνώμονα την εξόρυξη δεδομένων μπορεί να επιτύχει περισσότερα αποτελέσματα από ότι η απλή εξόρυξη δεδομένων.

8.3.1 Μοντέλα MI

Από τους εξαγόμενους κανόνες για τα μοντέλα του εμφράγματος μυοκαρδίου στο μοντέλο πριν και μετά από ένα επεισόδιο έχουμε κανόνες με τα ψηλότερα ποσοστά κινδύνου πάθησης ενός επεισοδίου. Στο μοντέλο πριν και μετά από ένα επεισόδιο έχουμε ασθενείς με χαμηλό κίνδυνο.

Οι πιο δυνατοί κανόνες όσον αφορά το ποσοστό κινδύνου είναι:

πριν το επεισόδιο (B)

SEX	AGE	SMBEF	HxDM	Class	Event_Risk
M	2	Y	Y	Y	30.75%

μετά το επεισόδιο (A)

SMAFT	SBP	GLU	TG	Class	Event_Risk
Y	H	Y	H	Y	25.30%

πριν και μετά το επεισόδιο (B+A)

SEX	AGE	HxHTN	GLU	Class	Event_Risk
M	2	Y	Y	Y	21.26%

8.3.2 Μοντέλα PCI

Από τους εξαγόμενους κανόνες για τα μοντέλα αγγειοπλαστικής στο μοντέλο πριν και μετά από ένα επεισόδιο έχουμε κανόνες με τα ψηλότερα ποσοστά κινδύνου πάθησης ενός επεισοδίου. Στο μοντέλο πριν και μετά από ένα επεισόδιο έχουμε ασθενείς με χαμηλό κίνδυνο.

Οι πιο δυνατοί κανόνες όσον αφορά το ποσοστό κινδύνου είναι:

πριν το επεισόδιο (B)

SEX	AGE	SMBEF	HxDM	Class	Event_Risk
M	2	Y	Y	Y	24.77%

μετά το επεισόδιο (A)

SBP	LDL	class	Event_Risk
H	H	Y	26.37%

πριν και μετά το επεισόδιο (B+A)

SEX	FH	class	Event_Risk
M	Y	Y	21.46%

8.3.3 Μοντέλα CABG

Οι πιο δυνατοί κανόνες όσον αφορά το ποσοστό κινδύνου είναι:

πριν το επεισόδιο (B)

SEX	AGE	HxHTN	HxDM	Class	Event_Risk
M	3	Y	Y	Y	32.73%

μετά το επεισόδιο (A)

SBP	LDL	class	Event_Risk
H	H	Y	23.14%

πριν και μετά το επεισόδιο (B+A)

FH	HxHTN	class	Event_Risk
Y	Y	Y	22.40%

8.4 Προτεινόμενο σύστημα

Έχουμε δημιουργήσει και εκτελέσει πολλά μοντέλα εξάγοντας χιλιάδες κανόνες. Αναφέραμε αναλυτικά την απόδοση και την ακρίβεια κάθε μοντέλου. Έχουμε παρουσιάσει επιλεγμένους κανόνες από κάθε μοντέλο. Χρησιμοποιήσαμε τις μεθόδους της συσχέτισης και των δέντρων απόφασης για την εξαγωγή των κανόνων. Στα δέντρα απόφασης χρησιμοποιήσαμε πέντε κριτήρια διαχωρισμού. Έχουμε υλοποιήσει διάφορα μέτρα για να μας βοηθήσουν να επιλέξουμε τους καλύτερους, πιο δυνατούς κανόνες. Έχουμε επίσης υπολογίσει το ποσοστό κινδύνου πάθησης ενός επεισοδίου.

Ξεκινώντας από τις μεθόδους, έχουμε δύο ζητήματα: 1. Στους κανόνες των δέντρων απόφασης γίνεται μια πολύ πιο σωστή ταξινόμηση όπου έχουμε μέσο όρο σωστής ταξινόμησης 65%, σε αντίθεση με τους κανόνες συσχέτισης με τον μέσο όρο σωστής ταξινόμησης να είναι 58%. 2. Στους κανόνες συσχέτισης έχουμε λίγους παράγοντες σε κάθε κανόνα, σε αντίθεση με τους κανόνες των δέντρων απόφασης που χρησιμοποιούνται σχεδόν όλοι οι παράγοντες σε κάθε κανόνα. Άρα λοιπόν αν ο καρδιολόγος θα επικεντρωθεί σε ελάχιστους παράγοντες σε έναν ασθενή, θα χρησιμοποιήσει μοντέλο με τη μέθοδο της συσχέτισης, διαφορετικά αν έχει όλους τους παράγοντες ενός ασθενή στη διάθεση του, θα χρησιμοποιήσει μοντέλο με δέντρα απόφασης.

Όσον αφορά τα κριτήρια διαχωρισμού στη μέθοδο της εξαγωγής κανόνων με δέντρα απόφασης, φάνηκε ότι όλα τα κριτήρια που υλοποιήθηκαν έχουν σχεδόν την ίδια απόδοση.

Στο σύνολο όλων των μοντέλων το Gini Index έδειξε να έχει ελάχιστα καλύτερη απόδοση από τα υπόλοιπα και γι' αυτό θα χρησιμοποιηθεί αυτό το κριτήριο.

Για τα μοντέλα πριν το επεισόδιο (B), μετά το επεισόδιο (A) και πριν και μετά το επεισόδιο (B+A), φάνηκε ότι το καλύτερο μοντέλο είναι αυτό πριν το επεισόδιο (B) και σαν δεύτερο το μοντέλο πριν και μετά το επεισόδιο (B+A). Εδώ μπορεί ο καρδιολόγος να κάνει χρήση του ενός ή του άλλου μοντέλου, ανάλογα με τους παράγοντες που έχει από τον νέο ασθενή.

Στον Πίνακα 8.1 φαίνεται το προτεινόμενο σύστημα.

Πίνακας 8.1: Προτεινόμενο σύστημα

Περίπτωση	**Μοντέλα**
Ασυμπτωματικός ασθενής	MI vs PCI ή CABG (B) PCI vs MI ή CABG (B) CABG vs MI ή PCI (B)
Συμπτωματικός ασθενής	MI vs PCI ή CABG (B+A) PCI vs MI ή CABG (B+A) CABG vs MI ή PCI (B+A)

8.5 Σύγκριση με άλλα εργαλεία εξόρυξης δεδομένων

Στον κλάδο εξόρυξης δεδομένων έχουν παρουσιαστεί διάφορα εργαλεία για εξόρυξη κανόνων συσχέτισης. Από τα πιο γνωστά είναι το WEKA [139], το DBMiner [140] και το Clementine [141].

Το εργαλείο που παρουσιάζεται σε αυτή τη μελέτη είναι ένα ευκολόχρηστο και απλό εργαλείο που μπορεί να εξάγει κανόνες συσχέτισης με βάση τους αλγόριθμους Apriori και AKAMAS, και κανόνες με δέντρα απόφασης με βάση τον αλγόριθμο C4.5. Ο χρήστης μπορεί εύκολα να εισάγει τα δεδομένα εισόδου, όπως επίσης και να κατανοήσει τους εξαγόμενους κανόνες συσχέτισης.

Μία σημαντική διαφορά του εργαλείου, με το εργαλείο WEKA είναι ότι έχουν υλοποιηθεί και υπολογίζονται για κάθε κανόνα αρκετά μέτρα αξιολόγησης κανόνων. Από αυτά τα μέτρα ο χρήστης μπορεί να επιλέξει ποία επιθυμεί να παρουσιάζονται, αλλά επίσης και να ορίσει ένα ελάχιστο όριο με το οποίο το εργαλείο επιλέγει κανόνες. Οι κανόνες παρουσιάζονται ταξινομημένοι με βάση τον αριθμό των χαρακτηριστικών των κανόνων, σε μορφή πίνακα, κι ο χρήστης μπορεί να τους ταξινομήσει και να τους επιλέξει με όποιο μέτρο αξιολόγησης επιθυμεί. Αυτό βοηθά τον χρήστη να αναλύσει εύκολα τους κανόνες και να εξάξει εύκολα και γρήγορα τη γνώση, με βάση τον σκοπό που επιθυμεί.

Το εργαλείο DBMiner έχει το μειονέκτημα ότι δεν μπορεί να χειριστεί πολλούς παράγοντες ταυτόχρονα. Ο χρήστης έχει τη δυνατότητα να συμπεριλάβει όσους παράγοντες θέλει, όμως το εργαλείο αυτό δεν βγάζει αποτελέσματα. Επίσης για να λειτουργήσει αυτό το εργαλείο χρειάζεται η εγκατάσταση άλλων προγραμμάτων όπως ο SQL Server. Στη δική μας περίπτωση ο χρήστης μπορεί να χρησιμοποιήσει όσους παράγοντες επιθυμεί.

Στο εργαλείο Clementine τα αποτελέσματα που βγαίνουν είναι σε μορφή δέντρου. Τους κανόνες πρέπει να τους δημιουργήσει ο χρήστης με βάση αυτό το δέντρο. Αυτό είναι πολύ χρονοβόρο.

Κεφάλαιο 9: Συμπεράσματα και μελλοντική εργασία

9.1 Συμπεράσματα

Σε αυτή τη διατριβή έχει μελετηθεί το πρόβλημα των καρδιαγγειακών παθήσεων χρησιμοποιώντας την εξόρυξη δεδομένων για την ανάπτυξη ενός ολοκληρωμένου συστήματος που θα υποστηρίζει την αξιολόγηση των παραγόντων κινδύνου σε καρδιαγγειακές βάσεις δεδομένων και την εξόρυξη κανόνων εκτίμησης κινδύνου βασισμένων σε αλγόριθμους δένδρων αποφάσεων και κανόνων συσχέτισης. Οι κανόνες αυτοί θα αποτελούν τη βάση για την αξιολόγηση ενός νέου ασθενή, συγκρίνοντας τις τιμές των παραγόντων του ασθενή με αυτές των κανόνων.

Οι καρδιακές παθήσεις είναι ένα πρόβλημα που μαστίζει τον κόσμο και που οι προβλέψεις των ειδικών δείχνουν ότι αυτή η κατάσταση θα χειροτερέψει. Η έγκαιρη πρόγνωση και διάγνωση των καρδιακών παθήσεων θα βοηθήσει τους ειδικούς και με αυτό τον τρόπο θα μειωθούν τα επεισόδια. Έχουν χρησιμοποιηθεί οι μέθοδοι της συσχέτισης και της ταξινόμησης, οι οποίες εξάγουν κανόνες. Αφού μελετήθηκε η βιβλιογραφία εστιάστηκε η προσοχή μας στα κριτήρια διαχωρισμού, στα μέτρα, τον υπολογισμό του κινδύνου πάθησης ενός επεισοδίου και στην παρουσίαση των αποτελεσμάτων.

Στη μέθοδο της ταξινόμησης χρησιμοποιήθηκε ο αλγόριθμος C4.5, ο οποίος τροποποιήθηκε έτσι ώστε να δέχεται σαν κριτήριο διαχωρισμού αυτό που θα ορίσει ο χρήστης. Έχουν υλοποιηθεί πέντε κριτήρια διαχωρισμού: το Information Gain, το Gain Ratio, το Gini Index, το Distance Measure και το Likelihood Ratio Chi-squared statistics. Από τη μέτρηση της σωστής ταξινόμησης (correct classification) φάνηκε ότι όλα τα κριτήρια διαχωρισμού έχουν περίπου την ίδια απόδοση.

Στη μέθοδο της συσχέτισης έχει υλοποιηθεί ένας νέος αλγόριθμος, ο AKAMAS, που σαρώνει μόνο μια φορά τη βάση για να εξάξει τους κανόνες. Σαν κατώφλι μπορεί να χρησιμοποιηθεί οποιοδήποτε από τα υλοποιημένα μέτρα.

Για κάθε μοντέλο και πάθηση που μελετήθηκαν είχαν εξαχθεί πολλοί κανόνες. Με βάση τη συχνότητα που είχαν οι παράγοντες σε αυτούς τους κανόνες, υπολογίστηκαν οι σημαντικότεροι παράγοντες σε κάθε μοντέλο. Με τα μέτρα της υποστήριξης και της εμπιστοσύνης δεν ήταν δυνατό να απομονωθούν οι πιο δυνατοί κανόνες για το λόγο ότι η απόκλιση των τιμών των δύο μέτρων ήταν στις περισσότερες περιπτώσεις πολύ μικρή. Με τη βοήθεια των άλλων μέτρων που έχουν υλοποιηθεί είχε καταστεί δυνατό να περιοριστούν οι κανόνες στους πιο σημαντικούς. Αναμένεται να διερευνηθεί η αξιολόγηση αυτών των κανόνων. Η επιλογή των μέτρων που χρησιμοποιήθηκαν στην κάθε περίπτωση έγινε με τη βοήθεια της μεθόδου της ταξινόμησης με δέντρα απόφασης, όπου σε κάθε μοντέλο τα μέτρα ήταν διαφορετικά. Επίσης, με τη χρήση του μέτρου χ^2 υπολογίστηκε το μέτρο p-value, που αν είναι μικρότερο από 0.05 δείχνει ότι ο κανόνας είναι στατιστικά σημαντικός. Εκτελώντας τον αλγόριθμο με τους σημαντικούς εξαγόμενους κανόνες, έδειξε ότι γίνεται πιο σωστή ταξινόμηση.

Για κάθε ασθενή της μελέτης έχει υπολογιστεί με τη βοήθεια της εξίσωσης του Framingham ο κίνδυνος πάθησης ενός επεισοδίου. Σε κάθε εξαγόμενο κανόνα λήφθηκαν υπόψη όλοι οι ασθενείς που συμπεριλαμβάνονται σε αυτόν τον κανόνα και υπολογίστηκε ο μέσος όρος του κινδύνου πάθησης ενός επεισοδίου για τον κανόνα. Αυτό θα βοηθήσει τον γιατρό να διαπιστώσει αμέσως σε έναν νέο ασθενή σε ποια κατηγορία κινδύνου είναι, αφού έχει απομονώσει τον κανόνα ή τους κανόνες που αντιπροσωπεύουν αυτό τον ασθενή. Θα μελετήσει τους μεταβαλλόμενους παράγοντες του ασθενή και θα του δώσει φαρμακευτική αγωγή για να πάρουν κανονικές τιμές οι παράγοντες αυτοί και έτσι να μειωθεί το ποσοστό κινδύνου πάθησης επεισοδίου.

Είναι σημαντικό με την εξαγωγή των κανόνων να μπορεί κάποιος άμεσα να επεξεργάζεται τα αποτελέσματα. Έχει δοθεί μεγάλη σημασία στην παρουσίαση των αποτελεσμάτων, έτσι ώστε να είναι εύκολη και γρήγορη η περεταίρω επεξεργασία τους.

Η μεθοδολογία έχει εφαρμοστεί στην βάση δεδομένων με καρδιαγγειακά επεισόδια, που στην ουσία ήταν τρεις βάσεις δεδομένων για τα τρία διαφορετικά επεισόδια που μελετούμε. Οι δύο βάσεις δεδομένων είχαν εντελώς διαφορετικούς παράγοντες κινδύνου και η τρίτη περιείχε τους παράγοντες που είχαν οι άλλες δύο μαζί. Έχει διαπιστωθεί ότι αυτή η μεθοδολογία μπορεί να εφαρμοστεί για οποιαδήποτε βάση δεδομένων χωρίς να χρειάζεται να γίνει οποιαδήποτε αλλαγή. Αυτό οφείλεται στο γεγονός ότι το σύστημα που αναπτύχθηκε με αυτή τη μεθοδολογία είναι δυναμικό και έτσι επιτρέπει τη χρήση οποιασδήποτε βάσης δεδομένων.

Όσον αφορά την απόδοση των αλγορίθμων, είχαμε στους κανόνες με δέντρα απόφασης 65% σωστή ταξινόμηση και στους κανόνες συσχέτισης 58% σωστή ταξινόμηση. Στους στατιστικά σημαντικούς κανόνες είχαμε στους κανόνες συσχέτισης 69% σωστή ταξινόμηση. Όσον αφορά τα δέντρα απόφασης, δεν ήταν εφικτό να γίνει αξιολόγηση για το λόγο ότι οι στατιστικά σημαντικοί κανόνες ήταν πολύ λίγοι. Άλλοι ερευνητές έχουν δείξει καλύτερα αποτελέσματα, όπως ο Ordonez [122] που το μοντέλο του είχε απόδοση γύρο στο 90%, όμως πάνω από το 80% των κανόνων που είχαν εξαχθεί αφορούσαν μόνον ένα ασθενή.

9.2 Μελλοντική εργασία

Το ολοκληρωμένο σύστημα εξόρυξης γνώσης με εξαγωγή κανόνων σε ιατρικά θέματα έχει εφαρμοστεί σε διάφορες βάσεις δεδομένων.

Με τη βοήθεια των κανόνων που έχουν εξαχθεί, είναι σε θέση κάποιος ειδικός να επισημάνει τους σημαντικούς παράγοντες πάθησης ενός επεισοδίου, όπως επίσης να δει το ποσοστό κινδύνου πάθησης ενός επεισοδίου σε ένα νέο ασθενή. Ξεκινώντας το σύστημα, ο καρδιολόγος έχει τη δυνατότητα να εισάγει τις τιμές των παραγόντων ενός νέου ασθενή. Με την εισαγωγή κάθε τιμής παράγοντα, αποκλείονται από το σύστημα τα πρότυπα που έχουν σε αυτούς τους παράγοντες διαφορετική τιμή. Έχοντας εισάγει όλες τις τιμές του νέου ασθενή, παρουσιάζονται τα πρότυπα που ταυτίζονται με αυτές τις τιμές. Έτσι μπορεί να διακρίνει ο καρδιολόγος αν ο νέος ασθενής θα πάθει επεισόδιο, ή και ακόμη σε τι ποσοστό κινδύνου

βρίσκεται ο ασθενής. Αλλάζοντας δοκιμαστικά τις τιμές των μεταβαλλόμενων παραγόντων και βλέποντας τα αντίστοιχα πρότυπα, μπορεί ο καρδιολόγος να επικεντρωθεί σε κάποιο ή κάποιους παράγοντες του νέου ασθενή και με κάποια φαρμακευτική αγωγή να μειώσει τον κίνδυνο πάθησης ενός επεισοδίου στον νέο ασθενή.

Ο αλγόριθμος AKAMAS είναι πιο αργός από τον αλγόριθμο Apriori για το λόγο ότι εξάγει περισσότερους κανόνες. Για το λόγο ότι ο AKAMAS βγάζει όλους τους πιθανούς κανόνες, θα μπορούσε να μην γίνεται στην αρχή η σάρωση της βάσης και να δημιουργούνται απ' ευθείας οι κανόνες. Μετά, με τη χρήση των μέτρων θα ελέγχονται και επιλέγονται οι κανόνες. Αποφεύγοντας τη σάρωση της βάσης ο αλγόριθμος θα γίνει σαφώς πιο γρήγορος.

Έχει υλοποιηθεί ένας αλγόριθμος επιλογής των μέτρων που είναι πιο σημαντικοί στο μοντέλο που χρησιμοποιείται. Με βάσει αυτό ή αυτά τα μέτρα θα μπορούσε να αυτοματοποιηθεί το σύστημα έτσι ώστε να βγαίνουν οι κανόνες αυτόματα.

Ο υπολογισμός του κινδύνου πάθησης ενός επεισοδίου έχει εξαχθεί με την εξίσωση του Framingham. Αυτή η εξίσωση έχει χρησιμοποιηθεί από πολλούς ειδικούς. Μερικοί από αυτούς ισχυρίζονται ότι η εξίσωση αυτή υπερεκτιμά τον κίνδυνο και άλλοι ότι τον υποεκτιμά. Έχοντας ένα μεγάλο δείγμα του πληθυσμού μπορούμε να εξάγουμε μια νέα εξίσωση που να είναι πιο αντιπροσωπευτική.

Όσον αφορά το σύστημα που εξάγει τους κανόνες συσχέτισης και ταξινόμησης μπορεί να διαφοροποιηθεί έτσι ώστε ο χρήστης να μπορεί να επιλέξει κάποιους από τους παράγοντες για να χρησιμοποιήσει για αυτές τις μεθόδους.

Το σύστημα αυτό θα εφαρμοστεί σε νέα ιατρικά προβλήματα όπως:

* Παράγοντες Καρδιακής ανεπάρκειας
* Μελέτη των προδιαθεσικών παραγόντων για την πάθηση του διαβήτη τύπου 2
* Παράγοντες παχυσαρκίας σε παιδιά

Στα τρία αυτά θέματα θα εφαρμοστεί το σύστημα μας για να εντοπιστούν οι κυριότεροι παράγοντες της κάθε πάθησης με τη βοήθεια των εξαγομένων κανόνων. Στο πρώτο θέμα, για την καρδιακή ανεπάρκεια, έχει ήδη εφαρμοστεί η μεθοδολογία όπου φάνηκε ότι δεν χρειάζεται καμιά αλλαγή στο σύστημα για να εξάγονται τα αποτελέσματα. Έχουν γίνει και

δοκιμαστικές εκτελέσεις με τη βάση δεδομένων για τους παράγοντες παχυσαρκίας. Παρόλο που η βάση δεν ήταν με ιατρικούς παράγοντες, η μεθοδολογία μας εφαρμόστηκε χωρίς κάποια τροποποίηση. Η ίδια μεθοδολογία θα εφαρμοστεί και στους παράγοντες για την πάθηση του διαβήτη τύπου 2, όπου μεταξύ άλλων περικλείονται και γενετικά δεδομένα.

Βιβλιογραφία

[1] C. Westphal, T. and Blaxton, 'Data mining solutions: Methods & tools for solving real-world problems,' John Wiley & Sons, 1998.

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

[3] Μ. Βαζιργιάννης και Μ. Χαλκίδη, 'Εξόρυξη γνώσης από βάσεις δεδομένων,' Τυπωθήτω, Αθήνα, 2003.

[4] P-N Tan, Introduction to Data Mining, Addison-Wesley, Boston, 2006.

[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. R. Quinlan, 'C4.5: Programs for Machine Learning,' Morgan Kaufmann, 1993.

[7] L. Breiman, J. Friedman, R. Olsen, C. Stone, 'Classification and Regression Trees,' Wadsworth Int. Group, Florida, 1984.

[8] R. Agrawal and R. Srikant, 'Fast algorithms for mining association rules,' In Proc. of the 20th Int'l Conf. on Very Large Databases (VLDB), Santiago, Chile, June 1994. SIGKDD Explorations. Vol. 2, Issue 1 – pp. 63, 1994.

[9] M. Fayyad Usama, Piatesky-Shapiro, Gregory, Padhraic Smuth and Ramasamy Uthurusamy, 'Advances in knowledge discovery and data mining,' AAAI Press, 1996.

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

[11] E. F. Codd, S. B. Codd, C.T., Salley, 'Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate,' E.F. Codd & Associates, 1993.

[12] S. Chaudhuri, and U. Dayal, 'Data warehousing and OLAP for Decision Support,' Tutorials of 22nd VLDB Conference, pp.: 507-508, 1996.

[13] J. Han and M. Kamber, 'Data Mining, Concepts and Techniques', Morgan Kaufmann Publishers, San Fransisco, 2006.

[14] Cardiovascular Diseases (CVDs), Fact Sheet No 317, Updated September 2009, http://www.who.int/mediacentre/factsheets/fs317/en/index.html, τελευταία επίσκεψη της ιστοσελίδας: 8.10.2010.

[15] Healthier World, 'Καρδιακά και εγκεφαλικά συχνότερη αιτία θανάτου,' http://www.healthyworld.gr, 2009, τελευταία επίσκεψη της ιστοσελίδας: 22/01/2010.

[16] Σ. Καρακατσάνη, «Οι θάνατοι από στεφανιαία νόσο μπορούν να μειωθούν στο ήμισυ», Εφημερίδα Σημερινή, 05/04/2009,

[17] http://www.sigmalive.com/simerini/news/health/140189, τελευταία επίσκεψη της ιστοσελίδας: 22/01/2010.

[18] Αγγειοπλαστική ή μπαλλονάκι στη στεφανιαία νόσο, http://www.incardiology.gr/pathiseis_stefaniaia/pc_ptca.htm, τελευταία επίσκεψη της ιστοσελίδας: 8.10.2010.

[19] J. R. Quinlan, 'Induction of Decision Trees,' Machine Learning. 1, pp.: 81-106, 1986.

[20] A. Agresti, *'Categorical Data Analysis,'* New York: Wiley-Interscience, 2002.

[21] H. Zhang, 'The Optimality of Naive Bayes,' FLAIRS2004 conference, 2004.

[22] S. Haykin, *'Neural Networks: A Comprehensive Foundation,'* Prentice Hall, 1999.

[23] K. P. Bennett and C. Campbell, 'Support Vector Machines: Hype or Hallelujah?,' SIGKDD Explorations, 2,2, pp.1-13, 2000.

[24] S. K. Murthy, 'Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery,' 2 (4), pp.:345-389, 1998.

[25] S. R. Safavian, and D. Landgrebe, 'A survey of decision tree classifier methodology. IEEE Trans. Systems, Man & Cybernetics,' 21 (3), pp.: 660-674, 1991.

[26] A. Fielding, 'Binary segmentation: the automatic interaction detector and related techniques for exploring data structures,' in O' Muircheartaigh, C.A. and Payne, C. (eds.). The analysis of survey data, volume I, pp.: 221-257, John Wiley & Sons. Chichester, UK, 1977.

[27] J. A. Sonquist, E. L. Baker, J. N. Morgan, 'Searching for Structure,' Institute for Social Research, University of Michigan, Ann Arbor, MI, 1971.

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

[29] 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.

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

[31] P. Swain and H. Hauska, 'The decision tree classifier design and potential,' IEEE Trans On Geoscience and Electronics. GE-15, pp.: 142-147, 1977.

[32] E. A. Feigenbaum, 'Expert Systems in the 1980s,' In Bond A. (ed.), State of the Art in Machine Intelligence. Pergamon- Infotech. Maidenhead, 1981.

[33] K. R. Pattipati and M. Alexandridis, 'Application of heuristic search and information theory to sequential fault diagnosis,' IEEE Trans On Systems, Man and Cybernetics. 20 (4), pp.: 872-887, 1990.

[34] P. K. Varshney, C. R. P. Hartmann, J. M. De Faria Jr., 'Application of information theory to sequential fault diagnosis,' IEEE Trans On Comp. C-31 (2), pp.: 164-170, 1982.

[35] E. B. Anderson, 'Asymptotic Properties of Conditional Maximum Likelihood Estimators,' Journal of the Royal Statistical Society B 32, pp.: 283-301, 1970.

[36] S. B. Patil and Y. S. Kumaraswamy,' Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network,' European Journal of Scientific Research, Vol.31 No.4, pp.642-656, 2009.

[37] N. Cristianini, J. Shawe-Taylor, 'An introduction to support vector machines and other kernel-based learning methods,' Cambridge: Cambridge University Press, 2000.

[38] G. Landeweerd, T. Timmers, E. Gelsema, M. Bins, and M. Halic, 'Binary trees versus single level tree classification of white blood cells,' Pattern Recognition. 19, pp.: 229-235, 1983.

[39] B. Kim, and A. Landgrebe, 'Hierarchical decision tree classifiers in high dimensional and large class data,' Ph.D. thesis and Technical Report TR-EE-90-47, School of EE, Purdue University, 1990.

[40] W. Buntine, 'Learning classification trees,' Statistics and Computing, vol. 2, pp. 63-73, 1992.

[41] C. E. Shannon, 'A mathematical theory of communication,' Bell System Technical Journal. 27, pp.: 379-423, 623-656, 1948.

[42] C.R.P. Hartmann, P. K. Varshney, K. G. Mehrotra, C. L. Gerberich, 'Application of information theory to the construction of efficient decision trees,' IEEE Trans. on Information Theory. IT-28 (4), pp.: 565-577, 1982.

[43] R. G. Casey, G. Nagy, 'Decision tree design using a probabilistic model,' IEEE Trans. on Information Theory. IT-30 (1), pp.: 93-99, 1984.

[44] W. Hanisch, 'Design and optimization of a hierarchical classifier,' Journal of new Generation Computer Systems. 3 (2), pp.: 159-173, 1990.

[45] S. B. Gelfand, C. S. Ravishankar, E. J. Delp, 'An iterative growing and pruning algorithm for classification tree design. IEEE Transactions on Pattern Analysis and Machine Intelligence. 13 (2), pp.: 163-174, 1991.

[46] P. C. Taylor, and B. W. Silverman, 'Block diagrams and splitting criteria for classification trees,' Statistics and Computing. 3 (4), pp.: 147-161, 1993.

[47] Y. K. Lin, and K.-S. Fu, 'Automatic classification of cervical cells using a binary tree classifier,' Pattern Recognition. 16 (1), pp.: 69-80, 1983.

[48] X. J. Zhou, T. S. Dillon, 'A statistical-heuristic feature selection criterion for decision tree induction,' IEEE Trans. on Pattern Analysis and machine Intelligence. PAMI-13 (8), pp.: 834-841, 1991.

[49] J. Mingers, 'Expert systems- rule induction with statistical data,' Journal of the Operational Research Society. 38 (1), pp.: 39-47, 1987.

[50] A. Hart, 'Experience in the use of an inductive system in knowledge eng,' In M. Bramer (ed.) Research and Development in Expert Systems. Cambridge University Press, Cambridge, MA, 1984.

[51] E. Rounds, 'A combined non-parametric approach to feature selection and binary decision tree design,' Pattern Recognition. 12, pp.: 313-317, 1980.

[52] J. H. Friedman, 'A recursive partitioning decision rule for nonparametric classifiers. IEEE Trans. on Comp. C- 26, pp.: 404-408, 1977.

[53] R. E. Haskell, and A.Noui- Mehidi, 'Design of hierarchical classifiers,' In N.A. Sherwani, E. de Doncker, J.A. Kapenga (eds.). Computing in the 90's: Proc. of the First Great Lakes Computer Science Conference, pp.: 118-124, 1989.

[54] R. C. Luo, R. S. Scherp, M. Lanzo, 'Object identification using automated decision tree construction approach for robotics applications,' Journal of Robotic Systems, 4 (3), pp.: 423-433, 1987.

[55] 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.

[56] W.-Y. Loh, and N. Vanichsetakul, 'Tree- structured classification via generalized discriminant analysis,' Journal of the American Statistical Association, 83 (403), pp.: 715-728, 1988.

[57] I. K. Sethi, 'Entropy nets: from decision trees to neural networks,' Proceedings of the IEEE. 78 (10), pp.: 1605-1613, 1990.

[58] H. Guo, and S. B. Gelfand, 'Classification trees with neural network feature extraction,' IEEE Trans on Neural Networks, 3 (6), pp.: 923-933, 1992.

[59] I. K. Sethi and B. Chatterjee, 'Efficient decision tree design for discrete variable pattern recognition problems,' Pattern Recognition, 9, pp.:197-206, 1977.

[60] P. H. Chou, 'Optimal partitioning for classification and regression trees,' IEEE Trans. on Pattern Analysis and Machine Intelligence, 13 (4): 340-354, 1991.

[61] I. Cleote and H. Theron, 'CID3: An extension of ID3 for attributes with ordered domains,' Journal of South African Computer, 4, pp.: 10-16, 1991.

[62] J. R. Quinlan, 'Simplifying decision trees,' International Journal of Man-, Machine Studies, 27, pp.: 221-234, 1987.

[63] J.-H. Lin, J. S. Vitter, 'Nearly optimal vector quantization via linear programming,' In J.A. Storer and M. Cohn (eds.), DCC 92: Data Compression Conference, pp.: 22-31, 1992.

[64] J. R. B. Cockett and J. A. Herrera, 'Decision tree reduction,' Journal of the ACM. 37 (4), pp.: 815-842, 1990.

[65] D. E. Brown and C. L. Pittard, 'Classification trees with optimal multivariate splits,' Proc. of the International Conference on Systems, Man and Cybernetics. 3, pp.: 475-477, 1993.

[66] J. R. Quinlan, 'Comparing connectionist and symbolic learning methods,'Proc. of the Conference on Computational learning theory, pp.: 445-456, 1990.

[67] T. G. Dietterich, H. Hild, G. Bakiri, 'A comparison of ID3 and backpropagation for English text-to-speech mapping,' Machine Learning, 18, pp.: 51-80, 1995.

[68] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining,' Addison Wesley, Boston, 2006.

[69] J. D. Callahan and S. W. Sorensen, 'Rule induction for group decisions with statistical data-an example,' Journal of the Operational Research Society, 42 (3), pp.: 227-234, 1991.

[70] W. J. Long, J. L. Griffith, H. P. Selker, R. B. D' Agostino, 'A comparison of logistic regression to decision tree induction in a medical domain,' Computers and Biomedical Research, 26 (1), pp.: 74-97, 1993.

[71] Κ. Χριστοφής, 'Σύστημα εξαγωγής γνώσης από κατανεμημένες και ετερογενείς βάσεις δεδομένων: Εφαρμογής σε ιατρικά πληροφοριακά συστήματα,' Μεταπτυχιακή εργασία, Πανεπιστήμιο Κρήτης, 2000.

[72] R. Agrawal, T. Imielinski, and A. Swami, 'Mining association rules between sets of items in large databases,' In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '93), Washington, USA, 1993.

[73] C. Marinica1 and F. Guillet1, 'Improving post-mining of association rules with ontologies,' The XIII International Conference "Applied Stochastic Models and Data Analysis", Vilnius, pp. 76–80, 2009.

[74] M. Houtsma and A. Swami, 'Set-oriented mining for association rules in relational databases,' Technical Report RJ 9567, IBM Almaden Research Center, San Jose, California, 1993.

[75] A. Savesere, E. Omiecinski, S. Navathe, 'An efficient algorithm for mining association rules in large databases,' Proc. of the 21st International Conference on Very Large Databases, pp.: 432-444, 1995.

[76] H. Mannila, H. Toivonen, A. I. Verkamo, 'Efficient Algorithms for Discovering Association Rules,' Proceedings of the AAAI Workshop on Knowledge Discovery in Databases (KDD-94), pp.: 181-192, 1994.

[77] H. Toivonen, 'Sampling Large Databases for Association Rules,' Proc. Of the 22nd International Conference on Very Large Databases, pp.: 134-145, 1996.

[78] J. S. Park, M.-S. Chen, P. S. Yu, 'An effective Hash-Based Algorithm for Mining Association Rules,' Proc.of the ACM SIGMOD International Conference on Management of Data, pp.: 175-186, 1995a.

[79] C. Hidber, 'Online Association Rule Mining,' Proceedings ACM SIGMOD International Conference on Management of Data, pp.: 145-156, 1999.

[80] R. Agrawal, and J. C. Shafer, 'Parallel Mining of Association Rules,' IEEE Transactions on Knowledge Discovery and Data Engineering. 8 (6), pp.: 962-969, 1996.

[81] 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.

[82] 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.

[83] E.-H. Han, G. Karypis, V. Kumar, 'Scalable parallel Data Mining for Association Rules,' Proc. of the ACM SIGMOD Conference, pp.: 277-288, 1997.

[84] T. Shintani and M. Kitsuregawa, 'Parallel Mining Algorithms for Generalized Association Rules with Classification Hierarchy,' Proceedings ACM SIGMOD International Conference on Management of Data. 27 (2), pp.: 25-36, 1998.

[85] 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.

[86] I. Cengiz, 'Mining Association Rules,' Bilkent University, Department of Computer Engineering and Information Sciences. Ankara, Turkey, 1997. http://www.cs.bilkent.edu.tr/~icegiz/datamone/mining.html.

[87] R. Srikant and R. Agrawal, 'Mining Generalized Association Rules,' Proc. of the 21st International Conference on Very Large Databases. 13 (2-3), pp.:161-180, 1995.

[88] T. Shintani and M. Kitsuregawa, 'Parallel Mining Algorithms for Generalized Association Rules with Classification Hierarchy,' Proceedings ACM SIGMOD International Conference on Management of Data. 27 (2), pp.: 25-36, 1998.

[89] J. Han and Y. Fu, 'Discovery of Multiple-Level Association Rules from Large Databases,' Proc. of the 21st International Conference on Very Large Databases, pp.: 420-431, 1995.

[90] K. Koperski and J. Han, 'Discovery of Spatial Association Rules in Geographic Information Databases,' Lecture Notes in Computer Science. 951, pp.:47- 66, 1995.

[91] C. E. Shannon, 'Communication Theory of Secrecy Systems,' MD Computing, 1998.

[92] R. Ld. Mantras, 'A distance-based attribute selection measure for decision tree induction,' Machine Learning, 6:81-92, 1991.

[93] F. Attneave, "Applications of Information Theory to Psychology," Holt, Rinehart, and Winston, 1959.

[94] M Kearns and Y. Mansour, ' On the boosting ability of the top-down decision tree learning algorithms,' Journal of Computer and Systems Sciences, 58(1): 109-128, 1999.

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

[96] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In The Proceedings of SIGMOD, pages 255-264, AZ,USA, 1997.

[97] R. Kohavi and F. Provost, 'On Applied Research in Machine Learning,' Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process vol. 30, Num. 2/3, 1998.

[98] Van Rijbergen, 'Information Retrieva,' Butterworths, C.J. (1979).

[99] M. Kubat, 'Decision Trees Can Initialize Radial-Basis-Function Networks,' IEEE Transactions on Neural Networks, 9 pp.: 813-821, 1998.

[100] J. A. Swets, R. M. Pickett, S. F. Whitehead, D. J. Getty, J. A. Schnur, J. B. Swets, and B. A. Freeman, 'Assessment of diagnostic technologies,' Science, Vol. 205. no. 4408, pp.: 753 – 759, 1979.

[101] F. Provost, T. Fawcett, and R. Kohavi, 'The Case Against Accuracy Estimation for Comparing Induction Algorithms,' Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, pp.:445-553, 1998.

[102] Euroaspire study group, "A European Society of Cardiology survey of secondary prevention of coronary heart disease: Principal results," European Heart Journal, vol. 18, pp. 1569-1582, 1997.

[103] Euroaspire II Study Group, "Lifestyle and risk factor management and use of drug therapies in coronary patients from 15 countries," European Heart Journal, vol. 22, pp. 554-572, 2002.

[104] T.D. Rea, S. R. Heckbert, R. C. Kaplan, N. L. Smith, R. N. Lemaitre, and B. M. Psaty., "Smoking Status and Risk for Recurrent Coronary Events after Myocardial Infraction," Ann Intern Med., vol. 137, pp. 494-500, 2002.

[105] B. Wuensche, 'The Visualization and Measurement of Left Ventricular Deformation,' First Asia_Pacific Bioinformatics Conference, 2003.

[106] 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.

[107] H. Bambrick, 'Relationships between BMI, waist circumference, hypertension and fasting glucose: Rethinking risk factors in Indigenous diabetes,' Australian Indigenous Health Bulletin Vol 5 No 4 September – December 2005.

[108] T. Marshall, "Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values," *BMC Public Health,,* vol. 8, pp. 25, 2008.

[109] Euroaspire study group, "Euroaspire III: a survey on the lifestyle, risk factors and use of cardioprotective drug therapies in coronary patients from 22 European countries," European Journal of Cardiovascular Prevention & Rehabilitation, vol. 16, no. 2, pp. 121-137, 2009.

[110] C. L. Tsien, H. S. F. Fraser, W. J. Long, and R. L. Kennedy, "Using classification trees and logistic regression methods to diagnose myocardial infraction," in Proc. 9th World Congr. Med. Inform., vol. 52, pp. 493-497, 1998.

[111] L. Goldman, M. Weinberg, M. Weisberg, R. Olshen, E. Cook, R. Sargent, G. Lamas, C. Dennis, C. Wilson, L. Deckelbaum, H. Fineberg, R. Stiratelli, 'A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain,' *N Engl J Med*; (307): 588-596, 1982.

[112] R. L. Kennedy, A. M. Burton, H. S. Fraser, L. N. McStay, R. F. Harrison, 'Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models,' *EHJ*; (17): 1181-1191, 1996.

[113] I. Colombet, A. Ruelland, G. Chatellier, F. Gueyffier, P. Degoulet, and M. C. Jaulent, 'Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression,' Proc AMIA Symp: 156–160, 2000.

[114] D. Gamberger, and R. Bošković Institute, Zarageb, Croatia, "Medical prevention: Targeting high-risk groups for coronary heart disease," Sol-EU-Net: Data Mining and Decision Support, [Online]. Available: http://soleunet.ijs.si/website/other/case_solutions/CHD.pdf.

[115] V. Podgorelec, P. Kokol, B. Stiglic and I. Rozman., "Decision Trees: an overview and their use in medicine," J. Med. Syst.; vol. 26, no. 5, pp. 445-63, 2002.

[116] R. Voss, P. Cullen, H. Schulte and G. Assmann, 'Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks,' International Journal of Epidemiology; 31:1253-1262, 2002.

[117] E. A. Madigan and O. L. Curet, 'A data mining approach in home healthcare: outcomes and service use,' BMC Health Services Research, **6:**18, 2006.

[118] S. Bayat, M. Cuggia, D. Rossille, M. Kessler, and L. Frimat, 'Comparison of Bayesian Network and Decision Tree Methods for Predicting Access to the Renal Transplant Waiting List,' Stud Health Technol Inform;150:600-4, 2009.

[119] C. Ordonez, E. Omiecinski, L. de Braal, C.A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczvnska, and E. V. Garcia, "Mining Constrained Association Rules to Predict Heart Disease," in Proc. International Conference on Data Mining, ICDM, IEEE, pp. 431-440, 2001.

[120] Th. Karban, J. Rauch, and M. Simunek, **'**SDS-rules and association rules**,'** Proceedings of the 2004 ACM symposium on Applied computing, Pp: 520 – 524, 2004.

[121] T.P. Exarchos, A.T. Tzallas, D.I. Fotiadis, S. Konitsiotis and S. Giannopoulos, 'A Data Mining based Approach for the EEG Transient Event Detection and Classification,' Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05), 2005.

[122] C. Ordonez, "Comparing Association Rules and Decision Trees for Disease Prediction," in Proc. Int. Conference on Information and Knowledge Management, workshop on Healthcare information and knowledge management, Arlington, Virginia, USA, pp. 17-24, 2006.

[123] Y.-T. Kuo, A. Lonie, L. Sonenberg, and K. Paizis, 'Domain Ontology Driven Data Mining', A Medical Case Study, ACM SIGKDD Workshop on Domain Driven Data Mining, (DDDM2007), 2007.

[124] S. Concaro, L. Sacchi, C. Cerra, R. Bellazzi, 'Mining Administrative and Clinical Diabetes Data with Temporal Association Rules,' *Medical Informatics in a United and Healthy Europe, Pp. 574-578, 2009.*

[125] P. Brindle, J. Emberson, F. Lampe, M. Walker, P. Whincup, T. Fahey, S. Ebrahim, 'Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study,' BMJ Vol. 327, 2003.

[126] P. E. Greenwood and M. S. Nikulin, 'A guide to chi-squared testing,' Wiley, New York. ISBN 047155779X, 1996.

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

[128] Δ. Χατζηπαναγή, 'Rule extraction of cardiovascular database using decision trees,' Μεταπτυχιακή εργασία, Πανεπιστήμιο Κύπρου, 2009.

[129] H. Hamilton, E. Gurak, L. Findlater, and W. Olive, 'Overview of Decision Trees,' Rudjer Boskovic Institute, 2001

[130] J. Gehrke, R. Ramakrishnan, V. Ganti, 'RainForest - A Framework for Fast Decision Tree Construction of Large Datasets,' In VLDB, p. 416-427, Morgan Kaufmann, 1998

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

[132] R.Rastogi, K.Shim, 'PUBLIC: A Decision Tree Classifier that Integrates Pruning and Building,' In Proc. of VLDB, 1998.

[133] Λ. Παπακωνσταντίνου, ' Εξόρυξη κανόνων από καρδιαγγειακή βάση με τη χρήση αλγορίθμων συσχέτισης,' Μεταπτυχιακή εργασία, Πανεπιστήμιο Κύπρου, 2009.

[134] Ν. Κουμπάρου, 'Αξιολόγηση κανόνων βάσει πολλαπλών μέτρων,' Μεταπτυχιακή εργασία, Πανεπιστήμιο Κύπρου, 2010.

[135] F. Wilcoxon, "Individual Comparisons by Ranking Methods," Biometrics, vol.1, pp. 80-83, 1945.

[136] W. J. Long, J. L. Griffith, H. P. Selker, and R. B. DAgostino, 'A comparison of logistic regression to decision-tree induction in a medical domain,' *Comput Biomed Res*; (26): 74-97, 1993.

[137] D. Michie, D. J. Spiegelhalter, C. C. Taylor, Machine Learning, Neural and Statistical Classification, West Susser, England:Ellis Horwood, 1994.

[138] L. Rokach and O. Maimon, Data Mining with decision trees, Theory and applications, World Scientific Publishing, Singapore, 2008.

[139] Weka 3: Data Mining Software in Java, *WEKA the University of Waikato*, http://www.cs.waikato.ac.nz/~ml/weka/

[140] DBMiner technology inc, http://www.dbminer.com

[141] The data mine, spss Clementine, http://www.the-data-mine.com/bin/view/Software

ΠΑΡΑΡΤΗΜΑ 1

Δημοσιεύσεις

A. Journals

- M. Karaolis, J. A. Moutiris, D. Hadjipanayi, C.S. Pattichis, 'Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining with Decision Trees,' IEEE Transactions on Information Technology in BioMedicine, accepted, to be published in 2010.
- E. Kyriacou, C.S. Pattichis , M. Karaolis ,C. Loizou, C. Christodoulou, M.S. Pattichis, S. Kakkos, A. Nicolaides, 'An Integrated System for Assessing Stroke Risk,' Special Issue on Image, Signal and Distributed Data Processing for Networked eHelath Applications, IEEE Engineering in Medicine and Biology Magazine, Vol. 26, No 8, pp. 43-50, Sept/Oct 2007.

B. Conference papers

- M. Karaolis, C. Pattichis, M. Pantzaris, A. Nicolaides, 'Data Mining in the Evaluation of Risk Factors fort the Assessment of Stroke,' Hellenic European Research on Computer Mathematics and its Applications, HERCMA 2003, pp. 284-289, September 25-27, Athens, Greece, 2003.
- E. Kyriacou, C. Pattichis, C. Christodoulou, M. Karaolis, A. Nicolaides, 'An Integrated Teleconsultation System for the Evaluation of the Risk of Stroke,' Proc. Of the 5th International Network Conference INC 2005, Samos, Greece, July 2005, 4 pages.
- M. Karaolis, J. A. Moutiris, C. S. Pattichis, 'Assessment of the risk of coronary heart event based on data mining,' BioInformatics and BioEngineering, BIBE 2008. 8th IEEE International Conference, pp.:1-5, 2008.

- M. Karaolis, J. A. Moutiris, L. Papaconstantinou, C. S. Pattichis, 'Association rule analysis for the assessment of the risk of coronary heart events,' Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, 3-6 Sept. 2009 Page(s):6238-6241.
- M. Karaolis, J.A. Moutiris, L. Papaconstantinou, C.S. Pattichis, 'AKAMAS: Mining Association rules using a new algorithm for the Assessment of the Risk of Coronary Heart Events,' 9th International Conference on Information Technology and Applications in Biomedicine, Larnaka, Cyprus, Nov. 5-7, 2009, 4 pages.

C. Abstracts

- C. Pattichis, C. Schizas, Y. Dimopoulos, G. Samaras, C. Christodoulou, M. Karaolis, M. Pantziaris, M. Pattichis, A. Nicolaides, 'Evaluation of the Risk of Stroke by Telemedicine,' World Congress on Medical Physics and Biomedical Engineering, Scientific Session: WE-A201-01 Telecom, Telemetry & High Speed Data Transmission, Track: 04 Medical Informatics and Biomedical Information Technology, Chicago, USA, July 23-28, 2000.
- M. Karaolis, C. Pattichis, M. Pantziaris and A. Nicolaides, 'Evaluation of the Risk of Stroke via Data Mining Analysis,' 6th Hellenic European Conference on Computer Mathematics and its Applications, September 25-27, 2003, Athens, Greece.
- M. Karaolis, C. Pattichis, M. Pantziaris, A. Nicolaides, 'Data Mining in the Evaluation of the Risk of Stroke,' First Mediterranean Congress of Neurology, 25-28 April 2002, Limassol, Cyprus.