ΠΕΡΙΛΗΨΗ

Η μελέτη αφορά την εξόρυξη κανόνων συσχέτισης από βάση δεδομένων με ασθενής με καρδιαγγειακά επεισόδια. Είναι γεγονός ότι στην Κύπρο παρατηρείται αυξημένη συχνότητα των καρδιαγγειακών επεισοδίων. Σκοπός της μελέτης μας, είναι μέσω της εφαρμογής αλγορίθμων εξόρυξης δεδομένων να προβούμε σε κάποια συμπεράσματα και κανόνες που συμβάλλουν στην πρόκληση καρδιαγγειακών επεισοδίων. Οι αλγόριθμοι που θα χρησιμοποιηθούν είναι αλγόριθμοι παραγωγής κανόνων συσχέτισης.

Για την εξόρυξη κανόνων συσχέτισης έχει χρησιμοποιηθεί ο αλγόριθμος Apriori, όπως επίσης παρουσιάζεται κι ένας καινούργιος αλγόριθμος, Akamas, ο οποίος για την εξόρυξη κανόνων δεν βασίζεται στο μέτρο αξιολόγησης κανόνων support. Για την εξόρυξη των κανόνων συσχέτισης έχει υλοποιηθεί ένα εργαλείο για εξόρυξη κανόνων, το οποίο βασίζεται στους αλγόριθμους Apriori και Akama. Στο εργαλείο έχουν υλοποιηθεί επίσης διάφορα σημαντικά μέτρα αξιολόγησης κανόνων, εκτός από το support confidence, από του οποίος ο χρήστης μπορεί να επιλέξει για να φιλτράρει και να ταξινομήσει τους εξαγόμενους κανόνες.

ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ ΑΠΟ ΚΑΡΔΙΑΓΓΕΙΑΚΗ ΒΑΣΗ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΣΥΣΧΕΤΙΣΗΣ

Παπακωνσταντίνου Λουκία

Η Διατριβή αυτή

Υποβλήθηκε προς Μερική Εκπλήρωση των

Απαιτήσεων για την Απόκτηση

Τίτλου Σπουδών Master

σε Προηγμένες Τεχνολογίες Πληροφορικής

στο

Πανεπιστήμιο Κύπρου

Συστήνεται προς Αποδοχή

από το Τμήμα Πληροφορικής

Ιούνης, 2009

ΣΕΛΙΔΑ ΕΓΚΡΙΣΗΣ

Διατριβή Master

ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ ΑΠΟ ΚΑΡΔΙΑΓΓΕΙΑΚΗ ΒΑΣΗ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΣΥΣΧΕΤΙΣΗΣ

Παρουσιάστηκε από

Παπακωνσταντίνου Λουκία

Ερευνητικός Σύμβουλος _____

Όνομα Ερευνητικού Συμβούλου

Μέλος Επιτροπής _____

Όνομα Μέλους Επιτροπής

Μέλος Επιτροπής _____

Όνομα Μέλους Επιτροπής

Πανεπιστήμιο Κύπρου

Ιούνης, 2009

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω όσους έχουν συμβάλει άμεσα ή έμμεσα στην διεκπεραίωση της διατριβής μου, που έχω αναλάβει προς εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου master. Αρχικά ευχαριστώ τον επιβλέποντα καθηγητή μου Δρ. Κωνσταντίνο Παττίχη όπως επίσης και τον κύριο Μηνά Καραολή για την βοήθεια και τη σωστή καθοδήγηση που μου πρόσφεραν οποιαδήποτε στιγμή την χρειάστηκα καθ' όλη την διάρκεια της εκπόνησης αυτής της μελέτης.

Τέλος θέλω να ευχαριστήσω τον σύζυγο, την οικογένεια μου και όλους τους φίλους που μου πρόσφεραν την υποστήριξη τους καθ' όλη τη διάρκεια εκπόνησης της διατριβής μου.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΜΕ ΠΙΝΑΚΕΣ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΞΙΣΩΣΕΩΝ

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Η κύρια αιτία θανάτου σε όλο τον κόσμο είναι τα καρδιαγγειακά νοσήματα. Αυτές οι παθήσεις ευθύνονται για τους περισσότερους από τους μισούς θανάτους στην Κύπρο αλλά και στην υπόλοιπη Ευρώπη, και σκοτώνουν περισσότερους ανθρώπους από ότι όλοι οι καρκίνοι μαζί. Σύμφωνα με διάφορες στατιστικές μελέτες που έγιναν σε όλη την Ευρώπη, τα καρδιοαγγειακά νοσήματα ευθύνονται για περισσότερους από 4,35 εκατομμύρια θανάτους κάθε έτος, στα 52 κράτη της Ευρωπαϊκής ηπείρου. Τα καρδιοαγγειακά νοσήματα προσβάλλουν εξίσου συχνά άνδρες και γυναίκες. Το 55% θανάτων των γυναικών οφείλεται σε αυτές τις νόσους, ενώ για τους άνδρες το 43% [1].

Σε όλες τις αναπτυγμένες χώρες του κόσμου παρατηρείται αύξηση των καρδιαγγειακών νοσημάτων, κι αυτό οφείλεται κυρίως στον σύγχρονο τρόπο ζωής, που χαρακτηρίζεται από καθιστική ζωή, ανθυγιεινή διατροφή, έλλειψη συστηματικής σωματικής άσκησης και παχυσαρκία. Οι επιπτώσεις αυτών των συνθηκών είναι η αύξηση των λιπιδίων του αίματος, η εμφάνιση σακχαρώδους διαβήτη και η αρτηριακή υπέρταση. Ένας άλλος σημαντικός παράγοντας καρδιαγγειακών νοσημάτων είναι το κάπνισμα, όπως επίσης και ο ψυχολογικός παράγοντας και το στρες [1].

Από διάφορες στατιστικές μελέτης που έχουν γίνει στη Κύπρο, 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%, το οποίο αυξάνεται κάθε χρόνο [2].

Όπως είχαμε αναφέρει πιο πάνω, υπάρχουν διάφοροι και πολλοί παράγοντες οι οποίο μπορεί να προκαλέσουν καρδιακό έμφραγμα. Έτσι είναι πολύ σημαντικό να μελετήσουμε και να βρούμε τους πιο κύριος παράγοντες στους οποίους προκαλείται καρδιακό έμφραγμα, και κυρίως για του λόγους που οδηγούν σε έμφραγμα του μυοκαρδίου.

Υπάρχουν διάφορα είδη καρδιακών παθήσεων, 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 απόφραξης λόγω δημιουργίας θρόμβου σε μια στεφανιαία αρτηρία. Ο θρόμβος διακόπτει τη κυκλοφορία του αίματος με αποτέλεσμα την νέκρωση του μυοκαρδίου.

1.2 Σχετική Εργασία

Βασικός σκοπός της μελέτης είναι η ανάπτυξη εργαλείου εξόρυξης δεδομένων για ανάλυση βάσεων δεδομένων. Έχουμε αναπτύξει ένα απλό και ευκολόχρηστο εργαλείο, το οποίο εξαγάγει κανόνες συσχέτισης. Στο κλάδο της εξόρυξης δεδομένων έχουν αναπτυχθεί διάφορα εργαλεία εξόρυξης δεδομένων, που είναι όμως πιο περίπλοκα και δύσκολο να χρησιμοποιηθούν από τους χρήστες.

Ένας άλλος σκοπός της μελέτης μας είναι η ανάλυση της βάσης δεδομένων με ασθενής με καρδιακά επεισόδια. Όπως έχει αναφερθεί και στην υποκεφάλαιο υπάρχουν διάφοροι παράγοντες για τους οποίους ένα καρδιακό επεισόδιο μπορεί να προκληθεί, όπως το κάπνισμα , χοληστερόλη, διαβήτης και άλλα. Επομένως, με την βοήθεια του εργαλείου εξόρυξης δεδομένων, θα βρούμε ποίοι είναι οι πιο σημαντικοί παράγοντες που προκαλούν κάποιο καρδιακό επεισόδιο, αλλά και σε πιο βαθμό που εξαρτάται από αυτό.

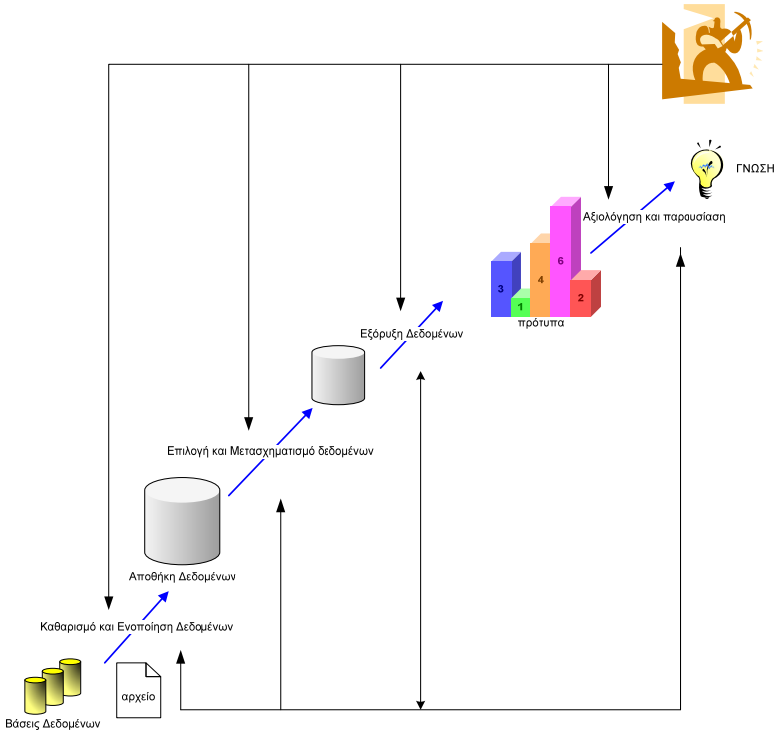
1.3 Εξόρυξη Δεδομένων (Data Mining)

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

Για την εξόρυξη δεδομένων από μια τεράστια βάση δεδομένων, θα πρέπει να ακολουθηθούν κάποια βήματα, που είναι ευρέως γνωστά με τον όρο, *Ανακάλυψη Γνώσης από Βάση Δεδομένων (Knowledge Discovery from Data ή KDD)*, και θα προετοιμάσουν τη βάση δεδομένων για την εξόρυξη χρήσιμης γνώσης. Η ανακάλυψη γνώσης είναι η διαδικασία που παρουσιάζεται στην Εικόνα 1 και αποτελείται από μια επαναληπτική σειρά των ακόλουθων βημάτων:

1. *Καθαρισμός Δεδομένων (Data Cleaning)*: Αφαίρεση θορύβου, δεδομένων με ελλιπείς τιμές και ασυνεπή δεδομένων.
2. *Ολοκλήρωση Δεδομένων (Data Intergation)*: Διάφορες πηγές δεδομένων μπορούν να ενωθούν μαζί.
3. *Επιλογή Δεδομένων (Data Selection)*: Ανακτώνται δεδομένα από τη βάση δεδομένων, που θεωρούνται χρήσιμα, σχετικά με το στόχο ανάλυσης
4. *Μετασχηματισμός Δεδομένων (Data Transformation)*: Τα δεδομένα μετασχηματίζονται ή μορφοποιούνται σε κατάλληλες μορφές για την εξόρυξη δεδομένων με την εκτέλεση διάφορων διαδικασιών όπως μέθοδος συνάθροισης (aggregation), σύνοψης (summary).
5. *Εξόρυξη Δεδομένων (Data Mining)*: Μια ουσιαστική διαδικασία κατά την οποία εφαρμόζονται ευφυείς μέθοδοι, αλγόριθμοι, έτσι ώστε να εξαχθούν πρότυπα δεδομένων.

6. *Αξιολόγηση Προτύπων (Pattern Evaluation)*: Για να καθορίσει τα πραγματικά ενδιαφέρων πρότυπα που να απεικονίζουν την γνώση με βάση διάφορων μέτρων.
7. *Παρουσίαση Γνώσης (Knowledge Representation)*: Χρησιμοποιούνται διάφορες τεχνικές απεικόνισης και παρουσίασης γνώσης για να παρουσιάσουν στο χρήστη την γνώση που εξάχθηκε. [3]



Εικόνα 1: Τα βήματα της Διαδικασίας KDD [3]

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

Για εξόρυξη δεδομένων έχουν προταθεί διάφορες τεχνικές ανάλυσης δεδομένων, όπως η εξόρυξη συχνών προτύπων (frequent pattern), κατηγοριοποίηση (classification), η

συσταδοποίηση (clustering), και εξόρυξη outlier. Η *κατηγοριοποίηση* (*classification*) είναι η διαδικασία για να βρεθούν μοντέλα – λειτουργίες που να περιγράφουν και να διακρίνουν τις κλάσεις των δεδομένων, με σκοπό να χρησιμοποιηθούν τα μοντέλα για να προβλέψουν την κλάση των κάποιου αντικειμένου που η κλάση του δεν είναι γνωστή.

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

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

Τα πιο *συχνά πρότυπα (frequent pattern)*, είναι τα πρότυπο που εμφανίζονται συχνά στα δεδομένα. Υπάρχουν διάφορα είδη συχνών προτύπων, τα συχνά σύνολα αντικειμένων (frequent itemset), τα ακολουθιακά πρότυπα (sequential patterns) και οι υποδομές (substructure). Το συχνό σύνολο αντικειμένων (frequent itemset) αναφέρεται σε ένα σύνολο από αντικείμενα που εμφανίζονται συχνά μαζί σε ένα σύνολο δεδομένων δοσοληψίας, παραδείγματος χάρη κάποιος που αγοράζει γάλα αγοράζει και ψωμί. Η συχνά εμφανιζόμενη ακολουθία, όπως το πρότυπο όπου ο πελάτης τείνει να αγοράζει πρώτα έναν υπολογιστή, μετά μία ψηφιακή φωτογραφική, και μετά μια κάρτα μνήμης, είναι ένα (συχνό) ακολουθιακό πρότυπο (sequential pattern). Μία υποδομή μπορεί να αναφερθεί σε διάφορες μορφές δομών, όπως γράφους, δέντρα που μπορούν να συνδυαστούν με τα σύνολα αντικειμένων ή με ακολουθίες. Εάν μια υποδομή εμφανίζεται συχνά, τότε ονομάζεται (συχνό) πρότυπο δομής. Η εξόρυξη συχνών προτύπων οδηγεί στην ανακάλυψη ενδιαφέρων σχέσεων και συσχετίσεων στα δεδομένα. [3]

1.4 Κανόνες Συσχέτισης (Association Rules)

Η εξόρυξη συχνών προτύπων οδηγεί στην ανακάλυψη ενδιαφέρων σχέσεων και συσχετίσεων στα δεδομένα. Συνεχώς, τεράστιες ποσότητες δεδομένων συλλέγονται και αποθηκεύονται, κι οι κανόνες συσχέτισης παρέχουν ένα συνοπτικό τρόπο για να εκφραστούν χρήσιμες πληροφορίες, που γίνονται εύκολα κατανοητές από τους χρήστες. Έτσι πολλές βιομηχανίες ενδιαφέρονται να εξαγάγουν κανόνες συσχέτισης από τις βάσεις δεδομένων τους. Η ανακάλυψη ενδιαφέρον συσχετίσεων μεταξύ τεραστίων ποσών από εγγραφές επιχειρησιακών δοσοληψιών μπορούν να βοηθήσουν σε πολλές διαδικασίες λήψης επιχειρησιακών αποφάσεων, όπως σχεδιασμό καταλόγου, ανάλυση της συμπεριφοράς αγορών των πελατών. [4]

Ένα τυπικό παράδειγμα είναι για εξόρυξη συχνών συνόλων αντικειμένων είναι η ανάλυση των καλαθιών αγοράς. Αυτή η διαδικασία αναλύει τις συνήθειες των αγορών των πελατών με το να βρίσκει τις συσχετίσεις μεταξύ διαφόρων αντικειμένων που οι πελάτες τοποθετούν στο καλάθι αγορών τους. Η ανακάλυψη συσχετίσεων μπορεί να βοηθήσει τον λιανοπωλητή να αναπτύξει εμπορικές στρατηγικές με το να γνωρίζει ποία αντικείμενα αγοράζονται συχνά μαζί από τους πελάτες. Για παράδειγμα, σε μια επίσκεψη στην υπεραγορά, εάν οι πελάτες αγοράζουν γάλα, πόσο πιθανό είναι να αγοράσουν και ψωμί; Τέτοιες πληροφορίες μπορούν να βοηθήσουν τους λιανοπωλητές να κάνουν εκλεκτικό μάρκετινγκ και να τοποθετήσουν σωστά τα αντικείμενα στα ράφια κι επομένως να αυξήσουν τις πωλήσεις τους.

Εάν θεωρήσουμε ότι όλα τα διαθέσιμα αντικείμενα στο κατάστημα έχουν μία δυαδική μεταβλητή που αντιπροσωπεύει την ύπαρξη ή όχι του αντικειμένου. Τότε κάθε καλάθι μπορεί να αντιπροσωπεύεται από ένα δυαδικό άνυσμα και να θέτει τιμές σε εκείνες τις μεταβλητές. Το δυαδικό άνυσμα μπορεί να αναλυθεί για πρότυπα αγοράς που εκφράζουν αντικείμενα που συχνά συσχετίζονται ή αγοράζονται μαζί. Αυτά τα πρότυπα μπορούν αν απεικονιστούν με τους κανόνες συσχέτισης (association rules). Για παράδειγμα ο κανόνας $\gamma\acute{\alpha}\lambda\alpha \Rightarrow \psi\omega\mu\acute{\imath}$

[support 2%, confidence 60%] αντιπροσωπεύει τους πελάτες που αγοράζουν γάλα και τείνουν επίσης να αγοράζουν και ψωμί.

Η υποστήριξη (support) και η εμπιστοσύνη (confidence) είναι μετρητές που καθαρίζουν κατά πόσο ένας κανόνας είναι ενδιαφέρον και χρησιμοποιούνται για αξιολόγηση των κανόνων. Απεικονίζουν αντίστοιχα τη χρησιμότητα και τη βεβαιότητα των εξαγόμενων κανόνων. Το 2% support (Εξίσωση 1) του πιο πάνω κανόνα συσχέτισης σημαίνει ότι το 2% όλων των πελατών της βάσης δεδομένων, αγοράζουν γάλα και ψωμί μαζί. Το confidence 60% σημαίνει ότι το 60% των πελατών που αγοράζουν γάλα, αγοράζουν και ψωμί. Ένας κανόνας συσχέτισης, θεωρείται ενδιαφέρων εάν ικανοποιεί το ελάχιστο κατώτατο όριο support και το ελάχιστο κατώτατο όριο confidence. Τα ελάχιστα κατώτατα όρια μπορούν να καθοριστούν από του χρήστες.[3]

1.5 Μέτρα Αξιολόγησης Κανόνων

Η αξιολόγηση των κανόνων και η μέτρηση του πόσο ενδιαφέρον είναι οι παραγόμενοι κανόνες είναι ένας ενεργός και σημαντικός τομέας στην έρευνα εξόρυξης δεδομένων. Μέχρι τώρα δεν υπάρχει καμία διαδεδομένη συμφωνία για κανένα επίσημο καθορισμό για αξιολόγηση των κανόνων. Με βάση την ποικιλομορφία των ορισμών που παρουσιάστηκαν μέχρι σήμερα, η αξιολόγηση μπορεί να οριστεί καλύτερα ως μια ευρεία έννοια που δίνει έμφαση στην περιεκτικότητα (conciseness), κάλυψη (coverage), αξιοπιστία (reliability), ιδιαιτερότητα (peculiarity), ποικιλομορφία (diversity), καινοτομία (novelty), πόσο απροσδόκητα (surprisingness), οφέλιμότητα (utility), και δυνατότητα εφαρμογής (actionability). Αυτά τα εννέα ειδικά κριτήρια χρησιμοποιούνται για να καθορίσουν εάν ένας κανόνας είναι ή όχι ενδιαφέρον. [5]

Περιεκτικότητα (Conciseness). Ένας κανόνας μπορεί να θεωρηθεί περιεκτικός εάν περιέχει σχετικά λίγα χαρακτηριστικά (attributes) - ζευγάρια τιμών, ενώ ένα σύνολο κανόνων

είναι περιεκτικό εάν περιέχει σχετικά λίγους κανόνες. Ένα περιεκτικός κανόνας ή ένα περιεκτικό σύνολο κανόνων μπορεί να γίνει εύκολα κατανοητό και να μπορεί να συγκρατηθεί στη μνήμη και έτσι προστίθεται ευκολότερα στη γνώση του χρήστη. Ένα μέτρο που μπορεί να ελαχιστοποιήσει το σύνολο των κανόνων είναι το confidence [6].

Γενικότητα/κάλυψη (Generality/Coverage). Ένας σύνολο κανόνων είναι γενικό (general) εάν καλύπτει ένα σχετικά μεγάλο υποσύνολο ενός συνόλου δεδομένων. Η γενικότητα (ή κάλυψη) μετρά την περιεκτικότητα ενός κανόνα, δηλαδή εάν καλύπτει όλη τη γκάμα, όλο το σύνολο των δεδομένων. Εάν ένα πρότυπο χαρακτηρίζει περισσότερες πληροφορίες στο σύνολο δεδομένων, τείνει να είναι πιο ενδιαφέρον. Τα συχνά σύνολα αντικειμένων (frequent itemsets) είναι τα πιο μελετημένα γενικά πρότυπα στη βιβλιογραφία της εξόρυξης δεδομένων. Ένα σύνολο αντικειμένων είναι συχνό εάν η υποστήριξή (support) του, το φράγμα των εγγραφών στο σύνολο δεδομένων που περιέχει το itemset, είναι πάνω από ένα δεδομένο κατώτατο όριο.

Αξιοπιστία (reliability). Ένας κανόνας είναι αξιόπιστος εάν η σχέση που περιγράφεται από τον κανόνα εμφανίζεται σε ένα υψηλό ποσοστό των εφαρμόσιμων περιπτώσεων. Παραδείγματος χάριν, ένας κανόνας συσχέτισης είναι αξιόπιστος εάν έχει υψηλή εμπιστοσύνη (confidence). Έχουν προταθεί πολλά μέτρα για εύρεση αξιοπιστίας των κανόνων συσχέτισης από διάφορους κλάδους όπως από τις πιθανότητες, τις στατιστική, ανάκτηση πληροφοριών.

Ιδιαιτερότητα (Peculiarity). Ένας κανόνας είναι ιδιαίτερος (peculiar) εάν είναι πολύ διαφορετικός από τους άλλους παραγόμενους κανόνες σύμφωνα με κάποιο κριτήριο απόστασης (distance). Οι ιδιαίτεροι κανόνες παράγονται από τα ιδιαίτερα δεδομένα (ή outliers), τα οποία είναι σχετικά λίγα σε αριθμό και σημαντικά διαφορετικά από τα υπόλοιπα δεδομένα [7]. Τα ιδιαίτερα πρότυπα μπορεί να είναι άγνωστα στο χρήστη, κι επομένως ενδιαφέρων.

Ποικιλομορφία (Diversity). Ένας κανόνας είναι ποικιλόμορφος εάν τα στοιχεία του διαφέρουν σημαντικά το ένα από το άλλο, ενώ ένα σύνολο κανόνων είναι ποικιλόμορφο εάν οι κανόνες του διαφέρουν σημαντικά ο ένας από τον άλλο. Η ποικιλομορφία είναι ένας κοινός

παράγοντας για τη μέτρηση του πόσο σημαντικές είναι οι περιλήψεις [8]. Σύμφωνα με απλή άποψη, μια περίληψη μπορεί να θεωρηθεί ποικιλόμορφη εάν η διανομή πιθανότητάς της είναι πολύ διαφορετική από την ομοιόμορφη διανομή. Μια ποικιλόμορφη περίληψη μπορεί να είναι ενδιαφέρουσα επειδή με την απουσία οποιασδήποτε σχετικής γνώσης, ένας χρήστης συνήθως υποθέτει ότι η ομοιόμορφη διανομή θα κρατήσει σε μια περίληψη. Σύμφωνα με αυτόν τον συλλογισμό, η πιο ποικιλόμορφη περίληψη είναι κι η πιο ενδιαφέρον.

Καινοτομία (Novelty). Ένας κανόνας είναι καινοφανής εάν δεν ήταν γνωστό πριν και δεν ήταν δυνατό να βγει ως συμπέρασμα από άλλους γνωστούς κανόνες. Κανένα σύστημα εξόρυξης δεδομένων δεν αντιπροσωπεύει όλα όσα ένας χρήστης ξέρει, και έτσι, η καινοτομία δεν μπορεί να μετρηθεί ρητά σε σχέση με τη γνώση του χρήστη. Ομοίως, κανένα γνωστό σύστημα εξόρυξης δεδομένων δεν αντιπροσωπεύει ότι ο χρήστης δεν ξέρει, και επομένως, η καινοτομία δεν μπορεί να μετρηθεί ρητά σε σχέση με την άγνοια του χρήστη. Αντί αυτού, η καινοτομία μπορεί να ανιχνευτεί από τον χρήστη, είτε ρητά να προσδιορίζει ένα κανόνα ως νέο, είτε παρατηρεί ότι ένας κανόνας δεν έρχεται σε αντιπαράθεση με προηγούμενους ανακαλυμμένους κανόνες. Στην τελευταία περίπτωση, τα ανακαλυμμένα πρότυπα χρησιμοποιούνται ως προσέγγιση στη γνώση του χρήστη.

Απροσδόκητοι (Surprisingness). Ένας κανόνας είναι έκπληξη (ή απροσδόκητος) εάν έρχεται σε αντίθεση με την υπάρχουσα γνώση ή τις προσδοκίες του χρήστη. Ένας κανόνας, που είναι εξαίρεση σε ένα πρότυπο που έχει ανακαλυφθεί ήδη, μπορεί επίσης να θεωρηθεί απροσδόκητος. Οι απροσδόκητοι κανόνες είναι επίσης ενδιαφέρον επειδή προσδιορίζουν τις αποτυχίες στην προηγούμενη γνώση και μπορούν να προτείνουν μια πτυχή δεδομένων που χρειάζεται περαιτέρω μελέτη. Η διαφορά μεταξύ του απροσδόκητου και της καινοτομίας είναι ότι ένας κανόνας είναι καινούργιος και δεν έρχεται σε αντίφαση με οποιοδήποτε άλλο κανόνα που ήταν ήδη γνωστός στο χρήστη, ενώ ένας απροσδόκητος κανόνας έρχεται σε αντίφαση με την προηγούμενη γνώση ή τις προσδοκίες του χρήστη.

Ωφελιμότητα (Utility). Ένας κανόνας είναι ωφέλιμος εάν η χρήση του, από τον χρήστη, συμβάλλει στην επίτευξη ενός στόχου. Διαφορετικοί χρήστες μπορεί να έχουν διαφορετικούς στόχους σχετικά με τη γνώση που μπορεί να εξαχθεί από ένα σύνολο δεδομένων.

Παραδείγματος χάριν, ένας χρήστης μπορεί να ενδιαφερθεί για την εύρεση όλων των πωλήσεων με ψηλό κέρδος από ένα σύνολο δεδομένων, ενώ άλλος μπορεί να ενδιαφέρεται για την εύρεση όλων το δοσοληψιών με μεγάλες αυξήσεις στις ακαθάριστες πωλήσεις. Αυτό το είδος ενδιαφέροντος είναι βασισμένο στις λειτουργίες χρησιμότητας που καθορίζονται από το χρήστη.

Δυνατότητα εφαρμογής(Actionability/ Applicability). Ένας κανόνας είναι εφαρμόσιμος σε κάποια περιοχή εάν επιτρέπει τη λήψη απόφασης για μελλοντικές ενέργειες σε αυτήν την περιοχή.

Αυτά τα εννέα κριτήρια μπορούν να ταξινομηθούν περαιτέρω σε τρεις κατηγορίες: αντικειμενικά, υποκειμενικά, και βασισμένα στη σημασιολογία. Ένα *αντικειμενικό* μέτρο είναι βασισμένο μόνο στα ακατέργαστα δεδομένα. Δεν απαιτείται καμία γνώση για το χρήστη ή την εφαρμογή. Τα περισσότερα αντικειμενικά μέτρα είναι βασισμένα στις θεωρίες πιθανοτήτων, τις στατιστικές, ή τη θεωρία πληροφοριών. Η περιεκτικότητα, η γενικότητα, η αξιοπιστία, η ιδιαιτερότητα, και η ποικιλομορφία εξαρτώνται μόνο από τα δεδομένα και τους κανόνες, και μπορούν έτσι να θεωρηθούν αντικειμενικά.

Ένα *υποκειμενικό* μέτρο λαμβάνει υπόψη και τα δεδομένα και το χρήστη των δεδομένων. Για να καθοριστεί ένα υποκειμενικό μέτρο, απαιτείται πρόσβαση στην περιοχή του χρήστη ή στο υπόβαθρο της γνώσης για τα δεδομένα. Αυτή η πρόσβαση μπορεί να ληφθεί με την αλληλεπίδραση με το χρήστη κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων ή με το να αντιπροσωπεύσει ρητά τη γνώση ή τις προσδοκίες του χρήστη. Η καινοτομία και τα απροσδόκητα πρότυπα εξαρτώνται από το χρήστη, καθώς επίσης και από τα δεδομένα και τους κανόνες, και ως εκ τούτου αυτά τα κριτήρια μπορούν να θεωρηθούν υποκειμενικά.

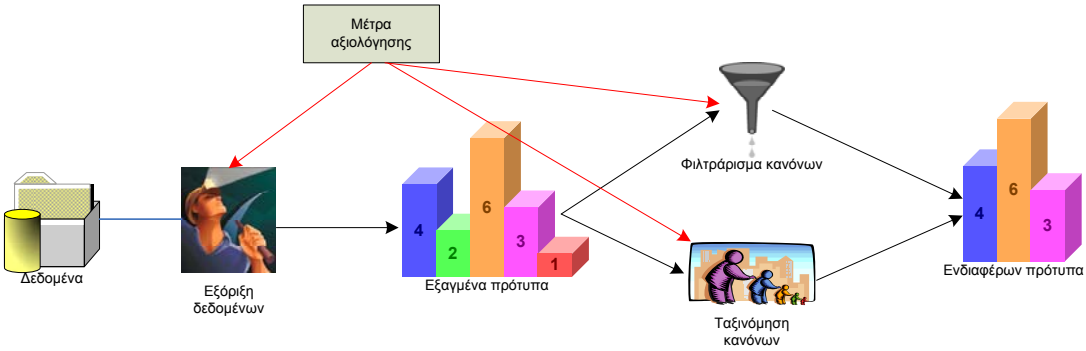
Ένα *σημασιολογικό* μέτρο εξετάζει τη σημασιολογία και τις εξηγήσεις των κανόνων. Επειδή τα σημασιολογικά μέτρα περιλαμβάνουν τη γνώση από το χρήστη, μερικοί ερευνητές τους θεωρούν σαν έναν ειδικό τύπο υποκειμενικού μέτρου. Η ωφελιμότητα και η δυνατότητα εφαρμογής εξαρτώνται από τη σημασιολογία των δεδομένων, και έτσι μπορούν να θεωρηθούν σημασιολογικά κριτήρια. Τα μέτρα που είναι βασισμένα στην ωφελιμότητα, όπου η σχετική σημασιολογία είναι οι χρησιμότητες των κανόνων, είναι ο πιο κοινός τύπος

σημασιολογικού μέτρου. Για να χρησιμοποιηθεί μια προσέγγιση βασισμένη στη χρησιμότητα, ο χρήστης πρέπει να διευκρινίσει πρόσθετη γνώση για την περιοχή. Αντίθετα από τα υποκειμενικά μέτρα, όπου η γνώση είναι για τα ίδια τα δεδομένα και αντιπροσωπεύεται συνήθως με ένα σχήμα παρόμοιο με αυτό του ανακαλυμμένου προτύπου, η γνώση που απαιτείται για τα σημασιολογικά μέτρα δεν σχετίζει τη γνώση ή τις προσδοκίες του χρήστη με τα δεδομένα. Αντί αυτού, αντιπροσωπεύει μια λειτουργία χρησιμότητας που απεικονίζει τους στόχους του χρήστη. Παραδείγματος χάριν, ένας διευθυντής καταστημάτων να προτιμήσει τους κανόνες συσχέτισης που αφορούν τα δεδομένα με ψηλό κέρδος παρά από εκείνους με την υψηλότερη στατιστική σημασία.

Για τον προσδιορισμό του πόσο ενδιαφέρον είναι ένας κανόνας (interestingness determination), υπάρχουν τρεις μέθοδοι εκτέλεσης. Κατ' αρχάς, μπορούμε να ταξινομήσουμε κάθε κανόνα από το εάν είναι ενδιαφέρων είτε όχι. Αφετέρου, μπορούμε να καθορίσουμε μια σχέση προτίμησης που καθορίζει εάν ένας κανόνας είναι πιο ενδιαφέρον από τον άλλο. Τρίτον, να βαθμολογήσουμε τους κανόνες. Για την πρώτη ή τρίτη προσέγγιση, μπορούμε να καθορίσουμε ένα μέτρο βασισμένο στα προαναφερθέντα εννέα κριτήρια και να το χρησιμοποιήσουμε για να ξεχωρίσουμε τους ενδιαφέρον και τους μη-ενδιαφέρον κανόνες στην πρώτη προσέγγιση ή για να βαθμολογήσουμε τα πρότυπα στην τρίτη προσέγγιση.

Κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων, τα μέτρα αξιολόγησης του πόσο ενδιαφέρον είναι οι κανόνες, μπορούν να χρησιμοποιηθούν με τρεις τρόπους, τους οποίους καλούμε ρόλους των μέτρων. Το Εικόνα 2 παρουσιάζει αυτούς τους τρεις ρόλους. Κατ' αρχάς, τα μέτρα μπορούν να χρησιμοποιηθούν για να κλαδέψουν τους κανόνες, που δεν είναι ενδιαφέρον, κατά τη διάρκεια της διαδικασίας εξόρυξης ώστε να ελαχιστοποιηθεί ο χώρος αναζήτησης και να βελτιωθεί έτσι η αποδοτικότητα της εξόρυξης. Παραδείγματος χάριν, να χρησιμοποιηθεί το κατώτατο όριο για το support για να φιλτράρει τους κανόνες, και να αφαιρέσει αυτούς που δεν το ικανοποιούν κατά τη διάρκεια της διαδικασίας εξόρυξης, και να βελτιωθεί έτσι η αποδοτικότητα [9]. Μπορεί επίσης να καθοριστεί ένα κατώτατο όριο για μέτρα βασισμένα στην χρησιμότητα, για περικοπή κανόνων σύμφωνα με τη σειρά των

αποτελεσμάτων. Τρίτον, τα μέτρα μπορούν να χρησιμοποιηθούν επίσης κατά τη διάρκεια επιλογής ενδιαφέρον κανόνων. [5]



Εικόνα 2: Ρόλοι των μέτρων αξιολόγησης κανόνων στη διαδικασία εξόρυξης δεδομένων.[5]

1.6 Δομή Μελέτης

Κεφάλαιο 2: θα γίνει περιγραφή των αλγόριθμων που έχουν υλοποιηθεί, Apriori και Akama και των μέτρων αξιολόγησης των κανόνων που χρησιμοποιήθηκαν.

Κεφάλαιο 3: Παρουσιάζονται τα χαρακτηριστικά της βάσης δεδομένων, και η επεξεργασία που έχει γίνει στα δεδομένων, πριν εφαρμοστούν οι αλγόριθμοι εξόρυξης κανόνων συσχέτισης.

Κεφάλαιο 4: Παρουσιάζονται τα αποτελέσματα εξόρυξης γνώσης.

Κεφάλαιο 5: Στο κεφάλαιο αυτό παρουσιάζονται τα συμπεράσματα μέσα από αυτή την έρευνα και γίνονται κάποιες εισηγήσεις για μελλοντικές μελέτες.

Κεφάλαιο 2

Περιγραφή εργαλείου και αλγορίθμων

2.1 Αλγόριθμος Apriori

Ο αλγόριθμος Apriori έχει προταθεί από τους R. Agrawal R. Srikant το 1994 [9]. 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 υιοθετά την τεχνική αναζήτηση, level-wise, η οποία είναι μια επαναλαμβανόμενη τεχνική που χρησιμοποιεί τα k-itemesets για να κτίσει τα (k+1)-itemsets.

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

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

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

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

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

Διαδικασία Κλαδέματος (prune): Κάποια από τα σύνολα αντικειμένων που ανήκουν στο Ck, μπορεί να είναι συχνά εμφανιζόμενα κι άλλα όχι, όμως όλα τα συχνά εμφανιζόμενα σύνολα k αντικειμένων (k-itemsets) συμπεριλαμβάνονται στο Ck. Θα πρέπει να γίνει μία αναζήτηση στη βάση δεδομένων για να μετρηθεί ο αριθμός όπου κάθε υποψήφιο σύνολο στο Ck, εμφανίζεται στη βάση δεδομένων. Όλα τα σύνολα αντικειμένων που περιλαμβάνονται στο Ck, και εμφανίζονται στη βάση δεδομένων όχι λιγότερό αριθμό από το ελάχιστο support, τότε αυτό το σύνολο αντικειμένων προστίθεται στο Lk. Αυτό γίνεται, όπως αναφέρει η Apriori ιδιότητα, οποιοδήποτε (k-1)-itemset σύνολο αντικειμένων δεν είναι συχνό τότε δεν

μπορεί να είναι υποσύνολο κάποιου k-itemset σύνολο αντικειμένων. Έτσι επειδή το σύνολο Ck, μπορεί να γίνει αρκετά μεγάλο, τα σύνολα αντικειμένων που δεν είναι συχνά αφαιρούνται.

2.1.1 Περιγραφή Ψευδοκώδικα Αλγόριθμου Apriori

Στην Εικόνα 3 παρουσιάζεται ο ψευδοκώδικας του αλγόριθμου Apriori και οι σχετικές διαδικασίες:

1. 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 όπου συνθέτουν την δοσοληψία.
2. Στο πρώτο βήμα βρίσκονται όλα τα συχνά σύνολα αντικειμένων με 1 χαρακτηριστικό και φυλάγονται στο σύνολο L1
3. Στο βήμα 3 είναι η διαδικασία όπου το σύνολο υποψηφίων Ck παράγεται από την ένωση του Lk-1 με τον εαυτό του. Η διαδικασία apriori_gen παράγει τα υποψήφια σύνολα αντικειμένων και μετά χρησιμοποιεί την ιδιότητα Apriori για να αφαιρέσει τα αυτά που δεν είναι συχνά.
4. Στο βήμα 4 – 10 γίνεται μία αναζήτηση στη βάση δεδομένων, για να βρεθεί ο αριθμός που τα σύνολα αντικειμένων εμφανίζονται στην βάση. Και στο βήμα 9 βρίσκει τα υποψήφια σύνολα αντικειμένων που έχουν μεγαλύτερο από το ελάχιστο support και τα προσθέτει στο σύνολο Lk.
5. Στο τελικό βήμα 11 γίνεται μια ένωση όλων των συχνών συνόλων αντικειμένων των συνόλων Lk στο L. Έτσι μετά μια διαδικασία για εξαγωγή κανόνων συσχέτισης μπορεί να χρησιμοποιήσει το σύνολο L.

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

Είσοδος:

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

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

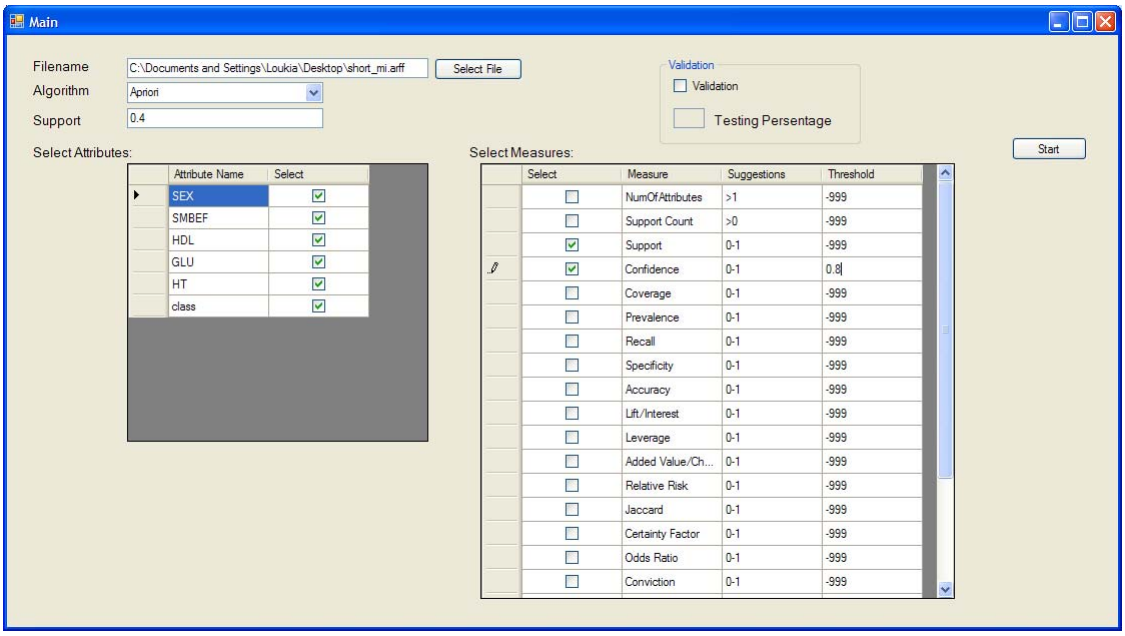
Μέθοδος:

```
(1) L1 =find_frequent_1-itemsets(D);
(2) for(k = 2;Lk-1 ≠ 0;k++){
(3)     Ck = apriori_gen(Lk-1);
(4)     for each transaction t∈D { //scan D for counts
(5)             Ct = subset(Ck,t);  //get the subsets of t that are candidates
(6)             for each candidate c ∈Ct
(7)                     c.count++;
(8)     }
(9)     Lk = {c ∈ Ck| c.count ≥ min-sup}
(10) }
(11) return L = UkLk;

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

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

Εικόνα 3: Ψευδοκώδικας Αλγόριθμου Apriori



Εικόνα 4: Οθόνη εισαγωγής δεδομένων στο εργαλείο

Πίνακας 1: Βάση δεδομένων δοσοληψιών για ασθενείς με καρδιαγγειακά επεισόδια (File: short_mi.arff)

SEX (Φύλο)	SM BEF (Καπνιστής)	HDL (Λιποπρωτεΐνες Υψηλής Πυκνότητας)	GLU (Γλυκόζη)	HT (Υπέρταση)	MI (Στεφανιαία Νόσος)
M	Y	M	N	N	Y
M	Y	M	N	Y	Y
M	Y	L	H	N	Y
M	Y	M	N	N	Y
F	Y	L	H	N	N
M	N	M	N	Y	N
F	N	M	N	Y	Y
M	Y	M	H	Y	N
M	Y	L	N	N	Y
M	Y	L	N	N	Y
M	Y	M	H	N	Y
M	Y	M	H	N	Y
M	Y	M	N	N	Y
M	Y	M	N	Y	Y
M	Y	M	H	N	Y
M	Y	H	H	N	Y
M	Y	L	N	N	N
M	Y	M	N	N	Y
M	Y	L	H	N	Y
M	Y	M	N	Y	Y
M	N	L	N	N	Y
M	N	M	N	N	Y
F	Y	M	N	N	Y
M	Y	M	N	N	Y
M	Y	L	N	Y	N
M	Y	H	N	N	N
F	N	H	N	Y	N
M	Y	M	N	Y	N

2.1.2 Παράδειγμα Εκτέλεσης αλγόριθμου Apriori

Έχοντας μία μικρή βάση δεδομένων με ασθενείς με έμφραγμα του μυοκαρδίου, που παρουσιάζονται στο Πίνακα 1, θα εφαρμόσουμε τον αλγόριθμο Apriori για να βρούμε τα συχνά σύνολα αντικειμένων. Θεωρούμε το ελάχιστο support 0,4 (40%), που αυτό σημαίνει ότι θα πρέπει ένα σύνολο αντικειμένων να εμφανίζεται τουλάχιστο 12 φορές (Στην Εικόνα 4 παρουσιάζεται η οθόνη του συστήματος για την εισαγωγή δεδομένων από το χρήστη, όπως το αρχείο με τα βάση δεδομένων και τα μέτρα αξιόλογης όπως support και confidence)

Βήμα 1: Καταρχάς γίνεται μια αναζήτηση στη βάση δεδομένων για να βρεθούν όλα τα σύνολα αντικειμένων με 1 χαρακτηριστικό και πόσες φορές αυτά εμφανίζονται στην βάση δεδομένων. Όλα τα σύνολα αντικειμένων αποθηκεύονται στο σύνολο C1. Σημειώνουμε ότι, το χαρακτηριστικό κλάσης δεν εισάγεται στα υποψήφια σύνολα αντικειμένων, θα προστεθεί στο τέλος της διαδικασίας για να βεβαιωθεί ότι δεν θα εξαχθούν κανόνες συσχέτισης που να μην συμπεριλαμβάνουν το χαρακτηριστικό για την κλάση. Τα αποτελέσματα του συνόλου C1 παρουσιάζονται στο πιο κάτω Πίνακα 2.

Πίνακας 2: Παραγόμενο σύνολο C1 (Παράδειγμα αλγόριθμου Apriori)

C1			
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M	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 αντικειμένων, που όταν ενωθούν και με κάποιο αντικείμενο της κλάσης έχουν support μεγαλύτερο από το ελάχιστο support και τα αποθηκεύουμε στο σύνολο L1. Δηλαδή θα αφαιρεθούν τα σύνολα αντικειμένων που εμφανίζονται λιγότερο από 12 φορές. Τα αποτελέσματα του συνόλου L1 παρουσιάζονται στο πιο κάτω Πίνακα 3.

Πίνακας 3: Παραγόμενο σύνολο L1 (Παράδειγμα αλγόριθμου Apriori)

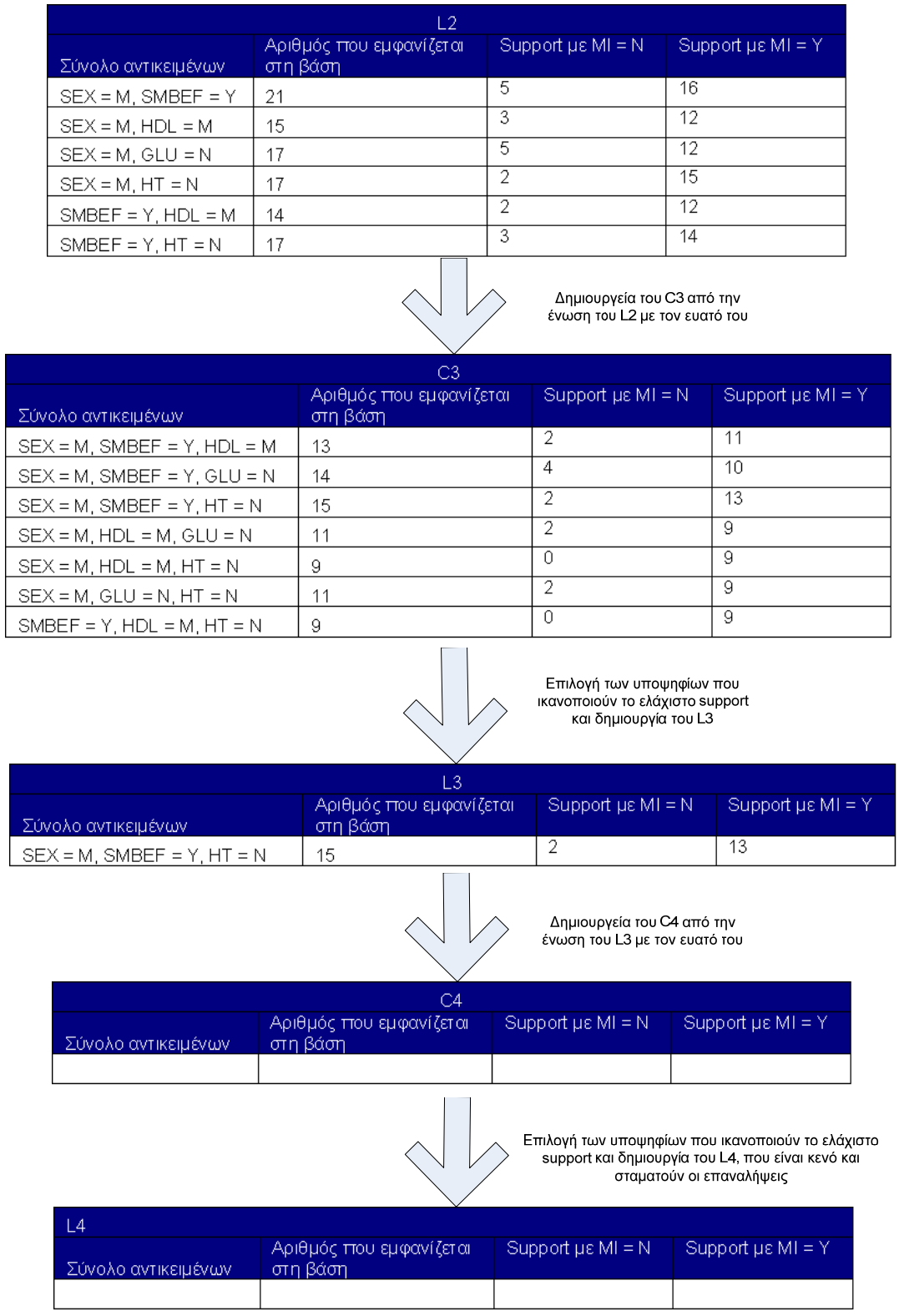
L1			
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M	24	6	18
SMBEF = Y	23	6	17
HDL = M	17	3	14
GLU = N	20	6	14
HT = N	19	3	16

Βήμα 3: Σε αυτό το βήμα θα γίνει η ένωση του L1 με τον εαυτό του για κτιστεί το σύνολο C2. Για κάθε σύνολο αντικειμένων του L1 γίνεται συνένωση με τα υπόλοιπα. Τα αποτελέσματα του συνόλου C2 παρουσιάζονται στο πιο κάτω Πίνακα 4.

Πίνακας 4: Παραγόμενο σύνολο C2 (Παράδειγμα αλγόριθμου Apriori)

C2			
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M, SMBEF = Y	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. Στην Εικόνα 5, παρουσιάζονται τα υπόλοιπα βήματα.



L2

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με 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, HT = N	17	3	14

Δημιουργεία του C3 από την ένωση του L2 με τον ευατό του

C3

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M, SMBEF = Y, HDL = M	13	2	11
SEX = M, SMBEF = Y, GLU = N	14	4	10
SEX = M, SMBEF = Y, HT = N	15	2	13
SEX = M, HDL = M, GLU = N	11	2	9
SEX = M, HDL = M, HT = N	9	0	9
SEX = M, GLU = N, HT = N	11	2	9
SMBEF = Y, HDL = M, HT = N	9	0	9

Επιλογή των υποψηφίων που ικανοποιούν το ελάχιστο support και δημιουργία του L3

L3

Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση	Support με MI = N	Support με MI = Y
SEX = M, SMBEF = Y, HT = N	15	2	13

Δημιουργεία του C4 από την ένωση του L3 με τον ευατό του

C4

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

Επιλογή των υποψηφίων που ικανοποιούν το ελάχιστο support και δημιουργία του L4, που είναι κενό και σταματούν οι επαναλήψεις

L4

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

Εικόνα 5: Παραγόμενα σύνολα αντικειμένων (Παράδειγμα εφαρμογής αλγόριθμου Apriori)

2.1.3 Διαδικασία εξόρυξης κανόνων συσχέτισης από τα εξαγόμενα συχνά σύνολα αντικειμένων

Αφού ο ολοκληρωθεί ο αλγόριθμος Apriori, το σύνολο L, με όλα τα συχνά σύνολα αντικειμένων, αποστέλλονται στην διαδικασία για δημιουργία των κανόνων συσχέτισης. Η διαδικασία αυτή λαμβάνει επίσης, από τον χρήστη, και ένα μέτρο αξιολόγησης των κανόνων. Για κάθε σύνολο αντικείμενων l που ανήκει στο σύνολο L, γίνεται ένωση του με κάποιο χαρακτηριστικό κλάσης, και ελέγχεται εάν το μέτρο ικανοποιείται. Εάν λάβουμε το όλα τα σύνολα L μη κενά σύνολα από το προηγούμενο παράδειγμα τότε μπορούμε να εξάξουμε τους ακόλουθους κανόνες συσχέτισης του Πίνακα 5. Για την αξιολόγηση των κανόνων θα χρησιμοποιήσουμε το μέτρο αξιολόγησης κανόνων, confidence (εμπιστοσύνη). Σε ένα κανόνα A $\rightarrow$ B, το confidence, είναι η πιθανότητα να ισχύει το B νοουμένου ότι ισχύει το A, και υπολογίζεται από την Εξίσωση 2 (Κεφάλαιο 2.3). (Το support_count είναι ο αριθμός εμφάνισης στην βάση δεδομένων) .

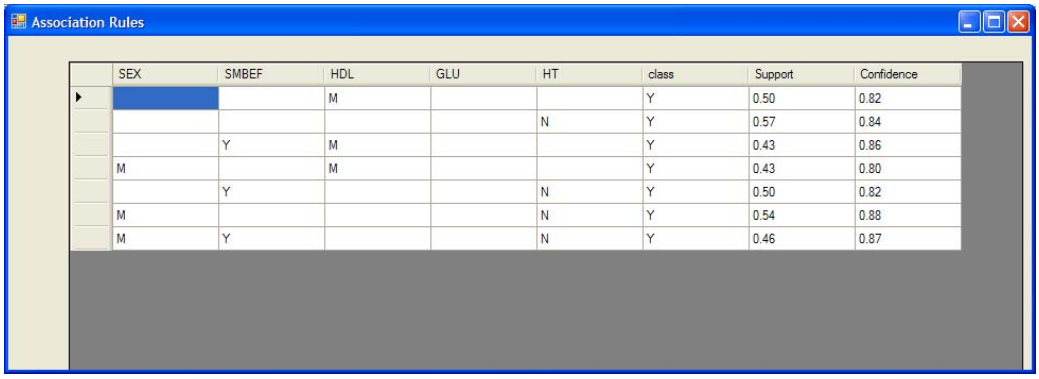
Οι κανόνες που ικανοποιούν το ελάχιστο όριο support και confidence, εξάγονται και παρουσιάζονται στο Πίνακα 6. Ενώ η Εικόνα 6 παρουσιάζει τους κανόνες όπως εξάγονται από το σύστημα.

Πίνακας 5: 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
SEX = M, SMBEF = 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: 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



Εικόνα 6: Παρουσίαση αποτελεσμάτων από το εργαλείο

2.2 Αλγόριθμος Akamas

Ο αλγόριθμος Ακάμας έχει προταθεί και υλοποιηθεί από τον Μηνά Καραολή και Λουκία Παπακωνσταντίνου. Ο αλγόριθμος 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 αντικειμένων με 1 μόνο χαρακτηριστικό). Ο αλγόριθμος αναζητά και συναθροίζει τον αριθμό που εμφανίζεται κάθε αντικείμενο – χαρακτηριστικό στη βάση δεδομένων, και μετά συλλέγει τα αντικείμενα που ικανοποιούν το ελάχιστο support, στο σύνολο L1.

Αφού συλλέξει όλα τα σύνολα αντικειμένων με 1 χαρακτηριστικό τα οποία ικανοποιούν το ελάχιστο όριο support, τότε κάνει όλους τους δυνατούς συνδυασμούς μεταξύ των συνόλων αντικειμένων του συνόλου L1 και χτίζει κανόνες συσχέτισης. Πρώτα για κάθε σύνολο αντικειμένων του L1, γίνεται ένωση του όλα τα υπόλοιπα σύνολα αντικειμένων, και δημιουργούνται έτσι σύνολα αντικειμένων με 2 χαρακτηριστικά (2-itemsets). Στην επόμενη επανάληψη για κάθε σύνολο αντικειμένων του L1 θα γίνει ένωση του με άλλα δύο σύνολα αντικειμένων 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 αντικειμένων με 1 χαρακτηριστικό.

Παράλληλα, ενώ οι δυνατοί συνδυασμοί των συνόλων αντικειμένων δημιουργούνται, κτίζονται παράλληλα και οι κανόνες συσχέτισης. Για κάθε καινούργιο σύνολο αντικειμένων γίνεται η ένωσή του και με ένα από τα χαρακτηριστικά κλάσης, κι έτσι δημιουργείται ένας κανόνας συσχέτισης. Για να γίνει κάποιο φιλτράρισμα 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 του support και του confidence. Επομένως, εάν σε κάποια επανάληψη, εάν 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.

Αλγόριθμος: Akamas. Εύρεση των συνόλων αντικειμένων (itemsets)

Είσοδος:

- D, βάση δεδομένων με δοσοληψίες
- min-sup, το ελάχιστο support.
- min_conf, το ελάχιστο confidence.

Έξοδος: Association_rules, σύνολο με όλους τους δυνατούς κανόνες που ικανοποιούν το ελάχιστο support και confidence.

Μέθοδος:

```
(1) L1 =find_frequent_1-itemsets(D);
(2) for(k = 2;Association_rules k-1 ≠ 0;k++){
(3)     for each itemset l1∈L1 {
(4)             new_rule = l1
(5)             for each itemset l2 ∈ L1 && l2 < l1{
(6)                     new_rule = new_rule U l2
(7)                     if (new_rule.length = k){
(8)                             if(support(new_rule)≥ min_supp &&
                                    confidence(new_rule)≥ min_supp)
(9)                                 Association_rulek = Association_rulek U new_rule
(10)                            new_rule = new_rule – l2
(11)                    }
(12)            }
(13)    }
(14) return Association_rules = Uk Association_rulesk;
```

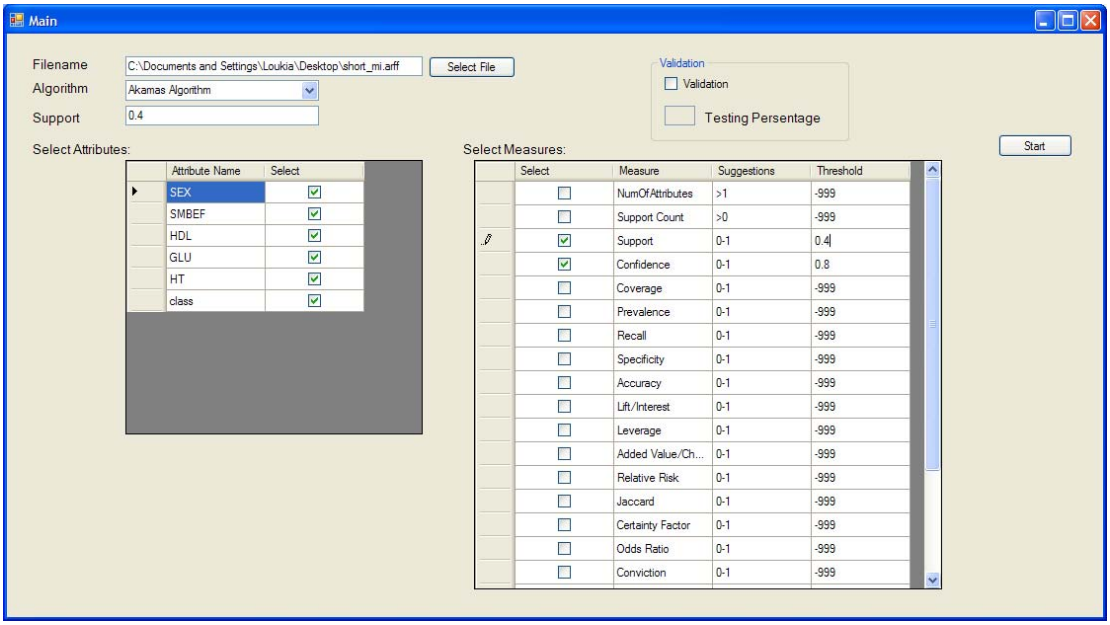
Εικόνα 7: Ψευδοκώδικας Αλγόριθμου Akama

2.2.1 Περιγραφή Ψευδοκώδικα Αλγόριθμου Akama

Στην Εικόνα 7 παρουσιάζεται ο ψευδοκώδικας του αλγόριθμου Akama:

1. Καταρχάς ο αλγόριθμος δέχεται μια βάση δεδομένων με δοσοληψίες. Όπως επίσης δέχεται από τον χρήστη ένα ελάχιστο όριο για support και confidence.
2. Στο πρώτο βήμα βρίσκονται όλα τα συχνά σύνολα αντικειμένων 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 support και confidence.

4. Κάθε σύνολο αντικειμένων που ανήκει στο L1 (Βήμα 3) θα γίνεται η ένωση του με άλλα k-1 σύνολα αντικειμένων 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 αντικειμένων l2 που ανήκει στο L1, και l2 είναι μικρότερου το l1, για να βεβαιωθούμε ότι δεν θα παραχθούν ίδιοι κανόνες (βήμα 5), προστίθεται στον κανόνα (Βήμα 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 το ελάχιστο support και confidence (βήμα 8) και θα προστεθεί στο σύνολο των κανόνων. Και μετά θα αφαιρεθεί το l2 από τον κανόνα για να γίνουν άλλη δυνατοί συνδυασμοί (βήμα 10) .
5. Στο τέλος ο αλγόριθμος θα επιστρέψει τους κανόνες συσχέτισης σε ένα σύστημα για παρουσίαση των κανόνων (Βήμα 14).



Εικόνα 8: Οθόνη εισαγωγή δεδομένων στο σύστημα για τον Αλγόριθμο Akama

2.2.2 Παράδειγμα Εκτέλεσης αλγόριθμου Akama

Έχοντας την βάση δεδομένων του παραδείγματος εκτέλεσης του αλγόριθμου Apriori, Πίνακα 1, θα εφαρμόσουμε τον αλγόριθμο Akama. Θεωρούμε το ελάχιστο support 0,4 (40%) και ελάχιστο confidence 0,8 (Η Εικόνα 8 παρουσιάζεται η οθόνη του συστήματος για την εισαγωγή δεδομένων από το χρήστη, όπως το αρχείο με τα βάση δεδομένων και τα μέτρα αξιόλογης όπως support και confidence).

Βήμα 1: Καταρχάς γίνεται μια αναζήτηση στη βάση δεδομένων για να βρεθούν όλα τα σύνολα αντικειμένων με 1 χαρακτηριστικό και πόσες φορές αυτά εμφανίζονται στην βάση δεδομένων. Όλα τα σύνολα αντικειμένων αποθηκεύονται στο σύνολο C1 (Πίνακας 7).

Πίνακας 7: Σύνολο αντικειμένων C1 (Παράδειγμα εφαρμογής αλγόριθμου Akama)

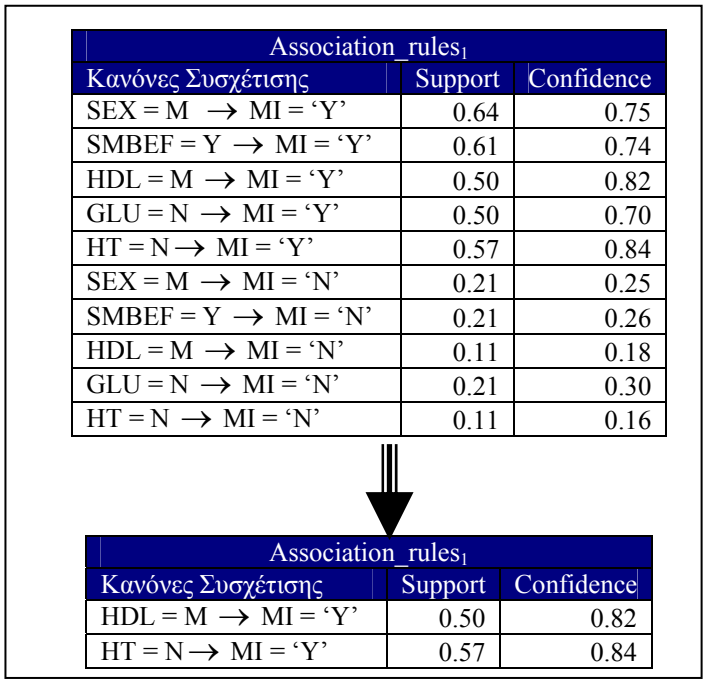
C1	
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση
SEX = M	24
SEX = F	4
SMBEF = N	5
SMBEF = Y	23
HDL = H	3
HDL = L	8
HDL = M	17
GLU = N	20
GLU = H	8
HT = N	19
HT = Y	9

Βήμα 2: Από το C1 επιλέγουμε τα σύνολα αντικειμένων, που όταν ενωθούν και με κάποιο αντικείμενο της κλάσης έχουν support μεγαλύτερο από το ελάχιστο support και τα αποθηκεύουμε στο σύνολο L1 (Πίνακας 8). Δηλαδή θα αφαιρεθούν τα σύνολα αντικειμένων που εμφανίζονται λιγότερο από 12 φορές.

Πίνακας 8: Σύνολο συχνών αντικειμένων L1 (Παράδειγμα εφαρμογής αλγόριθμου Akamas)

L1	
Σύνολο αντικειμένων	Αριθμός που εμφανίζεται στη βάση
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 support και confidence. Στην πιο κάτω εικόνα (Εικόνα 9) παρουσιάζονται οι κανόνες με 1-χαρακτηριστικό που επιλέγονται.

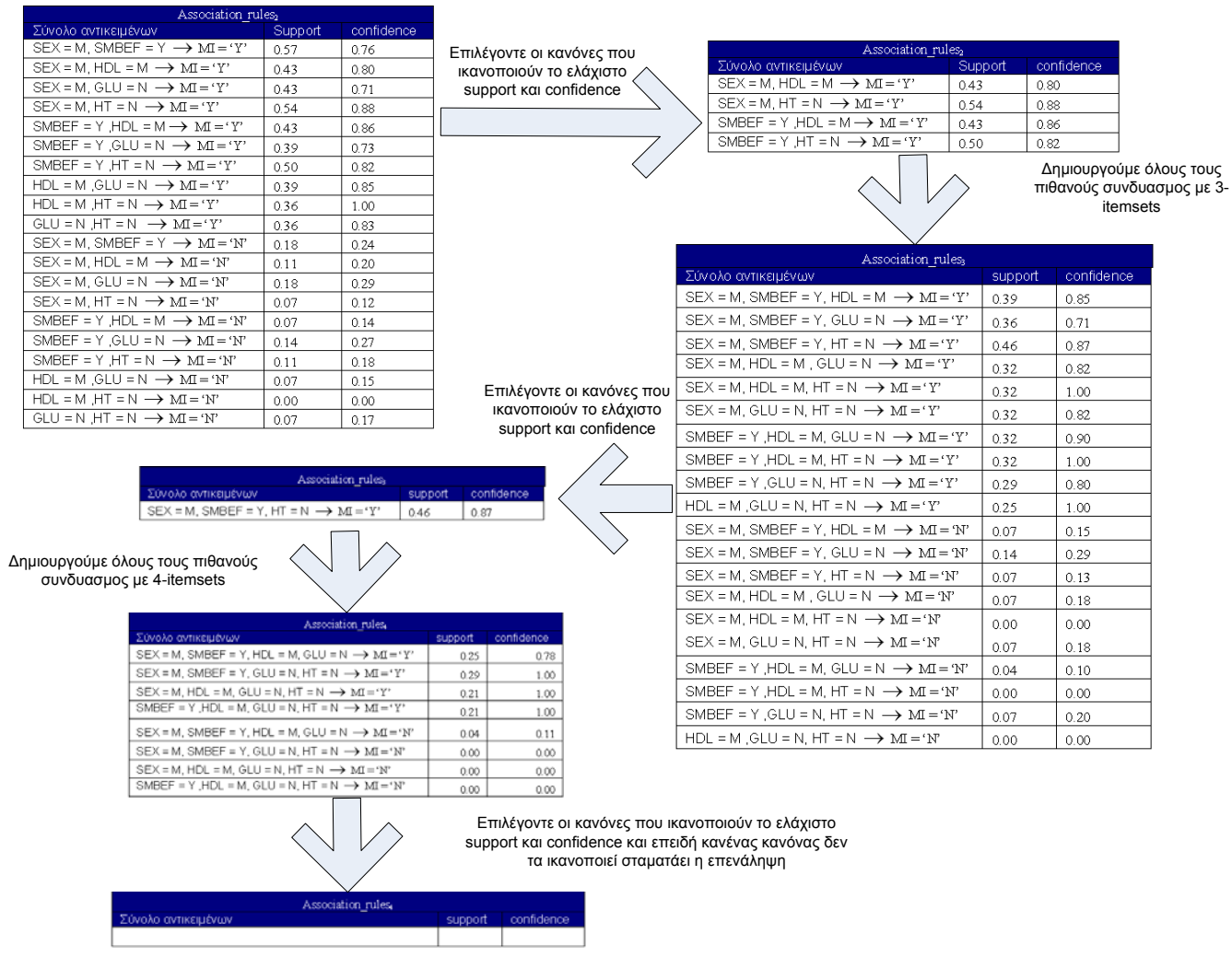


Εικόνα 9: Επιλογή κανόνων με 1-χαρακτηριστικό

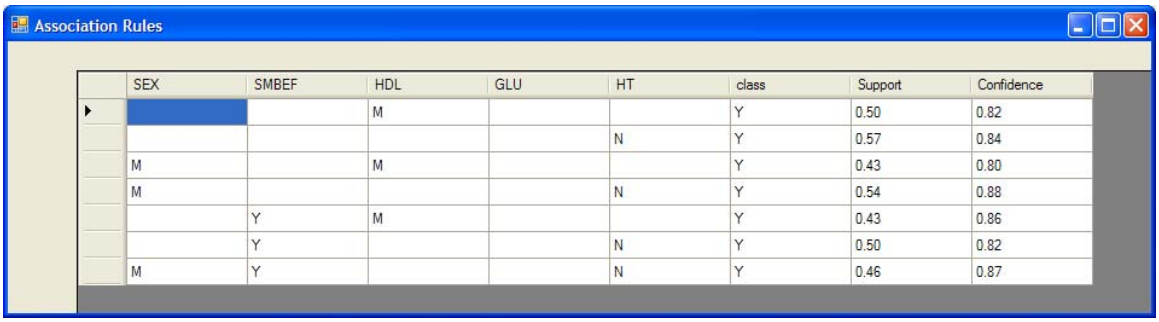
Βήμα 4: Σε αυτό το βήμα θα γίνει η ένωση του κάθε συνόλου αντικειμένων του L1 με όλα τα υπόλοιπα και δημιουργία των κανόνων συσχέτισης με 2-itemsets. Από τους κανόνες συσχέτισης επιλέγονται μόνο αυτοί που ικανοποιούν το ελάχιστο support και confidence. Αυτό το βήμα επαναλαμβάνεται μέχρι να μην μπορούν να

δημιουργηθούν άλλοι κανόνες που να ικανοποιούν τις συνθήκες. (στην Εικόνα 10 είναι η συνέχεια της εκτέλεσης)

Βήμα 5: Τέλος γίνεται παρουσίαση όλων των παραγόμενων κανόνων που ικανοποιούν τα ελάχιστα όρια support και confidence. Στην Εικόνα 11 παρουσιάζονται οι κανόνες που έχουν εξαχθεί από το σύστημα.



Εικόνα 10: Παραγόμενοι κανόνες συσχέτισης (Παράδειγμα εφαρμογής αλγόριθμου Akama)



Εικόνα 11: Παρουσίαση αποτελεσμάτων από το εργαλείο (παράδειγμα εφαρμογής αλγόριθμου Akamas)

2.3 Αξιολόγηση Κανόνων

Χρησιμοποιώντας τους αλγόριθμους χωρίς να χρησιμοποιήσουμε κάποιο μέτρο αξιολόγησης των υποψήφιων κανόνων, τότε το σύστημα θα εξαγάγει άπειρους κανόνες, που δεν θα βοηθήσουν το χρήστη να λάβει κάποια σημαντική γνώση χρήστη. Έτσι θα πρέπει να γίνει αξιολόγηση των κανόνων, με κάποια μέτρα. Αυτά τα μέτρα καλούνται αντικειμενικά μέτρα και είναι βασισμένα στις πιθανότητες. Είναι συνήθως λειτουργίες από 2×2 πίνακα ενδεχομένων. Ένας πίνακας ενδεχομένων αποθηκεύει τις συχνότητας που ικανοποιούν τους δεδομένους όρους. Ο Πίνακας 9 είναι ένας πίνακας ενδεχομένων για τον κανόνα A → B, όπου το n(AB) δείχνει τον αριθμό εγγραφών που ικανοποιούν και το A και το B, και το N δείχνει τον συνολικό αριθμό εγγραφών [5].

Πίνακας 9: Πίνακας ενδεχομένων 2×2 για τον κανόνα A → B [5]

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

Όπως έχει αναφερθεί στο Κεφάλαιο 1.6, κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων, τα μέτρα αξιολόγησης μπορούν να χρησιμοποιηθούν με τρεις τρόπους:

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

Η πρώτη προσέγγιση, χρησιμοποιείται από τους αλγόριθμους Apriori και Akama για να αφαιρεθούν σύνολα αντικειμένων που δεν είναι συχνά εμφανιζόμενα. Το μέτρο που χρησιμοποιείται είναι το support, το οποίο, για ένα κανόνα συσχέτισης $A \to B$, δείχνει την πιθανότητα εμφάνισης του Α και του Β μαζί στα δεδομένα (Εξίσωση 1).

$$P(AB) = \frac{n(AB)}{N}$$ **(Εξίσωση 1:** Support**)**

Το σύστημα που έχει υλοποιηθεί, χρησιμοποιεί διάφορα άλλα αντικειμενικά μέτρα αξιολόγησης κανόνων για φιλτράρισμα των εξαγόμενων κανόνων ή ταξινόμησης τους. Τα μέτρα που έχουν υλοποιηθεί είναι:

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

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

(**Εξίσωση 2**: Confidence)

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

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

(Εξίσωση 3: Coverage**)**

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

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

(**Εξίσωση 4:** Prevalence)

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

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

(**Εξίσωση 5**: Recall)

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

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

(**Εξίσωση 6**: Specificity)

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

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

(**Εξίσωση 7:** Accuracy)

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

$$Lift = \frac{P(B/A)}{P(B)} \quad \text{(}\textbf{Εξίσωση 8}\text{: Lift)}$$

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

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

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

(**Εξίσωση 9**: Leverage)

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

9. *Προστιθέμενη Αξία (Added Value)*: Ορίζει την διαφορά μεταξύ της τελικής απάντησης με την άμεση και έμμεση απάντηση [12]. Και είναι η διαφορά του confidence με την πιθανότητα να ισχύει το B.

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

(**Εξίσωση 10**: Added Value)

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

10. *Σχετικός κίνδυνος (Relative Risk):* Είναι υπολογισμό του κινδύνου γεγονότος (πχ μίας ασθένειας) στη βάση δεδομένων. Υπολογίζεται από την αναλογία της πιθανότητας να ισχύει το B δεδομένου ότι ισχύει το A (εκτεθειμένης ομάδας), με την πιθανότητα να ισχύει το B δεδομένου ότι δεν ισχύει το A (μη εκτεθειμένης ομάδας) [13].

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

(**Εξίσωση 11:** Relative Risk)

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

11. *Jaccard*: Χρησιμοποιείται για την σύγκριση ομοιότητας και ποικιλομορφίας συνόλων δειγμάτων. Το Jaccard ορίζεται ως το μέγεθος της τομής που διαιρείται με την ένωση των δειγμάτων [14]:

$$Jaccard = \frac{P(AB)}{(P(A) + P(B) - P(AB))}$$

(**Εξίσωση 12:** Jaccard)

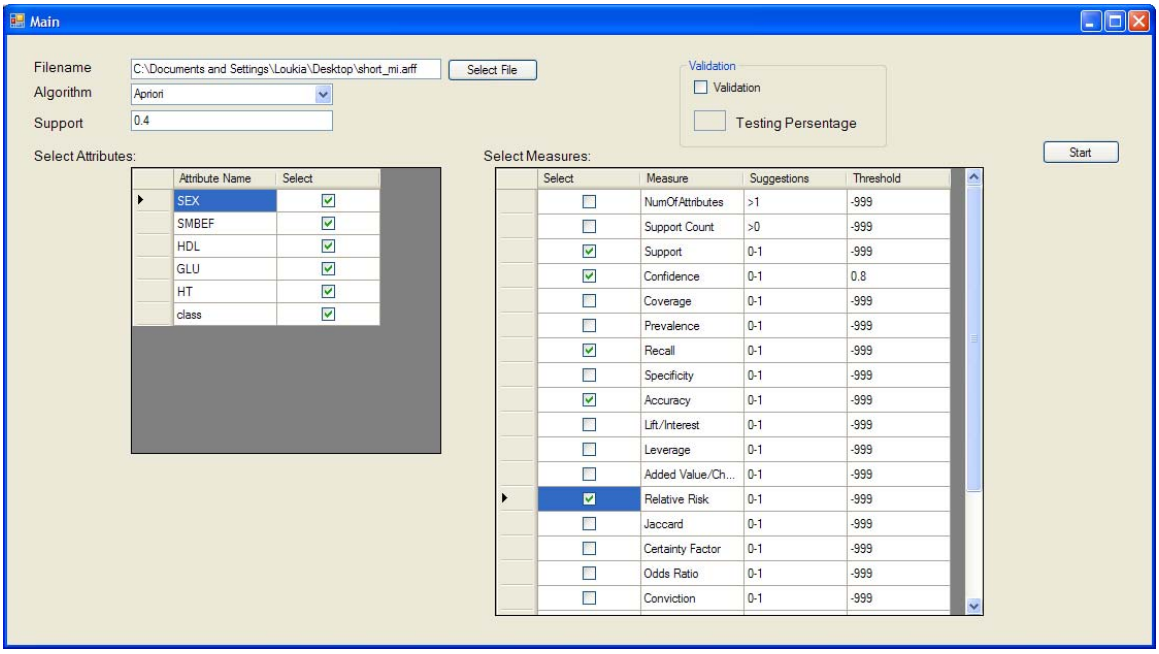
Παίρνει πραγματικές θετικές τιμές (>0). Όταν η τιμή τείνει στο ένα, τα δύο μέρη είναι ανεξάρτητα και έτσι ο κανόνας δεν παρουσιάζει κάποιο ενδιαφέρον. Όταν η τιμή τείνει στο +∞ αυτό σημαίνει ότι τα δύο μέρη είναι εξαρτώμενα και δείχνει ότι ο κανόνας είναι σημαντικός.

12. *Αναλογία Πιθανοτήτων (Odds Ratio)*: Είναι η αναλογία της πιθανότητας ενός γεγονότος να εμφανίζεται σε μια ομάδα, με τις πιθανότητες να εμφανίζεται σε μια άλλη ομάδα [15]. Υπολογίζεται από την πιο κάτω εξίσωση.

$$OddsRatio = \frac{P(AB)P(\overline{A}\overline{B})}{P(A\overline{B})P(\overline{A}B)}$$ (**Εξίσωση 13**: Odds Ratio)

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

Ο χρήστης μπορεί να επιλέξει ποία από τα μέτρα θέλει να παρουσιάζονται μαζί με τους εξαγόμενους κανόνες συσχέτισης. Για παράδειγμα ο χρήστης στην Εικόνα 12 επιλέγει να παρουσιάζονται το support, confidence, accuracy, relative risk, και τα αποτέλεσμα παρουσιάζονται στο Εικόνα 13. Επομένως ο χρήστης μπορεί να ταξινομήσει τους κανόνες με βάσει την υπολογισμένη τιμή των μέτρων αξιολόγησης.

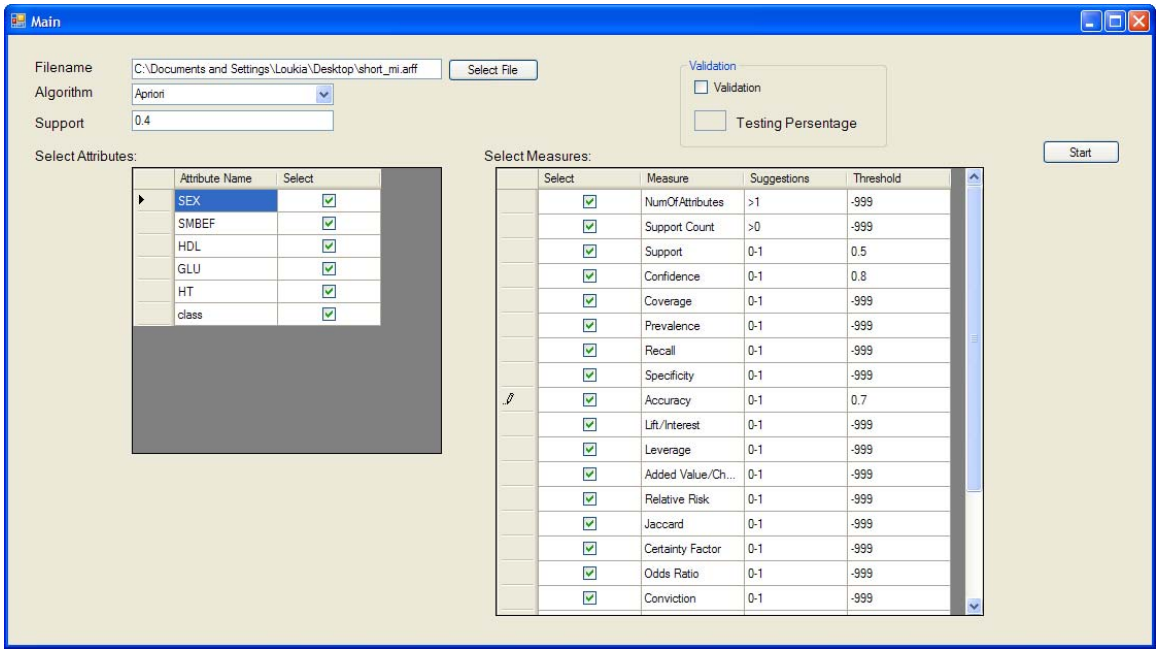


Εικόνα 12: Οθόνη για εισαγωγή δεδομένων στο σύστημα (Επιλογή μέτρων αξιολόγησης για παρουσίαση)

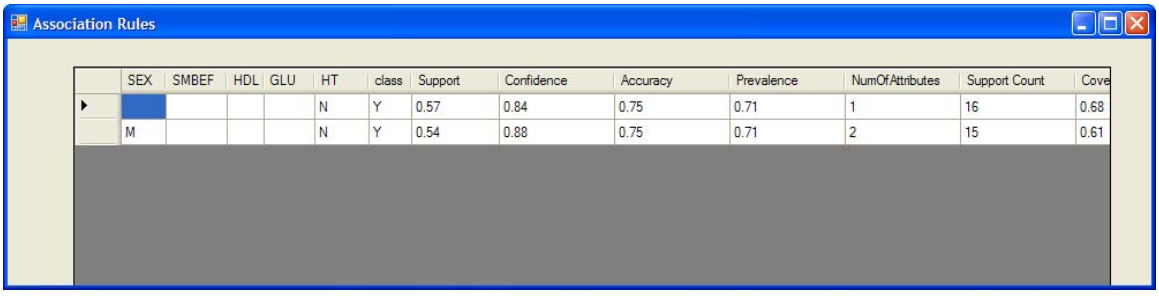
SEX	SMBEF	HDL	GLU	HT	class	Support	Confidence	Recall	Accuracy	Relative Risk
		M			Y	0.50	0.82	0.70	0.68	1.51
				N	Y	0.57	0.84	0.80	0.75	1.89
	Y	M			Y	0.43	0.86	0.60	0.64	1.50
M		M			Y	0.43	0.80	0.60	0.61	1.30
	Y			N	Y	0.50	0.82	0.70	0.68	1.51
M				N	Y	0.54	0.88	0.75	0.75	1.94
M	Y			N	Y	0.46	0.87	0.65	0.68	1.61

Εικόνα 13: Οθόνη παρουσίασης κανόνων μαζί με τα αποτελέσματα των μέτρων αξιολόγησης

Η τρίτη προσέγγιση εφαρμόζεται επίσης για αυτά τα μέτρα που έχουν αναφερθεί πιο πάνων. Ο χρήστης μπορεί να εισάγει κάποιο ελάχιστο όριο για οποιοδήποτε μέτρο αξιολόγησης κανόνων. Τότε το σύστημα θα παρουσιάσει μόνος τους κανόνες που τα ικανοποιούν. Για παράδειγμα ο χρήστης στην Εικόνα 13 βάζει κάποιο ελάχιστο όριο το support (0,5), confidence (0,8) και accuracy (0,7), και τα αποτέλεσμα παρουσιάζονται στην Εικόνα 14.



Εικόνα 14: Οθόνη για εισαγωγή δεδομένων στο σύστημα (επιλογή μέτρων αξιολόγησης για φιλτράρισμα κανόνων)



Εικόνα 15: Οθόνη παρουσίασης κανόνων που ικανοποιούν τα ελάχιστα όρια των μέτρων αξιολόγησης

2.4 Αξιολόγηση Προτύπου

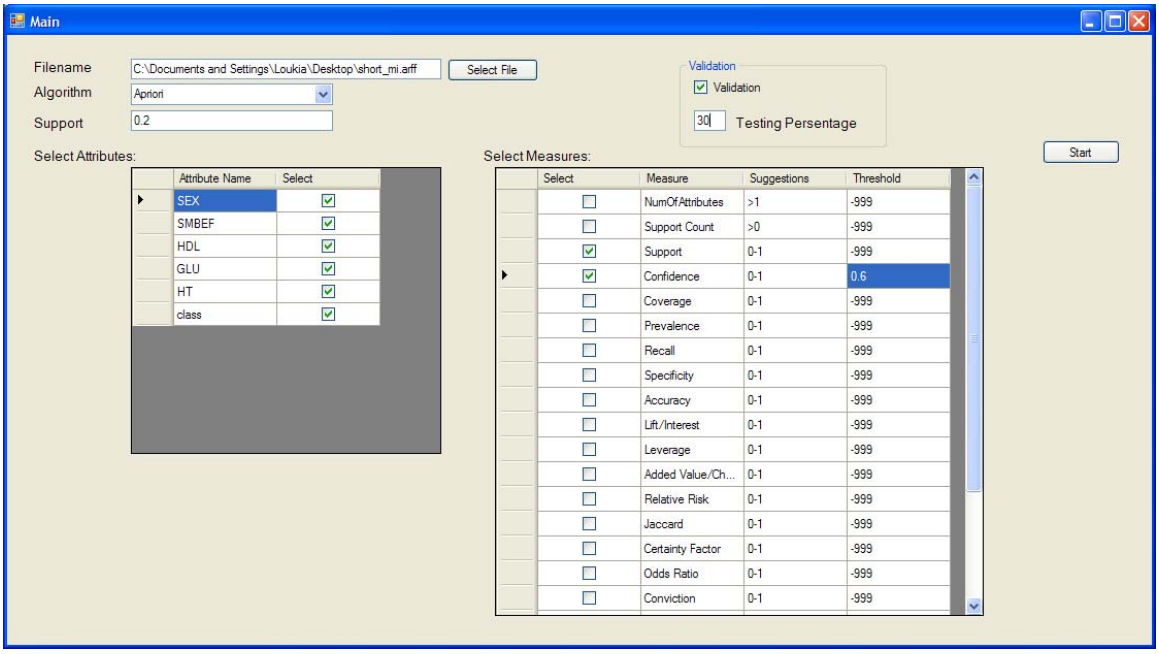
Έκτος από την αξιολόγηση των κανόνων, που χρησιμοποιείται για να επιλεγούν οι πιο σημαντικοί κανόνες, θα πρέπει επίσης να αξιολογήσουμε την αξιοπιστία των παραγόμενων κανόνων, κατά πόσο αυτοί οι κανόνες συμπεριφέρονται το ίδιο σε μια άλλη βάση δεδομένων, που χρησιμοποιείται για έλεγχο (testing). Για την αξιολόγησης αξιοπιστίας του προτύπου και των κανόνων, θα πρέπει να γίνει εισαγωγή δύο βάσεων δεδομένων. Η μία βάση θα χρησιμοποιηθεί για εκπαίδευση (training), στην οποία θα εφαρμοστούν οι αλγόριθμοι και θα εξαχθούν οι ενδιαφέρον κανόνες. Η άλλη βάση δεδομένων θα χρησιμοποιηθεί για τον έλεγχο (testing) αξιοπιστίας όλων των εξαγόμενων κανόνων. Σε αυτή την βάση δεδομένων ελέγχεται για κάθε εξαγόμενο κανόνα, της βάσης εκπαίδευσης, το ποσοστό επιτυχίας του στην βάσης ελέγχου.

Για την υλοποίηση του ελέγχου αξιοπιστίας στο σύστημα, η εισερχόμενη βάση δεδομένων μοιράζεται τυχαία το ένα ποσοστό σε βάση εκπαίδευσης και το άλλο ποσοστό σε βάση ελέγχου. Το ποσοστό για το οποίο η βάση δεδομένων χωρίζεται καθορίζεται από τον χρήστη. Κατά την εισαγωγή των δεδομένων του χρήστη στο σύστημα, μπορεί να επιλέξει εάν θέλει να γίνεται έλεγχος αξιοπιστίας των κανόνων, και το ποσοστό για το οποίο επιθυμεί να είναι η βάση ελέγχου. Αφού γίνεται η εισαγωγή αυτού του ποσοστού, παραδείγματος χάριν εάν ο χρήστης επιλέξει το 20%, τότε το σύστημα χωρίζει τυχαία την βάσης δεδομένων στο 80% για εκπαίδευση και στο 20% για έλεγχο.

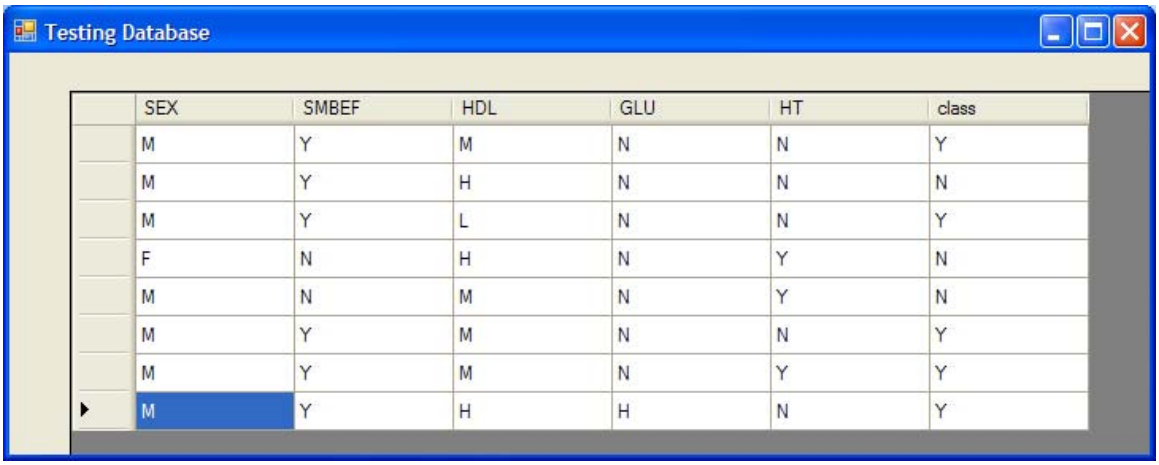
Αφού χωριστεί η βάση δεδομένων, τότε για κάθε εξαγόμενο κανόνα, γίνεται ο έλεγχος επιτυχίας του στην βάση ελέγχου. Για κάποιο κανόνα συσχέτισης $A \to B$ αναζητούνται στη βάση ελέγχου οι δοσοληψίες στις οποίες ισχύει το Α, και μετά βρίσκεται η πιθανότητα να ισχύει το Β σε αυτές της δοσοληψίες. Στην πραγματικότητα αυτός ο υπολογισμός είναι το μέτρο confidence (Εξίσωση 2).

Για παράδειγμα εάν στη βάση δεδομένων του Πίνακα 1, ορίσουμε το ποσοστό της βάσης ελέγχου (Testing) 30% (Εικόνα 16), τότε η βάση δεδομένων θα χωριστεί σε δύο βάσεις. Επομένως η βάση ελέγχου θα αποτελείται από 8 δοσοληψίες (Εικόνα 17 παρουσιάζεται η βάση ελέγχου) οι βάση εκπαίδευσης (Training) θα αποτελείται από τις υπόλοιπες 20 δοσοληψίες.

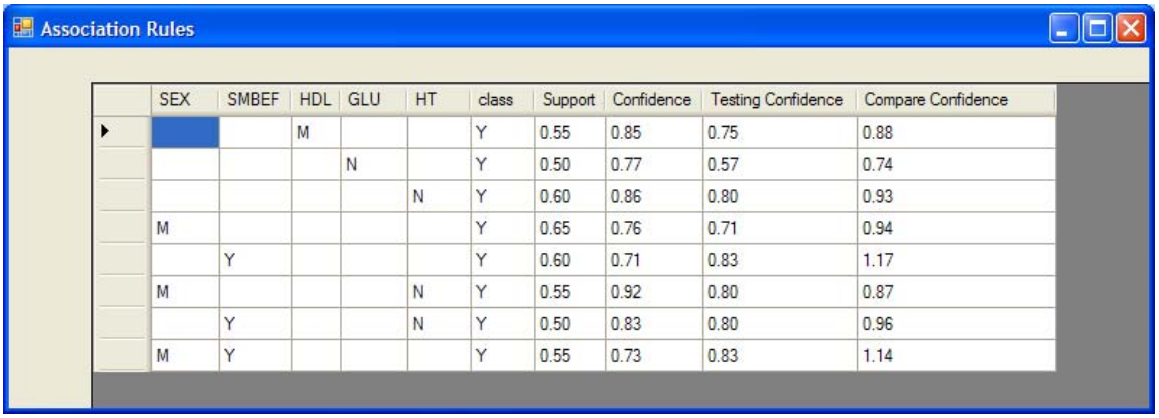
Οι κανόνες που εξάγονται από την εισαγωγή των δεδομένων του πιο πάνω παραδείγματος παρουσιάζονται στην Εικόνα 18. Το Testing Confidence είναι το ποσοστό επιτυχίας του κάθε κανόνα στην βάση ελέγχου. Όσο πιο κοντά είναι αυτό το αποτέλεσμα στο confidence που υπολογίστηκε στη βάση εκπαίδευσης, τότε οι κανόνες είναι πιο αξιόπιστοι.



Εικόνα 16: Οθόνη για εισαγωγής δεδομένων στο σύστημα (Επιλογή αξιολόγησης προτύπου)



Εικόνα 17: Επιλεγμένη βάση δεδομένων ελέγχου



Εικόνα 18: Παραγόμενοι κανόνες (Παράδειγμα αξιολόγησης προτύπου)

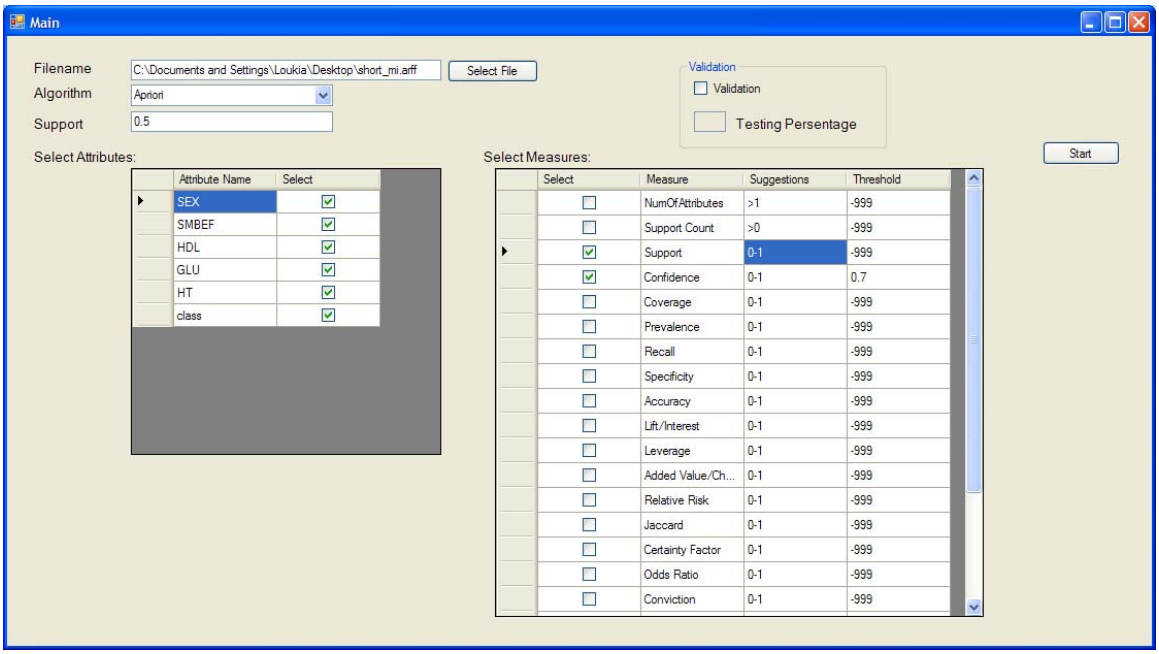
2.5 Σύγκριση Αλγόριθμων Apriori και Akama

Όπως έχει αναφερθεί πιο πάνω κι οι δύο αλγόριθμοι χρησιμοποιούνται για εξαγωγή κανόνων συσχέτισης.

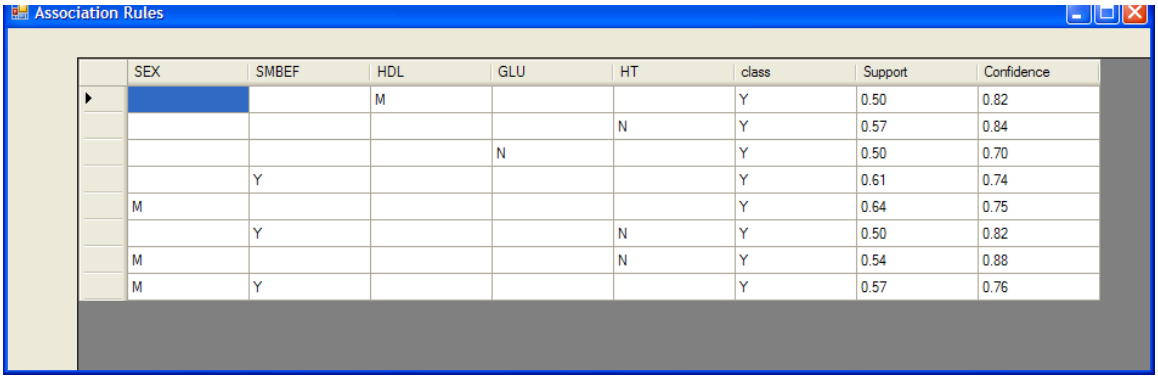
Ο αλγόριθμός Apriori, χρησιμοποιεί τα συχνά σύνολα αντικείμενων με κ-1 αντικείμενα, αυτά που ικανοποιούν το ελάχιστο όριο support, για παραχθούν τα συχνά σύνολα αντικειμένων με κ αντικείμενα. Επομένως, τα αποτελέσματα του αλγόριθμου Apriori εξαρτώνται από το support, και θα παραχθούν κανόνες με κ αντικείμενα, μόνο εάν στην προηγούμενη επανάληψη, τα κ-1 αντικείμενα ικανοποιούσαν το support. Ενώ αυτό δεν ισχύει στον αλγόριθμο Akama.

Ο αλγόριθμος Akamas στηρίζεται στο support μόνο στην πρώτη επανάληψη, όπου βρίσκει τα συχνά σύνολα αντικειμένων με 1 χαρακτηριστικό, τα οποία ικανοποιούν το ελάχιστο support. Μετά κτίζει όλους του δυνατούς συνδυασμούς συνόλων αντικειμένων, για να βρει κανόνες συσχέτισης χωρίς να εξαρτάται από το πόσο συχνά αυτά τα σύνολα αντικειμένων εμφανίζονται 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 το support.

Για παράδειγμα εφαρμόζεται ο αλγόριθμος Apriori στην βάση δεδομένων του Πίνακα 1 με ελάχιστο όριο support 0,4 και confidence 0,7 (Εικόνα 19 τα δεδομένων εισόδου) τότε θα παραχθούν οι κανόνες οι κανόνες της Εικόνας 20.

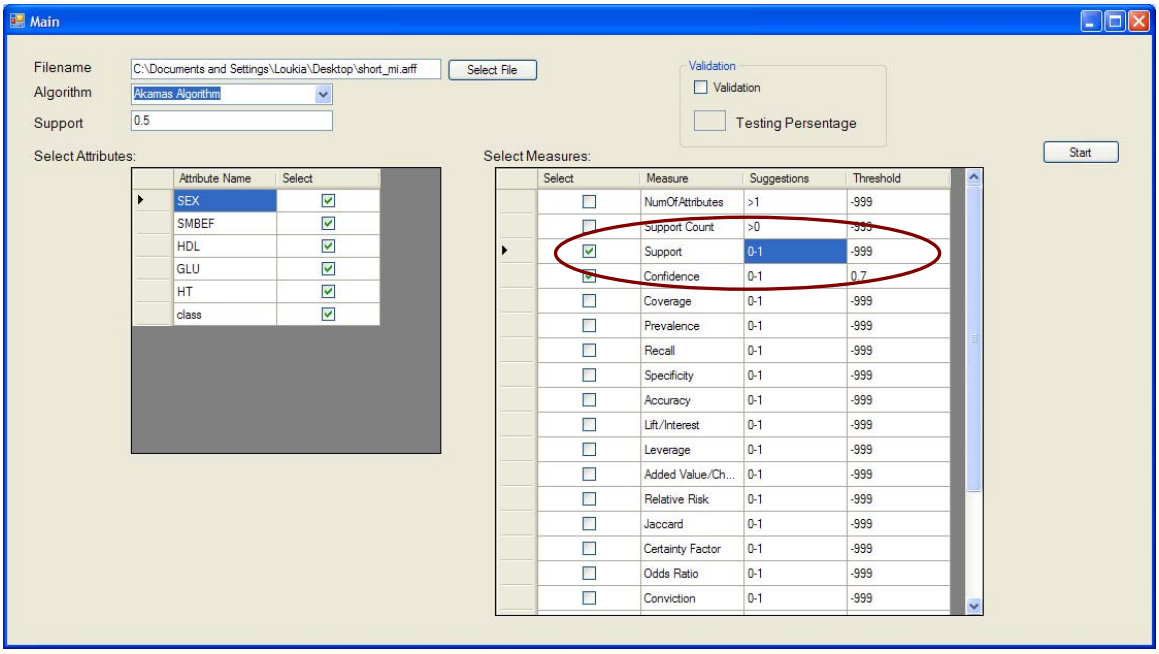


Εικόνα 19: Οθόνη για εισαγωγής δεδομένων στο σύστημα (Αλγόριθμου Apriori)

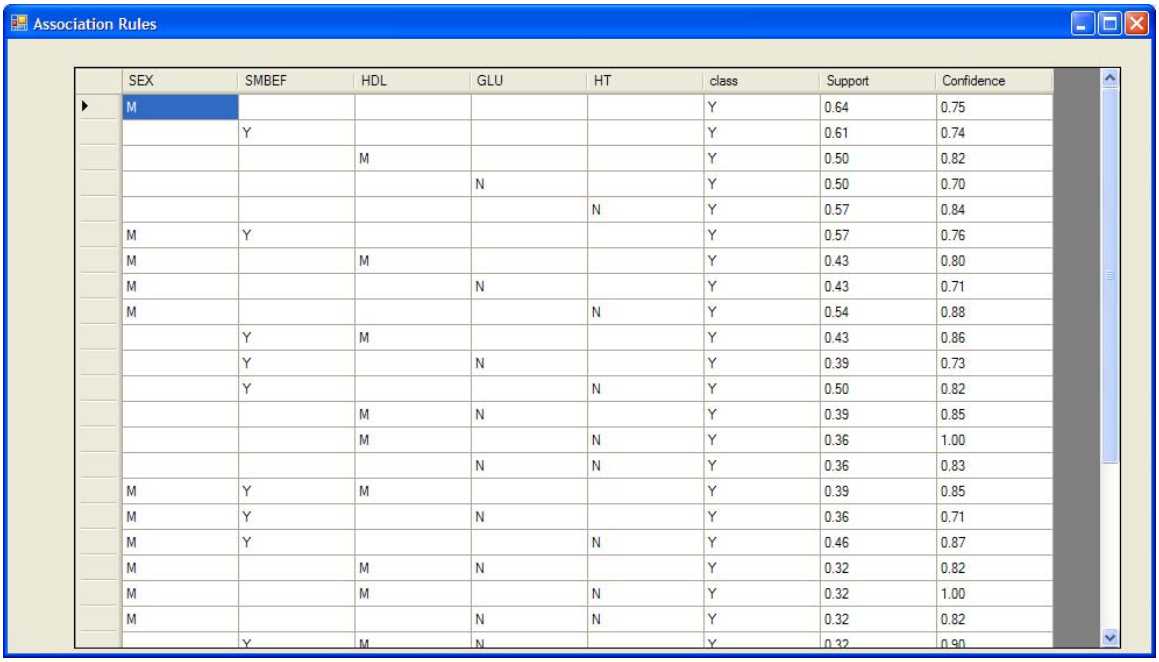


Εικόνα 20: Παραγόμενοι κανόνες αλγόριθμου Apriori

Εάν όμως εφαρμόσουμε τα ίδια δεδομένα εισόδου (Εικόνα 21) (χωρίς να επιλέξουμε κάποιο όριο support στα μέτρα αξιολόγησης των κανόνων) στον αλγόριθμο Akama, τότε θα παραχθούν οι κανόνες συσχέτισης της Εικόνας 22. Παρατηρείται ότι παράγονται πολύ περισσότεροι κανόνες από ότι παράγονται από τον αλγόριθμο Apriori. Όλοι οι κανόνες συσχέτισης του αλγόριθμου Akama παρουσιάζονται στον Πίνακα 4.



Εικόνα 21: Οθόνη για εισαγωγής δεδομένων στο σύστημα (Αλγόριθμου Akama)



SEX	SMBEF	HDL	GLU	HT	class	Support	Confidence
M					Y	0.64	0.75
	Y				Y	0.61	0.74
		M			Y	0.50	0.82
			N		Y	0.50	0.70
				N	Y	0.57	0.84
M	Y				Y	0.57	0.76
M		M			Y	0.43	0.80
M			N		Y	0.43	0.71
M				N	Y	0.54	0.88
	Y	M			Y	0.43	0.86
	Y		N		Y	0.39	0.73
	Y			N	Y	0.50	0.82
		M	N		Y	0.39	0.85
		M		N	Y	0.36	1.00
			N	N	Y	0.36	0.83
M	Y	M			Y	0.39	0.85
M	Y		N		Y	0.36	0.71
M	Y			N	Y	0.46	0.87
M		M	N		Y	0.32	0.82
M		M		N	Y	0.32	1.00
M			N	N	Y	0.32	0.82
	Y	M	N		Y	0.32	0.90

Εικόνα 22: Παραγόμενοι κανόνες αλγόριθμου Akama

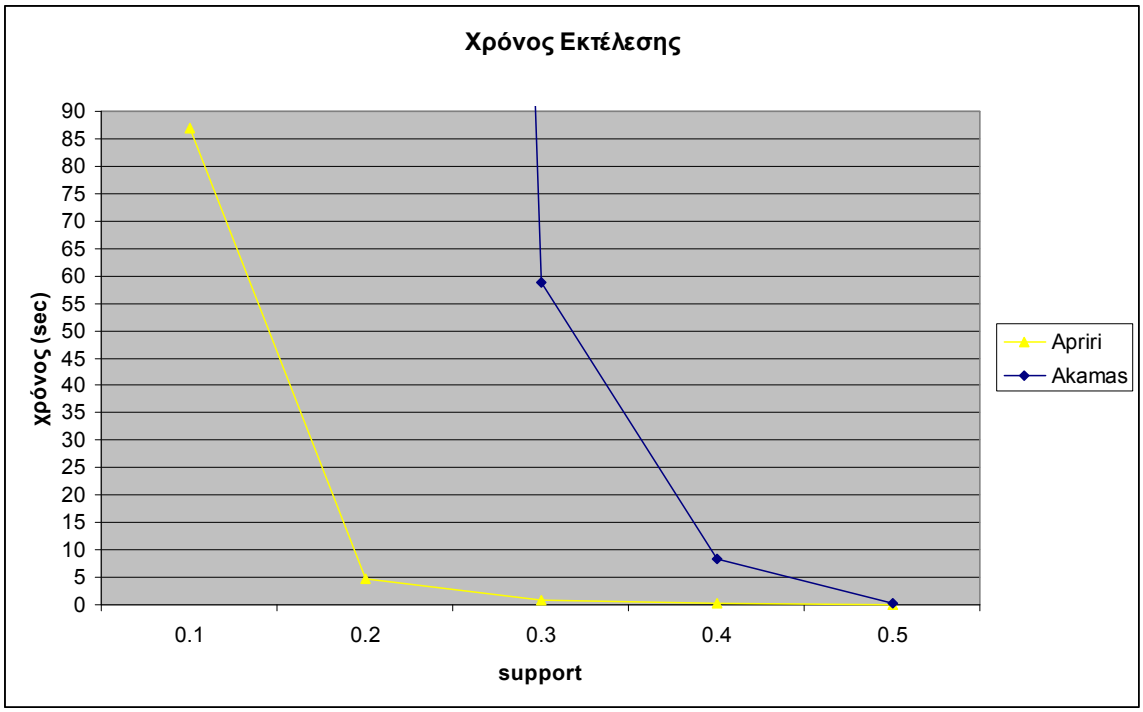
Πίνακας 10: Παραγόμενοι κανόνες αλγόριθμου Akama

SEX	SMBEF	HDL	GLU	HT	class	Support	Confidence
M					Y	0.64	0.75
	Y				Y	0.61	0.74
		M			Y	0.50	0.82
			N		Y	0.50	0.70
				N	Y	0.57	0.84
M	Y				Y	0.57	0.76
M		M			Y	0.43	0.80
M			N		Y	0.43	0.71
M				N	Y	0.54	0.88
	Y	M			Y	0.43	0.86
	Y		N		Y	0.39	0.73
	Y			N	Y	0.50	0.82
		M	N		Y	0.39	0.85
		M		N	Y	0.36	1.00
			N	N	Y	0.36	0.83
M	Y	M			Y	0.39	0.85
M	Y		N		Y	0.36	0.71
M	Y			N	Y	0.46	0.87
M		M	N		Y	0.32	0.82
M		M		N	Y	0.32	1.00
M			N	N	Y	0.32	0.82
	Y	M	N		Y	0.32	0.90
	Y	M		N	Y	0.32	1.00
	Y		N	N	Y	0.29	0.80
		M	N	N	Y	0.25	1.00
M	Y	M	N		Y	0.29	0.89
M	Y	M		N	Y	0.29	1.00
M	Y		N	N	Y	0.25	0.78
M		M	N	N	Y	0.21	1.00
	Y	M	N	N	Y	0.21	1.00
M	Y	M	N	N	Y	0.18	1.00

Επομένως, συγκρίνοντας τους παραγόμενους κανόνες συσχέτισης των δύο αλγορίθμων (Εικόνα 19 και Πίνακα 10), ο αλγόριθμος Akamas παράγει κανόνες που ενώ έχουν σχετικά χαμηλό support, έχουν ψηλό confidence. Αυτοί οι κανόνες απορρίπτονται από τον αλγόριθμο Apriori, ενώ για τους χρήστες μπορεί αυτοί οι κανόνες είναι ενδιαφέρον. Όπως για παράδειγμα ο κανόνας $SMBEF = Y, HDL = M, HT = N \rightarrow MI = Y$, έχει support 32%, δηλαδή παρουσιάζεται 9 φορές στη βάση δεδομένων, ενώ και για τις εννέα δοσοληψίες ισχύει ότι MI = Y (confidence 100%). Από αυτό μπορεί να βγει το συμπέρασμα ότι σε μεγάλες βάσεις δεδομένων μπορούν να βρεθούν σπάνιες περιπτώσεις στις οποίες να ισχύει ο κανόνας.

Να σημειωθεί ότι εάν ο χρήστης επιλέξει να αξιολογούνται οι κανόνες συσχέτισης στον αλγόριθμο Akama και με κάποιο όριο για support, όπως και στον Apriori, τότε οι αλγόριθμοι θα παραγάγουν τους ίδιους κανόνες.

Μία άλλη σημαντική διαφορά των δύο αλγόριθμων είναι ότι, λόγω του ότι ο αλγόριθμος Akamas κτίζει όλους τους δυνατούς συνδυασμούς, χωρίς να απορρίπτει κάποιους κανόνες, για να ελαττώνει τον χώρο αναζήτησης, τον καθιστό πιο αργό από τον αλγόριθμο Apriori. Για μια βάση δεδομένων με 369 δοσοληψίες έχουν εφαρμοστεί οι δύο αλγόριθμοι με τα ίδια δεδομένα εισόδου. Κρατώντας σταθερό το όριο confidence 0,5 κι αλλάζοντας το support έχει μετρηθεί ο χρόνος εκτέλεσης για τον κάθε αλγόριθμο. Στο σχεδιάγραμμα της Εικόνας 23, παρουσιάζεται η σύγκριση των δύο αλγορίθμων όσον αφορά το χρόνο εκτέλεσης. Παρατηρείται ότι όσο μικραίνει το ελάχιστο όριο support, οι κανόνες συσχέτισης αυξάνονται και οι δύο αλγόριθμοι αυξάνουν το χρόνο εκτέλεσης τους κατακόρυφα. Επίσης παρατηρείται ότι ο χρόνος εκτέλεσης του Akama είναι πολύ πιο αργός από τον Apriori.



Εικόνα 23: Χρόνος εκτέλεσης Αλγορίθμων Apriori και Akama

2.6 Σύγκριση με άλλα εργαλεία εξόρυξης δεδομένων

Στον κλάδο εξόρυξης δεδομένων έχουν παρουσιαστεί διάφορα εργαλεία για εξόρυξη κανόνων συσχέτισης. Ένα από τα πιο γνωστά είναι το WEKA [10].

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 Akama. 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 που παρουσιάζονται στο Κεφάλαιο 2.3. Από αυτά τα μέτρα ο χρήστης μπορεί να επιλέξει ποία επιθυμεί να παρουσιάζονται, αλλά επίσης και να ορίσει ένα ελάχιστο όριο με το οποία το εργαλείο απορρίπτει 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 λάβει εύκολα και γρήγορα τη γνώση, με βάση τον σκοπό που επιθυμεί.

Κεφάλαιο 3

Περιγραφή Βάσης Δεδομένων

3.1 Χαρακτηριστικά της βάσης δεδομένων

Στον Πίνακα 11 που ακολουθεί, παρουσιάζεται μια γενική αναφορά και περιγραφή των πεδίων της Βάσης Δεδομένων:

Πίνακας 11: Πεδία Βάσης Δεδομένων

ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	ΠΕΡΙΓΡΑΦΗ
ACS/MI Στεφανιαία Νόσος	Ένδειξη εάν ο ασθενής πάσχει από στεφανιαία νόσο. Παίρνει τιμές Y (ο ασθενής πάσχει από στεφανιαία νόσο) και N (ο ασθενής δεν πάσχει από στεφανιαία νόσο).
PCI Αγγειοπλαστική	Δείχνει κατά πόσο ο ασθενής έχει υποστεί σε αγγειοπλαστική εγχείρηση. Παίρνει τιμές Y(έχει υποστεί) και N (δεν έχει υποστεί).
CABG Παράκαμψη (bypass)	Δείχνει κατά πόσο ο ασθενής έχει κάνει παράκαμψη (bypass). Παίρνει τιμές Y(έχει κάνει) και N (δεν έχει κάνει).
AGE Ηλικία	Αντιπροσωπεύει την ηλικία του ασθενή.
GEN Φύλο	Δείχνει το φύλο του ασθενή. Παίρνει τιμές M (MALE) και F (FEMALE).
W Βάρος	Αντιπροσωπεύει το βάρος του ασθενή.
H Ύψος	Αντιπροσωπεύει το ύψος του ασθενή.
BMI Δείκτης Μάζας Σώματος	Ο δείκτης μάζας σώματος υπολογίζει το βάρος ενός ασθενή βάσει του ύψους του.
AS Ενεργός Καπνιστής	Δείχνει εάν ο ασθενής είναι ενεργός καπνιστής. Παίρνει τιμές N (δεν είναι ενεργός καπνιστής) και Y (είναι ενεργός καπνιστής).
PS Παθητικός Καπνιστής	Δείχνει εάν ο ασθενής είναι παθητικός καπνιστής. Παίρνει τιμές N (δεν είναι παθητικός καπνιστής) και Y (είναι παθητικός καπνιστής).
S-R Stop – Restart smoking	Δείχνει εάν ο ασθενής είχε σταματήσει το κάπνισμα και μετά το ξεκίνησε ξανά. Παίρνει τιμές N (ο ασθενής δεν σταμάτησε κα ξεκίνησε ξανά το κάπνισμα) και Y (ο ασθενής σταμάτησε κα ξεκίνησε ξανά το κάπνισμα).
EX-SM Πρώην Καπνιστής	Δείχνει εάν ο ασθενής είναι πρώην καπνιστής. Παίρνει τιμές N (ο ασθενής δεν είναι πρώην καπνιστής), Y (ο ασθενής είναι πρώην καπνιστής) και NA (δεν γνωρίζουμε).

POS FH Ιστορικό Οικογένειας	Παρουσιάζει το ιστορικό της οικογένειας του ασθενή σε καρδιακά επεισόδια. Παίρνει τιμές Y (κάποιος από την οικογένεια του ασθενή έχει πάθει καρδιακό επεισόδιο) και N (κανένας από την οικογένεια του ασθενή δεν έχει πάθει καρδιακό επεισόδιο).
HT Υπέρταση	Η υπέρταση είναι η υψηλή πίεση αίματος. Δείχνει αν ο ασθενής πάσχει από υπέρταση. Παίρνει τιμές N (ο ασθενής δεν πάσχει από υπέρταση) και Y (ο ασθενής πάσχει από υπέρταση).
DM Διαβήτης	Ο διαβήτης είναι μια ανίατη ασθένεια που 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. Παίρνει τιμές N (ο ασθενής δεν πάσχει από διαβήτη) και Y (ο ασθενής πάσχει από διαβήτη).
STAT Άγχος	Δείχνει κατά πόσο ο ασθενής καταβάλλεται από άγχος. Παίρνει τιμές N (ο ασθενής καταβάλλεται από άγχος) και Y (ο ασθενής δεν καταβάλλεται από άγχος).
EXER Άσκηση	Δείχνει κατά πόσο ο ασθενής γυμνάζεται ή όχι. Παίρνει τιμές Y (ο ασθενής γυμνάζεται), N(ο ασθενής δεν γυμνάζεται), και NA (ο ασθενής δεν έχει δώσει οποιαδήποτε πληροφορία).
HR Παλμοί καρδίας	Δείχνει τους παλμούς της καρδίας του ασθενούς.
SBP Ψηλή Πίεση (Συστολική Πίεση)	Η συστολική πίεση αναπαριστά τη μέγιστη πίεση που εξασκείται όταν η καρδιά συστέλλεται.
DBP Χαμηλή Πίεση (Διαστολική Πίεση)	Η διαστολική πίεση αναπαριστά την πίεση στις αρτηρίες όταν η καρδιά είναι ξεκούραστη.
TC Ολική Χοληστερόλη	Δείχνει την ολική ποσότητα χοληστερόλης στο αίμα.
HDL Λιποπρωτεΐνες Υψηλής Πυκνότητας	Δείχνει την ποσότητα των λιποπρωτεϊνών υψηλής πυκνότητας στο αίμα.
LDL Λιποπρωτεΐνες Χαμηλής Πυκνότητας	Δείχνει την ποσότητα των λιποπρωτεϊνών χαμηλής πυκνότητας στο αίμα.
TG Τριγλυκερίδια	Δείχνει την ποσότητα των τριγλυκεριδίων στο αίμα.
GLU Γλυκόζη	Δείχνει την ποσότητα της γλυκόζης στο αίμα.
UA Ουρία – Ουρικό Οξύ	Δείχνει την ποσότητα της ουρίας στο αίμα.
FIBR Fibrinogen	Ινοδογόνο
CRP	Οξείας φάσης πρωτεΐνης του πλάσματος που παράγει το συκώτι κατά την διάρκεια της φλεγμονής

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 3.2, τα οποία κωδικοποιήσαμε με βάση τις οδηγίες των γιατρών και των διεθνών και ευρωπαϊκών προδιαγραφών στο κεφάλαιο 3.4. Η αρχική βάση δεδομένων περιείχε 1200 ασθενείς με τριών ειδών καρδιοαγγειακά νοσήματα: α) Στεφανιαία νόσο, β) Αγγειοπλαστική και γ) Παράκαμψη (bypass). 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 δύο, ή 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.

3.2 Επιλογή των χαρακτηριστικών που έχουν μελετηθεί

Για την επιλογή των χαρακτηριστικών είχαν ληφθεί υπ' όψη οι ακόλουθες προϋποθέσεις:

α) 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.

β) 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 (BMI).

γ) Παράγοντες που είχαν πολλές ελλιπείς τιμές και δεν υπήρχε η ευχέρεια ανάκτησης αυτών των τιμών, έχουν αφαιρεθεί.

3.3 Συμπλήρωση ελλιπών τιμών

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.

3.4 Κωδικοποίηση των χαρακτηριστικών

Η κωδικοποίηση των χαρακτηριστικών έγινε βάσει των οδηγιών που μας έδωσαν οι γιατροί και των διεθνών και ευρωπαϊκών προδιαγραφών. Συγκεκριμένα έχουμε:

Πίνακας 12: Κωδικοποίση χαρακτηριστικών

	Risk Factor	Code 1	Code 2	Code 3	Code 4
Clinical factors					
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		
Biochemical factors					
9	TC **	D <200	N:201 –240	H>240	
10	HDL**				
	Women	L<50	M:50-60	H>60	
	Men	L<40	M:40-60		
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

3.5 Στατιστική ανάλυση των χαρακτηριστικών που έχουν χρησιμοποιηθεί

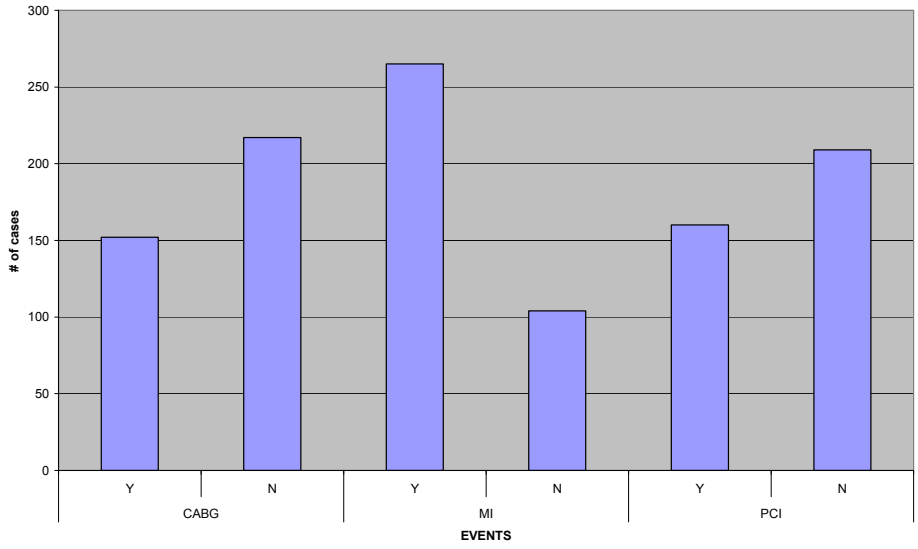
Στο Πίνακα 13 παρουσιάζεται ο αριθμός των ασθενών που παρουσιάζουν το κάθε χαρακτηριστικό.

Πίνακας 13: Αριθμός ασθενών ανά χαρακτηριστικό

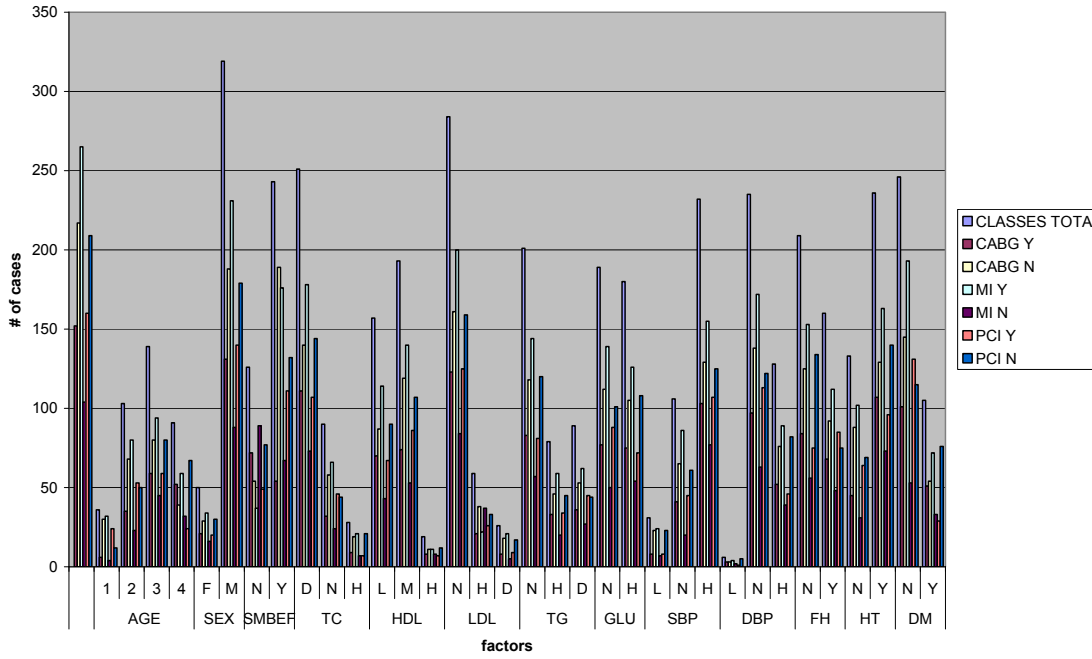
CLASSES			CABG		MI		PCI	
	VALUE	TOTAL	Y	N	Y	N	Y	N
			152	217	265	104	160	209
AGE	1	36	6	30	32	4	24	12
	2	103	35	68	80	23	53	50
	3	139	59	80	94	45	59	80
	4	91	52	39	59	32	24	67
SEX	F	50	21	29	34	16	20	30
	M	319	131	188	231	88	140	179
SMBEF	N	126	72	54	37	89	49	77
	Y	243	54	189	176	67	111	132
TC	D	251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

Η Εικόνα 24 αναπαριστά τον αριθμό ασθενών σε σχέση με τις διάφορες τιμές των κλάσεων, MI, CABG και PCI. Παρατηρούμε ότι για τις κλάσεις CABG και PCI, ο αριθμός των ασθενών για τις τιμές 'Y' και 'N' των κλάσεων είναι περίπου ο ίδιος, ενώ για την κλάση MI παρουσιάζονται περισσότερες δοσοληψίες με τιμή 'Y' παρά με 'N'.

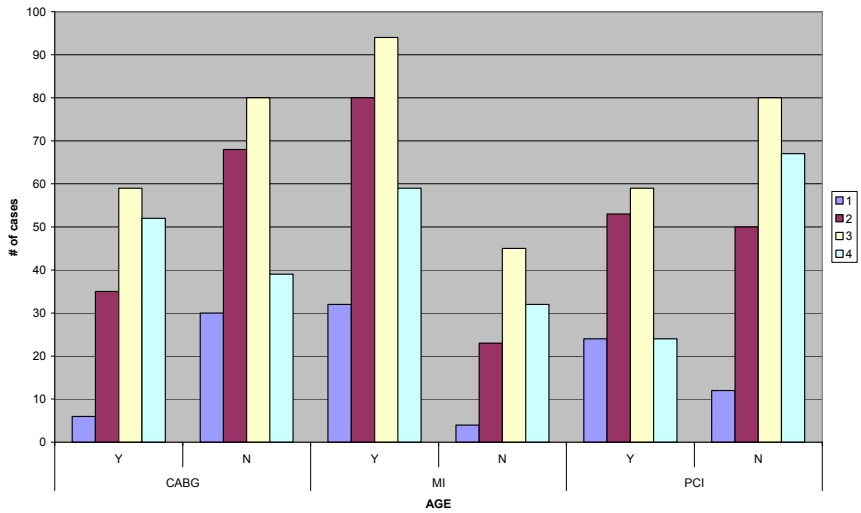
Ενώ η Εικόνα 25 παρουσιάζει τον αριθμό των ασθενών για κάθε χαρακτηριστικό. Παρατηρούμε ότι παρουσιάζονται περισσότερες περιπτώσεις να είναι άνδρας (SEX = M), και πολλές περιπτώσεις με κανονική ποσότητα των λιποπρωτεϊνών χαμηλής πυκνότητας στο αίμα (LDL = N).



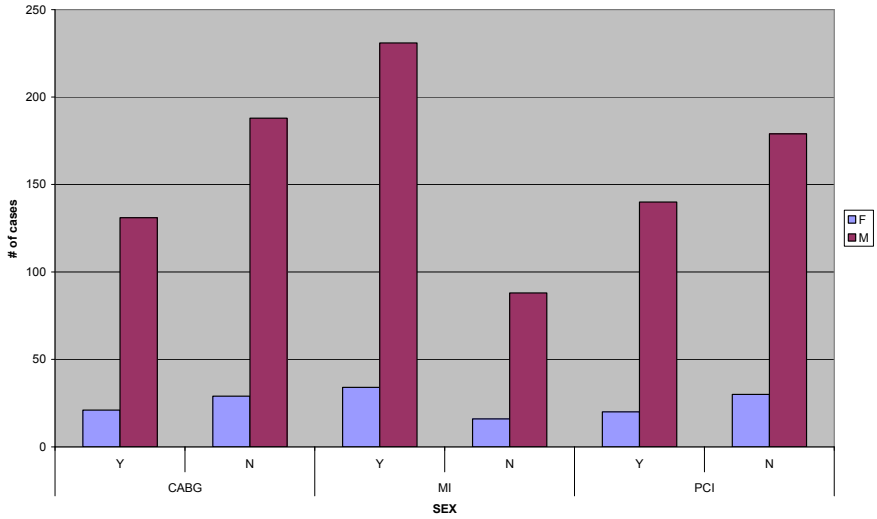
Εικόνα 24: Αριθμός ασθενών για κάθε κλάση (MI, CABG, PCI)



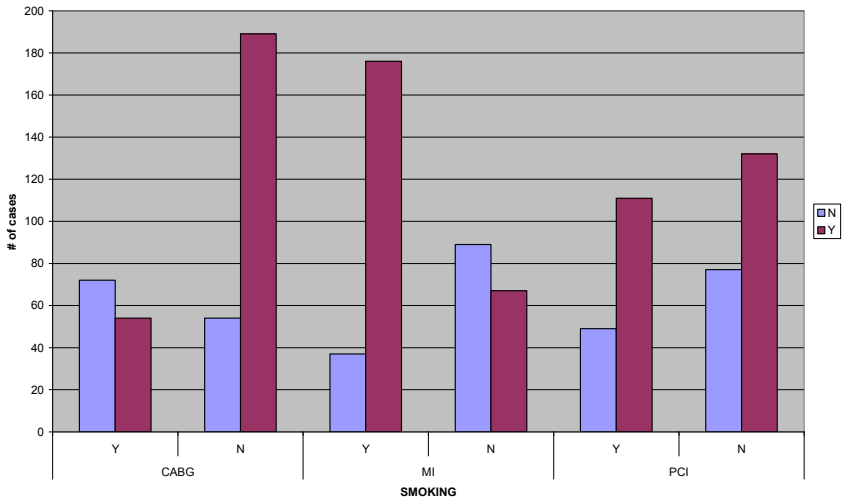
Εικόνα 25: Αριθμός ασθενών για κάθε χαρακτηριστικό της βάσης δεδομένων



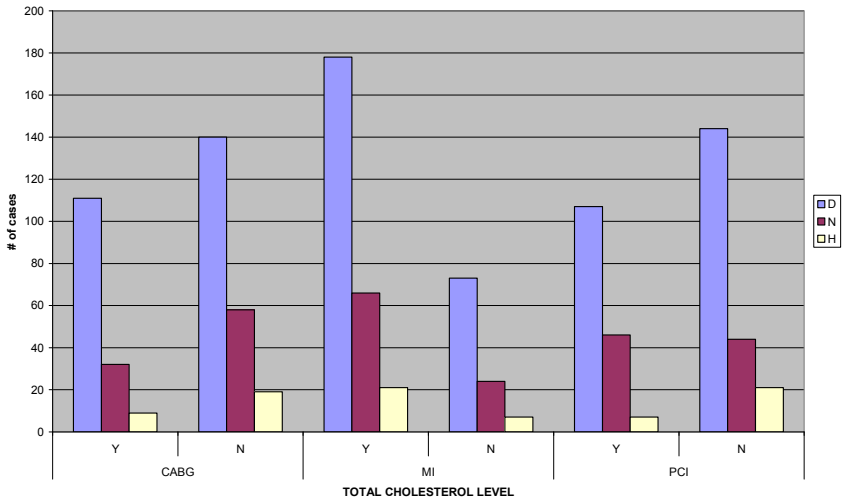
Εικόνα 26: Κατανομή ασθενών κάθε ηλικίας στις κλάσεις



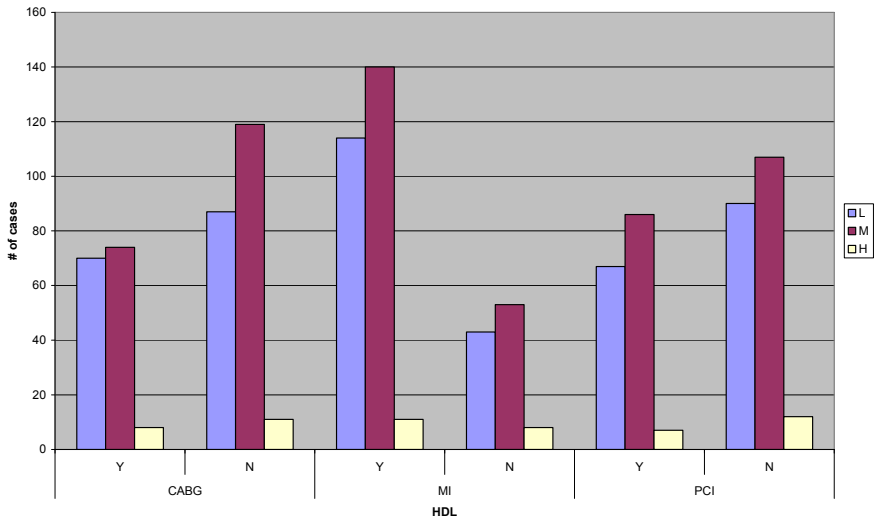
Εικόνα 27: Κατανομή ασθενών κάθε φύλου στις κλάσεις



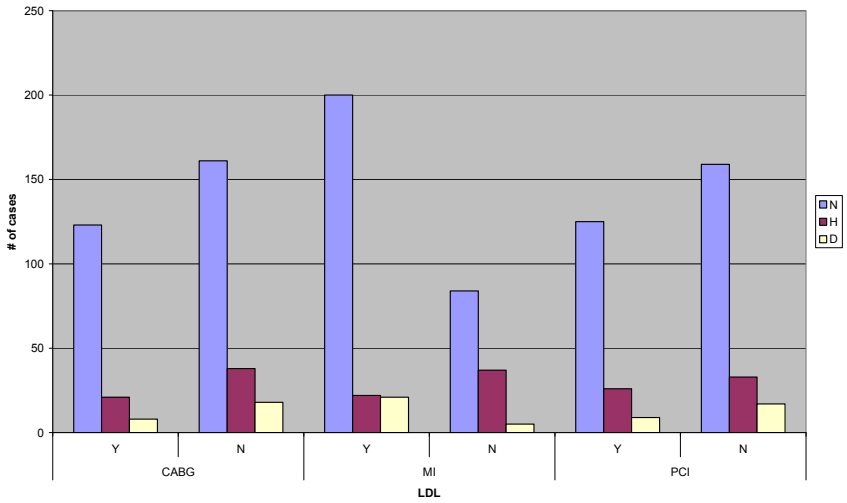
Εικόνα 28: Κατανομή ασθενών με το χαρακτηριστικό κάπνισμα στις κλάσεις



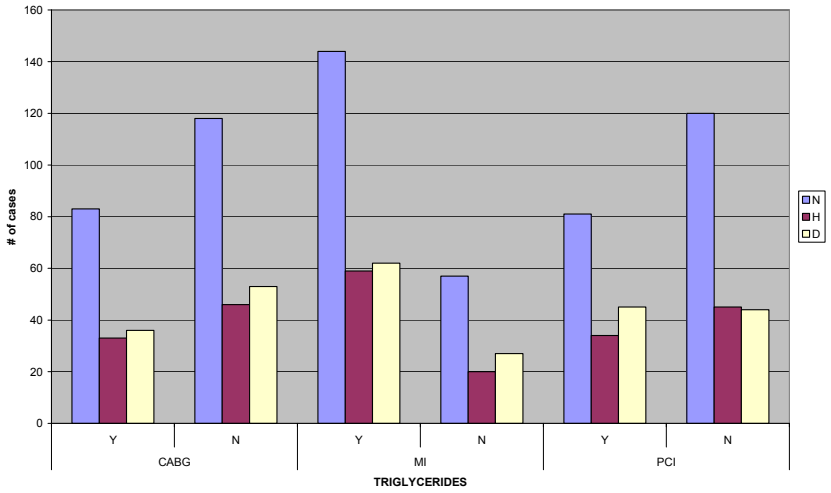
Εικόνα 29: Κατανομή ασθενών με το χαρακτηριστικό ολικής χοληστερόλης (TC) στις κλάσεις



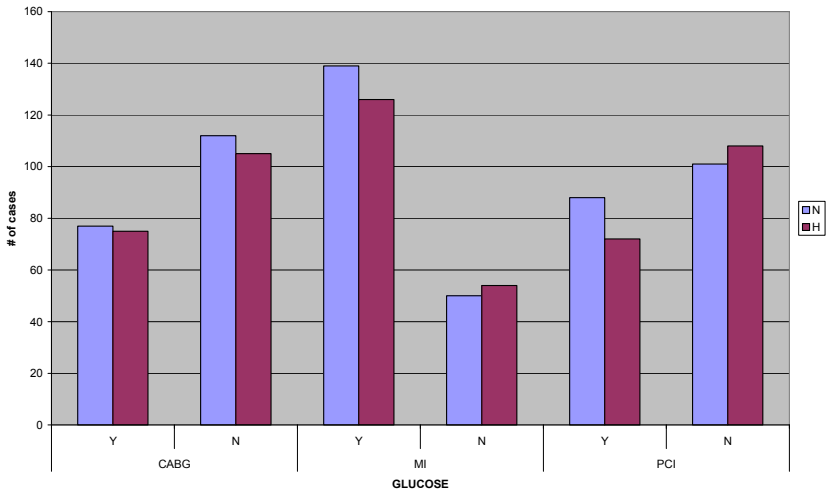
Εικόνα 30: Κατανομή ασθενών με το χαρακτηριστικό HDL στις κλάσεις



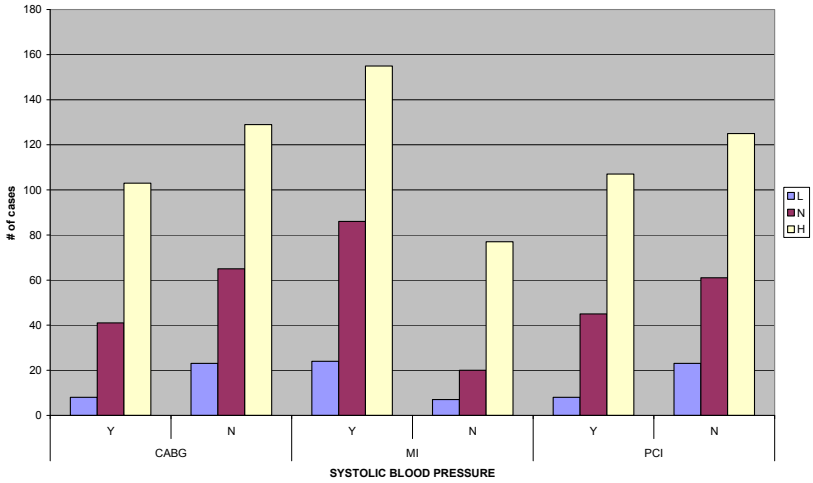
Εικόνα 31: Κατανομή ασθενών με το χαρακτηριστικό LDL στις κλάσεις



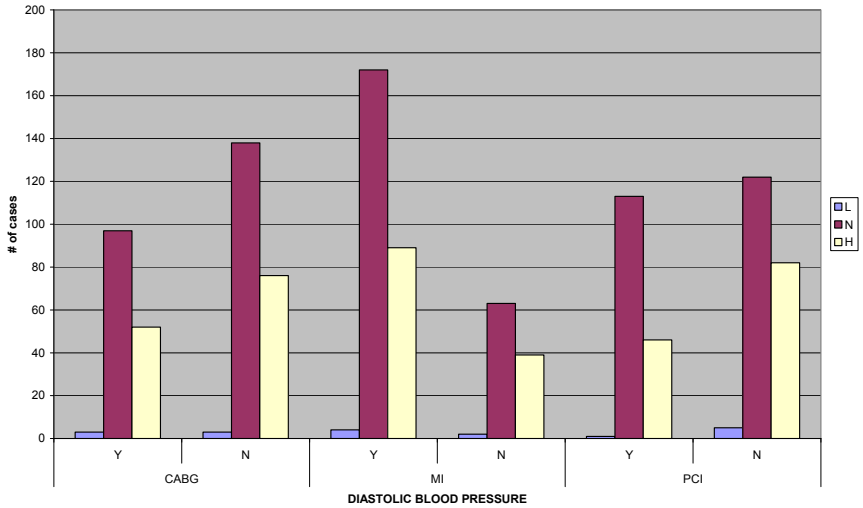
Εικόνα 32: Κατανομή ασθενών με το χαρακτηριστικό TG στις κλάσεις



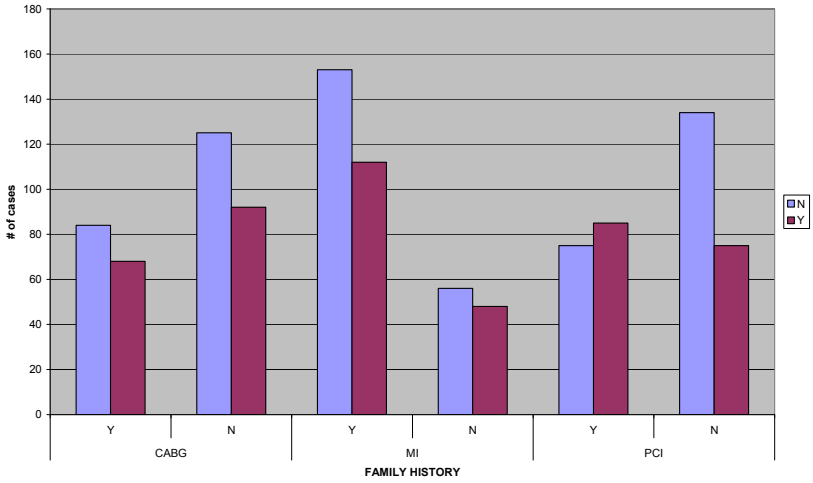
Εικόνα 33: Κατανομή ασθενών με το χαρακτηριστικό GLU στις κλάσεις



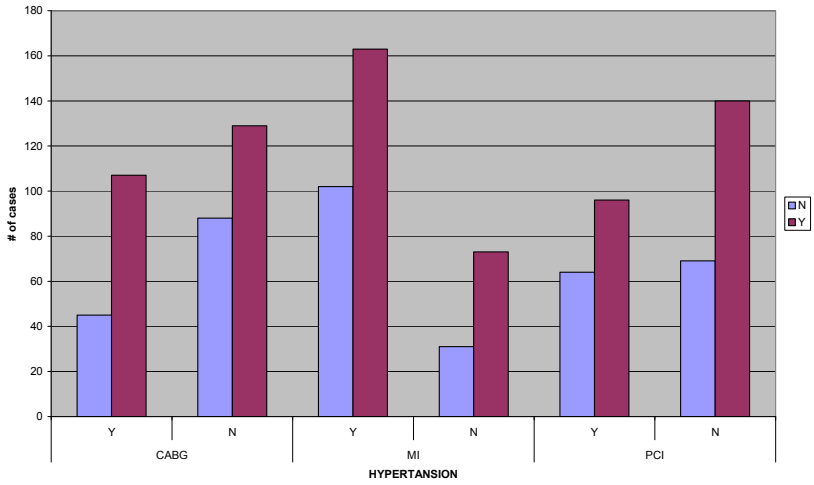
Εικόνα 34: Κατανομή ασθενών με το χαρακτηριστικό SBP στις κλάσεις



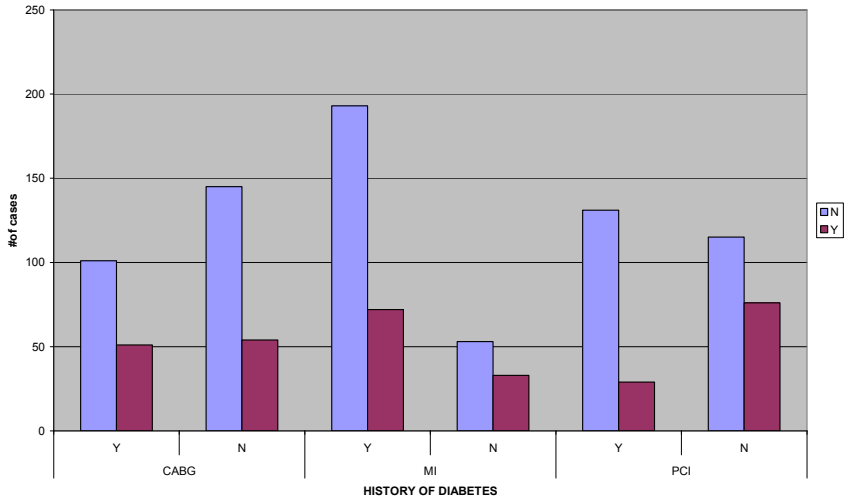
Εικόνα 35: Κατανομή ασθενών με το χαρακτηριστικό DBP στις κλάσεις



Εικόνα 36: Κατανομή ασθενών με το χαρακτηριστικό FH στις κλάσεις



Εικόνα 37: Κατανομή ασθενών με το χαρακτηριστικό HT στις κλάσεις



Εικόνα 38: Κατανομή ασθενών με το χαρακτηριστικό DM στις κλάσεις

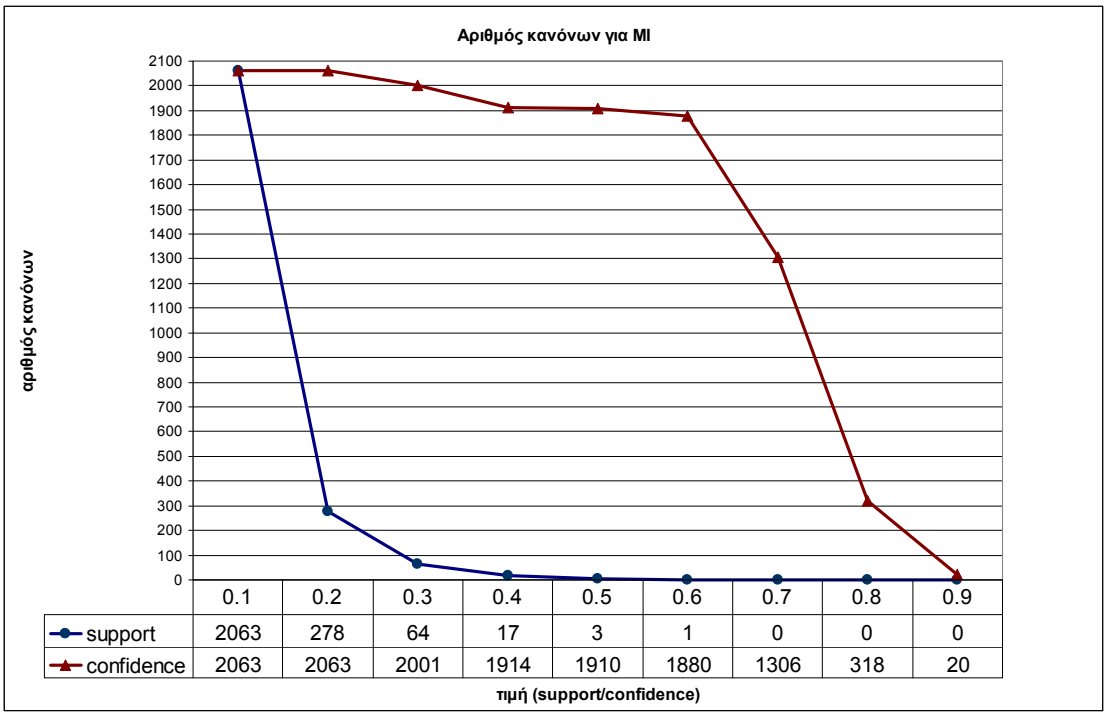
Κεφάλαιο 4

Παρουσίαση Αποτελεσμάτων

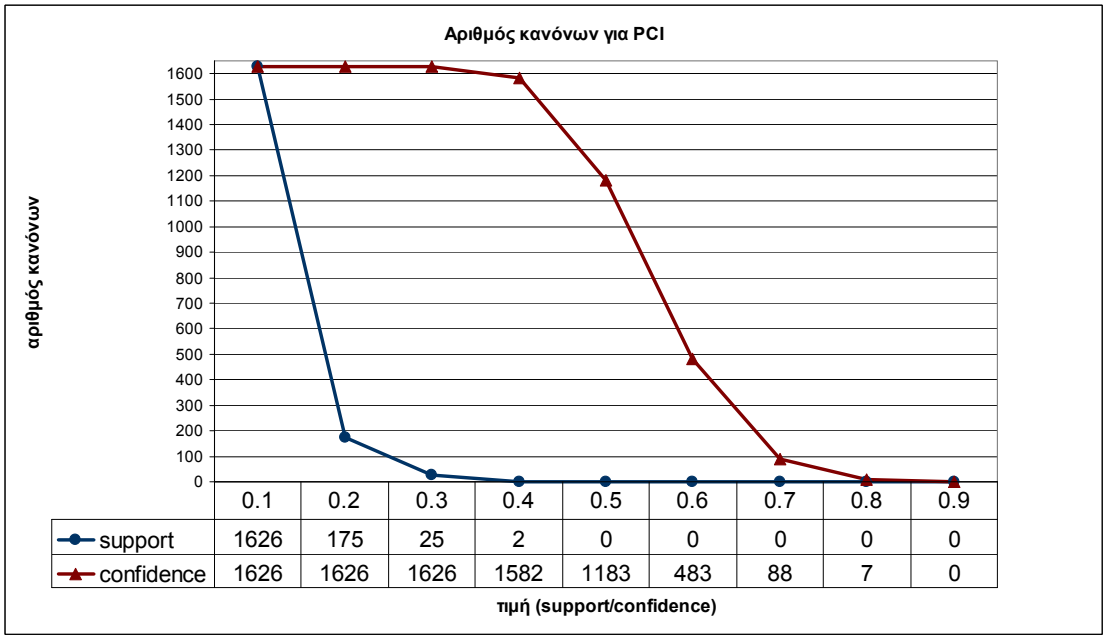
4.1. Επιλογή ελάχιστων ορίων για τα μέτρα αξιολόγησης κανόνων

Για να μπορέσουμε να πάρουμε μια γενική εικόνα της βάσης δεδομένων, αλλά και να αποφασίσουμε ποίες τιμές είναι ιδανικές για τα ελάχιστα όρια του support αλλά και των υπόλοιπων μέτρων αξιολόγησης των κανόνων συσχέτισης, εφαρμόσαμε τον αλγόριθμο Apriori με διάφορες τιμές support και confidence. Αφήνοντας σταθερό, και σε μια χαμηλή τιμή το ένα από τα δύο τρέχουμε το σύστημα με διάφορες τιμές.

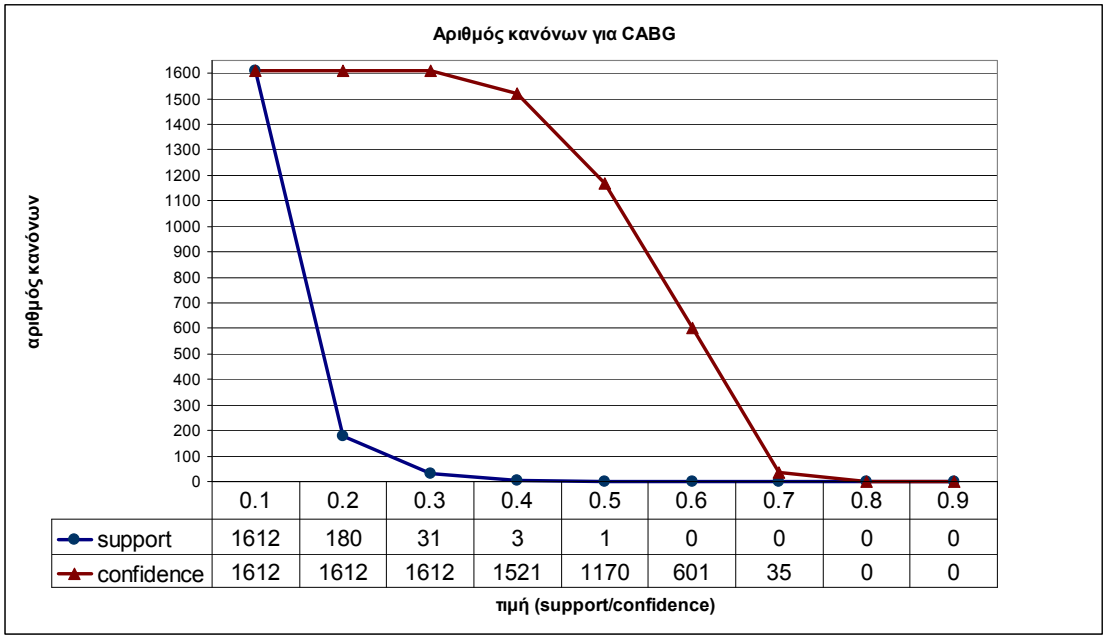
Στην Εικόνα 39 παρουσιάζονται οι αριθμοί των κανόνων για την κλάση MI. Παρατηρούμε ότι για πολύ μικρό αριθμό support και confidence, εξάγονται περίπου 2050 κανόνες, ενώ για την εξαγωγή ελάχιστον κανόνων θα πρέπει το support να αυξηθεί μέχρι και 50%. Για μεγαλύτερο από 50% κανένας κανόνας δεν εξάγεται. Ενώ παρατηρώντας των αριθμό κανόνων που εξάγονται σε σχέση με το confidence, για εξαγωγή μεγάλου αριθμού κανόνων μπορεί αν χρησιμοποιηθεί μέχρι και 60% confidence. Παρόμοια συμπεριφορά παρουσιάζεται και στις άλλες κλάσεις PCI και CABG που φαίνονται στις Εικόνες 40 και 41 αντίστοιχα. Επομένως, ανάλογα με το πόσο αριθμό κανόνων θέλει ο χρήστης να παραχθούν, θα διαλέξει το κατάλληλο όριο αριθμό support και confidence. Ένα καλό όριο για παραγωγή ελάχιστον κανόνων συσχέτισης, για όλες τις κλάσεις είναι 0.3 support και 0.5 confidence, με τα οποία μπορούν να παραχθούν περίπου 30 με 60 κανόνες.



Εικόνα 39: Αριθμός κανόνων για την κλάση MI σε σχέση με το ελάχιστο όριο support confidence



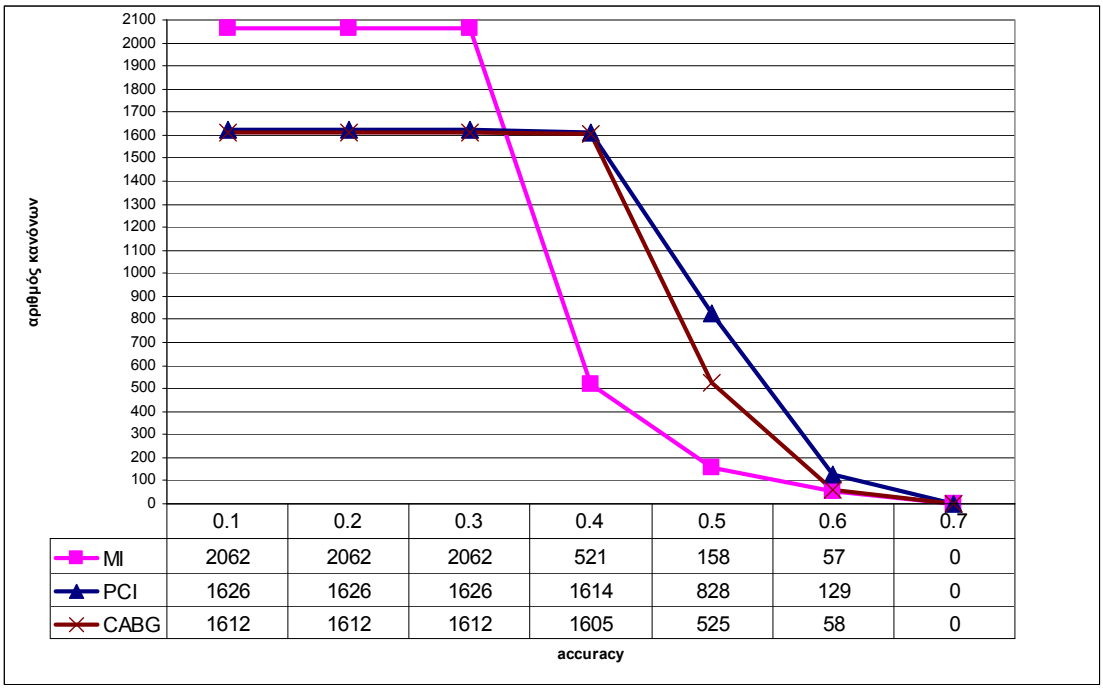
Εικόνα 40: Αριθμός κανόνων για την κλάση PCI σε σχέση με το ελάχιστο όριο support confidence



Εικόνα 41: Αριθμός κανόνων για την κλάση CABG σε σχέση με το ελάχιστο όριο support confidence

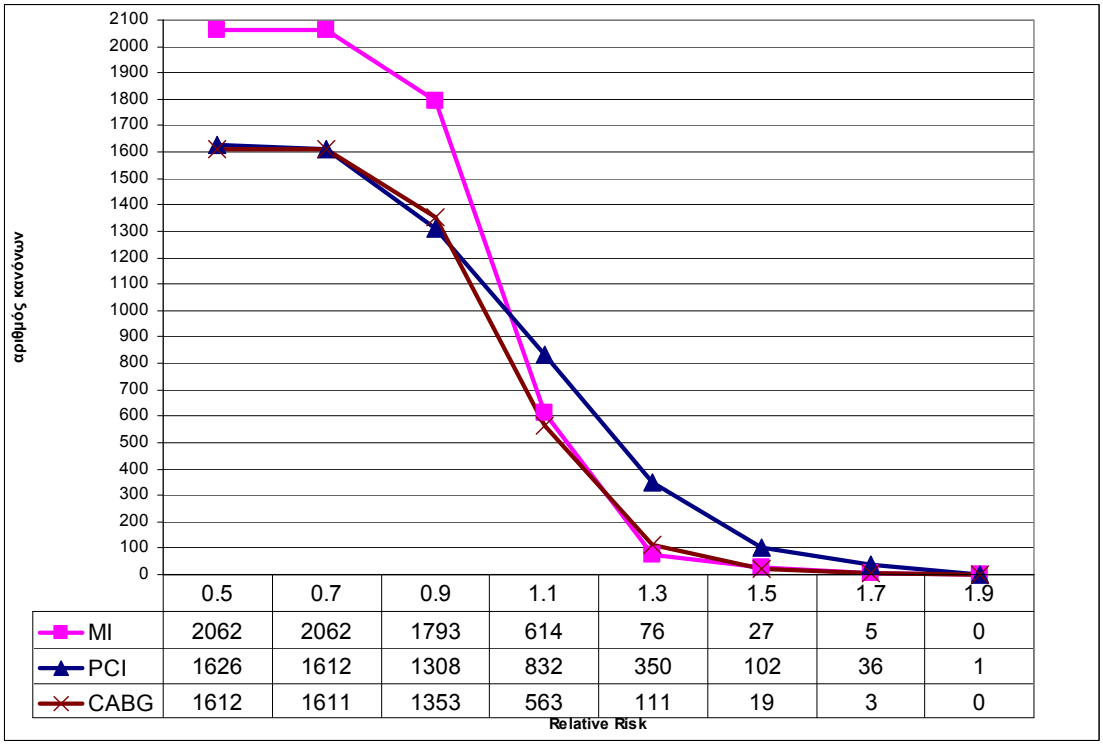
Θα εξετάσουμε επίσης, κι άλλα σημαντικά μέτρα αξιολόγησης κανόνων, όπως την Ακρίβεια (Accuracy) (Εξίσωση 7), Σχετικό Ρίσκο (Relative Risk) (Εξίσωση 11) και το Odds Ratio (Εξίσωση 13). Στην Εικόνα 41 παρουσιάζεται η κατανομή του αριθμού των κανόνων συσχέτισης σε σχέση με τις τιμές του accuracy για κάθε κλάση. Από την εικόνα μπορούμε να θεωρήσουμε ένα καλό όριο να το 0,4 για την κλάση MI, για το λόγω ότι μπορούν να παραχθούν μέχρι και 500 κανόνες, ενώ για τις άλλες κλάσεις το 0,5 θα ήταν ένα καλό όριο το οποίο αποκλείει αρκετούς κανόνες.

Στην Εικόνα 42, παρουσιάζεται η κατανομή του αριθμού των παραγόμενων κανόνων σε σχέση με το μέτρο Relative Risk για κάθε κλάση, ενώ στην Εικόνα 43, παρουσιάζεται η κατανομή του αριθμού των παραγόμενων κανόνων σε σχέση με το μέτρο Odds Ratio για κάθε κλάση. Από τις εικόνες μπορούμε να θεωρήσουμε ένα καλό ελάχιστο όριο για το Relative Risk το 1.1. Ενώ για το μέτρο αξιολόγησης, Odds Ratio μπορεί να οριστεί το 1.



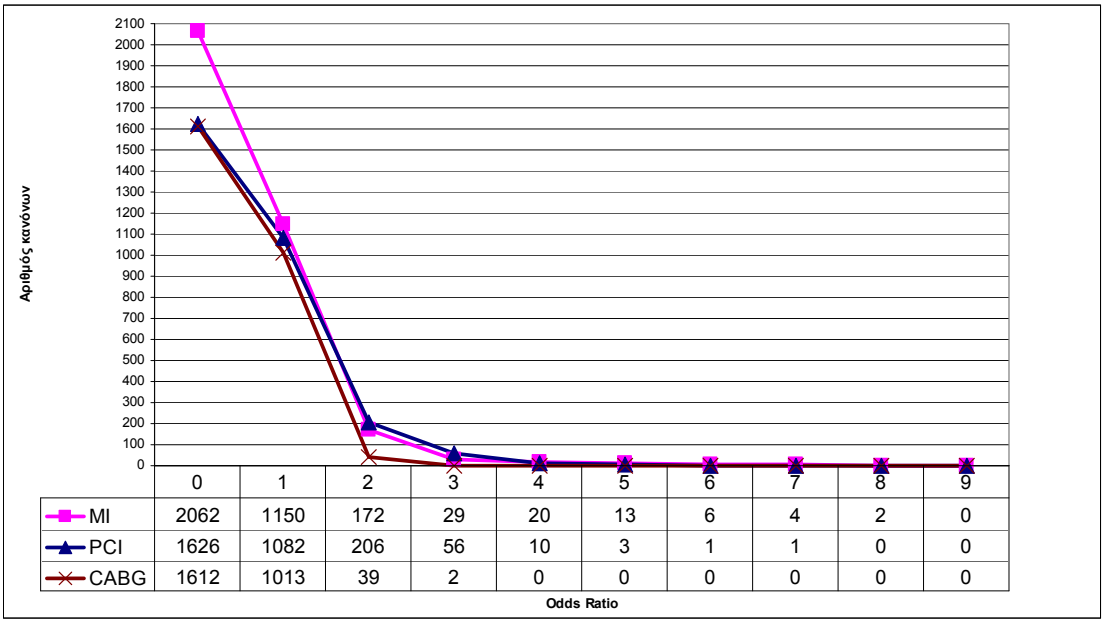
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
MI	2062	2062	2062	521	158	57	0
PCI	1626	1626	1626	1614	828	129	0
CABG	1612	1612	1612	1605	525	58	0

Εικόνα 42: Αριθμός κανόνων για τις κλάσεις σε σχέση με το ελάχιστο όριο accuracy



	0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
MI	2062	2062	1793	614	76	27	5	0
PCI	1626	1612	1308	832	350	102	36	1
CABG	1612	1611	1353	563	111	19	3	0

Εικόνα 43: Αριθμός κανόνων για τις κλάσεις σε σχέση με το ελάχιστο όριο Relative Risk



	0	1	2	3	4	5	6	7	8	9
MI	2062	1150	172	29	20	13	6	4	2	0
PCI	1626	1082	206	56	10	3	1	1	0	0
CABG	1612	1013	39	2	0	0	0	0	0	0

Εικόνα 44: Αριθμός κανόνων για τις κλάσεις σε σχέση με το ελάχιστο όριο Odds Ratio

4.2. Αποτελέσματα για την κλάση MI (έμφραγμα του μυοκαρδίου)

Ακολουθώντας την μεθοδολογία της Ανακάλυψης Γνώσης από Δεδομένα (Knowledge Discovery from Data ή KDD) (Κεφάλαιο 1.3), στην αρχή γίνεται καθαρισμός της βάσης, κι επιλογή των χαρακτηριστικών (Κεφάλαιο 3). Μετά για την εξόρυξη των κανόνων εφαρμόζεται ο αλγόριθμος Apriori, και για επιλογή των κανόνων ακολουθούνται τα πιο κάτω βήματα.

Βήμα 1: Εξόρυξη κανόνων με ψηλό support και confidence

Εφαρμόζοντας τον αλγόριθμο Apriori στην βάση δεδομένων για την κλάση MI και με ελάχιστο όριο support 0.3 και ελάχιστο όριο confidence 0.5. Επίσης θέτουμε σαν ένα όριο και τον αριθμό των χαρακτηριστικών, που θα επιθυμούμε να έχουν οι κανόνας, και το ορίζουμε να είναι τουλάχιστο 3. Οι κανόνες που εξάγονται παρουσιάζονται στο Πίνακα 14.

Πίνακας 14: Εξόρυξη Κανόνων για κλάση MI, με 0.3 support και 0.5 confidence

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
Κανόνες με τα χαρακτηριστικά φύλου και καπνίσματος SEX = M, SMBEF=Y																					
1		M	Y							N				Y	3	112	0.3	0.76	0.49	1.09	1.38
2		M	Y	D										Y	3	117	0.32	0.71	0.47	0.98	0.92
3		M	Y										N	Y	3	131	0.36	0.75	0.52	1.1	1.39
4		M	Y			N								Y	3	123	0.33	0.7	0.47	0.95	0.83
5		M	Y	D		N								Y	4	111	0.3	0.7	0.45	0.95	0.84
Κανόνες με το χαρακτηριστικό συνολικής χοληστερόλης HDL=D																					
6				D		N	N							Y	3	111	0.3	0.71	0.46	0.98	0.95
7			Y	D		N								Y	3	112	0.3	0.7	0.45	0.95	0.82
8				D		N				N				Y	3	122	0.33	0.75	0.5	1.09	1.37
9		M		D									N	Y	3	120	0.33	0.72	0.48	1.01	1.04
10				D		N							N	Y	3	131	0.36	0.73	0.51	1.04	1.14
11		M		D		N								Y	3	148	0.4	0.7	0.51	0.95	0.82
12		M		D		N							N	Y	4	115	0.31	0.72	0.47	1	1.01
Κανόνες με τα χαρακτηριστικά φύλου και καπνίσματος SEX = M																					
13		M						N					N	Y	3	116	0.31	0.75	0.49	1.07	1.3
14		M								N			N	Y	3	119	0.32	0.75	0.5	1.09	1.36
15		M				N				N				Y	3	122	0.33	0.74	0.5	1.05	1.21
16		M				N							N	Y	3	131	0.36	0.72	0.5	1	1.02

Οι κανόνες που έχουν εξαχθεί και παρουσιάζονται στον πίνακα 6, είναι τα συχνά εμφανιζόμενα σύνολα αντικειμένων με ψηλό support και confidence. Αυτά τα σύνολα αντικειμένων μπορούν να θεωρηθούν ως οι κύριοι παράγοντες που οδηγούν σε ένα καρδιαγγειακό επεισόδιο. Παρατηρούμε ότι τα πιο συχνά σύνολα αντικειμένων είναι το φύλο να είναι άνδρας (SEX = M), να είναι καπνιστής (SMBEF = Y) είτε να έχει χαμηλή ολική χοληστερόλη, και τα υπόλοιπα να είναι φυσιολογικά. Επομένως με βάση τα αποτελέσματα δεν μπορούμε να βγάλουμε κάποιο συμπέρασμα για τους παράγοντες που μπορεί να οδηγήσουν σε ένα καρδιαγγειακό επεισόδιο και χρειάζεται περισσότερη ανάλυση, με χαμηλότερο support.

Βήμα 2: Εξόρυξη κανόνων με χαμηλό support και confidence

Θέτοντας χαμηλό ελάχιστο όριο support και confidence, το εργαλείο εξάγει περίπου 2050 κανόνες (όπως φαίνεται στην Εικόνα 39). Με αυτό το τρόπο μπορεί να γίνει επιλογή κανόνων, με βάσει άλλων μέτρων αξιολόγησης κανόνων, με τα οποία μπορούν να επιλεγούν κανόνες που να δίνουν σημαντική γνώση.

Βήμα 3: Επιλογή κανόνων

Όπως έχει παρατηρηθεί στις Εικόνες 43 και 44, άλλα σημαντικά μέτρα αξιολόγησης κανόνων είναι Relative Risk με ελάχιστο όριο 1.1 και το Odds Ratio με ελάχιστο όριο 1. Επομένως, επιλέγονται οι κανόνες με τουλάχιστο 5 χαρακτηριστικά, τα οποία ικανοποιούν αυτά τα μέτρα αξιολόγησης. Οι κανόνες που εξάγονται είναι περίπου 150 κι από αυτούς επιλέγονται οι κανόνες που εξαγάγουν διαφορετική γνώση από αυτή του Πίνακα 14.

Πίνακας 15: Εξόρυξη Κανόνων για κλάση MI, με χαμηλό support

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
Κανόνες με τα χαρακτηριστικά φύλου και καπνίσματος SEX = M, SMBEF=Y																					
1		M	Y	D	M						N			Y	5	38	0.1	0.83	0.36	1.18	2.01
2		M	Y		M							N	N	Y	5	38	0.1	0.83	0.36	1.18	2.01
3		M	Y		M					N	N			Y	5	41	0.11	0.85	0.37	1.22	2.54
4		M	Y	D	M					N				Y	5	39	0.11	0.78	0.36	1.1	1.46
5		M	Y							N	Y		N	Y	5	39	0.11	0.78	0.36	1.1	1.46
6		M	Y		M	N					N			Y	5	43	0.12	0.8	0.37	1.13	1.64
7		M	Y	D	L								N	Y	5	45	0.12	0.78	0.37	1.1	1.43
8		M	Y		M						N		N	Y	5	50	0.14	0.79	0.38	1.13	1.63
9		M	Y		M					N			N	Y	5	51	0.14	0.78	0.38	1.11	1.53
10		M	Y	D		N			H					N	5	39	0.11	0.39	0.66	1.59	1.96
11		M	Y	D	M	N					N			Y	6	37	0.1	0.82	0.36	1.17	1.95
12		M	Y		M			N		N			N	Y	6	38	0.1	0.79	0.36	1.12	1.57
Κανόνες με το χαρακτηριστικό συνολικής χοληστερόλης HDL=D																					
13			Y	D	M	N					N			Y	5	37	0.1	0.82	0.36	1.17	1.95
14			Y	D	M			N					N	Y	5	37	0.1	0.79	0.36	1.11	1.53
15		M		D	L	N		N						Y	5	38	0.1	0.79	0.36	1.12	1.57
16				D	M	N	N				N			Y	5	39	0.11	0.8	0.36	1.13	1.62
17			Y	D	M	N				N				Y	5	39	0.11	0.78	0.36	1.1	1.46
18				D	L	N		N					N	Y	5	40	0.11	0.8	0.36	1.13	1.67
19		M		D	L					N			N	Y	5	42	0.11	0.81	0.37	1.15	1.77
20				D	M	N					N		N	Y	5	42	0.11	0.79	0.37	1.12	1.59
21				D	L	N				N			N	Y	5	47	0.13	0.81	0.38	1.16	1.82
22		M		D	M	N					N			Y	5	47	0.13	0.78	0.37	1.11	1.51
23		M		D		N			H			Y		N	5	39	0.11	0.39	0.66	1.64	2.05
24		M		D	L	N				N			N	Y	6	41	0.11	0.8	0.37	1.14	1.72

Κανόνες με το χαρακτηριστικό φύλου, SEX = M																				
25		M			M	N				N	N			Y	5	37	0.1	0.8	0.36	1.14	1.71
26		M			L	N		N					N	Y	5	39	0.11	0.8	0.36	1.13	1.62
27		M						N		N	Y		N	Y	5	40	0.11	0.8	0.36	1.13	1.67
28		M			L	N				N			N	Y	5	44	0.12	0.8	0.37	1.14	1.68
Κανόνες με το χαρακτηριστικό καπνίσματος SMBEF= Y																				
29			Y		M			N			N		N	Y	5	37	0.1	0.8	0.36	1.14	1.71
30			Y		M			N		N			N	Y	5	39	0.11	0.8	0.36	1.13	1.62

Πίνακας 16: Κανόνες για κλάση MI, όπου ισχύει ο παράγοντας ψηλή πίεση (SBP = H)

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
Κανόνες με το χαρακτηριστικό ψηλή πίεση SBP = H																					
1		M	Y	D		N			H					N	5	39	0.11	0.39	0.66	1.59	1.96
2		M	Y	D		N			H					Y	5	62	0.17	0.61	0.34	0.81	0.51
3		M		D		N			H			Y		N	5	39	0.11	0.39	0.66	1.64	2.05
4		M		D		N			H			Y		Y	5	60	0.16	0.61	0.34	0.8	0.49

Στον Πίνακα 15 παρουσιάζονται οι κανόνες που έχουν επιλεγεί. Από τους κανόνες, παρατηρείται ότι ένας άλλος σημαντικός παράγοντας που μπορεί να οδηγήσει σε έμφραγμα του μυοκαρδίου είναι η καλή χοληστερόλη (HDL) που είτε αν είναι χαμηλή (L) είτε σε κανονικά επίπεδα (M) μπορεί να επηρεάσει αρνητικά. Για τους άνδρες, επηρεάζει επίσης, και το οικογενειακό ιστορικό (FH).

Ενώ οι ασθενείς της βάσης, που δεν υπέστησαν έμφραγμα του μυοκαρδίου, όπως έχει παρατηρηθεί από τον Πίνακα 16, εμφάνιζαν υψηλή συστολική πίεση (SBP = H). Αυτός ο παράγοντας επηρεάζεται το ίδιο είτε ο ασθενής είναι καπνιστής (SMBEF = Y) είτε παρουσιάζει ψηλή υπέρταση (HT=Y).

4.3. Αποτελέσματα για την κλάση PCI (αγγειοπλασική) και CABG (παράκαμψη (bypass))

Ακολουθώντας τα βήματα του προηγούμενου κεφαλαίου 4.2, εφαρμόζεται ο αλγόριθμος Apriori για τις κλάσεις PCI και CABG.

Βήμα 1: Εξόρυξη κανόνων με ψηλό support και confidence

Σύμφωνα την γραφική παράσταση της Εικόνας 40 και 41, εφαρμόζεται ο αλγόριθμος Apriori στην βάση δεδομένων για την κλάση PCI και CABG, με ελάχιστο όριο support 0.3 και ελάχιστο όριο confidence 0.5. Οι κανόνες που εξάγονται οι κανόνες παρουσιάζονται στους Πίνακες 17 και 18 αντίστοιχα.

Πίνακας 17: Εξόρυξη Κανόνων για κλάση PCI, με 0.3 support και 0.5 confidence

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
1							N							N	1	120	0.33	0.6	0.54	1.13	1.32
2											N			N	1	134	0.36	0.64	0.59	1.37	2.02
3									H					N	1	125	0.34	0.54	0.48	0.88	0.74
4										N				N	1	122	0.33	0.52	0.46	0.8	0.58
5												Y		N	1	140	0.38	0.59	0.55	1.14	1.35
6			Y											N	1	132	0.36	0.54	0.49	0.89	0.76
7				D										N	1	144	0.39	0.57	0.53	1.04	1.1
8													N	N	1	133	0.36	0.5	0.44	0.7	0.39
9						N								N	1	159	0.43	0.56	0.53	0.95	0.89
10		M												N	1	179	0.49	0.56	0.54	0.94	0.85
11		M									N			N	2	116	0.31	0.64	0.57	1.31	1.87
12		M										Y		N	2	113	0.31	0.59	0.53	1.1	1.24
13		M		D										N	2	124	0.34	0.57	0.51	1.01	1.02
14		M											N	N	2	117	0.32	0.5	0.44	0.74	0.48
15		M	Y											N	2	129	0.35	0.54	0.49	0.9	0.78
16		M				N								N	2	133	0.36	0.55	0.5	0.93	0.84
17				D		N								N	2	140	0.38	0.58	0.53	1.05	1.12
18		M		D		N								N	3	120	0.33	0.57	0.51	1.01	1.02

Πίνακας 18: Εξόρυξη Κανόνων για κλάση CABG, με 0.3 support και 0.5 confidence

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
1								N						N	1	112	0.3	0.59	0.51	1.02	1.04
2					M									N	1	119	0.32	0.62	0.53	1.11	1.28
3							N							N	1	118	0.32	0.59	0.51	1	0.99
4											N			N	1	125	0.34	0.6	0.52	1.04	1.1
5									H					N	1	129	0.35	0.56	0.48	0.87	0.7
6										N				N	1	138	0.37	0.59	0.52	1	0.99
7												Y		N	1	129	0.35	0.55	0.47	0.83	0.62
8			Y											N	1	145	0.39	0.6	0.54	1.04	1.11
9				D										N	1	140	0.38	0.56	0.49	0.85	0.67
10													N	N	1	163	0.44	0.62	0.58	1.2	1.52
11						N								N	1	161	0.44	0.57	0.51	0.86	0.68
12		M												N	1	188	0.51	0.59	0.57	1.02	1.04
13		M			M									N	2	111	0.3	0.62	0.53	1.12	1.33
14			Y										N	N	2	114	0.31	0.64	0.55	1.19	1.52
15				D									N	N	2	115	0.31	0.62	0.53	1.11	1.29
16		M								N				N	2	121	0.33	0.58	0.51	0.99	0.97
17						N							N	N	2	129	0.35	0.61	0.54	1.11	1.28
18		M		D										N	2	119	0.32	0.55	0.47	0.84	0.65
19		M											N	N	2	145	0.39	0.62	0.57	1.18	1.46
20		M	Y											N	2	141	0.38	0.59	0.53	1.03	1.08
21		M				N								N	2	137	0.37	0.57	0.5	0.91	0.79
22				D		N								N	2	136	0.37	0.56	0.49	0.87	0.71
23		M	Y										N	N	3	111	0.3	0.64	0.54	1.17	1.48
24				D		N							N	N	3	111	0.3	0.62	0.53	1.11	1.29
25		M				N							N	N	3	113	0.31	0.62	0.53	1.12	1.31
26		M		D		N								N	3	116	0.31	0.55	0.47	0.86	0.69

Βήμα 2: Εξόρυξη κανόνων με χαμηλό support και confidence

Θέτοντας χαμηλό ελάχιστο όριο support και confidence, το εργαλείο εξάγει περίπου 1600 κανόνες για κάθε κλάση (όπως παρουσιάζονται στις Εικόνες 43 και 44). Έτσι μπορεί να γίνει επιλογή κανόνων με άλλα μέτρα αξιολόγησης κανόνων, με τα οποία μπορούν να επιλεγούν κανόνες που να δίνουν σημαντική γνώση.

Βήμα 3: Επιλογή κανόνων

Επιλέγονται οι κανόνες που ικανοποιούν το ελάχιστο όριο 1.1 για Relative Risk, 1 για το Odds Ratio και το 0.6 για το accuracy. Επίσης επιλέγονται οι κανόνες με τουλάχιστο 5 χαρακτηριστικά, τα οποία ικανοποιούν αυτά τα μέτρα αξιολόγησης. Ο Πίνακας 19 παρουσιάζει τα αποτελέσματα για την κλάση PCI, κι ο πίνακας 13 τα αποτελέσματα της κλάσης CABG.

Από τα αποτελέσματα του Πίνακα 19 παρατηρείται ότι ένας από τους πιο σημαντικούς παράγοντες που μπορεί να οδηγήσουν σε αγγειοπλαστική είναι η υψηλή συστολική πίεση (SBP = H), ενώ ο ασθενής δεν θα πρέπει να πάσχει από διαβήτη (DM = N), όπως επίσης κι από χαμηλή. Παρατηρώντας τους ασθενής που δεν έχουν υποβληθεί σε αγγειοπλαστική (Πίνακας 20) παρατηρείται ότι παρουσίαζαν ψηλή ποσότητας γλυκόζης (GLU = H) όπως επίσης και υπέρταση (HT = Y).

Ενώ παρατηρείται ότι, οι ασθενείς που έχουν υποβληθεί σε εγχείρηση παράκαμψη (bypass), σύμφωνα με τα αποτελέσματα του πίνακα 21, παρουσίαζαν ψηλή συστολική πίεση (SBP = Y) και υπέρταση.

Πίνακας 19: Εξόρυξη Κανόνων για κλάση PCI, με χαμηλό support

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
Κανόνες με τα χαρακτηριστικά φύλου και καπνίσματος SEX = M, SMBEF=Y																					
1		M	Y			N					Y		N	Y	5	38	0.1	0.67	0.62	1.7	3.11
2		M	Y		M					N			N	Y	5	41	0.11	0.63	0.61	1.61	2.66
3		M	Y						H	N			N	Y	5	39	0.11	0.59	0.6	1.48	2.17
4		M	Y			N		N		N				Y	5	40	0.11	0.59	0.6	1.48	2.15
5		M	Y					N		N			N	Y	5	45	0.12	0.58	0.6	1.46	2.09
6		M	Y			N				N			N	Y	5	54	0.15	0.58	0.61	1.51	2.22
7		M	Y			N		N		N			N	Y	6	40	0.11	0.6	0.6	1.5	2.25
Κανόνες με το χαρακτηριστικό συνολικής χοληστερόλης HDL=D																					
8				D		N		N			Y		N	Y	5	37	0.1	0.73	0.63	1.88	4.19
9				D		N			H		Y		N	Y	5	37	0.1	0.66	0.62	1.68	3.01
10				D		N				N	Y		N	Y	5	39	0.11	0.66	0.62	1.69	3.05
11		M		D	M					N			N	Y	5	39	0.11	0.61	0.6	1.54	2.37
12				D	M	N				N			N	Y	5	41	0.11	0.61	0.61	1.55	2.43
13		M		D					H	N			N	Y	5	40	0.11	0.58	0.6	1.45	2.07
14		M		D		N					Y		N	Y	5	47	0.13	0.65	0.63	1.72	3.06
15				D		N			H	N			N	Y	5	45	0.12	0.58	0.6	1.48	2.16
16		M		D		N				N			N	Y	5	63	0.17	0.55	0.6	1.43	1.96
17		M		D	M	N				N			N	Y	6	38	0.1	0.61	0.6	1.54	2.4
18		M		D		N			H	N			N	Y	6	40	0.11	0.58	0.6	1.45	2.07
Κανόνες με τα χαρακτηριστικά φύλου και καπνίσματος SEX = M																					
19		M				N		N			Y		N	Y	5	39	0.11	0.72	0.63	1.88	4.17
20		M				N				N	Y		N	Y	5	41	0.11	0.68	0.63	1.77	3.45
21		M			M			N		N			N	Y	5	37	0.1	0.59	0.6	1.46	2.12
22		M							H	N		Y	N	Y	5	37	0.1	0.59	0.6	1.46	2.12
23		M				N	N			N			N	Y	5	41	0.11	0.59	0.6	1.47	2.14
24		M			M	N				N			N	Y	5	46	0.12	0.64	0.62	1.66	2.84
25		M				N			H	N			N	Y	5	45	0.12	0.58	0.6	1.46	2.09
Κανόνες με το χαρακτηριστικό καπνίσματος SMBEF= Y																					
26			Y			N		N		N			N	Y	5	41	0.11	0.6	0.6	1.53	2.32

Πίνακας 20: Εξόρυξη Κανόνων για κλάση PCI = N, με χαμηλό support

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
Κανόνες με τα χαρακτηριστικά υψηλή γλυκόζη και υπερταση GLU= H, HT=Y																					
1				D		N	N	H				Y		N	5	38	0.1	0.73	0.5	1.35	2.32
2			Y	D		N		H				Y		N	5	38	0.1	0.72	0.5	1.32	2.15
3		M	Y			N		H				Y		N	5	38	0.1	0.68	0.49	1.24	1.75
4		M	Y	D				H				Y		N	5	40	0.11	0.71	0.5	1.32	2.13
5		M		D				H	H			Y		N	5	39	0.11	0.66	0.49	1.21	1.61
6		M				N		H	H			Y		N	5	42	0.11	0.65	0.49	1.18	1.5
7		M	Y					H	H			Y		N	5	43	0.12	0.65	0.49	1.19	1.54
8		M		D		N		H				Y		N	5	45	0.12	0.67	0.5	1.24	1.72
9				D		N		H	H			Y		N	5	45	0.12	0.66	0.49	1.21	1.63
10		M	Y	D		N		H				Y		N	6	38	0.1	0.72	0.5	1.32	2.15
11		M		D		N		H	H			Y		N	6	38	0.1	0.67	0.49	1.22	1.65
Κανόνες με το χαρακτηριστικό υψηλή συστολική πίεση SBP = H																					
12		M			L	N			H			Y		N	5	38	0.1	0.64	0.48	1.17	1.47
13		M	Y						H		N	Y		N	5	39	0.11	0.66	0.49	1.21	1.61
14			Y	D		N		H	H					N	5	38	0.1	0.63	0.48	1.14	1.39
15		M	Y			N		H	H					N	5	38	0.1	0.62	0.47	1.12	1.32
16				D	L	N			H			Y		N	5	38	0.1	0.61	0.47	1.1	1.26
17		M	Y	D				H	H					N	5	39	0.11	0.63	0.48	1.14	1.37
18				D		N	N		H			Y		N	5	43	0.12	0.62	0.48	1.13	1.34
19		M		D		N			H			Y		N	5	60	0.16	0.61	0.49	1.1	1.25
20		M	Y	D		N		H	H					N	6	38	0.1	0.63	0.48	1.14	1.39
21		M	Y	D		N			H			Y		N	6	48	0.13	0.62	0.48	1.11	1.29
Κανόνες με το χαρακτηριστικό υπέρταση HT=Y																					
22		M		D		N					N	Y		N	5	40	0.11	0.71	0.5	1.32	2.13
23		M	Y	D			N					Y		N	5	37	0.1	0.65	0.48	1.18	1.51
24		M		D	L	N						Y		N	5	42	0.11	0.64	0.48	1.15	1.43
25		M		D		N	N					Y		N	5	48	0.13	0.63	0.49	1.15	1.41
26		M	Y	D		N						Y		N	5	61	0.17	0.62	0.5	1.14	1.37
Κανόνες με το χαρακτηριστικό γλυκόζη GLU = H																					

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
27		M		D		N	N	H						N	5	42	0.11	0.67	0.49	1.22	1.66
28		M	Y	D		N		H						N	5	50	0.14	0.66	0.5	1.21	1.62

Πίνακας 21: Εξόρυξη Κανόνων για κλάση CABG, με χαμηλό support

Rule no.	AGE	SEX	SMBEF	TC	HDL	LDL	TG	GLU	SBP	DBP	FH	HT	DM	Class	Num Of Attributes	Support Count	Support	Confidence	Accuracy	Relative Risk	Odd Ratio
Κανόνες με τα χαρακτηριστικά υψηλή συστολική πίεση και υπερταση SBP= H, HT=Y																					
1		M				N			H	N		Y		Y	5	37	0.1	0.52	0.6	1.35	1.73
2			Y	D		N			H			Y		Y	5	43	0.12	0.54	0.61	1.45	1.98
3		M	Y	D					H			Y		Y	5	45	0.12	0.56	0.61	1.5	2.11
4		M	Y			N			H			Y		Y	5	44	0.12	0.52	0.6	1.36	1.75
5		M		D		N			H			Y		Y	5	54	0.15	0.55	0.61	1.5	2.11
6		M	Y	D		N			H			Y		Y	6	43	0.12	0.55	0.61	1.47	2.05
Κανόνες με το χαρακτηριστικό υψηλή συστολική πίεση SBP = H																					
7		M		D	L	N			H					Y	5	37	0.1	0.53	0.6	1.37	1.79
8		M		D		N		H	H					Y	5	40	0.11	0.53	0.6	1.38	1.8
9		M		D		N			H	N				Y	5	43	0.12	0.49	0.58	1.26	1.51
10		M	Y	D		N			H					Y	5	52	0.14	0.51	0.6	1.38	1.78
Κανόνες με το χαρακτηριστικό υπέρταση HT=Y																					
11		M		D	L	N						Y		Y	5	37	0.1	0.56	0.61	1.48	2.09
12		M		D	L	N						Y		Y	5	37	0.1	0.56	0.61	1.48	2.09
13		M		D		N	N					Y		Y	5	37	0.1	0.49	0.58	1.24	1.47
14		M		D		N				N		Y		Y	5	42	0.11	0.51	0.59	1.32	1.64
15		M	Y	D		N						Y		Y	5	52	0.14	0.53	0.6	1.44	1.93
16		M		D		N			H			Y		Y	5	54	0.15	0.55	0.61	1.5	2.11
Κανόνες με τα χαρακτηριστικά φύλο, κάπνισμα και ολική χοληστερόλη SEX = M, SMBEF = Y , TC = D																					
17		M	Y	D	L	N								Y	5	39	0.11	0.49	0.59	1.27	1.53
18		M	Y	D		N	N							Y	5	44	0.12	0.47	0.57	1.19	1.36

Κεφάλαιο 5

Συμπεράσματα και Μελλοντική Μελέτη

5.1. Συμπεράσματα για την βάση δεδομένων

Εξετάζοντας τα αποτελέσματα του Κεφαλαίου 4, παρατηρήσαμε τους πιο σημαντικούς παράγοντες που μπορεί να οδηγήσουν σε καρδιαγγειακά επεισόδια. Σύμφωνα με τους ασθενείς της βάσης που μας έχει δοθεί, ο πιο σημαντικός παράγοντας που μπορεί να έχει προκαλέσει έμφραγμα, είναι το κάπνισμα. Ασθενείς με χαμηλή περιεκτικότητα καλής χοληστερόλης στο αίμα, είναι επίσης κρίσιμο για πρόκληση εμφράγματος. Ενώ ξέροντας ότι ένας σημαντικός παράγοντας είναι η ψηλή περιεκτικότητα σε τριγλυκερίδια, αυτό δεν επαληθεύεται σύμφωνα με τα αποτελέσματα.

Είναι επίσης σημαντικό να τονιστεί ότι, ο κάθε παράγοντας που μπορεί να οδηγήσει σε καρδιαγγειακό επεισόδιο, όπως το κάπνισμα και χαμηλή καλή χοληστερόλη είναι ανεξάρτητοι μεταξύ τους. Δηλαδή καρδιαγγειακό επεισόδιο μπορεί να προκληθεί έστω κι αν μόνο ένας παράγοντας παρουσιάζεται. Αυτό μπορεί να το συμπεράνουμε από το ότι, οι κανόνες με που εμφανίζονται συχνά στην βάση, με ψηλό support, δεν παρουσίαζαν κάποιο πρόβλημα (Πίνακας 14). Αυτό δείχνει ότι όλοι οι παράγοντες είναι σημαντικοί, κι ανεξάρτητοι μεταξύ τους.

Οι ασθενείς με καρδιαγγειακά επεισόδια που υποβάλλονται σε αγγειοπλαστική παρουσιάζουν υψηλή συστολική πίεση και δεν πάσχουν από διαβήτη. Δεν υποβάλλονται σε αγγειοπλαστική οι ασθενής που παρουσιάζουν ψηλή γλυκόζη και υπέρταση. Αυτοί οι ασθενής υποβάλλονται σε εγχείρηση παράκαμψη (bypass), αφού παρατηρώντας τα

αποτελέσματα για τους ασθενείς που έχουν υποβληθεί σε εγχείρησης παράκαμψη (bypass), οι παράγοντες που οδηγούν σε τέτοια εγχείρηση είναι να έχει ψηλή συστολική πίεση και υπέρταση.

5.2. Συμπεράσματα για τους αλγόριθμους εξόρυξης κανόνων συσχέτισης

Με βάση την θεωρία του αλγόριθμου Apriori, οι σημαντικοί κανόνες είναι αυτοί που εμφανίζονται πιο συχνά, δηλαδή έχουν ψηλό support. Στην βάση δεδομένων που μας έχει δοθεί, κι έχει εφαρμοστεί ο αλγόριθμος Apriori, έχει παρατηρηθεί ότι οι κανόνες με ψηλό support δεν μας πρόσφεραν κάποια σημαντική γνώση. Επομένως ανάλογα με την φύση της βάσης δεδομένων θα πρέπει να επιλεγούν οι κανόνες συσχέτισης, με το κατάλληλο μέτρο αξιολόγησης κανόνων. Ο αλγόριθμος Akamas, δεν στηρίζεται στο μέτρο αξιολόγησης support, επομένως σε μια τέτοια βάση δεδομένων, είναι ένας καλός αλγόριθμος εξαγωγής κανόνων συσχέτισης.

5.3. Μελλοντική Μελέτη

Για το σκοπό της μελέτης έχει αναπτυχθεί ένα εργαλείο εξόρυξης δεδομένων το οποίο στηρίζεται μόνο σε αλγόριθμους εξόρυξης κανόνων συσχέτισης (αλγόριθμο Apriori και Akama). Το φιλτράρισμα και ταξινόμηση των κανόνων συσχέτισης βασίζεται σε οποιοδήποτε μέτρο αξιολόγησης κανόνων κι όχι μόνο στο support και το confidence, που έχει αποδειχτεί ότι είναι πολύ σημαντικό, ειδικά στη βάση δεδομένων που αναλύσαμε.

Όπως έχουμε αναφέρει στο Κεφάλαιο 1.5 (Εικόνα 2), τα διάφορα μέτρα αξιολόγηση κανόνων μπορούν να χρησιμοποιηθούν για το φιλτράρισμα, την ταξινόμηση των κανόνων, αλλά και στους αλγόριθμους εξόρυξης δεδομένων. Στο μέλλον θα μπορούσε το εργαλείο να

αναβαθμιστεί έτσι ώστε οι αλγόριθμοι εξόρυξης κανόνων, να χρησιμοποιούν διάφορα μέτρα αξιολόγησης, για την εξαγωγή των κανόνων.

Ο αλγόριθμος Akamas έχει αποδειχθεί ένας καλός αλγόριθμος εξόρυξης κανόνων συσχέτισης επειδή δεν βασίζεται μόνο στο support. Το πρόβλημα του αλγόριθμου είναι ότι κτίζει όλους του δυνατούς συνδυασμούς, που τον καθιστούν πολύ αργό. Έτσι θα πρέπει να γίνει βελτίωση στην απόδοση του αλγόριθμου Akama.

Βιβλιογραφία

[1] ΣΥΛΒΙΑ ΚΑΡΑΚΑΤΣΑΝΗ, «Οι θάνατοι από στεφανιαία νόσο μπορούν να μειωθούν στο ήμισυ», *Εφημερίδα Σημερινή, 05/04/2009*, http://www.sigmalive.com/simerini/news/health/140189.

[2] ΣΥΛΒΙΑ ΚΑΡΑΚΑΤΣΑΝΗ, «Οκτακόσιοι αιφνίδιοι καρδιακοί θάνατοι», *Εφημερίδα Σημερινή, 19/11/2008*, http://www.sigmalive.com/news/local/87157.

[3] JIAWEI HAN, MICHELINE KAMBER, "Data Mining, Concepts and Techniques", *Morgan Kaufman Publishers, Academic Press 2001.*

[4] Μ. ΒΑΖΙΡΓΙΑΝΝΗΣ, Μ. ΧΑΛΚΙΔΙΚΗ, "Εξόρυξη Γνώσης από Βάσης Δεδομένων", *Γιώργος Δαρδάνος, Αθήνα 2003.*

[5] LIQIANG GENG, HOWARD J. HAMILTON, "Interestingness Measures for Data Mining: A Survey", *ACM Computing Surveys, Vol. 38, No. 3, Article 9, September 2006.*

[6] Y. BASTIDE, N. PASQUIER, R. TAOUIL, G. STUMME, L. LAKHAL, "Mining minimal non-redundant association rules using frequent closed itemsets." In *Proceedings of the First international Conference on Computational Logic, Lecture Notes In Computer Science, vol. 1861. Springer-Verlag, London, 2000.*

[7] E. KNORR, NG RAYMOND, V. TUCAKOV, "Distance based outliers: Algorithms and applications", *J. Very Large Databases 8, 237–253. 2000.*

[8] R. HILDERMAN, J. HAMILTON, “Knowledge Discovery and Measures of Interest”, *Kluwer Academic, Boston, MA. 200.*

[9] R. AGRAWAL, R. SRIKANT, “Fast Algorithms for Mining Association Rules”, *IBM Almaden Research Center 650 Harry Road, San Jose, CA 9512, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994*

[10] Weka 3: Data Mining Software in Java, *WEKA the University of Waikato*, http://www.cs.waikato.ac.nz/~ml/weka/

[11] Leverage (statistics): From Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Leverage_(statistics), 01/06/2009

[12] Added value: From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Added_Value, 01/06/2009

[13] Relative risk: From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Relative_risk, 01/06/2009

[14] Jaccard index: From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Jaccard_index, 01/06/2009

[15] Odds ratio: From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Odds_ratio, 01/06/2009

ΠΑΡΑΡΤΗΜΑ Ι

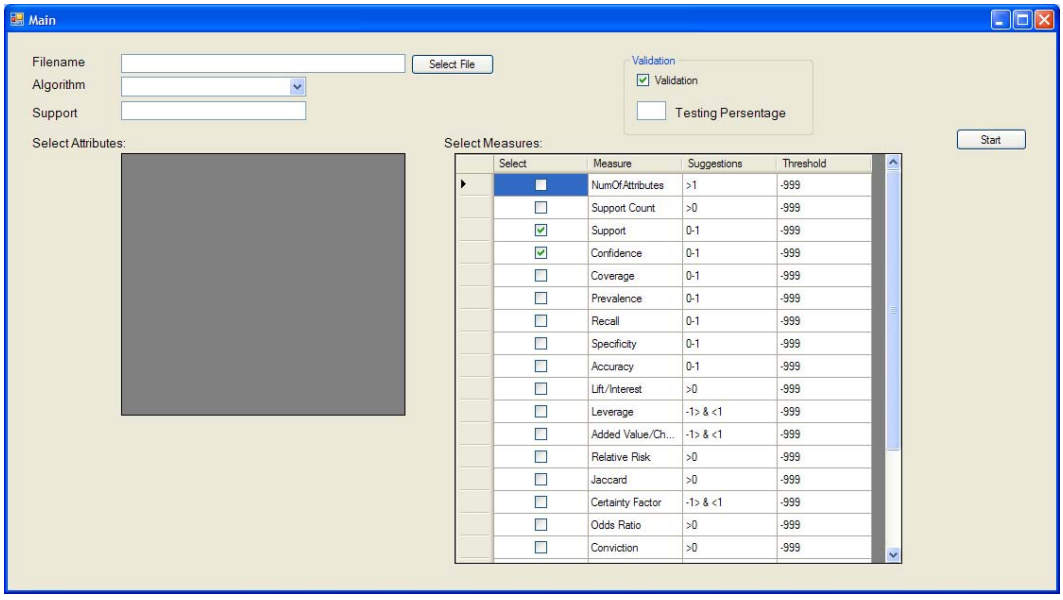
Περιγραφή Εργαλείου

1. Πλατφόρμα Υλοποίησης

Το εργαλείο εξόρυξης κανόνων συσχέτισης έχει υλοποιηθεί σε πλατφόρμα Microsoft Visual C#. Έτσι το εργαλείο μπορεί εύκολα να εγκατασταθεί και να λειτουργήσει σε οποιοδήποτε λειτουργικό σύστημα.

2. Περιγραφή Οθόνης Εισόδου

Στην Εικόνα 45 παρουσιάζεται η Οθόνη Εισόδου του συστήματος. Ο χρήστης καλείται να συμπληρώσει τα κατάλληλα πεδία για να μπορεί να γίνει το η ανάλυση της βάσης δεδομένων.



Εικόνα 45: Οθόνη για εισαγωγής δεδομένων στο σύστημα

Ο χρήστης θα πρέπει να ακολουθήσει τα πιο κάτω βήματα.

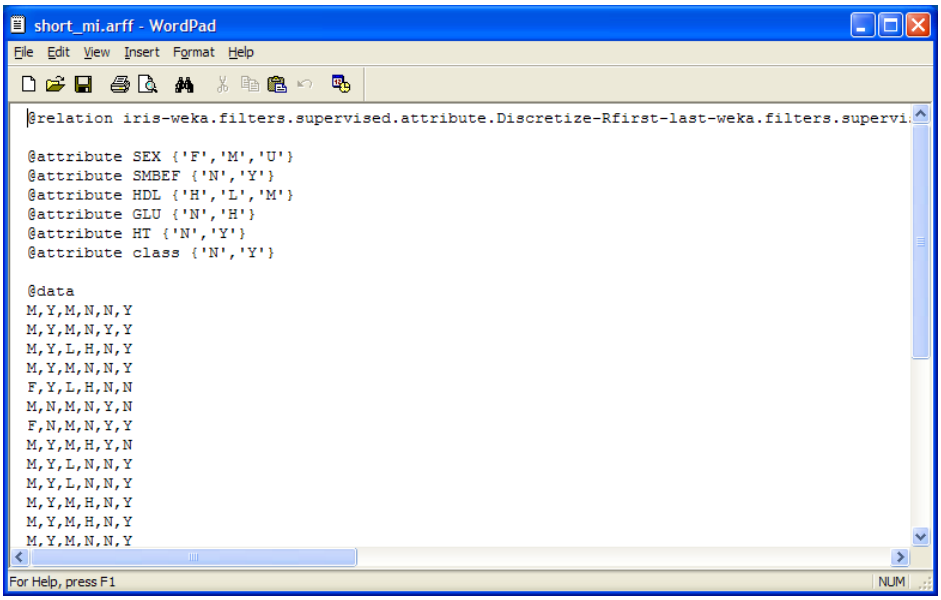
2.1. Είσοδος Αρχείο

Το αρχείο που περιέχει την βάση δεδομένων, θα πρέπει να είναι σε μορφή .arff. Θα πρέπει να έχει την μορφή όπως φαίνεται στην Εικόνα 46.

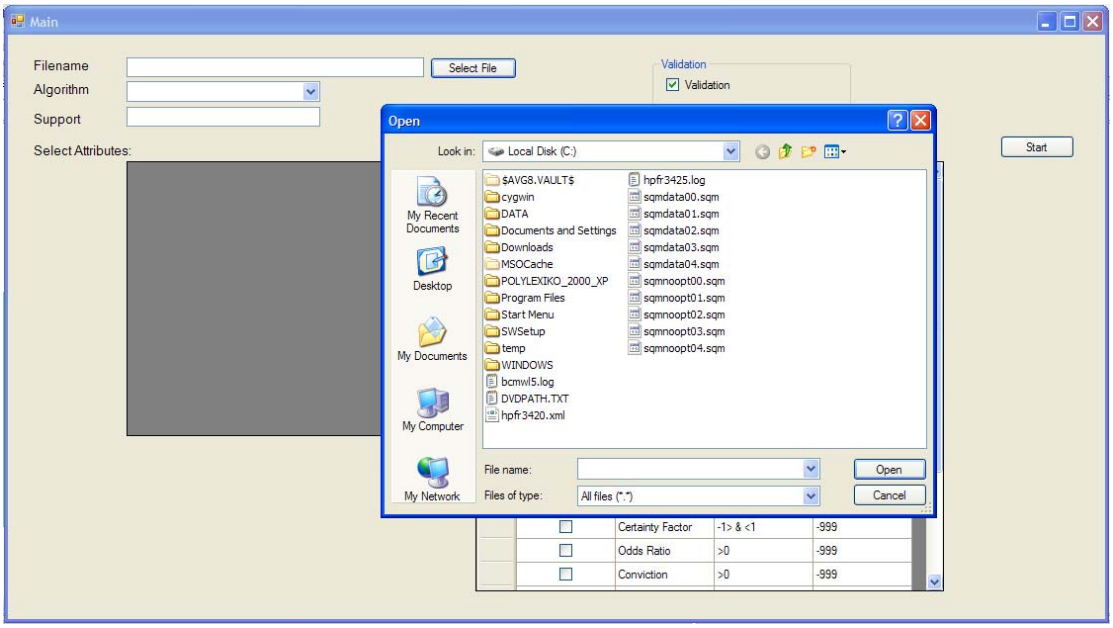
Τα πεδία της βάσης δεδομένων θα πρέπει να καθορίζονται στις πρώτες γραμμές. Ορίζονται από την αρχική λέξη @attribute, μετά ακολουθεί το όνομα τους και στο τέλος μέσα σε αγκύλες «{» οι διάφορες τιμές τους.

Μετά ακολουθούν τα δεδομένα (οι δοσοληψίες) της βάσης δεδομένων. Για να καθοριστεί ότι αρχίζουν τα δεδομένα θα πρέπει να γίνει εισαγωγή της λέξης @data στην προηγούμενη γραμμή. Η δοσοληψίες αντιπροσωπεύονται από μία πλειάδα τιμών για κάθε πεδίο, που θα πρέπει να είναι στη σωστή σειρά.

Για την εισαγωγή του αρχείου ο χρήστης θα πρέπει να επιλέξει το «Select» και μετά να επιλέξει το αρχείο με την βάση δεδομένων (Εικόνα 47).



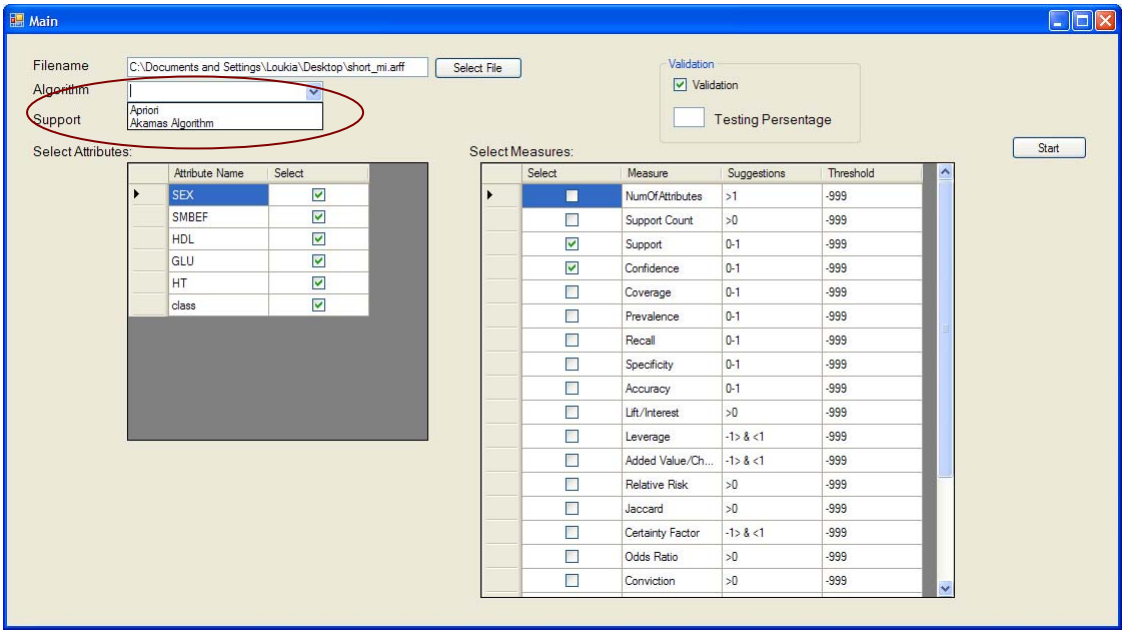
Εικόνα 46: Αρχείο σε μορφή .arff



Εικόνα 47: Επιλογή αρχείου

2.2. Επιλογή αλγόριθμου

Αφού γίνει η εισαγωγή του αρχείου ο χρήστης μπορεί να επιλέξει ένα από τους αλγόριθμους Apriori και Akama, από την λίστα του πεδίου «Algorithm» (Εικόνα 48).

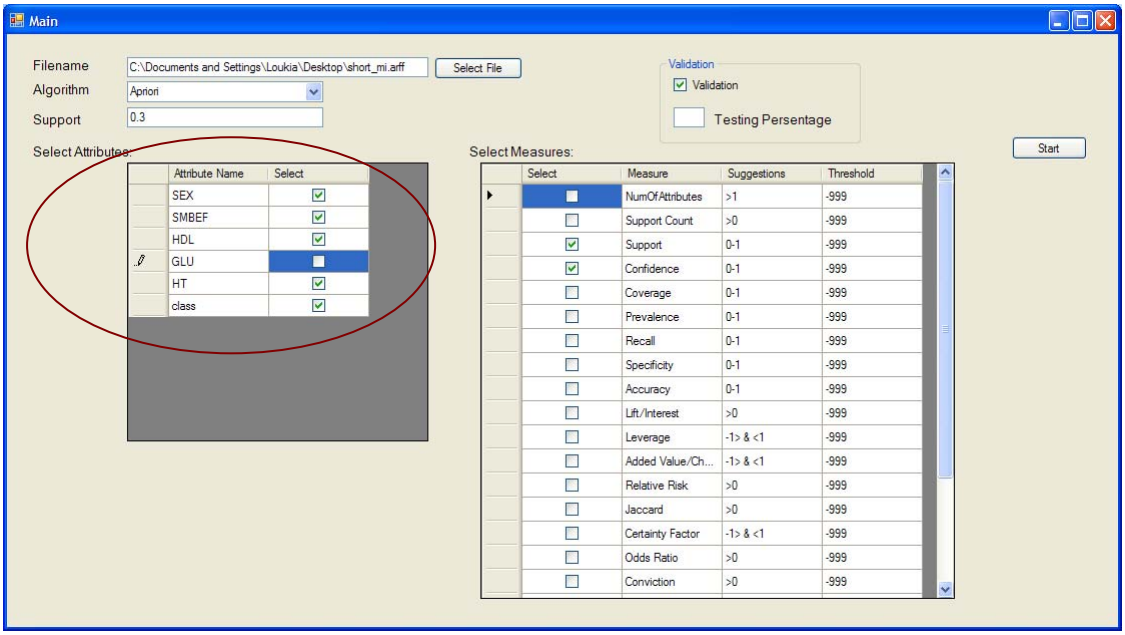


Εικόνα 48: Επιλογή Αλγόριθμου

2.3. Συμπλήρωση του ελάχιστου ορίου support

2.4. Επιλογή χαρακτηριστικών της βάσης

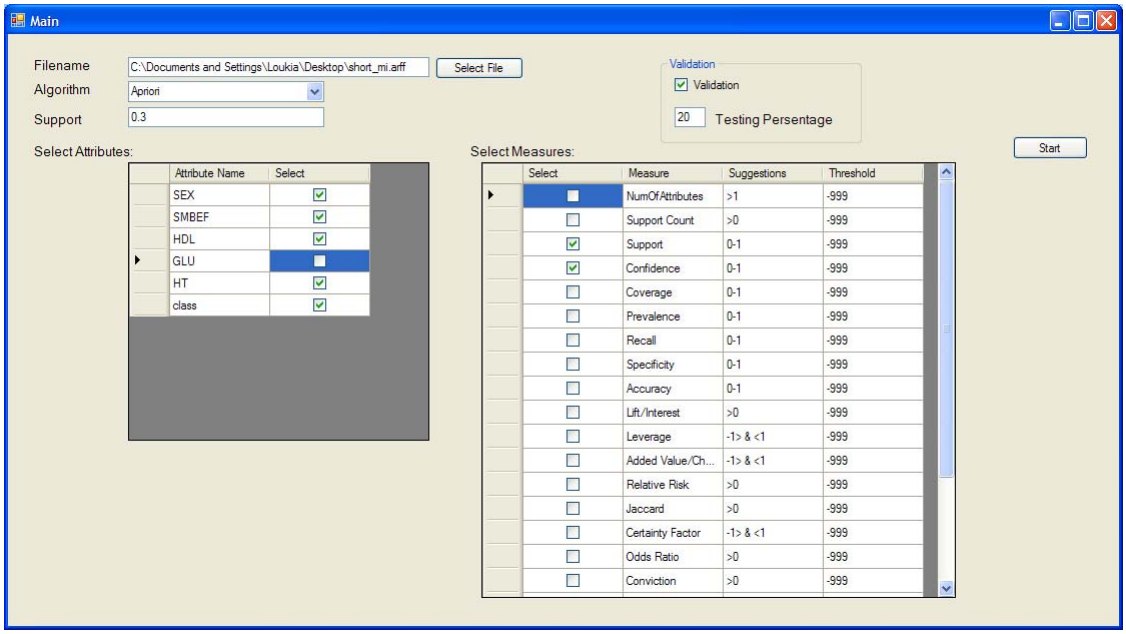
Αφού γίνει η εισαγωγή του αρχείου, το σύστημα αυτόματα παρουσιάζει τα χαρακτηριστικά της βάσης δεδομένων στο πεδίο «Select Attribute». Ο χρήστης μπορεί εάν επιθυμεί να αφαιρέσει κάποιο χαρακτηριστικό να μην περιλαμβάνεται στην ανάλυση της βάσης. Αυτό μπορεί να γίνει με το να κάνει «uncheck» (να αφαιρέσει το ✓) από το χαρακτηριστικό. (Εικόνα 49)



Εικόνα 49: Αφαίρεση Χαρακτηριστικού

2.5. Επιλογή για αξιολόγηση προτύπου

Ο χρήστης μπορεί να επιλέξει εάν επιθυμεί να γίνεται αξιολόγηση του προτύπου με το να κάνει «check» το πεδίο «validation». Αφού το επιλέξει να πρέπει επίσης να συμπληρώσει και το ποσοστό της που θα πρέπει να έχει η βάση ελέγχου. (Εικόνα 50)



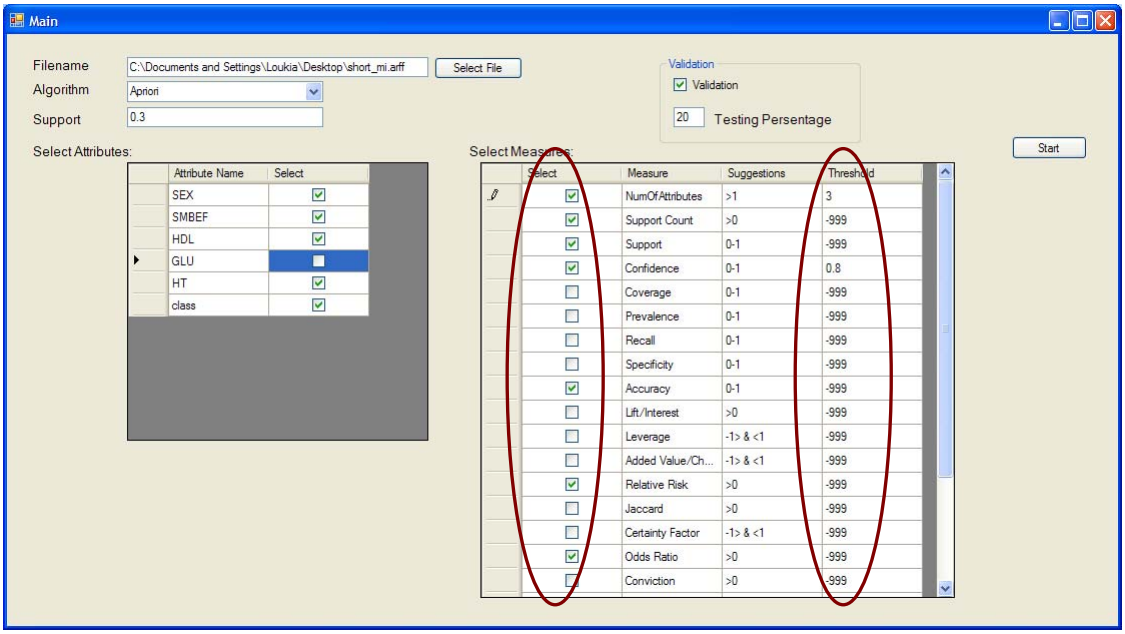
Εικόνα 50: Επιλογή Αξιολόγησης προτύπου

2.6. Επιλογή για μέτρου αξιολόγησης κανόνων

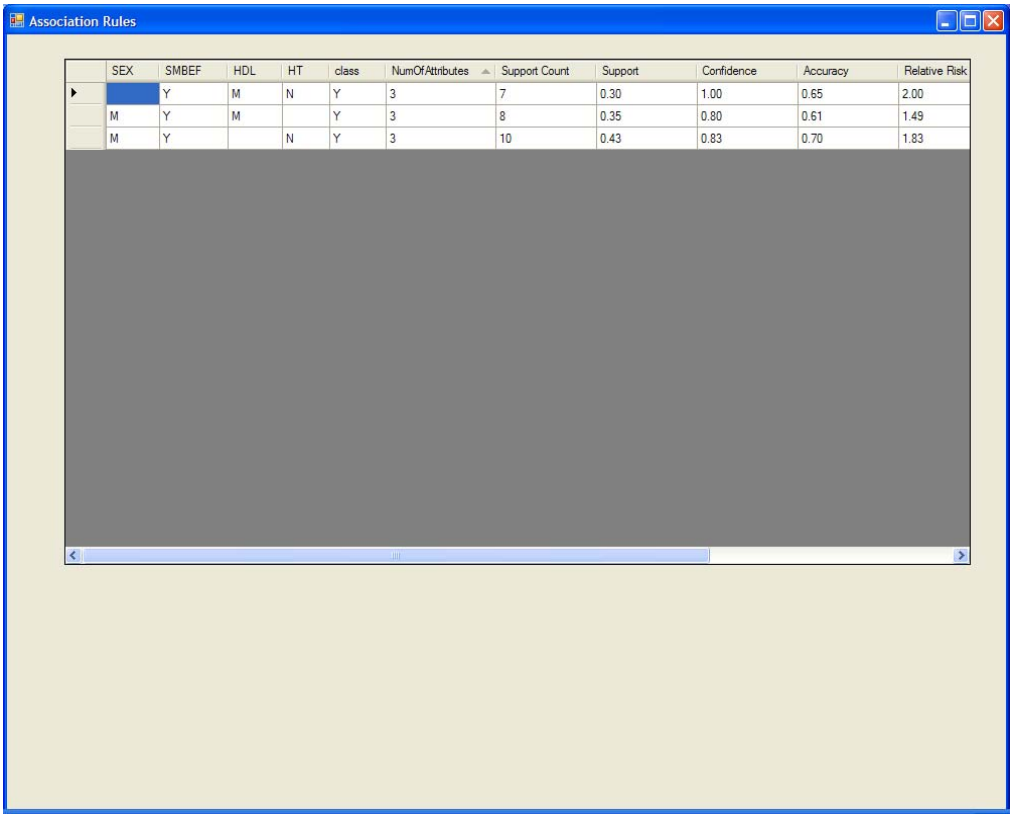
Ο χρήστης μπορεί να επιλέξει κάποιο μέτρο αξιολόγησης κανόνων, μόνο για παρουσίαση του αποτελέσματος του για τους κανόνες με το να κάνει «check» την στήλη «Select» από την λίστα «Select Measure». Επίσης μπορεί να θέσει για οποιαδήποτε μέτρα, κάποιο ελάχιστο όριο, (στήλη «Threshold») για τα οποία οι κανόνες θα φιλτράρονται. Εάν το Threshold είναι -999 αυτό σημαίνει ότι δεν έχει οριστεί κάποιο όριο. (Εικόνα 51)

2.7. Εξόρυξη κανόνων

Αφού γίνει η εισαγωγή των κατάλληλων πεδίων, ο χρήστης θα πρέπει να επιλέξει το «start» για να αρχίσει η εξόρυξη και θα παρουσιαστούν οι κανόνες συσχέτισης σε μια καινούργια οθόνη (Εικόνα 52). Οι κανόνες παρουσιάζονται σε μορφή πίνακα και έτσι μπορούν εύκολα να κατανοηθούν από τον χρήστη. Ο χρήστης μπορεί να ταξινομήσει του κανόνες, με το να επιλέξει οποιαδήποτε στήλη. Επίσης μπορεί να επιλέξει τους κανόνες που επιθυμεί για να τους μεταφέρει σε άλλη εφαρμογή (πχ. Microsoft Excel).



Εικόνα 51: Επιλογή μέτρων αξιολόγησης



Εικόνα 52: Παρουσίαση κανόνων συσχέτησης

ΠΑΡΑΡΤΗΜΑ ΙΙ

Σύγκριση Αποτελεσμάτων Αλγορίθμων Εξαγωγής Κανόνων Συσχέτισης με Αλγόριθμους Κατηγοριοποίησης

1. Σύγκριση για την κλάση MI (Έμφραγμα του μυοκαρδίου)

Εφαρμόζοντας τον αλγόριθμο Apriori στην βάση δεδομένων για την κλάση MI, και θέτοντας επίσης ένα ελάχιστο όριο support 0.2 και 0.6 για confidence παράγονται συνολικά 247 κανόνες. Εφαρμόζοντας στην ίδια βάση δεδομένων αλγόριθμους κατηγοριοποίησης, όπως παραδείγματος χάριν Δέντρα Απόφασης (Decision Trees) με κριτήρια χωρισμού (splitting criteria) Gini Gain παράγονται 147 περίπου κανόνες. Συγκρίνοντας τους κανόνες παρατηρούμε ότι δεν οι κανόνες που παράγονται δεν είναι κοινοί. Ο κάθε αλγόριθμος παράγει ξεχωριστούς κανόνες.

Έτσι συγκρίνονται τα σημαντικά χαρακτηριστικά. Σημαντικά χαρακτηριστικά από τους κανόνες συσχέτισης λαμβάνονται τα χαρακτηριστικά που εμφανίζουν κάποια τιμή συχνά στους κανόνες, και λαμβάνονται τα έξι πιο συχνά χαρακτηριστικά. Αυτά για την κλάση MI είναι:

1) SEX (Φύλο)
2) LDL (Λιποπρωτεΐνες Χαμηλής Πυκνότητας)
3) TC (Ολική Χοληστερόλη)
4) DM (Διαβήτης)
5) DBP (Χαμηλή Πίεση (Διαστολική Πίεση))
6) SMBEF (Καπνιστής)

7) HT (Υπέρταση)

Ενώ από τον αλγόριθμό κατηγοριοποίησης, σημαντικά χαρακτηριστικά θεωρούνται τα χαρακτηριστικά που παρουσιάζονται στους κόμβους των ψηλότερων επιπέδων του δέντρου. Αυτά τα χαρακτηριστικά είναι:

1) SBP (Ψηλή - Συστολική Πίεση)
2) DBP (Χαμηλή - Διαστολική Πίεση)
3) AGE (Ηλικία)
4) HDL (Λιποπρωτεΐνες Υψηλής Πυκνότητας (Καλή χοληστερόλη))
5) HT (Υπέρταση)
6) TC (Ολική Χοληστερόλη)
7) TG (Τριγλυκερίδια)

Παρατηρούμε ότι οι σημαντικοί παράγοντες, που εμφανίζονται συχνότερα στους κανόνες είναι επίσης διαφορετικοί. Οι μόνοι κοινοί παράγοντες είναι τα χαρακτηριστικά DBP (Χαμηλή - Διαστολική Πίεση), HT (Υπέρταση) και TC (Ολική οληστερόλη).

2. Σύγκριση για την κλάση PCI (Αγγειοπλαστική)

Εφαρμόζοντας τον αλγόριθμο Apriori στην βάση δεδομένων για την κλάση PCI, και θέτοντας επίσης ένα ελάχιστο όριο support 0.2 και 0.5 για confidence παράγονται συνολικά 115 κανόνες. Εφαρμόζοντας στην ίδια βάση δεδομένων αλγόριθμους κατηγοριοποίησης, όπως παραδείγματος χάριν Δέντρα Απόφασης (Decision Trees) με κριτήρια χωρισμού (splitting criteria) το Gain Ratio παράγονται 195 περίπου κανόνες. Συγκρίνοντας τους κανόνες παρατηρούμε ότι δεν οι κανόνες που παράγονται δεν είναι κοινοί. Ο κάθε αλγόριθμος παράγει ξεχωριστούς κανόνες.

Έτσι συγκρίνονται τα σημαντικά χαρακτηριστικά. Σημαντικά χαρακτηριστικά από τους κανόνες συσχέτισης λαμβάνονται τα χαρακτηριστικά που εμφανίζουν κάποια τιμή συχνά στους κανόνες, και λαμβάνονται τα έξι πιο συχνά χαρακτηριστικά. Αυτά για την κλάση MI είναι:

1) SEX (Φύλο)
2) LDL (Λιποπρωτεΐνες Χαμηλής Πυκνότητας)
3) TC (Ολική Χοληστερόλη)
4) DM (Διαβήτης)
5) SMBEF (Καπνιστής)
6) HT (Υπέρταση)
7) FH (Οικογενειακό Ιστορικό)

Ενώ από τον αλγόριθμό κατηγοριοποίησης, σημαντικά χαρακτηριστικά θεωρούνται τα χαρακτηριστικά που παρουσιάζονται στους κόμβους των ψηλότερων επιπέδων του δέντρου. Αυτά τα χαρακτηριστικά είναι:

1) DM (Διαβήτης)
2) DBP (Χαμηλή - Διαστολική Πίεση)
3) AGE (Ηλικία)
4) SBP (Ψηλή - Συστολική Πίεση)
5) LDL (Λιποπρωτεΐνες Χαμηλής Πυκνότητας)
6) TG (Τριγλυκερίδια)
7) HDL (Λιποπρωτεΐνες Υψηλής Πυκνότητας (Καλή χοληστερόλη))

Παρατηρούμε ότι οι σημαντικοί παράγοντες, που εμφανίζονται συχνότερα στους κανόνες είναι επίσης διαφορετικοί. Οι μόνοι κοινοί παράγοντες είναι τα χαρακτηριστικά DM (Διαβήτης) και LDL (Λιποπρωτεΐνες Χαμηλής Πυκνότητας).

3. Σύγκριση για την κλάση CABG (Παράκαμψη)

Εφαρμόζοντας τον αλγόριθμο Apriori στην βάση δεδομένων για την κλάση CABG, και θέτοντας επίσης ένα ελάχιστο όριο support 0.2 και 0.5 για confidence παράγονται συνολικά 130 κανόνες. Εφαρμόζοντας στην ίδια βάση δεδομένων αλγόριθμους κατηγοριοποίησης, όπως παραδείγματος χάριν Δέντρα Απόφασης (Decision Trees) με κριτήρια χωρισμού (splitting criteria) το μέτρο απόστασης (Distance Measure) παράγονται 166 περίπου κανόνες. Συγκρίνοντας τους κανόνες παρατηρούμε ότι δεν οι κανόνες που παράγονται δεν είναι κοινοί. Ο κάθε αλγόριθμος παράγει ξεχωριστούς κανόνες.

Έτσι συγκρίνονται τα σημαντικά χαρακτηριστικά. Σημαντικά χαρακτηριστικά από τους κανόνες συσχέτισης λαμβάνονται τα χαρακτηριστικά που εμφανίζουν κάποια τιμή συχνά στους κανόνες, και λαμβάνονται τα έξι πιο συχνά χαρακτηριστικά. Αυτά για την κλάση MI είναι:

1) SEX (Φύλο)
2) LDL (Λιποπρωτεΐνες Χαμηλής Πυκνότητας)
3) DM (Διαβήτης)
4) TC (Ολική Χοληστερόλη)
5) SMBEF (Καπνιστής)
6) DBP (Χαμηλή - Διαστολική Πίεση)
7) HT (Υπέρταση)
8) GLU (γλυκόζη)
9) FH (Οικογενειακό Ιστορικό)
10) SBP (Ψηλή - Συστολική Πίεση)

Ενώ από τον αλγόριθμό κατηγοριοποίησης, σημαντικά χαρακτηριστικά θεωρούνται τα χαρακτηριστικά που παρουσιάζονται στους κόμβους των ψηλότερων επιπέδων του δέντρου. Αυτά τα χαρακτηριστικά είναι:

1) AGE (Ηλικία)

2) GLU (γλυκόζη)
3) HDL (Λιποπρωτεΐνες Υψηλής Πυκνότητας (Καλή χοληστερόλη))
4) SBP (Ψηλή - Συστολική Πίεση)
5) SMBEF (Καπνιστής)
6) TG (Τριγλυκερίδια)
7) SEX (Φύλο)
8) DM (Διαβήτης)
9) HT (Υπέρταση)
10) LDL (Λιποπρωτεΐνες Χαμηλής Πυκνότητας)

Παρατηρούμε ότι οι σημαντικοί παράγοντες, που εμφανίζονται συχνότερα στους κανόνες είναι επίσης διαφορετικοί, όμως υπάρχουν περισσότεροι κοινοί από ότι στις άλλες κλάσεις. Οι κοινοί παράγοντες είναι τα χαρακτηριστικά:

1) SEX (Φύλο)
2) LDL (Λιποπρωτεΐνες Χαμηλής Πυκνότητας)
3) DM (Διαβήτης)
4) SMBEF (Καπνιστής)
5) HT (Υπέρταση)
6) GLU (γλυκόζη)
7) SBP (Ψηλή - Συστολική Πίεση)

4. Συμπέρασμα

Διαφορετικές μεθοδολογίες παράγουν διαφορετικά αποτελέσματα, εφαρμόζοντάς τις στις ίδιες βάσεις δεδομένων. Επομένως ανάλογα με την βάση δεδομένων θα πρέπει να εφαρμόζεται η κατάλληλη μεθοδολογία.