



ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ
Π.Μ.Σ.: «ΠΛΗΡΟΦΟΡΙΚΗ & ΤΗΛΕΜΑΤΙΚΗ»
Κατεύθυνση: Πληροφοριακά Συστήματα στη Διοίκηση
Επιχειρήσεων
MSc in Advanced Information Systems in Business

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ ΔΙΑΓΝΩΣΕΩΝ/ ΧΟΡΗΓΗΣΕΩΝ

Επιβλέπων Καθηγητής
Βασιλάκης Ηρακλής, Επίκουρος Καθηγητής

Σανιδά Θεοδώρα
Α.Μ. 14314

Αθήνα, Φεβρουάριος 2016

Ευχαριστίες

Η παρούσα μεταπτυχιακή διπλωματική εργασία εκπονήθηκε στο Τμήμα Πληροφορικής & Τηλεματικής του Χαροκοπείου Πανεπιστημίου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας, κ. Ηρακλή Βαρλάμη, Επίκουρο Καθηγητή του Τμήματος Πληροφορικής & Τηλεματικής του Χαροκοπείου Πανεπιστημίου, για τη δυνατότητα που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα. Θα ήθελα επίσης να τον ευχαριστήσω για την πολύτιμη καθοδήγηση, την εξαιρετική συνεργασία, καθώς και για τον πολύτιμο χρόνο που διέθεσε για την διεκπεραίωση της παρούσας έρευνας.

Επίσης, θα ήθελα να ευχαριστήσω τον Γενικό Ιατρό που μου παρείχε τα δεδομένα, καθώς και τις χρήσιμες ιατρικές συμβουλές που μου πρόσφερε καθ' όλη τη διάρκεια της διπλωματικής μου έρευνας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, την ανιψιά μου Μαρία – Βασιλική Σανιδά και όλους εκείνους που με στηρίζουν, με ενθαρρύνουν και πιστεύουν στην κάθε μου προσπάθεια.

Σανιδά Θεοδώρα

Αθήνα, Φεβρουάριος 2016

Στην ανιψιά μου

Μαρία – Βασιλική Σανιδά

Πίνακας Περιεχομένων

1. Εισαγωγή.....	17
1.1. Πεδίο διατριβής	17
1.2. Σκοπός	19
1.3. Σύνοψη αποτελεσμάτων	19
2. Επιστημονικό υπόβαθρο	21
2.1. Βασικές έννοιες και βασικές βιβλιογραφικές αναφορές	21
2.1.1. Data Mining - Εξόρυξη Γνώσης.....	21
2.1.2. Association Rules - Κανόνες συσχέτισης	24
2.1.3. Sequential Patterns – Ακολουθιακά Πρότυπα	25
2.1.4. Ιδιοπαθής (πρωτοπαθής) υπέρταση – Essential (primary) hypertension – ICD10: I10.....	26
2.1.5. Αμιγής υπερχοληστερολαιμία – Pure hypercholesterolaemia – ICD10: E78	30
2.1.6. ICD10 - International statistical classification of diseases and related health problems.....	31
2.2. Σχετικές επιστημονικές εργασίες.....	33
2.3. Περιγραφή συστημάτων που χρησιμοποιήθηκαν στη διατριβή	38
3. Ανάλυση της προσέγγισης	39
4. Σχεδίαση	51
4.1. Αρχιτεκτονικό διάγραμμα	51
5. Υλοποίηση.....	55
5.1. Λεπτομέρειες υλοποίησης FP-Growth.....	55
5.2. Λεπτομέρειες υλοποίησης GSP	57
6. Αποτελέσματα/Αξιολόγηση	59

6.1. Dataset - Σύνολο δεδομένων	59
6.2. Μετρικές αξιολόγησης	62
6.3. Αποτελέσματα.....	64
6.3.1. Αποτελέσματα των datasets ανά ασθενή	65
6.3.2. Αποτελέσματα των datasets ανά επίσκεψη.....	70
7. Συμπεράσματα.....	75
7.1. Σύνοψη της προτεινόμενης προσέγγισης.....	75
7.2. Μελλοντικές επεκτάσεις	77
8. Παραρτήματα	79
8.1. Παραρτήματα I – Αποτελέσματα εκτέλεσης των datasets ανά ασθενή	79
8.1.1. FP-Growth.....	79
8.1.2. GSP.....	81
8.2. Παραρτήματα II – Αποτελέσματα εκτέλεσης των datasets ανά επίσκεψη.....	85
8.2.1. FP-Growth.....	85
8.2.2. GSP.....	87
9. Βιβλιογραφία	91

Κατάλογος Σχημάτων

Σχήμα 1: Κατανομή ανδρών – γυναικών στις 100.000.....	42
Σχήμα 2: Κατανομή ανδρών – γυναικών ασθενών ανά ηλικιακή ομάδα στις 100.000	43
Σχήμα 3: Κατανομή επικρατέστερων διαγνώσεων στις 100.000.....	44
Σχήμα 4: Κατανομή επικρατέστερων διαγνώσεων ανδρών-γυναικών στις 100.000	44
Σχήμα 5: Κατανομή I10 ανδρών-γυναικών στις 100.000	45
Σχήμα 6: Κατανομή E78 ανδρών-γυναικών στις 100.000	45
Σχήμα 7: Κατανομή I10 & E78 ανδρών-γυναικών στις 100.000.....	46
Σχήμα 8: Κατανομή K21 ανδρών-γυναικών στις 100.000	46
Σχήμα 9: Κατανομή J06 ανδρών-γυναικών στις 100.000	47
Σχήμα 10: Κατανομή E10 ανδρών-γυναικών στις 100.000	47
Σχήμα 11: Κατανομή I25.5 ανδρών-γυναικών στις 100.000	48
Σχήμα 12: Κατανομή ICD10: E78 ανδρών-γυναικών στις 100.000.....	61
Σχήμα 13: Κατανομή ICD10:I10 ανδρών-γυναικών στις 100.000.....	61
Σχήμα 14: Κατανομή ICD10:I10 & E78 ανδρών-γυναικών στις 100.000	62

Κατάλογος Εικόνων

Εικόνα 1: Τα βήματα της διαδικασίας KDD	23
Εικόνα 2: Η σχέση της εξόρυξης γνώσης με άλλους τομείς.	23
Εικόνα 3: WHO - Σύνολο θανάτων, υπερτασικές νόσοι, άνδρες, όλες οι ηλικίες για ICD10 από I10 – I15.....	28
Εικόνα 4: WHO - Σύνολο θανάτων, υπερτασικές νόσοι, γυναίκες, όλες οι ηλικίες για ICD10 από I10 – I15.....	29
Εικόνα 5: WHO - Σύνολο θανάτων, υπερτασικές νόσοι, άνδρες - γυναίκες, όλες οι ηλικίες για ICD10 από I10 – I15	29
Εικόνα 6: Το σχεσιακό σχήμα της βάσης.....	41
Εικόνα 7: Αρχιτεκτονικό Διάγραμμα	51
Εικόνα 8: Μοντέλο για τον αλγόριθμο FP-Growth στο Rapid Miner	56
Εικόνα 9: Μοντέλο για τον αλγόριθμο GSP στο Rapid Miner	57

Κατάλογος Πινάκων

Πίνακας 1: Ομάδα κωδικών ICD10 από I10 έως I15.....	26
Πίνακας 2: Ταξινόμηση ICD10	32
Πίνακας 3: Μελέτες εξόρυξης γνώσης για υπέρταση και δυσλιπιδαιμίες	37
Πίνακας 4: Ηλικιακές ομάδες βάσει του World Health Organization (WHO).....	40
Πίνακας 5: Ηλικιακές ομάδες βάσει της ΕΛ.ΣΤΑΤ.	42
Πίνακας 6: Επικρατέστερες Διαγνώσεις.....	43
Πίνακας 7: Δραστικές ουσίες για χοληστερόλη (E78)	64
Πίνακας 8: Δραστικές ουσίες για υπέρταση (I10)	64
Πίνακας 9: Δραστικές ουσίες για διαβήτη (E10)	64
Πίνακας 10: Δραστικές ουσίες για στεφανιαία νόσο (I25.5)	64
Πίνακας 11: Δραστικές ουσίες για γαστροοισοφαγική παλινδρόμηση (K21)	64

Κατάλογος Εξιώσεων

Εξίσωση 1: Support.....	63
Εξίσωση 2: Confidence.....	63
Εξίσωση 3: Lift.....	63

Περίληψη

Οι ραγδαίες εξελίξεις στην τεχνολογία των υπολογιστών και των βάσεων δεδομένων, έχουν ως συνέπεια, ο όγκος των δεδομένων που συλλέγονται και αποθηκεύονται, να αυξάνεται συνεχώς. Αυτό έχει ως αποτέλεσμα την επιτακτική ανάγκη για αποδοτικές μεθόδους ανάλυσης και την εξαγωγή χρήσιμης πληροφορίας.

Οι ιατρικές βάσεις δεδομένων περιέχουν τεράστιο, ετερογενή και πολύπλοκο όγκο δεδομένων. Έτσι, η εξόρυξη γνώσης στον κλάδο της ιατρικής γίνεται όλο και περισσότερο απαραίτητη, επειδή αποτελεί σημαντικό παράγοντα για νέες μεθόδους και ανακαλύψεις.

Η καρδιαγγειακή νόσος, βάσει του Παγκόσμιου Οργανισμού Υγείας, αποτελεί το κύριο αίτιο για το ένα τρίτο των θανάτων ανά έτος, τόσο σε παγκόσμιο επίπεδο, όσο και στην Ελλάδα. Βασικούς παράγοντας κινδύνου της καρδιαγγειακής νόσου και του εγκεφαλικού επεισοδίου είναι ο συνδυασμός υπέρτασης και αυξημένων επιπέδων χοληστερόλης, με αποτέλεσμα την πρόωρη θνησιμότητα και αναπηρία. Αυτός ο συνδυαστικός παράγοντας κινδύνου αποτελεί παγκόσμιο πρόβλημα δημόσιας υγείας και πρέπει να ληφθούν μέτρα για να μειωθεί ο αριθμός των περιστατικών. Διάφορες τεχνικές εξόρυξης γνώσης χρησιμοποιούνται για τον προσδιορισμό αυτού του συνδυαστικού παράγοντα κινδύνου, με σκοπό την πρόβλεψη ή την εξαγωγή χρήσιμων συσχετίσεων.

Το αντικείμενο της παρούσας έρευνας αφορά την εξόρυξη γνώσης σε ιατρικά δεδομένα, που προέρχονται από την ηλεκτρονική συνταγογράφηση στην περιοχή της Αθήνας. Οι στόχοι της έρευνας είναι η παραγωγή κανόνων συσχέτισης

και ακολουθιακών προτύπων, τόσο νόσων, βάσει του προτύπου ICD10, όσο και δραστικών ουσιών, από ασθενείς που έχουν I10 (ιδιοπαθής (πρωτοπαθής) υπέρταση), E78 (αμιγής υπερχοληστερολαιμία) και I10 & E78 ταυτόχρονα.

Λέξεις-κλειδιά: αυξημένη αρτηριακή πίεση, αυξημένη χοληστερόλη, υπέρταση, δυσλιπιδαιμία, ιδιοπαθής (πρωτοπαθής) υπέρταση, υπερχοληστερολαιμία, εξόρυξη δεδομένων, τεχνικές εξόρυξης δεδομένων, εφαρμογές εξόρυξης δεδομένων, κανόνες συσχέτισης, ακολουθιακά πρότυπα.

Abstract

Rapid developments in computer technology and databases will constantly increase the data volume which are collected and stored. This results in the urgent need for efficient analysis methods to extract useful information.

Medical databases contain huge, heterogeneous and complexed data volumes. Thus, knowledge extraction in the medical industry is becoming increasingly necessary as an important factor for new methods and discoveries.

Cardiovascular disease, according to the World Health Organization, is the main cause for one third of deaths per year, both globally and in Greece. The main risk factor for cardiovascular disease and stroke is the combination of elevated levels of hypertension and cholesterol, leading to premature mortality and disability. This combinational risk factor is a global public health problem and actions should be taken to reduce the number of incidents. Various data mining techniques are used to distinguish the combined risk factor in predicting and extracting useful relationships.

The objective of this research relates to data mining medical data from electronic prescribing in the Athens area. The objectives of the survey is to produce association rules and sequential patterns, both diseases under standard ICD10, and active substances, patients I10 (essential (primary) hypertension), E78 (pure hypercholesterolaemia) and I10 & E78 simultaneously.

Keywords: raised blood pressure, raised total cholesterol, hypertension, dyslipidaemia, pure hypercholesterolemia, essential (primary) hypertension, data mining, data mining techniques, data mining applications, association rules, sequential patterns.

1. Εισαγωγή

1.1. Πεδίο διατριβής

Οι ραγδαίες εξελίξεις στην τεχνολογία και η συστηματική διείσδυση της πληροφορικής σε όλα τα επιστημονικά πεδία, σε συνδυασμό με τη χρήση των πληροφοριακών συστημάτων, τόσο στον ιδιωτικό όσο και στο δημόσιο τομέα, έχει ως αποτέλεσμα τη συσσώρευση μεγάλων ποσοτήτων δεδομένων, που οργανώνονται συνήθως μέσα από ένα σύστημα διαχείρισης βάσης δεδομένων. Οι βάσεις δεδομένων είναι η τεχνολογία που μας επιτρέπει την αποθήκευση και την επεξεργασία τεράστιου όγκου δεδομένων. Η εξόρυξη γνώσης από αυτά τα δεδομένα, το data mining, αποτελεί ένα από τα βασικότερα ερευνητικά πεδία και δημιουργεί προκλήσεις σε πολλούς τομείς. (Tan et al., 2006).

Η αξιοποίηση των δεδομένων και η εξαγωγή γνώσης από αυτά, έχει αποκτήσει ιδιαίτερα σημαντική θέση στον σύγχρονο κόσμο, καθώς χρησιμοποιείται για τη διαμόρφωση προτύπων. Ιδιαίτερα σημαντική θέση έχει η ιατρική πληροφορία, όπου η εξόρυξη γνώσης από αυτήν, γίνεται όλο και περισσότερο απαραίτητη, εφόσον αποτελεί σημαντικό παράγοντα εύρεσης νέων μεθόδων και ανακαλύψεων. Έτσι, με την εφαρμογή της στην ιατρική, μπορούν να εντοπιστούν μοντέλα πρόβλεψης και πρόγνωσης νόσων, αλλά και παράλληλα να γίνει αξιολόγηση της αποτελεσματικότητας της θεραπείας που χρησιμοποιήθηκε. (Koh & Tan, 2011) Επίσης, μπορεί να αποτελέσει ένα πολύτιμο εργαλείο φαρμακοεπαγρύπνησης, ώστε να αποφευχθούν ή να περιοριστούν οι ανεπιθύμητες ενέργειες των δραστικών ουσιών. (Wilson et al., 2004)

Οι ιατρικές βάσεις δεδομένων περιέχουν τεράστιο και πολύπλοκο όγκο δεδομένων. Η επεξεργασία και ανάλυση αυτών με μη αυτοματοποιημένο τρόπο, απαιτεί υψηλό κόστος διαχείρισης. (Simonici, 2011) Παράλληλα, η εξόρυξη γνώσης μέσα από ιατρικά δεδομένα, αντιμετωπίζει αρκετές δυσκολίες λόγω της ιδιαιτερότητας της πληροφορίας που προκύπτει, αφενός από το αντικείμενο της ιατρικής και αφετέρου από τη διαφορετική δομή των συσχετίσεων τους. Μια δυσκολία είναι η ετερογένεια των ιατρικών δεδομένων, όπως παρακλινικές εξετάσεις (ακτινολογικά και εργαστηριακά ευρήματα), συνεντεύξεις και σημειώσεις, ο συνδυασμός των οποίων όμως, πολλές φορές είναι απαραίτητος για την πρόγνωση και τη διάγνωση της νόσου. Από την άλλη, εγείρονται ηθικά, νομικά και κοινωνικά ζητήματα, που απορρέουν από το απόρρητο που σύμφωνα με τον κώδικα ιατρικής δεοντολογίας, προστατεύει τα ιατρικά δεδομένα. (Cios & Moore, 2002)

Η καρδιαγγειακή νόσος, βάσει του Παγκόσμιου Οργανισμού Υγείας, αποτελεί το κύριο αίτιο θανάτων, τόσο σε παγκόσμιο επίπεδο, όσο και στην Ελλάδα. Βασικός παράγοντας κινδύνου, τόσο της καρδιαγγειακής νόσου, όσο και του εγκεφαλικού επεισοδίου – μεταξύ των άλλων – είναι ο συνδυασμός υπέρτασης και αυξημένων επιπέδων χοληστερόλης και τριγλυκεριδίων (δυσλιπιδαιμία). Για τον προσδιορισμό αυτού του συνδυαστικού παράγοντα κινδύνου χρησιμοποιούνται διάφορες τεχνικές εξόρυξης γνώσης, με σκοπό την ανακάλυψη νέων προτύπων για την εξαγωγή χρήσιμων πληροφοριών.

1.2. Σκοπός

Το αντικείμενο της παρούσας έρευνας αφορά την εξόρυξη γνώσης σε ιατρικά δεδομένα, που προέρχονται από την ηλεκτρονική συνταγογράφηση στην περιοχή της Αθήνας. Οι στόχοι της έρευνας είναι η παραγωγή κανόνων συσχέτισης και ακολουθιακών προτύπων, τόσο σε σχέση με τη νόσο, όσο και των δραστικών ουσιών που χρησιμοποιούνται από ασθενείς που πάσχουν είτε από υψηλή χοληστερόλη (E78), είτε από υπέρταση (I10), αλλά και από ασθενείς που ταυτόχρονα πάσχουν από υπέρταση (I10) και υψηλή χοληστερόλη (E78).

1.3. Σύνοψη αποτελεσμάτων

Η παρούσα διπλωματική εργασία, είχε ως στόχο την παραγωγή κανόνων συσχέτισης και ακολουθιακών προτύπων, τόσο νόσων, βάσει του προτύπου ICD10, όσο και δραστικών ουσιών, από ασθενείς που πάσχουν από υψηλή χοληστερόλη (E78), υπέρταση (I10) αλλά και από υψηλή χοληστερόλη (E78) και υπέρταση (I10) ταυτόχρονα. Τα ιατρικά δεδομένα, προέρχονται από την ηλεκτρονική συνταγογράφηση στην περιοχή της Αθήνας, από Γενικό Ιατρό, για το χρονικό διάστημα από 7/1/2012 έως 31/7/2015.

Δημιουργήθηκαν έξι σύνολα δεδομένων και σύμφωνα με τα αποτελέσματα που προέκυψαν από την παρούσα έρευνα, βλέπουμε ότι υπάρχει θετική συσχέτιση υψηλής χοληστερόλης (E78), υπέρτασης (I10) και στεφανιαίας νόσου (I25.5). Εμφανίζονται διάφοροι παράγοντες κινδύνου υπέρτασης και υψηλής χοληστερόλης, όπως αγχώδεις διαταραχές (F41.9), ινσουλινοεξαρτώμενος

σακχαρώδης διαβήτης (E10), μη ινσουλινοεξαρτώμενος σακχαρώδης διαβήτης (E11) και καταθλιπτική διαταραχή (F33).

Όσον αφορά την γαστροοισοφαγική παλινδρόμηση με οισοφαγίτιδα (K21), ένας παράγοντας κινδύνου είναι τα φάρμακα για την υπέρταση ή τις καρδιακές νόσους. Παρατηρούμε σε πολλούς κανόνες θετική συσχέτιση υψηλής χοληστερόλης (E78), υπέρτασης (I10), στεφανιαίας νόσου (I25.5) και γαστροοισοφαγικής παλινδρόμησης με οισοφαγίτιδα (K21). Τέλος, βλέπουμε τις δραστικές ουσίες που χορηγούνται ανά νόσο.

2. Επιστημονικό υπόβαθρο

2.1. Βασικές έννοιες και βασικές βιβλιογραφικές αναφορές

2.1.1. Data Mining - Εξόρυξη Γνώσης

Η εξόρυξη γνώσης αναφέρεται στη διαδικασία αυτόματης ανακάλυψης χρήσιμης πληροφορίας από μεγάλες αποθήκες δεδομένων. Οι τεχνικές που χρησιμοποιεί η εξόρυξη γνώσης έχουν ως σκοπό και στόχο την εύρεση νέων και χρήσιμων προτύπων, με τη χρήση αλγορίθμων από μεγάλες βάσεις δεδομένων, που σε διαφορετική περίπτωση θα παρέμεναν άγνωστα. (Tan et al., 2006)

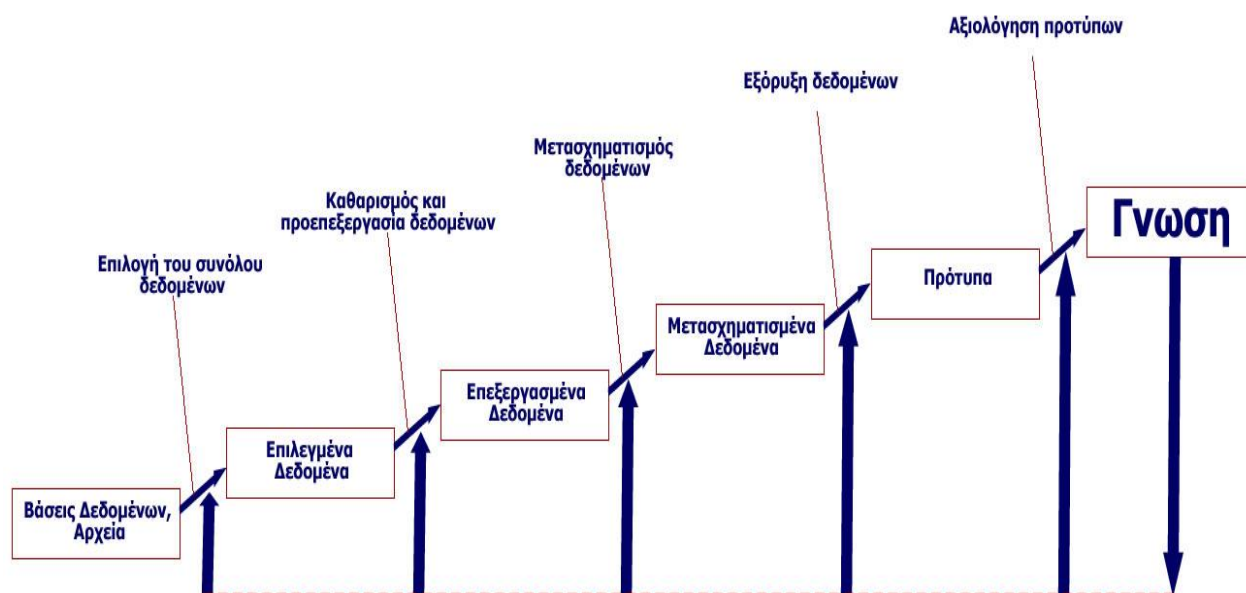
Η εξόρυξη γνώσης αποτελεί ένα αναπόσπαστο κομμάτι της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD). Η εξόρυξη δεδομένων είναι το πιο κρίσιμο βήμα στην διαδικασία KDD και αποτελεί την κύρια διεργασία στην διαδικασία της εξόρυξης γνώσης. Η KDD αναφέρεται στη συνολική διαδικασία μετατροπής των ακατέργαστων δεδομένων σε χρήσιμες πληροφορίες. (Tan et al., 2006; Han et al., 2011)

Η KDD αποτελεί μια μη τετριμμένη διαδικασία εύρεσης πρωτότυπων, έγκυρων και χρήσιμων κατανοητών μοτίβων μέσα από τα δεδομένα. Η KDD διαδικασία είναι μια επαναληπτική διαδικασία, η οποία απαρτίζεται από τα ακόλουθα βήματα: (Fayyad et al., 1996)

- ❖ **Επιλογή του συνόλου των δεδομένων:** Σ' αυτό το βήμα καθορίζεται το σύνολο των δεδομένων στα οποία θα εκτελεστεί η διαδικασία της εξόρυξης δεδομένων.
- ❖ **Καθαρισμός και προεπεξεργασία των δεδομένων :** Σ' αυτό το βήμα περιλαμβάνονται διαδικασίες, όπως η συλλογή των απαραίτητων

πληροφοριών για τον εντοπισμό του θορύβου, καθώς και η αφαίρεση του, η απαλοιφή ακραίων δεδομένων (outliers) και η λήψη αποφάσεων σχετικά με τις στρατηγικές για τον χειρισμό των ελλειπόντων τιμών στα πεδία των δεδομένων.

- ❖ **Μετασχηματισμός των δεδομένων:** Σ' αυτό το βήμα τα δεδομένα πρέπει να μετασχηματιστούν σε κατάλληλη μορφή, έτσι ώστε να εφαρμοστούν οι διάφορες τεχνικές εξόρυξης δεδομένων. Εδώ χρησιμοποιούνται διάφορες τεχνικές, όπως η μείωση των διαστάσεων και των μεταβλητών που θα εξεταστούν στην εξόρυξη δεδομένων.
- ❖ **Εξόρυξη δεδομένων:** Σ' αυτό το βήμα περιλαμβάνεται η επιλογή των μεθόδων που θα χρησιμοποιηθούν για την εξαγωγή χρήσιμων προτύπων γνώσης. Τα πρότυπα που εξάγονται και αποτελούν ενδιαφέρον, μπορεί να είναι κανόνες συσχέτισης, συσταδοποίηση, παλινδρόμηση κτλ.
- ❖ **Αξιολόγηση των προτύπων:** Σ' αυτό το βήμα τα πρότυπα που έχουν εξαχθεί από το προηγούμενο βήμα πρέπει να αξιολογηθούν με κάποια κριτήρια – μετρικές, προκειμένου να παραμείνουν μόνον τα πρότυπα που έχουν ενδιαφέρον και προσφέρουν γνώση.
- ❖ **Παρουσίαση γνώσης:** Σ' αυτό το βήμα η εξορυγμένη γνώση απεικονίζεται στον χρήστη με τη βοήθεια των τεχνικών οπτικοποίησης.



Εικόνα 1: Τα βήματα της διαδικασίας KDD

Η εξόρυξη γνώσης περιλαμβάνει μεθοδολογίες και αλγόριθμους, οι οποίοι έχουν προέλθει από άλλους τομείς, όπως η στατιστική, η τεχνητή νοημοσύνη, η μηχανική μάθηση, η παράλληλη επεξεργασία, τα καταναμημένα συστήματα, η αναγνώριση προτύπων και η τεχνολογία βάσεων δεδομένων. (Han et al., 2011).



Εικόνα 2: Η σχέση της εξόρυξης γνώσης με άλλους τομείς.

2.1.2. Association Rules - Κανόνες συσχέτισης

Η εξαγωγή κανόνων συσχέτισης (Association Rules) αποτελεί μία από τις πιο σημαντικές τεχνικές εξόρυξης δεδομένων, με στόχο της να ανακαλυφθούν ενδιαφέρουσες σχέσεις και συσχετίσεις μεταξύ του συνόλου των πεδίων της βάσης δεδομένων. (Agrawal, 1993) Η εξαγωγή κανόνων συσχέτισης έχει ένα ευρύ φάσμα εφαρμογής, όπως την ανάλυση του καλαθιού αγοράς (market basket analysis), την ιατρική διάγνωση, τα τηλεπικοινωνιακά δίκτυα και τη διαχείριση ρίσκου (Kotsiantis & Kanellopoulos, 2006)

Οι κανόνες συσχέτισης είναι μια έκφραση της μορφής $X \rightarrow Y$ όπου X και Y είναι στοιχειοσύνολα (itemset). Κάθε κανόνας εκφράζεται με δύο μετρικές, την υποστήριξη κανόνα (support) και την εμπιστοσύνη κανόνα (confidence). Όσο πιο μεγάλος είναι ο αριθμός της υποστήριξης και της εμπιστοσύνης, τόσο πιο ισχυρός είναι ο κανόνας που εξάγεται.

Το βασικό πρόβλημα στην εξαγωγή κανόνων συσχέτισης είναι ότι εξάγεται μεγάλος αριθμός κανόνων. Για να μειωθεί ο αριθμός των κανόνων που εξάγονται, έχουμε ως περιορισμούς την ελάχιστη υποστήριξη (minimum support) και την ελάχιστη εμπιστοσύνη (minimum confidence). (Agrawal and Srikant 1994)

Έτσι, γίνεται χωρισμός του προβλήματος σε δύο υποπροβλήματα (Agrawal and Srikant, 1994):

- ❖ στην εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Itemset Generation) με υποστήριξη (support) πάνω από την ελάχιστη υποστήριξη (minimum support)

- ❖ στη δημιουργία κανόνων (Rule Generation) με εμπιστοσύνη (confidence) πάνω από την ελάχιστη εμπιστοσύνη (minimum confidence)

Για την εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Itemset Generation) ο πιο διαδεδομένος αλγόριθμος είναι ο FP-growth.

Για τη μέτρηση του ενδιαφέροντος των κανόνων συσχέτισης, εκτός από την υποστήριξη (Support) και την εμπιστοσύνη (Confidence), υπάρχουν μέτρα όπως η περιεκτικότητα (Conciseness), η κάλυψη (Coverage), η αξιοπιστία (Reliability), η ιδιαιτερότητα (Peculiarity), το Lift κ.α (Geng & Hamilton, 2006)

2.1.3. Sequential Patterns – Ακολουθιακά Πρότυπα

Τα ακολουθιακά πρότυπα επεκτείνουν τη μέθοδο των κανόνων συσχέτισης, ώστε να εντοπιστούν πρότυπα – μοτίβα μεταξύ των γεγονότων, τα οποία επαναλαμβάνονται στα δεδομένα. Ο στόχος είναι μέσα από ένα σύνολο δεδομένων να ανακαλυφθούν ακολουθίες στο μέγιστο μήκος, με υποστήριξη (support) πάνω από την ελάχιστη υποστήριξη (minimum support). (Agrawal & Srikant, 1995)

Ένα παράδειγμα αποτελεί η κακή χρήση φαρμάκων που μπορεί να προκύψει, όταν σ' έναν ασθενή χορηγούνται δύο ή περισσότερα φάρμακα μέσα σε μια δεδομένη χρονική στιγμή. Τα φάρμακα με τις ανεπιθύμητες αντιδράσεις καταγράφονται μαζί με τη χρονική στιγμή ως ένα πρότυπο – μοτίβο, το οποίο μπορεί να βρίσκεται και μέσα στο φάκελο του ασθενή. (Roddick & Spiliopoulou, 2002)

Κάθε ακολουθιακό πρότυπο εκφράζεται με μια μετρική, την υποστήριξη κανόνα (support). Το βασικό πρόβλημα στην εξαγωγή ακολουθιακών προτύπων είναι ότι εξάγεται μεγάλος αριθμός προτύπων. Για να μειωθεί ο αριθμός ακολουθιακών προτύπων που εξάγονται, έχουμε ως περιορισμό την ελάχιστη υποστήριξη (minimum support). Ο πιο διαδεδομένος αλγόριθμος είναι ο GSP (Agrawal & Srikant, 1995)

2.1.4. Ιδιοπαθής (πρωτοπαθής) υπέρταση - Essential (primary) hypertension - ICD10: I10

Το ICD10 I10 (ιδιοπαθής (πρωτοπαθής) υπέρταση ή υψηλή - αυξημένη αρτηριακή πίεση ή απλά υπέρταση) ανήκει στις παθήσεις του κυκλοφορικού συστήματος (Diseases of the circulatory system) Η ομάδα κωδικών ICD10 από I10 έως I15 ανήκει στην υπερτασική νόσο (WHO, 2004)

ICD10	Περιγραφή
I10	Ιδιοπαθής (πρωτοπαθής) υπέρταση
I11	Υπερτασική καρδιοπάθεια
I12	Υπερτασική νεφροπάθεια
I13	Υπερτασική καρδιοπάθεια και νεφροπάθεια
I15	Δευτεροπαθής υπέρταση

Πίνακας 1: Ομάδα κωδικών ICD10 από I10 έως I15

Η υπέρταση, αποτελεί παγκόσμιο πρόβλημα δημόσιας υγείας και σπάνια προκαλεί συμπτώματα στα πρώιμα στάδια. (WHO, 2015) Έχει αποδειχτεί ότι

συμβάλλει σημαντικά στην επιβάρυνση της καρδιακής νόσου και του εγκεφαλικού επεισοδίου, με αποτέλεσμα την πρόωρη θνησιμότητα και αναπηρία. (Lewington et al., 2002; Cutler et al., 2008) Επιπλέον, ο συνδυασμός υπέρτασης και υψηλής χοληστερόλης αυξάνουν τον κίνδυνο της στεφανιαίας νόσου και του εγκεφαλικού επεισοδίου. (Ezzati et al., 2002; WHO & UNAIDS, 2007)

Παράλληλα αποδείχτηκε, ότι η υπέρταση συμβάλλει στην καρδιακή ανεπάρκεια, στην αγγειακή νόσο, στην νεφρική ανεπάρκεια, στην άνοια, στην πρωτεϊνουρία, στην αιμορραγία του αμφιβληστροειδούς και στις διαταραχές οράσεως. (Williams et al., 2004; Mendis et al., 2014) Τέλος, ο έλεγχος και η πρόληψη της υπέρτασης σε διαβητικούς έχει αποδειχθεί ότι μειώνει τον κίνδυνο των καρδιαγγειακών επιπλοκών, της νεφροπάθειας και της αμφιβληστροειδοπάθειας. (Alwan, 2011)

Παράγοντες κινδύνου υπέρτασης είναι (Williams et al., 2004) :

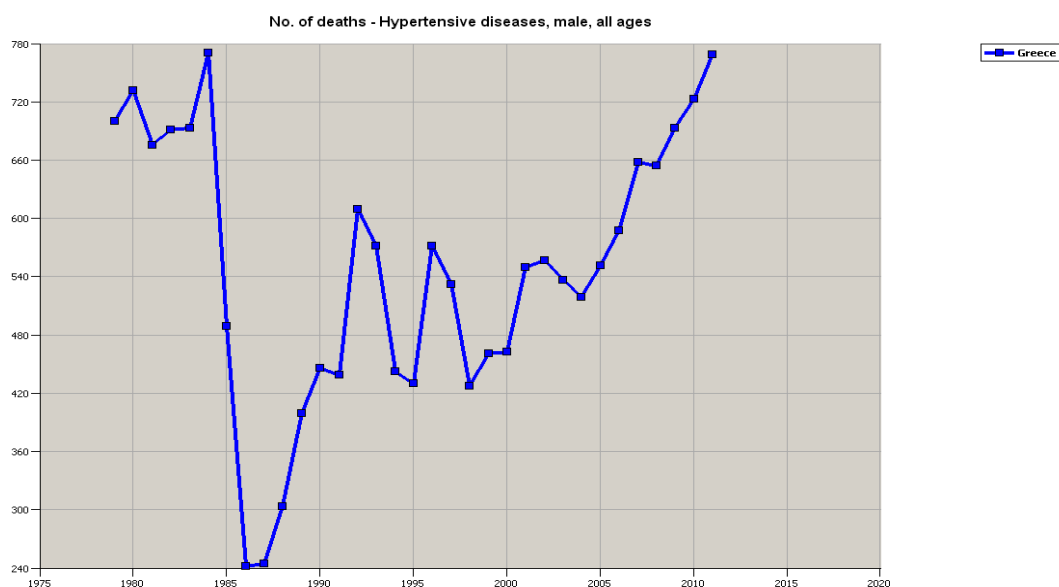
- ❖ η παχυσαρκία
- ❖ το άγχος
- ❖ η ανθυγιεινή διατροφή που περιέχει πολύ αλάτι και λίπος, ενώ δεν περιέχει φρούτα και λαχανικά
- ❖ η κατανάλωση αλκοόλ
- ❖ η χρήση καπνού
- ❖ οι γενετικοί παράγοντες
- ❖ η έλλειψη σωματικής άσκησης.

Αν δεν ληφθούν μέτρα για να μειωθεί η έκθεση σ' αυτούς τους παράγοντες έχουμε αύξηση της υπέρτασης. (WHO, 2015)

Η καρδιαγγειακή νόσος αποτελεί το κύριο αίτιο για το ένα τρίτο του συνόλου των θανάτων ανά έτος σε παγκόσμιο επίπεδο. Η υπέρταση εκτιμάται ότι προκαλεί 7,5 εκατομμύρια θανάτους, δηλαδή περίπου το 12,8% όλων των θανάτων. (WHO, 2009; Alwan, 2011)

Ο Παγκόσμιος Οργανισμός Υγείας (World Health Organization – WHO) έχει μια ιατρική βάση δεδομένων που περιέχει στατιστικά στοιχεία για όλες τις αιτίες θανάτου και νόσων. Στις παρακάτω εικόνες φαίνεται το σύνολο των θανάτων στην Ελλάδα από υπερτασικές νόσους για ICD10 I10 έως I15.

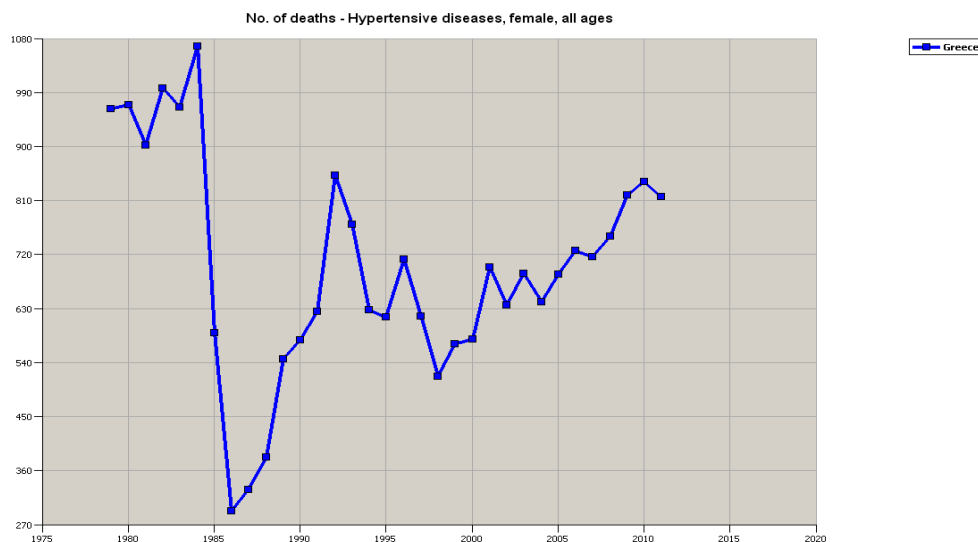
Στους άνδρες παρατηρούμε ότι από το 1998 υπάρχει άνοδος στους θανάτους που οφείλονται σε υπερτασικές νόσους με ICD10 I10 έως I15.



Εικόνα 3: WHO - Σύνολο θανάτων, υπερτασικές νόσοι, άνδρες, όλες οι ηλικίες για ICD10 από I10 – I15¹

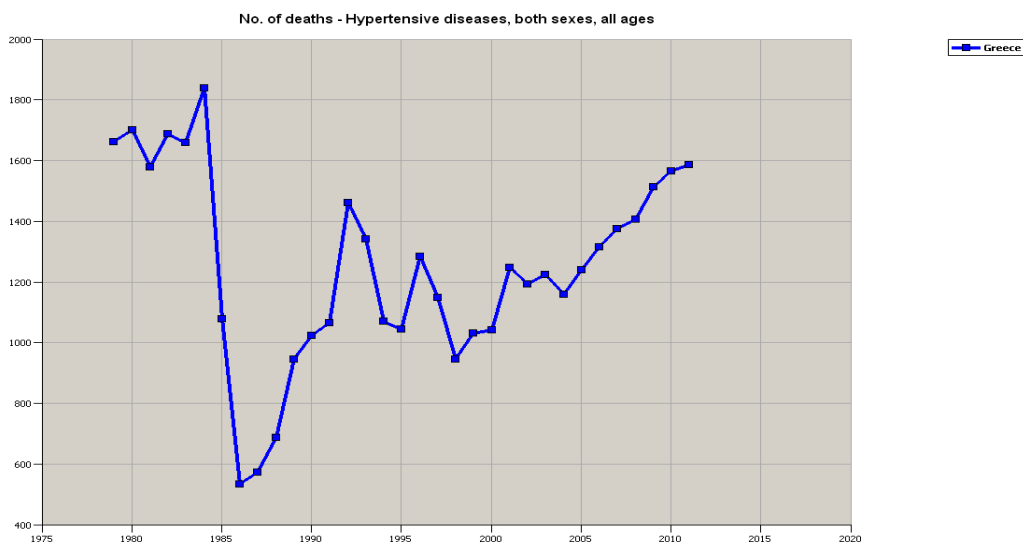
¹ <http://apps.who.int/healthinfo/statistics/mortality/whodpms/>

Στις γυναίκες παρατηρούμε ότι από το 1998 υπάρχει άνοδος στους θανάτους που οφείλονται σε υπερτασικές νόσους με ICD10 I10 έως I15.



Εικόνα 4: WHO - Σύνολο θανάτων, υπερτασικές νόσοι, γυναίκες, όλες οι ηλικίες για ICD10 από I10 – I15²

Και στα δυο φύλλα παρατηρούμε ότι από το 1998 υπάρχει άνοδος στους θανάτους που οφείλονται σε υπερτασικές νόσους με ICD10 I10 έως I15.



Εικόνα 5: WHO - Σύνολο θανάτων, υπερτασικές νόσοι, άνδρες - γυναίκες, όλες οι ηλικίες για ICD10 από I10 – I15³

² <http://apps.who.int/healthinfo/statistics/mortality/whodpms/>

2.1.5. Αμιγής υπερχοληστερολαιμία – Pure hypercholesterolaemia – ICD10: E78

Το ICD10 E78 (αμιγής υπερχοληστερολαιμία – pure hypercholesterolaemia) ανήκει στις ενδοκρινικές, διατροφικές και μεταβολικές παθήσεις (Endocrine, nutritional and metabolic diseases) Η ομάδα κωδικών ICD10 από E70 έως E90 ανήκει στις μεταβολικές διαταραχές (metabolic disorders) (WHO, 2004)

Η υπερχοληστερολαιμία ή δυσλιπιδαιμία είναι η παρουσία υψηλών επιπέδων χοληστερόλης στο αίμα. (Durrington, 2003; Bhatnagar et al., 2008)

Η πρωτογενής υπερχοληστερολαιμία εκτός των διατροφικών συνηθειών όπως διατροφή πλούσια σε κορεσμένα λίπη και φτωχή σε φυτικές ίνες οφείλεται και σε άλλους προδιαθεσιακούς παράγοντες όπως η παχυσαρκία , η κατανάλωση αλκοόλ, η έλλειψη άσκησης, το κάπνισμα, και η κληρονομικότητα. (Bhatnagar et al., 2008)

Παράγοντες κινδύνου δευτερογενούς υπερχοληστερολαιμίας είναι διάφορες συστηματικές νόσοι ή καταστάσεις όπως το νεφρωσικό σύνδρομο, ο υποθυρεοειδισμός, η ηπατική ανεπάρκεια, η νευρική ανορεξία, ή ορισμένα φάρμακα όπως τα θειαζιδικά διουρητικά, η κυκλοσπορίνη κ.α. (Bhatnagar et al., 2008)

Παράγοντες κινδύνου μεικτής υπερχοληστερολαιμίας είναι νοσήματα που επηρεάζουν τον μεταβολισμό όπως ο σακχαρώδης διαβήτης και η νεφρική ανεπάρκεια, η παχυσαρκία, οι διατροφικές συνήθειες και ορισμένα φάρμακα όπως οι β αποκλειστές, τα γλυκοκορτικοειδή κ.α. (Bhatnagar et al., 2008)

³ <http://apps.who.int/healthinfo/statistics/mortality/whodpms/>

Μέτρα για να μειωθεί η υπερχοληστερολαιμία είναι (Bhatnagar et al., 2008):

- ❖ η διατροφή που δίνει έμφαση σε τρόφιμα με χαμηλή περιεκτικότητα χοληστερόλης (Tang et al., 1998)
- ❖ η διακοπή του καπνίσματος
- ❖ η μείωση του αλκοόλ
- ❖ η σωματική άσκηση
- ❖ η διατήρηση φυσιολογικού βάρους

Τα αυξημένα επίπεδα χοληστερόλης αυξάνουν τους κινδύνους καρδιαγγειακής νόσου, αθηροσκληρωτικής νόσου και εγκεφαλικού επεισοδίου. (Ezzati et al., 2002)

2.1.6. ICD10 - International statistical classification of diseases and related health problems

Το πρότυπο ICD10 (International statistical classification of diseases and related health problems - Διεθνής Στατιστική Ταξινόμηση των Νόσων και των Σχετικών Προβλημάτων Υγείας) χρησιμοποιείται για την ταξινόμηση των νόσων, και ορίζεται ως ένα σύστημα κατηγοριών, στο οποίο καταγράφονται οι νοσολογικές οντότητες σύμφωνα με προκαθορισμένα κριτήρια. (WHO, 2004)

Ο σκοπός του προτύπου ICD10 είναι να επιτρέπει τη συστηματική καταγραφή, ανάλυση, ερμηνεία και σύγκριση των δεδομένων θνησιμότητας και νοσηρότητας, που συλλέγονται στις ιατρικές βάσεις δεδομένων σε διάφορες χώρες. (WHO, 2004)

Το πρότυπο ICD10 χρησιμοποιείται προκειμένου να μετατρέψει τη διάγνωση των νόσων από λέξεις, σ' ένα αλφαριθμητικό κωδικό, ο οποίος επιτρέπει την εύκολη αποθήκευση, ανάκτηση και ανάλυση των δεδομένων. (WHO, 2004)

ICD10	Περιγραφή
A00-B99	Ορισμένα λοιμώδη και παρασιτικά νοσήματα
C00-D48	Νεοπλάσματα
D50-D89	Παθήσεις του αίματος, των αιμοποιητικών οργάνων και ορισμένες διαταραχές του ανοσολογικού μηχανισμού
E00-E90	Ενδοκρινικές, διατροφικές και μεταβολικές παθήσεις
F00-F99	Ψυχικές διαταραχές και διαταραχές συμπεριφοράς
G00-G99	Παθήσεις του νευρικού συστήματος
H00-H59	Παθήσεις του οφθαλμού και των εξαρτημάτων του
H60-H95	Παθήσεις του ωτός και της μαστοειδούς απόφυσης
I00-I99	Παθήσεις του κυκλοφορικού συστήματος
J00-J99	Παθήσεις του αναπνευστικού συστήματος
K00-K93	Παθήσεις του πεπτικού συστήματος
L00-L99	Παθήσεις του δέρματος και του υποδόριου ιστού
M00-M99	Παθήσεις του μυοσκελετικού συστήματος και του συνδετικού ιστού
N00-N99	Παθήσεις του ουροποιογεννητικού συστήματος
O00-O99	Κύηση, τοκετός και λοχεία
P00-P96	Καταστάσεις που έχουν την αρχή τους στην περιγεννητική περίοδο
Q00-Q99	Συγγενείς ανωμαλίες, διαμαρτίες της διάπλασης και χρωμοσωμικές ανωμαλίες
R00-R99	Συμπτώματα, σημεία και παθολογικά κλινικά και εργαστηριακά ευρήματα που δεν ταξινομούνται αλλού
S00-T98	Συνέπειες τραυματισμού, δηλητηρίασης και ορισμένες άλλες εξωγενείς αιτίες
V01-Y98	Εξωγενή αίτια νοσηρότητας και θνησιμότητας
Z00-Z99	Παράγοντες που επηρεάζουν την κατάσταση της υγείας και την επικοινωνία με τις υπηρεσίες υγείας
U00-U99	Κωδικοί για ειδικούς σκοπούς

Πίνακας 2: Ταξινόμηση ICD10⁴

⁴ <http://apps.who.int/classifications/icd10/browse/2016/en>

2.2. Σχετικές επιστημονικές εργασίες

Όπως αναφέραμε, βασικός παράγοντας κινδύνου της καρδιαγγειακής νόσου και του εγκεφαλικού επεισοδίου είναι ο συνδυασμός υπέρτασης και αυξημένων επιπέδων χοληστερόλης. Αυτός ο συνδυαστικός παράγοντας κινδύνου αποτελεί παγκόσμιο πρόβλημα δημόσιας υγείας και πρέπει να ληφθούν μέτρα για να μειωθεί ο αριθμός των περιστατικών. Διάφορες τεχνικές εξόρυξης γνώσης χρησιμοποιούνται για τον προσδιορισμό αυτού του συνδυαστικού παράγοντα κινδύνου, με σκοπό την πρόβλεψη ή την εξαγωγή χρήσιμων συσχετίσεων.

Οι Ture et al. (Ture et al., 2005) σύγκριναν τις επιδόσεις εννέα τεχνικών ταξινόμησης για την πρόβλεψη της ιδιοπαθούς υπέρτασης. Έγινε συγκριτική μελέτη στις επιδόσεις από τρία δέντρα απόφασης, τέσσερεις στατιστικούς αλγόριθμους και δύο νευρωνικά δίκτυα με ιεραρχική ανάλυση συστάδων. Η μελέτη έγινε σε 694 άτομα όπου, οι 452 ήταν ασθενείς και οι 242 ήταν μάρτυρες δηλαδή υγιείς.

Μεταβλητές για την πρόβλεψη της ιδιοπαθούς υπέρτασης ήταν: η ηλικία, το φύλο, το οικογενειακό ιστορικό υπέρτασης, η συνήθεια του καπνίσματος, το επίπεδο των τριγλυκεριδίων στον ορό του αίματος, το επίπεδο της λιποπρωτεΐνης (α) στον ορό του αίματος, το επίπεδο του ουρικού οξέος στον ορό του αίματος, το επίπεδο της ολικής χοληστερόλης στον ορό του αίματος και ο δείκτης μάζας σώματος (BMI).

Τα τρία δέντρα απόφασης ήταν: Chi-squared Automatic Interaction Detector (CHAID), Classification and Regression Tree (CART) και Quick, Unbiased, Efficient Statistical Tree (QUEST). Οι τέσσερεις στατιστικοί αλγόριθμοι ήταν: Logistic

Regression Analysis (LR), Flexible Discriminant Analysis (FDA), Multivariate Additive Regression Splines (MARS) και FDA / MARS. Τα δύο νευρωνικά δίκτυα ήταν: Multi-Layer Perceptron (MLP) και Radial Basis Function (RBF).

Η αξιολόγηση όλων των μοντέλων έγινε βάσει: της ευαισθησίας (sensitivity), της ειδικότητας (specificity) και της προγνωστικής τιμής (PR).

Το αποτέλεσμα της έρευνας ήταν ότι τα νευρωνικά δίκτυα Radial Basis Function (RBF) και Multi-Layer Perceptron (MLP) απέδωσαν καλύτερα, απ' ό,τι άλλες τεχνικές στην πρόβλεψη της υπέρτασης. Το δέντρο απόφασης Quick, Unbiased, Efficient Statistical Tree (QUEST) είχε τη χαμηλότερη απόδοση.

Οι Akdag et al. (Akdag et al., 2006) έκαναν μια έρευνα με σκοπό τον προσδιορισμό των παραγόντων κινδύνου για την υπέρταση, με δέντρα ταξινόμησης. Η μελέτη αυτή χρησιμοποιεί δεδομένα από τον Ιανουάριο του 2002 έως και τον Ιούλιο του 2004, εξωτερικών ασθενών από μια κλινική στο Ντενιζλί στην Δυτική Τουρκία και περιλαμβάνει 1761 άτομα.

Το αποτέλεσμα της έρευνας ήταν ότι βασικοί παράγοντες κινδύνου της υπέρτασης είναι: η υψηλή πίεση σε συγγενείς πρώτου βαθμού, το όριο της περιμέτρου της μέσης προς την περίμετρο των γοφών, το επίπεδο της ολικής χοληστερόλης στον ορό του αίματος, το επίπεδο των τριγλυκεριδίων στον ορό του αίματος, ο δείκτης μάζας σώματος, το φύλο και η κατανάλωση κεκορεσμένων λιπαρών.

Οι Dogan & Turkoglu (Dogan & Turkoglu, 2008) έκαναν μια μελέτη όπου με την χρήση κανόνων συσχέτισης γινόταν η πρόβλεψη της υπερλιπιδαιμίας. Η μελέτη αυτή περιλαμβάνει 492 ασθενείς.

Ως παράμετροι εισόδου δινόταν: ολική χοληστερόλη, LDL, τριγλυκερίδια, HDL και VLDL. Ως έξοδο, το σύστημα έλεγε αν κάποιος είχε ή δεν είχε υπερλιπιδαιμία. Τα αποτελέσματα κανόνων συσχέτισης ταυτίζονταν πλήρως με τις αποφάσεις των ιατρών.

Οι Chang et al. (Chang et al., 2011) έκαναν μια έρευνα χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων για τις πολυ-ασθένειες (multi-diseases), με σκοπό την πρόβλεψη υπέρτασης και υπερλιπιδαιμίας βάσει κοινών παραγόντων κινδύνου. Η μελέτη αυτή χρησιμοποιεί δεδομένα από ένα εξεταστικό κέντρο στην Ταϊβάν και περιλαμβάνει 2048 άτομα.

Η μελέτη χωρίστηκε σε 2 στάδια. Στο πρώτο στάδιο γίνεται επιλογή των παραγόντων κινδύνου της υπέρτασης και της υπερλιπιδαιμίας με έξι προσεγγίσεις εξόρυξης δεδομένων: logistic regression analysis, C5.0 decision tree, Classification And Regression Tree (CART), Chi-squared Automatic Interaction Detector (CHAID), exhaustive CHAID και discriminant analysis.

Στο δεύτερο στάδιο με βάση τους κοινούς παράγοντες κινδύνου της υπέρτασης και της υπερλιπιδαιμίας χρησιμοποιεί την προσέγγιση Multivariate Adaptive Regression Splines (MARS) για την κατασκευή ενός μοντέλου πρόβλεψης υπέρτασης και υπερλιπιδαιμίας.

Η αξιολόγηση έγινε βάσει: της ευαισθησίας (sensitivity), της ειδικότητας (specificity), της ακρίβειας (accuracy), του τύπου σφάλματος I (type I error) και του τύπου σφάλματος II (type II error).

Το αποτέλεσμα της έρευνας ήταν, ότι κοινοί παράγοντες κινδύνου της υπέρτασης και της υπερλιπιδαιμίας είναι: η συστολική αρτηριακή πίεση, τα τριγλυκερίδια, το ουρικό οξύ, η γλουταμινική πυροσταφυλική τρανσαμινάση και το φύλο.

Ο παρακάτω πίνακας περιγράφει τις μελέτες που αναφέραμε, καθώς και τα αποτελέσματα που έχουν εξαχθεί.

Πίνακας 3: Μελέτες εξόρυξης γνώσης για υπέρταση και δυσλιπιδαιμίες

Συγγραφέας	Χρόνος	Research Group	Data Set	Data Mining Methods / Algorithms	Αποτελέσματα
Ture et al.	2005	πρόβλεψη της ιδιοπαθούς υπέρτασης	694 άτομα	CHAID, CART , QUEST, LR, FDA, MARS, FDA / MARS, MLP, RBF	Απόδοση των μοντέλων με ιεραρχική ανάλυση συστάδων. Τα νευρωνικά δίκτυα (RBF) και (MLP) απέδωσαν καλύτερα ενώ το δέντρο απόφασης (QUEST) είχε τη χαμηλότερη απόδοση.
Akdag et al.	2006	πρόβλεψη παραγόντων κινδύνου υπέρτασης	1761 άτομα	Classification	Οι βασικοί παράγοντες κινδύνου της υπέρτασης είναι: η υψηλή πίεση σε συγγενείς πρώτου βαθμού, το όριο της περιμέτρου της μέσης προς την περίμετρο των γοφών, το επίπεδο της ολικής χοληστερόλης στον ορό του αίματος, το επίπεδο των τριγλυκεριδίων στον ορό του αίματος, ο δείκτης μάζας σώματος, το φύλο και η κατανάλωση κεκορεσμένων λιπαρών
Dogan & Turkoglu	2008	πρόβλεψη υπερλιπιδαιμίας	492 ασθενείς	Association Rules	Ως παράμετροι εισόδου δινόταν: ολική χοληστερόλη, LDL, τριγλυκερίδια, HDL και VLDL. Ως έξοδο, το σύστημα έλεγε αν κάποιος είχε ή δεν είχε υπερλιπιδαιμία.
Chang et al.	2011	πρόβλεψη υπέρτασης και υπερλιπιδαιμίας βάσει κοινών παραγόντων κινδύνου	2048 άτομα	logistic regression analysis, C5.0 , CART, CHAID, exhaustive CHAID, discriminant analysis	Οι κοινοί παράγοντες κινδύνου της υπέρτασης και της υπερλιπιδαιμίας είναι: η συστολική αρτηριακή πίεση, τα τριγλυκερίδια, το ουρικό οξύ, η γλουταμινική πυροσταφυλική τρανσαμινάση και το φύλο.

2.3. Περιγραφή συστημάτων που χρησιμοποιήθηκαν στη διατριβή

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε ως βάση δεδομένων η access, και με το Neatbeans γράφτηκε κώδικας Java για να δημιουργήσουμε τα έξι σύνολα δεδομένων που χρησιμοποιήσαμε. Τέλος, χρησιμοποιήθηκε το λογισμικό Rapid Miner, το οποίο είναι ένα από τα πιο αναγνωρισμένα εργαλεία στο πεδίο της εξόρυξης γνώσης.

Το Rapid Miner είναι ένα open source λογισμικό και αποτελεί ένα ολοκληρωμένο περιβάλλον για μηχανική εκμάθηση, εξόρυξη γνώσης, προβλεπτική και επιχειρησιακή ανάλυση. Χρησιμοποιείται στην ιατρική, στη βιομηχανία, στο εμπόριο και σε διάφορους άλλους τομείς. Πρόκειται για ένα σύστημα που υποστηρίζει τον σχεδιασμό και την τεκμηρίωση μιας συνολικής διαδικασίας εξόρυξης δεδομένων, καθώς προσφέρει τις δομές για τη ροή και τον έλεγχο της διαδικασίας. (Hofmann & Klinkenberg, 2013)

3. Ανάλυση της προσέγγισης

Τα ιατρικά δεδομένα προέρχονται από την ηλεκτρονική συνταγογράφηση στην περιοχή της Αθήνας, από Γενικό Ιατρό, για το χρονικό διάστημα από 7/1/2012 έως 31/7/2015. Συνολικά η βάση έχει 4.473 ασθενείς από τους οποίους οι 2.677 είναι Female – γυναίκες και οι 1.796 είναι Male – άνδρες. Υπάρχουν 225 μοναδικές διαγνώσεις, ενώ το αρχείο συνταγών ανά επίσκεψη έχει 21.843 εγγραφές.

Από τις 21.843 εγγραφές, στις 195 δεν μπορέσαμε να δώσουμε ICD10 κωδικό, γιατί κυρίως αφορούσαν παρακλινικές εξετάσεις όπως:

- ❖ ακτινογραφίες
- ❖ αξονική τομογραφία (C.T)
- ❖ μαγνητική τομογραφία (M.R.I.)
- ❖ Χημειοθεραπείες (Χ.Μ.Θ)
- ❖ ακτινοβολίες
- ❖ γενικές εξετάσεις αίματος και ούρων.

Αυτές τις 195 εγγραφές τις διαγράψαμε.

Το αρχείο δραστικών ουσιών περιέχει 45.928 εγγραφές, όπου ανά συνταγή μπορούσαν να γραφτούν έως 3 δραστικές ουσίες. Στο αρχείο υπάρχουν 768 μοναδικές δραστικές ουσίες. Οι διαγνώσεις στο αρχείο συνταγών ήταν σε ένα πεδίο, χωρισμένες με κόμματα και ελληνική περιγραφή. Χωρίσαμε τις διαγνώσεις σε πεδία και δώσαμε το ICD10 που αντιστοιχούσε στην κάθε διάγνωση.

Ο χωρισμός των ηλικιακών ομάδων στην βάση δεδομένων έγινε βάσει του Παγκόσμιου Οργανισμού Υγείας (World Health Organization – WHO) (Boschi-Pinto et al., 2008)

Ηλικιακές ομάδες βάσει του WHO
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+

Πίνακας 4: Ηλικιακές ομάδες βάσει του World Health Organization (WHO)

Στο αρχείο ασθενών δώσαμε έναν μοναδικό αύξοντα κωδικό (PatID) αντί του Α.Μ.Κ.Α. που υπήρχε, κρατήσαμε το φύλο (Sex) και χωρίσαμε τους ασθενείς σε ηλικιακές ομάδες (AgeGroup).

Αφαιρέσαμε από το αρχείο ασθενών:

- το Α.Μ.Κ.Α.
- το Α.Μ.Α.
- το ασφαλιστικό ταμείο από το οποίο προέρχονται

- το ονοματεπώνυμο
- τη διεύθυνση
- το τηλέφωνο
- αν ήταν έμμεσο ή άμεσο μέλος

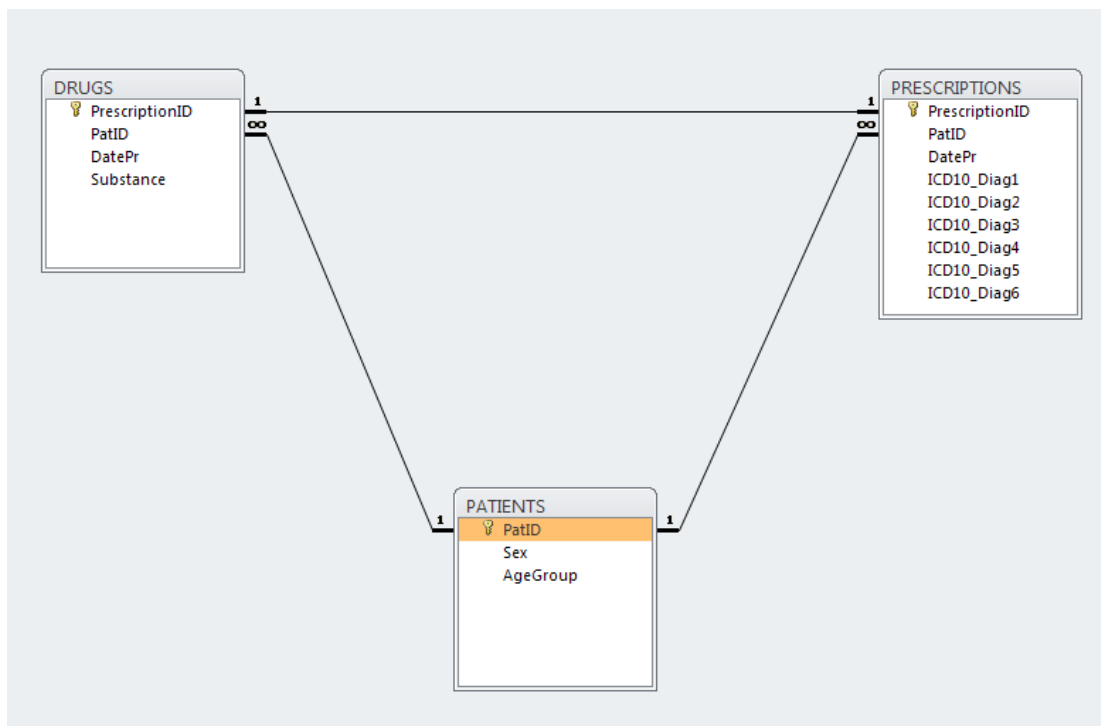
Αφαιρέσαμε από το αρχείο δραστικών ουσιών:

- το barcode
- την ποσότητα
- την τιμή εκτέλεσης

Αφαιρέσαμε από το αρχείο συνταγών:

- την ημερομηνία εκτέλεσης
- αν ήταν επαναλαμβανόμενη
- την κατάσταση

Το τελικό σχεσιακό σχήμα της βάσης φαίνεται στο παρακάτω σχήμα:



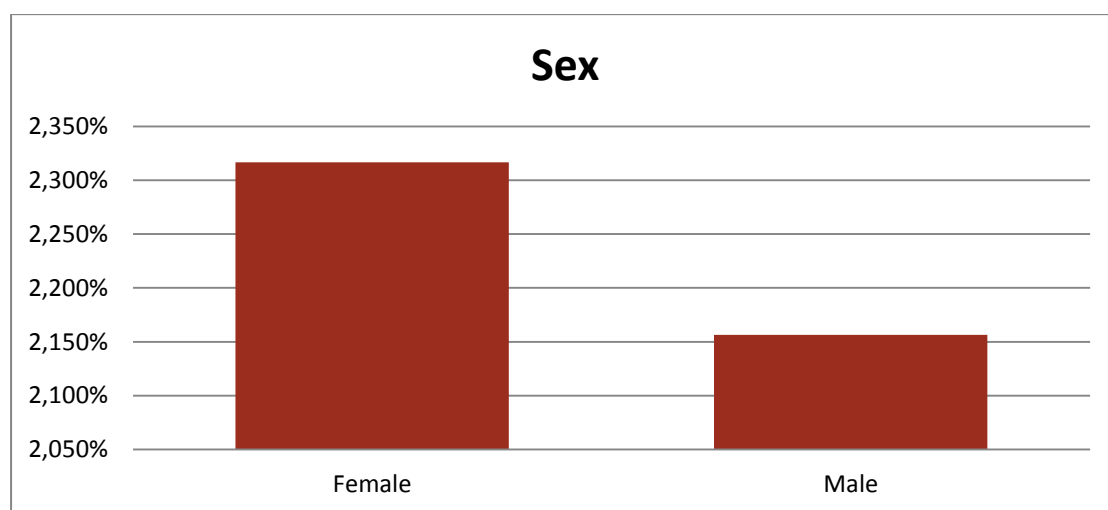
Εικόνα 6: Το σχεσιακό σχήμα της βάσης

Στα παρακάτω σχήματα η κατανομή έγινε βάσει των στοιχείων της ΕΛ.ΣΤΑΤ. όπου ο πληθυσμός της Ελλάδας στην απογραφή του 2011 ήταν 10.815.197 και ο πληθυσμός της Αττικής ήταν 3.828.434.⁵ Οι ηλικιακές ομάδες που χρησιμοποιεί η ΕΛ.ΣΤΑΤ. ήταν λιγότερες από τον World Health Organization (WHO) όποτε κάναμε συγχώνευση των ηλικιακών ομάδων για τα παρακάτω σχήματα.

Ηλικιακές ομάδες βάσει της ΕΛ.ΣΤΑΤ.
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80+

Πίνακας 5: Ηλικιακές ομάδες βάσει της ΕΛ.ΣΤΑΤ.

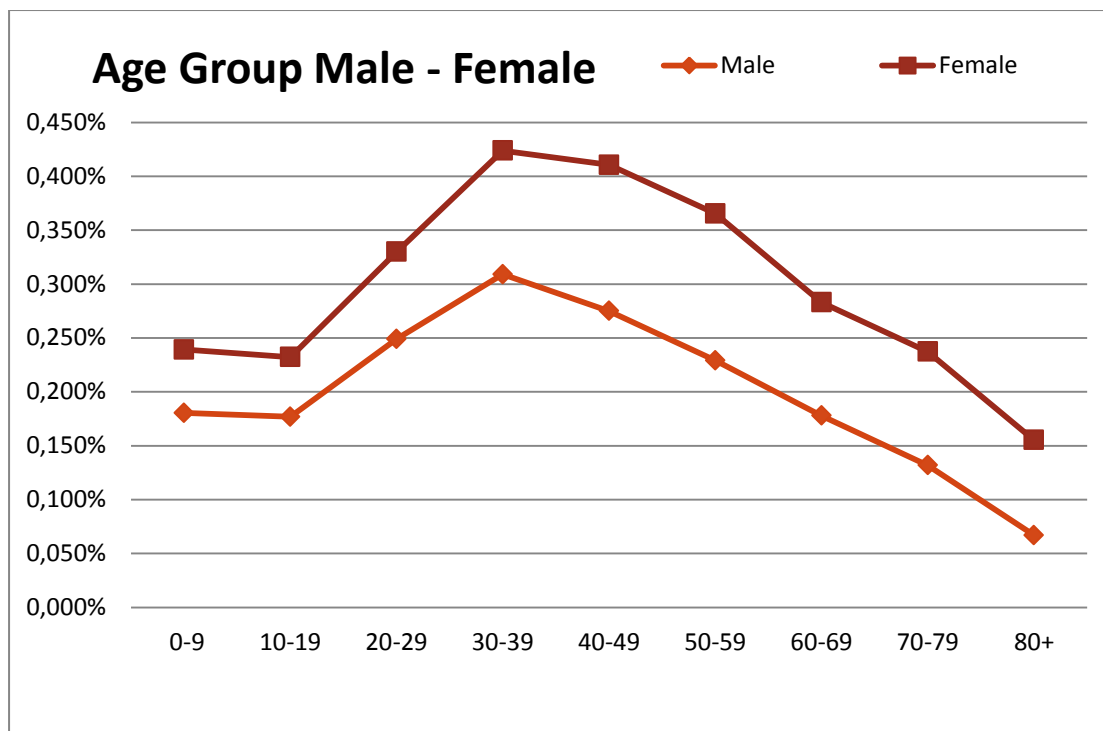
Η κατανομή των ανδρών και των γυναικών στις 100.000 φαίνεται στην παρακάτω εικόνα.



Σχήμα 1: Κατανομή ανδρών – γυναικών στις 100.000

⁵ <http://www.statistics.gr/demographic-data>

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς, ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



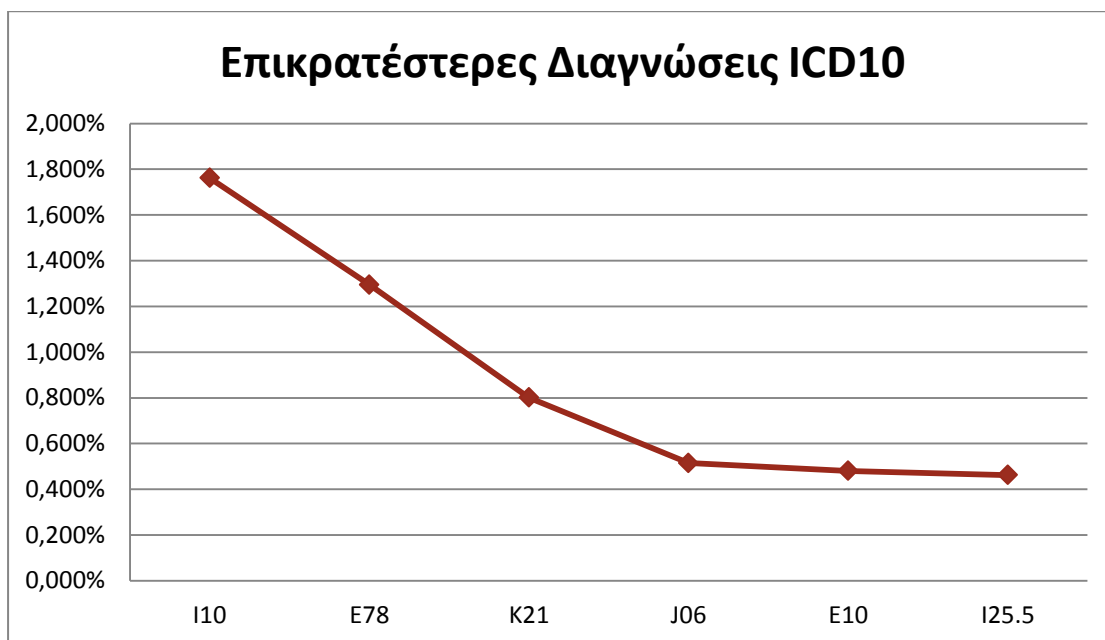
Σχήμα 2: Κατανομή ανδρών – γυναικών ασθενών ανά ηλικιακή ομάδα στις 100.000

Οι επικρατέστερες διαγνώσεις με πάνω από 400 ασθενείς είναι

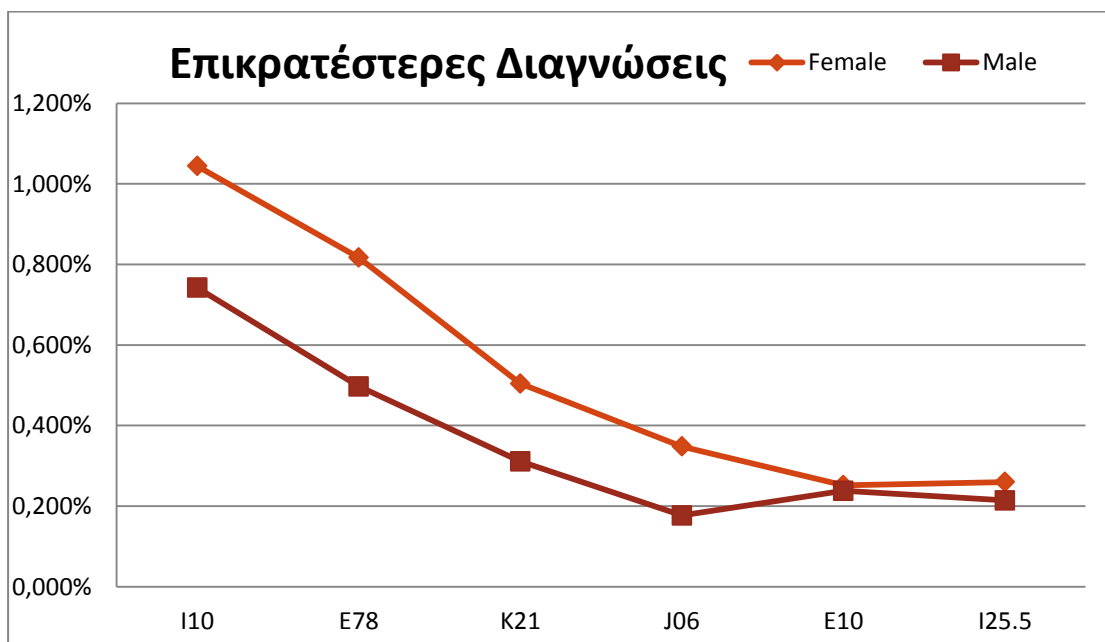
ICD10	Περιγραφή
I10	Ιδιοπαθής (πρωτοπαθής) υπέρταση
E78	Αμιγής υπερχοληστερολαιμία
K21	Γαστροοισοφαγική παλινδρόμηση με οισοφαγίτιδα
J06	Λοίμωξη αναπνευστικού
E10	Ινσουλινοεξαρτώμενος σακχαρώδης διαβήτης
I25.5	Στεφανιαία νόσος

Πίνακας 6: Επικρατέστερες Διαγνώσεις

Από τους 4.473 ασθενείς, οι 1.762 έχουν υπέρταση και οι 1.295 έχουν υψηλή χοληστερόλη. Από τους ασθενείς που έχουν υπέρταση (I10) ή υψηλή χοληστερόλη (E78), οι 729 ασθενείς έχουν και τις 2 παθήσεις.

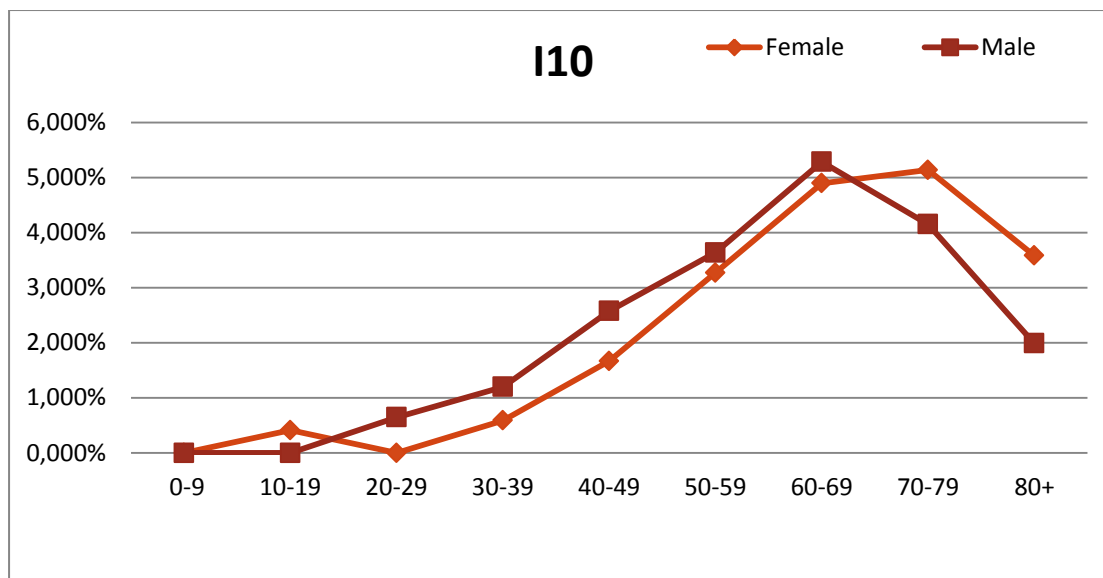


Σχήμα 3: Κατανομή επικρατέστερων διαγνώσεων στις 100.000



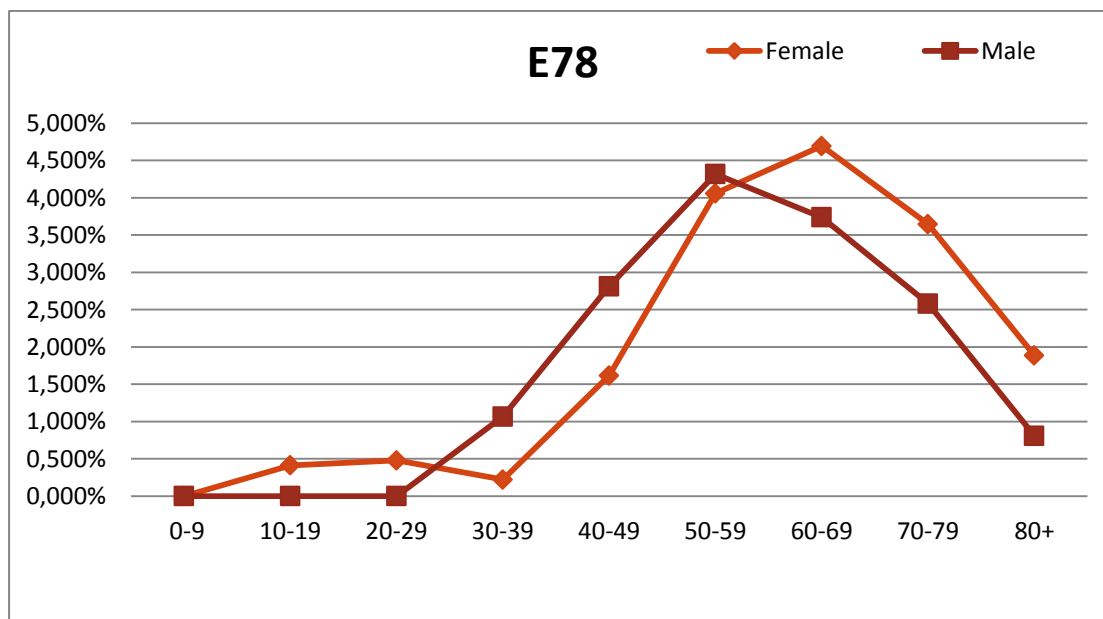
Σχήμα 4: Κατανομή επικρατέστερων διαγνώσεων ανδρών-γυναικών στις 100.000

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς που εμφανίζουν υπέρταση (I10), ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



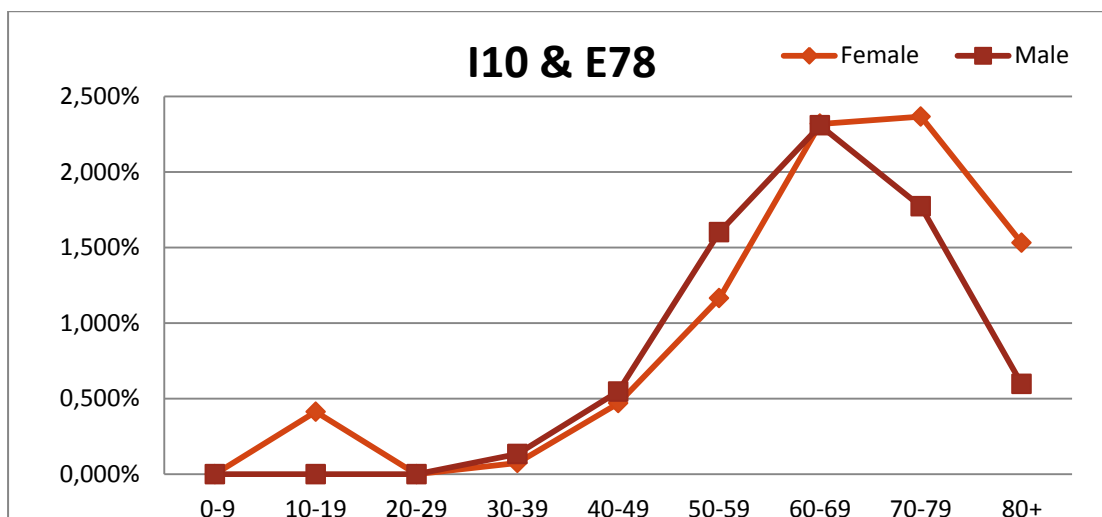
Σχήμα 5: Κατανομή I10 ανδρών-γυναικών στις 100.000

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς που εμφανίζουν υψηλή χοληστερόλη (E78), ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



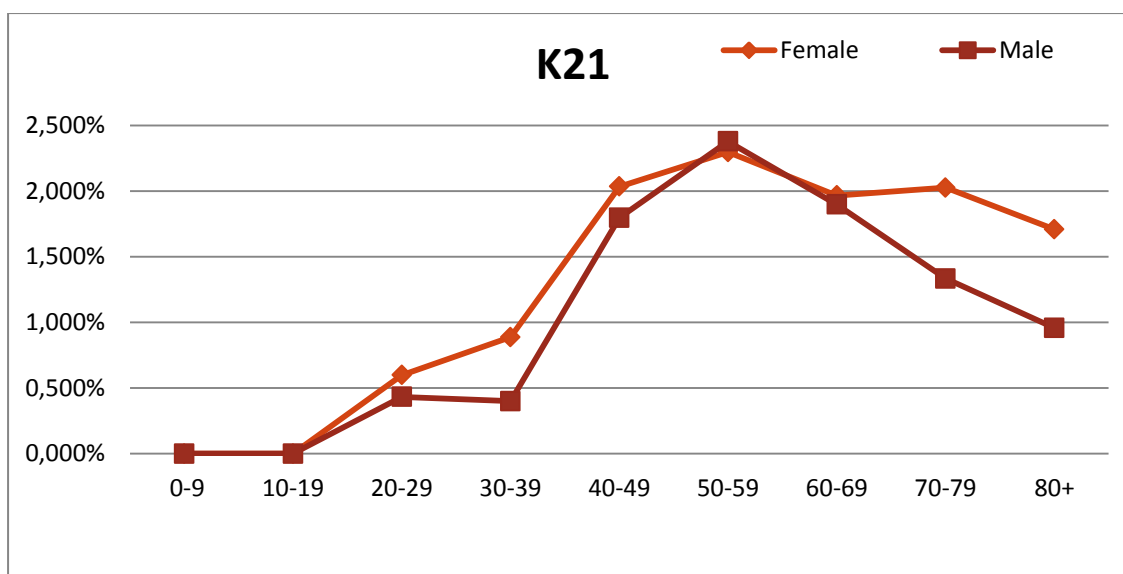
Σχήμα 6: Κατανομή E78 ανδρών-γυναικών στις 100.000

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς που εμφανίζουν υπέρταση (I10) και υψηλή χοληστερόλη (E78), ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



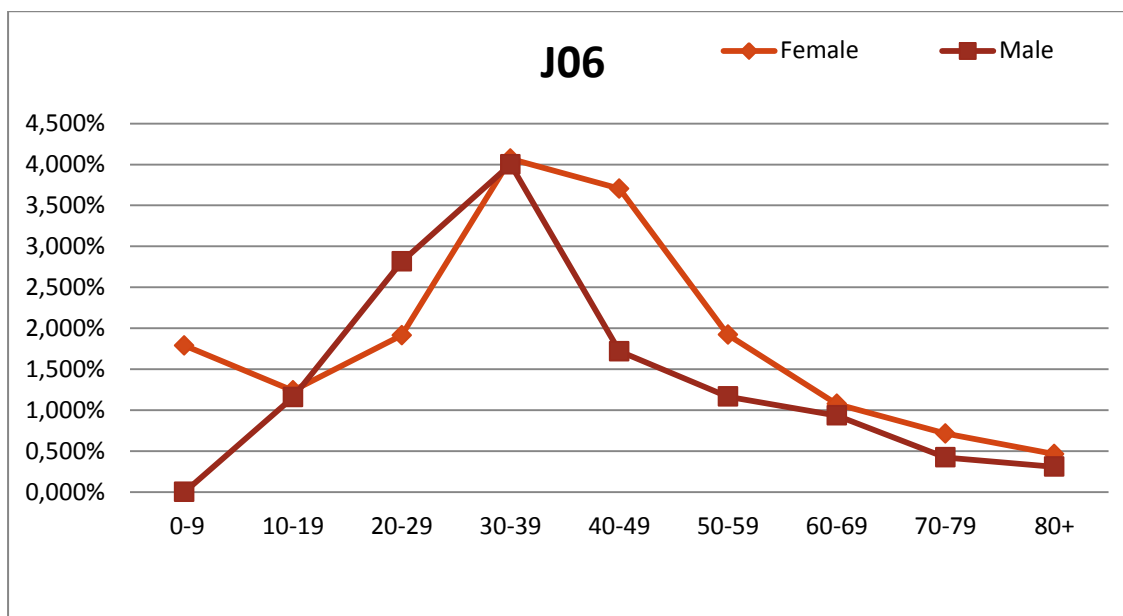
Σχήμα 7: Κατανομή I10 & E78 ανδρών-γυναικών στις 100.000

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς που εμφανίζουν γαστροοισοφαγική παλινδρόμηση με οισοφαγίτιδα (K21), ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



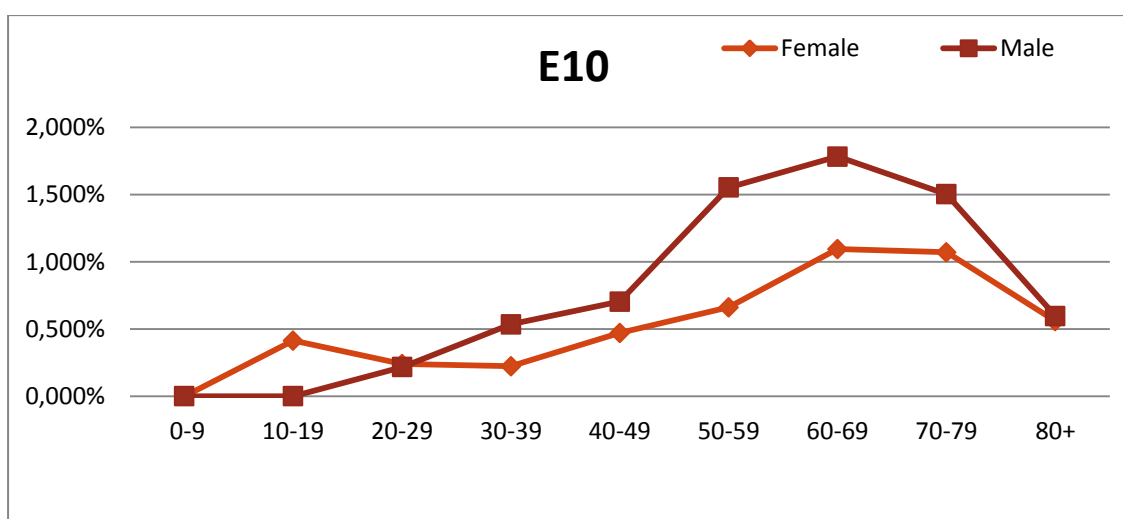
Σχήμα 8: Κατανομή K21 ανδρών-γυναικών στις 100.000

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς που εμφανίζουν λοίμωξη του αναπνευστικού (J06), ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



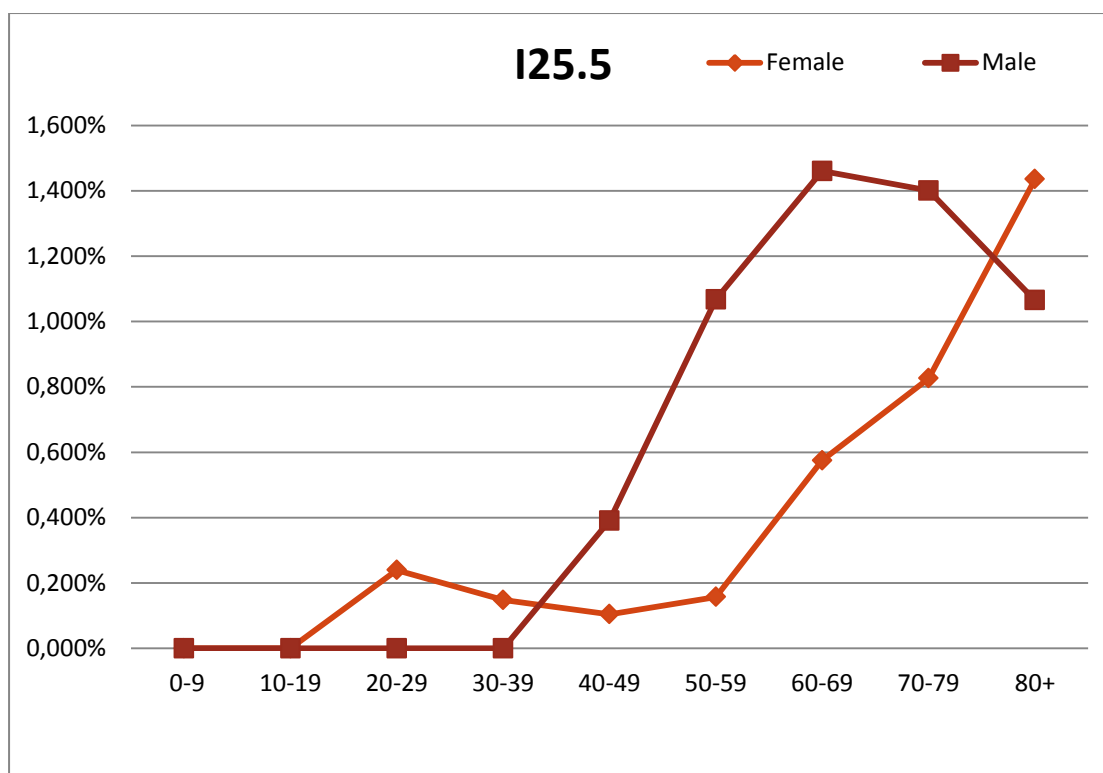
Σχήμα 9: Κατανομή J06 ανδρών-γυναικών στις 100.000

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς που εμφανίζουν (E10) ινσουλινοεξαρτώμενο σακχαρώδη διαβήτη, ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



Σχήμα 10: Κατανομή E10 ανδρών-γυναικών στις 100.000

Οι γυναίκες ασθενείς και οι άνδρες ασθενείς που εμφανίζουν στεφανιαία νόσο (I25.5), ανά ηλικιακή ομάδα, στις 100.000 φαίνεται στην παρακάτω εικόνα.



Σχήμα 11: Κατανομή I25.5 ανδρών-γυναικών στις 100.000

Στην συνέχεια δημιουργηθήκαν έξι σύνολα δεδομένων με το Neatbeans:

- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ημερομηνία επίσκεψης
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά ημερομηνία επίσκεψης

- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10 & E78, ανά ημερομηνία επίσκεψης
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ασθενή, στο σύνολο όλων των επισκέψεων του
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά ασθενή, στο σύνολο όλων των επισκέψεων του
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10 & E78, ανά ασθενή, στο σύνολο όλων των επισκέψεων του

Τα έξι σύνολα δεδομένων έχουν ως πεδία:

- ❖ PatID (μοναδικός κωδικός ασθενή)
- ❖ DatePr (ημερομηνία συνταγής)
- ❖ Diagnosis (διάγνωση σε ICD10)
- ❖ Substance (δραστική ουσία)

Επιλέξαμε τις μεθόδους που θα χρησιμοποιηθούν στα έξι σύνολα δεδομένων για την εξαγωγή χρησίμων πληροφοριών, που θα μας οδηγήσουν σε γνώση. Επιλέξαμε ως μεθόδους τους αλγόριθμους FP-Growth για την εξαγωγή κανόνων συσχέτισης και GSP για την εξαγωγή ακολουθιακών προτύπων.

Η εξαγωγή κανόνων συσχέτισης έχει ως στόχο να ανακαλυφθούν ενδιαφέρουσες συσχετίσεις και πρότυπα μεταξύ του συνόλου των δεδομένων. Κάθε κανόνας εκφράζεται με δύο μέτρα, την υποστήριξη κανόνα (support) και την εμπιστοσύνη κανόνα (confidence). Το βασικό πρόβλημα στην εξαγωγή κανόνων συσχέτισης είναι ότι εξάγεται μεγάλος αριθμός κανόνων. Για να μειωθεί ο αριθμός των κανόνων που εξάγονται, έχουμε ως περιορισμούς την ελάχιστη υποστήριξη (minimum support) και την ελάχιστη εμπιστοσύνη (minimum confidence). (Agrawal & Srikant, 1994) Ο πιο διαδεδομένος αλγόριθμος είναι ο FP-Growth.

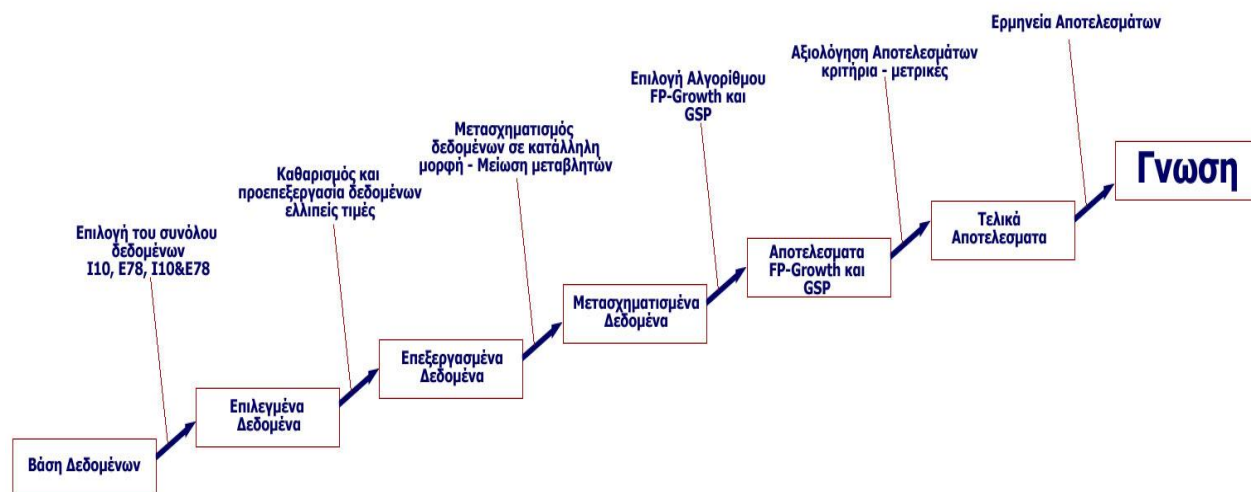
Τα ακολουθιακά πρότυπα επεκτείνουν τη μέθοδο των κανόνων συσχέτισης και έχουν ως στόχο, μέσα από ένα σύνολο δεδομένων να ανακαλυφθούν ακολουθίες στο μέγιστο μήκος. Κάθε ακολουθιακό πρότυπο εκφράζεται με ένα μέτρο, την υποστήριξη κανόνα (support). Το βασικό πρόβλημα στην εξαγωγή ακολουθιακών προτύπων είναι ότι εξάγεται μεγάλος αριθμός προτύπων. Για να μειωθεί ο αριθμός ακολουθιακών προτύπων που εξάγονται, έχουμε ως περιορισμό την ελάχιστη υποστήριξη (minimum support). (Agrawal & Srikant, 1995) Ο πιο διαδεδομένος αλγόριθμος είναι ο GSP.

Τέλος, χρησιμοποιήθηκε το λογισμικό Rapid Miner και από τα αποτελέσματα έγινε καθαρισμός των κανόνων και επιλέχθηκαν κανόνες που περιείχαν διαφορετική νόσο και δραστικές ουσίες.

4. Σχεδίαση

4.1. Αρχιτεκτονικό διάγραμμα

Το αρχιτεκτονικό διάγραμμα σχεδίασης, το οποίο ακολουθήσαμε, αποτελείται από τα εξής βήματα:



Εικόνα 7: Αρχιτεκτονικό Διάγραμμα

❖ **Επιλογή του συνόλου των δεδομένων:** Σ' αυτό το βήμα, επιλέξαμε ως σύνολο δεδομένων στα οποία θα εκτελεστεί η διαδικασία της εξόρυξης δεδομένων:

- ασθενείς που έχουν υπέρταση (I10),
- ασθενείς που έχουν υψηλή χοληστερόλη (E78) και
- ασθενείς που έχουν και τα δυο ταυτόχρονα.

❖ **Καθαρισμός και προεπεξεργασία των δεδομένων :** Σ' αυτό το βήμα, για τον χειρισμό των ελλειπόντων τιμών στα πεδία δεδομένων κάναμε το εξής:

- σε 195 εγγραφές δεν μπορέσαμε να δώσουμε ICD10 κωδικό, γιατί αφορούσαν παρακλινικές εξετάσεις, όπως ακτινογραφίες, C.T, M.R.I., Χ.Μ.Θ (χημειοθεραπείες), ακτινοβολίες, γενικές εξετάσεις αίματος και ούρων. Αυτές τις 195 εγγραφές τις διαγράψαμε.

❖ **Μετασχηματισμός των δεδομένων:** Σ' αυτό το βήμα, μειώσαμε τον αριθμό των μεταβλητών που θα εξεταστούν στην εξόρυξη δεδομένων και μετασχηματίσαμε τα δεδομένα.

- Οι διαγνώσεις ήταν σε ένα πεδίο, χωρισμένες με κόμματα και ελληνική περιγραφή. Χωρίσαμε τις διαγνώσεις σε πεδία και δώσαμε το ICD10 που αντιστοιχούσε στην κάθε διάγνωση.
- Ο χωρισμός των ηλικιακών ομάδων έγινε βάσει του Παγκόσμιου Οργανισμού Υγείας (World Health Organization – WHO)
- Αφαιρέσαμε από το αρχείο ασθενών: το Α.Μ.Κ.Α., το Α.Μ.Α., το ασφαλιστικό ταμείο από το οποίο προέρχονται, το ονοματεπώνυμο, τη διεύθυνση, το τηλέφωνο, αν ήταν έμμεσο ή άμεσο μέλος
- Αφαιρέσαμε από το αρχείο δραστικών ουσιών: το barcode, την ποσότητα, την τιμή εκτέλεσης

- Αφαιρέσαμε από το αρχείο συνταγών: την ημ/νία εκτέλεσης, αν ήταν επαναλαμβανόμενη, την κατάσταση

❖ **Εξόρυξη δεδομένων:** Σ' αυτό το βήμα περιλαμβάνεται η επιλογή των μεθόδων που θα χρησιμοποιηθούν για την εξαγωγή χρήσιμων προτύπων γνώσης. Επιλέξαμε τους αλγόριθμους FP-Growth και GSP. Η εξαγωγή κανόνων συσχέτισης έχει ως στόχο να εξαχθούν ενδιαφέρουσες συσχετίσεις και πρότυπα, μεταξύ του συνόλου των δεδομένων. Ο FP-growth αποτελεί έναν από τους πιο διαδεδομένους αλγόριθμους κανόνων συσχέτισης και είναι ταχύτερος από τον Apriori. (Han et al., 2000) Τα ακολουθιακά πρότυπα επεκτείνουν τη μέθοδο των κανόνων συσχέτισης, ώστε να εντοπιστούν πρότυπα – μοτίβα μεταξύ των γεγονότων, τα οποία επαναλαμβάνονται στα δεδομένα. Ο GSP αλγόριθμος βρίσκει τα πιο συχνά διαδοχικά μοτίβα και είναι πιο γρήγορος από τον AprioriAll. (Agrawal & Srikant, 1995)

❖ **Αξιολόγηση αποτελεσμάτων:** Σ' αυτό το βήμα, τα αποτελέσματα που έχουν εξαχθεί από τους αλγόριθμους FP-Growth και GSP θα αξιολογηθούν με τα κριτήρια – μετρικές, προκειμένου να παραμείνουν μόνον τα πρότυπα που έχουν ενδιαφέρον και προσφέρουν γνώση.

- Κάθε κανόνας συσχέτισης εκφράζεται με δύο μετρικές, την υποστήριξη κανόνα (support) και την εμπιστοσύνη κανόνα (confidence). Για να μειωθεί ο αριθμός των κανόνων συσχέτισης που εξάγονται, έχουμε ως περιορισμούς την ελάχιστη υποστήριξη (minimum support) και την ελάχιστη εμπιστοσύνη (minimum confidence)

- Κάθε ακολουθιακό πρότυπο εκφράζεται με ένα μέτρο, την υποστήριξη κανόνα (Support). Για να μειωθεί ο αριθμός ακολουθιακών προτύπων που εξάγονται, έχουμε ως περιορισμό την ελάχιστη υποστήριξη (minimum support)

❖ **Ερμηνεία αποτελεσμάτων:** Σ' αυτό το βήμα τα αποτελέσματα των ακολουθιακών προτύπων και των κανόνων συσχέτισης θα πρέπει να ερμηνευθούν, για να εξαχθούν τα συμπεράσματα και η γνώση. Από τα αποτελέσματα έγινε καθαρισμός των κανόνων και επιλέχθηκαν κανόνες που περιείχαν διαφορετική νόσο και δραστικές ουσίες.

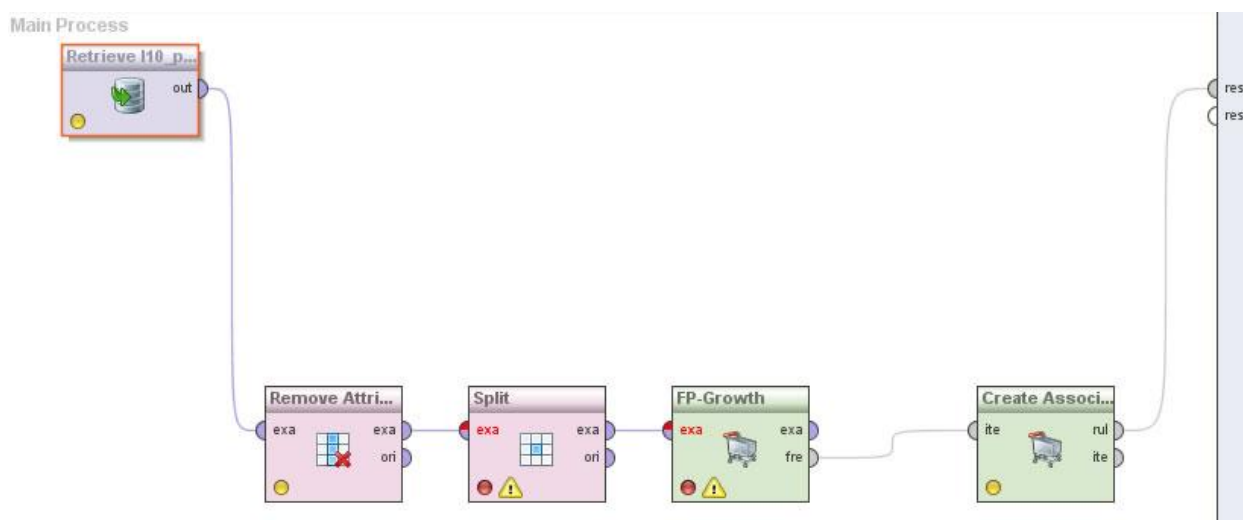
5. Υλοποίηση

5.1. Λεπτομέρειες υλοποίησης FP-Growth

Όπως αναφέραμε και παραπάνω, δημιουργήσαμε έξι σύνολα δεδομένων με το Neatbeans: ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ημερομηνία επίσκεψης, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά ημερομηνία επίσκεψης, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10 & E78, ανά ημερομηνία επίσκεψης, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ασθενή, στο σύνολο όλων των επισκέψεων του, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά ασθενή, στο σύνολο όλων των επισκέψεων του, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10 & E78, ανά ασθενή, στο σύνολο όλων των επισκέψεων του.

Τα έξι σύνολα δεδομένων (datasets) είχαν ως πεδία: PatID (μοναδικός κωδικός ασθενή), DatePr (ημερομηνία συνταγής), Diagnosis (διάγνωση σε ICD10), Substance (δραστική ουσία).

Το μοντέλο που αναπτύχθηκε για τον αλγόριθμο FP-Growth στο Rapid Miner είναι:



Εικόνα 8: Μοντέλο για τον αλγόριθμο FP-Growth στο Rapid Miner

Εισάγεται το κάθε dataset και με τον operator Remove Attribute Range αφαιρείται η ημερομηνία συνταγής. Στη συνέχεια, με τον operator Split δημιουργούμε νέα χαρακτηριστικά / πεδία επιλέγοντας τη διάγνωση και τη δραστική ουσία σε τμήματα ως μη ταξινομημένη λίστα (Unordered Splits).

Η λειτουργία της μη ταξινομημένης λίστας (Unordered Splits) είναι χρήσιμη, όταν η σειρά δεν είναι σημαντική και ο στόχος είναι να δημιουργήσουμε ένα καλάθι ως σύνολο δεδομένων, που περιέχει όλες τις τιμές που εμφανίζονται στο dataset.

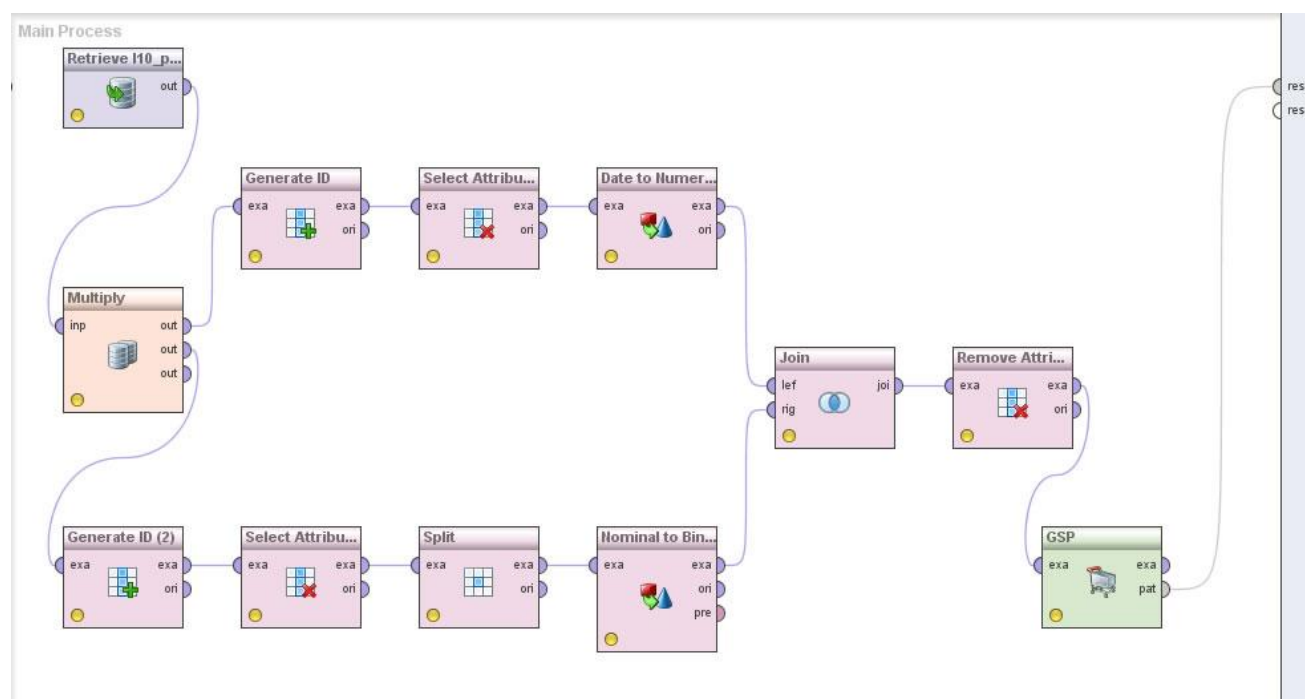
Εισάγεται ο operator FP-Growth όπου όλα τα πεδία πρέπει να είναι διωνυμικά (binominal) δηλαδή να έχουν μόνο δύο δυνατές τιμές. Ο FP-Growth υπολογίζει όλα τα συχνά στοιχειosύνολα από το σύνολο δεδομένων (dataset).

Προκειμένου να μειωθεί ο αριθμός των συχνών στοιχειοσυνόλων που εξάγονται, ως περιορισμό έχουμε την ελάχιστη υποστήριξη (minimum support)

Τέλος εισάγεται ο operator Create Association Rules όπου δημιουργεί ένα σύνολο κανόνων συσχέτισης από το σύνολο των συχνών στοιχειοσυνόλων (Frequent Itemset) του FP-Growth. Προκειμένου να μειωθεί ο αριθμός των κανόνων που εξάγονται, ως περιορισμό έχουμε την ελάχιστη εμπιστοσύνη (minimum confidence).

5.2. Λεπτομέρειες υλοποίησης GSP

Το μοντέλο που αναπτύχθηκε για τον αλγόριθμο GSP στο Rapid Miner είναι:



Εικόνα 9: Μοντέλο για τον αλγόριθμο GSP στο Rapid Miner

Αφού εισάγαμε το κάθε dataset, με τον operator Multiply δημιουργήσαμε δυο ίδια dataset και με την προσθήκη του operator Generate ID προσθέσαμε ένα νέο χαρακτηριστικό/πεδίο με το ρόλο id.

Στο πάνω μέρος του σχήματος με τον operator Select Attributes επιλέξαμε τα πεδία PatID, DatePr και το id του operator Generate ID. Στην συνέχεια με τον operator Date to Numerical επιτύχαμε την αλλαγή της ημερομηνίας σε αριθμητικό τύπο.

Στο κάτω μέρος του σχήματος με τον operator Select Attributes επιλέξαμε τα πεδία PatID, Diagnosis, DatePr, Substance και το id του operator Generate ID. Στην συνέχεια, με τον operator Split δημιουργήσαμε νέα χαρακτηριστικά / πεδία, επιλέγοντας τη διάγνωση και τη δραστική ουσία σε τμήματα ως μη ταξινομημένη λίστα (Unordered Splits). Η λειτουργία της μη ταξινομημένης λίστας (Unordered Splits) είναι χρήσιμη όταν η σειρά δεν είναι σημαντική και ο στόχος είναι να δημιουργήσουμε ένα καλάθι ως σύνολο δεδομένων που περιέχει όλες τις τιμές που εμφανίζονται στο dataset. Με τον operator Nominal to Binominal επιτύχαμε την αλλαγή του τύπου των επιλεγμένων πεδίων σε διωνυμικό (binominal) τύπο, δηλαδή να έχουν μόνο δύο δυνατές τιμές. Στην συνέχεια με τον operator Join ενώσαμε τα δύο datasets. Με τον operator Remove Attribute Range αφαιρέσαμε την ημερομηνία συνταγής.

Τέλος έγινε εισαγωγή του operator GSP. Ο GSP αναζητεί διαδοχικά patterns σε ένα σύνολο δεδομένων σε συνάρτηση με το χρόνο. Προκειμένου να μειωθεί ο αριθμός των κανόνων που εξάγονται, ως περιορισμό έχουμε την ελάχιστη υποστήριξη (minimum support).

6. Αποτελέσματα/Αξιολόγηση

6.1. Dataset - Σύνολο δεδομένων

Όπως αναφέραμε, τα δεδομένα προέρχονται από την ηλεκτρονική συνταγογράφηση ενός Γενικού Ιατρού για το χρονικό διάστημα από 7/1/2012 έως 31/7/2015. Δημιουργήσαμε έξι σύνολα δεδομένων με το Neatbeans:

- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ασθενή, στο σύνολο όλων των επισκέψεων του. Το σύνολο δεδομένων είχε 1.762 ασθενείς. Γυναίκες: 1028, άνδρες: 734.
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά ασθενή, στο σύνολο όλων των επισκέψεων του. Το σύνολο δεδομένων είχε 1.295 ασθενείς. Γυναίκες: 805, άνδρες: 490.
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10&E78, ανά ασθενή, στο σύνολο όλων των επισκέψεων του. Το σύνολο δεδομένων είχε 729 ασθενείς. Γυναίκες: 448, άνδρες: 281.
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ημερομηνία επίσκεψης. Το σύνολο δεδομένων είχε 10.136 εγγραφές. Γυναίκες: 1.028, άνδρες: 734.
- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά

ημερομηνία επίσκεψης. Το σύνολο δεδομένων είχε 7.455 εγγραφές.

Γυναίκες: 805, άνδρες: 490.

- ❖ ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10&E78, ανά ημερομηνία επίσκεψης. Το σύνολο δεδομένων είχε 5.370 εγγραφές. Γυναίκες: 448, άνδρες: 281.

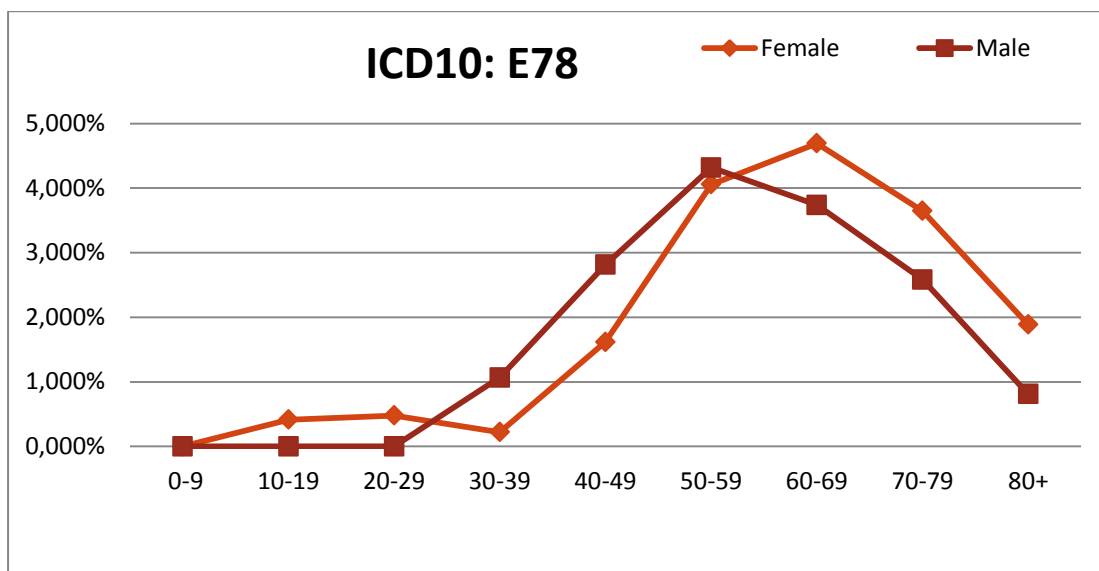
Τα έξι datasets είχαν ως πεδία:

- ❖ PatID (μοναδικός κωδικός ασθενή)
- ❖ DatePr (ημερομηνία συνταγής)
- ❖ Diagnosis (διάγνωση σε ICD10)
- ❖ Substance (δραστική ουσία)

Ο χωρισμός των ηλικιακών ομάδων στην βάση δεδομένων έγινε βάσει του Παγκόσμιου Οργανισμού Υγείας (World Health Organization – WHO) (Boschi-Pinto et al., 2008) Στα παρακάτω σχήματα η κατανομή έγινε βάσει των στοιχείων της ΕΛ.ΣΤΑΤ. όπου ο πληθυσμός της Αττικής στην απογραφή του 2011 ήταν 3.828.434.⁶ Οι ηλικιακές ομάδες που χρησιμοποιεί η ΕΛ.ΣΤΑΤ. ήταν λιγότερες από τον Παγκόσμιο Οργανισμό Υγείας (WHO) όποτε κάναμε συγχώνευση των ηλικιακών ομάδων για τα παρακάτω σχήματα.

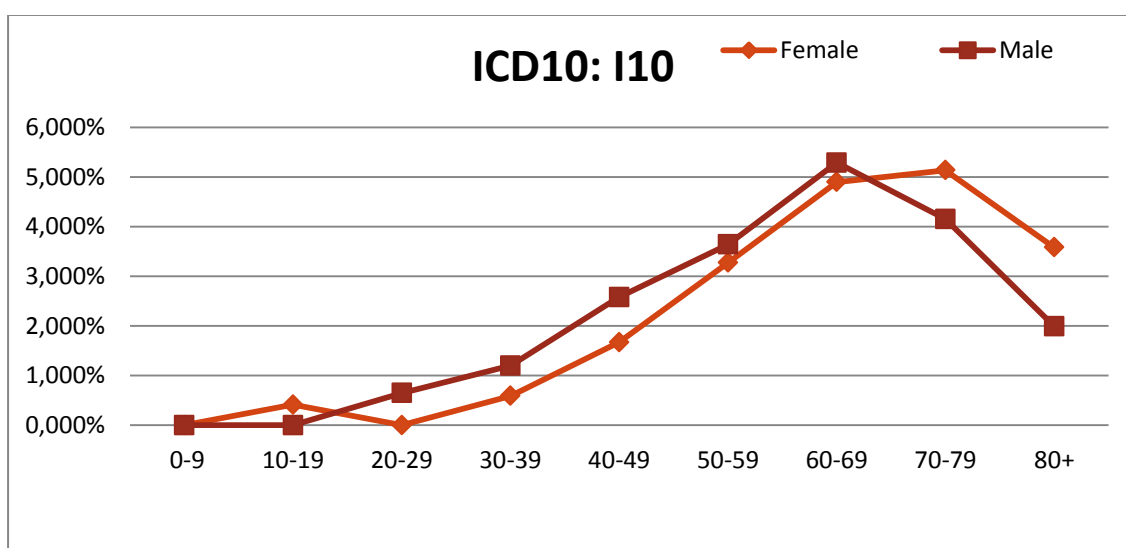
⁶ <http://www.statistics.gr/demographic-data>

Στην παρακάτω εικόνα φαίνονται οι άνδρες ασθενείς και οι γυναίκες ασθενείς που εμφανίζουν υψηλή χοληστερόλη (E78), ανά ηλικιακή ομάδα, στις 100.000.



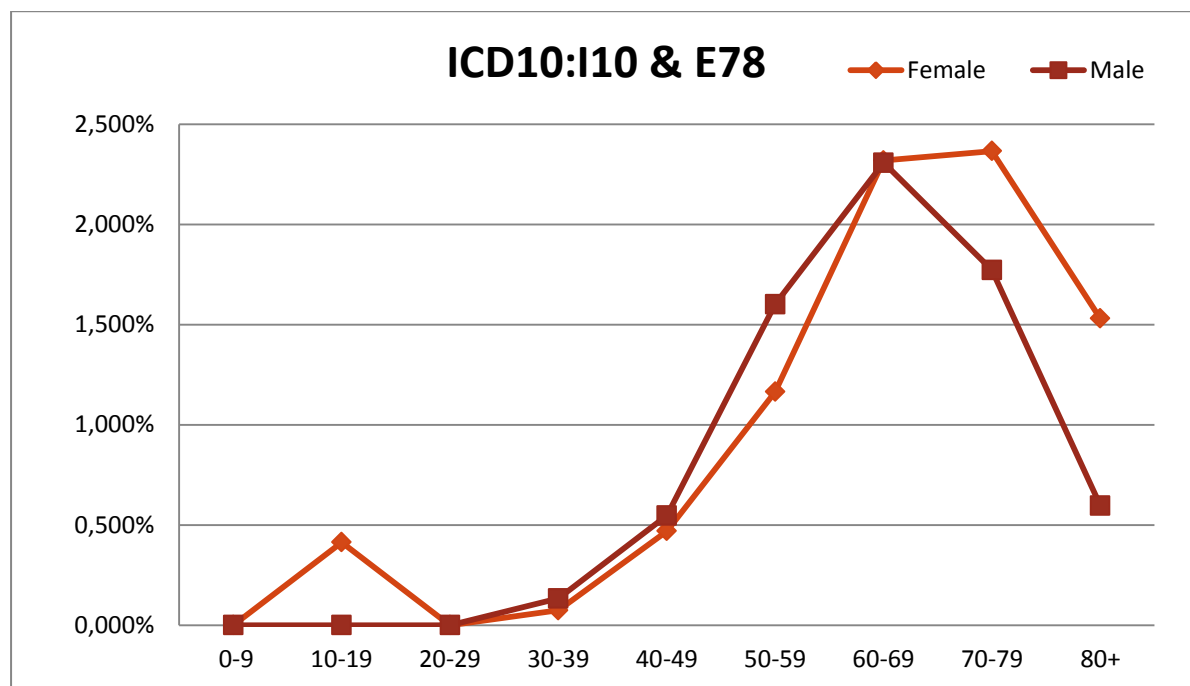
Σχήμα 12: Κατανομή ICD10: E78 ανδρών-γυναικών στις 100.000

Στην παρακάτω εικόνα φαίνονται οι άνδρες ασθενείς και οι γυναίκες ασθενείς που εμφανίζουν υπέρταση (I10), ανά ηλικιακή ομάδα, στις 100.000.



Σχήμα 13: Κατανομή ICD10:I10 ανδρών-γυναικών στις 100.000

Στην παρακάτω εικόνα φαίνονται οι άνδρες ασθενείς και οι γυναίκες ασθενείς που εμφανίζουν υπέρταση (I10) και υψηλή χοληστερόλη (E78), ανά ηλικιακή ομάδα, στις 100.000.



Σχήμα 14: Κατανομή ICD10:I10 & E78 ανδρών-γυναικών στις 100.000

6.2. Μετρικές αξιολόγησης

Οι μετρικές ενδιαφέροντος των κανόνων συσχέτισης εκφράζονται με την υποστήριξη κανόνα (support) και την εμπιστοσύνη κανόνα (confidence).

Η υποστήριξη κανόνα support (s) εκφράζει το s % των δοσοληψιών που περιέχουν και το X και το Y ($X \cup Y$), προς το σύνολο όλων των δοσοληψιών N. (Agrawal & Srikant, 1994)

$$\text{Support}, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

Εξίσωση 1: Support

Η εμπιστοσύνη κανόνα confidence (c) εκφράζει το c % των δοσοληψιών που περιέχουν το X, περιέχουν επίσης και το Y, προς το πλήθος των δοσοληψιών που περιέχουν το X. (Agrawal & Srikant, 1994)

$$\text{Confidence}, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Εξίσωση 2: Confidence

Στα σύνολα δεδομένων με ασθενείς που πάσχουν από υπέρταση (I10) και υψηλή χοληστερόλη (E78) ταυτόχρονα, ο αριθμός των κανόνων ήταν πάρα πολύ μεγάλος, όποτε εφαρμόσαμε σαν πρόσθετη μετρική το στατιστικό μέτρο Lift. Το lift ορίζεται, ως η εμπιστοσύνη διαιρούμενη με το ποσοστό όλων των περιπτώσεων. Αν η τιμή του Lift είναι μικρότερη ή ίση με ένα, τότε ο κανόνας δεν παρουσιάζει κάποιο ενδιαφέρον. Αν η τιμή του Lift είναι μεγαλύτερη από ένα, τότε υπάρχει θετική συσχέτιση μεταξύ των X και Y και ο κανόνας παρουσιάζει ενδιαφέρον. (Geng & Hamilton, 2006)

$$\text{Lift}(X \rightarrow Y) = \frac{\sup p(X \cup Y)}{\sup p(X) \cdot \sup p(Y)}$$

Εξίσωση 3: Lift

Τα ακολουθιακά πρότυπα εκφράζονται με μια μετρική ενδιαφέροντος, την υποστήριξη κανόνα (support).

Τέλος, από τα αποτελέσματα έγινε καθαρισμός των κανόνων και επιλέχθηκαν κανόνες που περιείχαν διαφορετική νόσο και δραστικές ουσίες.

6.3. Αποτελέσματα

Στα παρακάτω αποτελέσματα εμφανίζεται μέρος των κανόνων, που αποτελούν χαρακτηριστικά παραδείγματα. Το σύνολο όλων των κανόνων βρίσκεται στα παραρτήματα. Από τα αποτελέσματα έγινε καθαρισμός των κανόνων και επιλέχθηκαν κανόνες που περιείχαν διαφορετική νόσο και δραστικές ουσίες. Επίσης στα αποτελέσματα, εμφανίζονται διάφορες δραστικές ουσίες. Οι παρακάτω πίνακες περιέχουν σε ποιά νόσο δίνεται η κάθε δραστική ουσία.

Δραστικές ουσίες για χοληστερόλη (E78)
Atorvastatin
Simvastatin
Rosuvastatin
Pitavastatin
Ezetimide

Πίνακας 7: Δραστικές ουσίες για χοληστερόλη (E78)

Δραστικές ουσίες για υπέρταση (I10)
Hydrochlorothiazide
Valsartan
Irbesartan
Olmasartan
Amlodipine

Πίνακας 8: Δραστικές ουσίες για υπέρταση (I10)

Δραστικές ουσίες για διαβήτη (E10)
Metformin

Πίνακας 9: Δραστικές ουσίες για διαβήτη (E10)

Δραστικές ουσίες για στεφανιαία νόσο (I25.5)
Clopidogrel

Πίνακας 10: Δραστικές ουσίες για στεφανιαία νόσο (I25.5)

Δραστικές ουσίες για γαστροοισοφαγική παλινδρόμηση (K21)
Omeprazole

Πίνακας 11: Δραστικές ουσίες για γαστροοισοφαγική παλινδρόμηση (K21)

6.3.1. Αποτελέσματα των datasets ανά ασθενή

ICD10: E78 – Αμιγής υπερχοληστερολαιμία – Pure hypercholesterolaemia

Ασθενείς: 1295. Γυναίκες: 805, άνδρες: 490

FP-Growth: min support 0.1

Create Association Rules: min confidence 0.7

1. [Substance_SIMVASTATIN, Substance_HYDROCHLOROTHIAZIDE] --> [Diagnosis_I10] (confidence: 0.985)
2. [Substance_VALSARTAN] --> [Diagnosis_I10] (confidence: 0.951)
3. [Substance_HYDROCHLOROTHIAZIDE] --> [Diagnosis_I10] (confidence: 0.943)
4. [Substance_METFORMIN HYDROCHLORIDE] --> [Diagnosis_E10] (confidence: 0.806)
5. [Diagnosis_E10] --> [Diagnosis_I10] (confidence: 0.737)
6. [Diagnosis_K21] --> [Diagnosis_I10] (confidence: 0.726)
7. [Diagnosis_I25.5] --> [Diagnosis_I10] (confidence: 0.705)

Από τους κανόνες 1,2,3 βλέπουμε συσχέτιση E78 και I10, επειδή με E78 και τη χορήγηση δραστικής ουσίας για την υπέρταση, τότε έχουμε υπέρταση. Από τους κανόνες 4 και 5 βλέπουμε συσχέτιση E10 με E78 και I10. Με τον κανόνα 6 βλέπουμε συσχέτιση I10 και K21. Με τον κανόνα 7 βλέπουμε συσχέτιση στεφανιαίας νόσου I25.5 με E78 και I10.

Ασθενείς: 1295. Γυναίκες: 805, άνδρες: 490

GSP: min support 0.1

1. 0.563: <Diagnosis_E78, Diagnosis_I10>
2. 0.471: <Diagnosis_E78, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
3. 0.347: <Diagnosis_E78, Substance_SIMVASTATIN>
4. 0.239: <Diagnosis_E78, Diagnosis_K21>

5. 0.162: <Diagnosis_E78, Diagnosis_I25.5>
6. 0.150: <Diagnosis_E10, Diagnosis_E78>
7. 0.114: <Diagnosis_E78, Diagnosis_I10, Diagnosis_I25.5>
8. 0.113: <Diagnosis_E78, Diagnosis_K21, Substance_ATORVASTATIN
CALCIUM TRIHYDRATE>
9. 0.110: <Diagnosis_E10, Diagnosis_E78, Diagnosis_I10>

Από τον κανόνα 1, βλέπουμε συσχέτιση E78 και I10. Από τους κανόνες 2 και 3 βλέπουμε την χορήγηση δραστικής ουσίας για την χοληστερόλη. Με τον κανόνα 4 βλέπουμε συσχέτιση E78 και K21. Με τον κανόνα 5 βλέπουμε συσχέτιση E78 και I25.5. Με τον κανόνα 6 βλέπουμε συσχέτιση E78 και E10. Με τον κανόνα 7 βλέπουμε συσχέτιση E78, I10 και I25.5. Με τον κανόνα 8 βλέπουμε συσχέτιση E78 και K21, με την χορήγηση δραστικής ουσίας για την χοληστερόλη. Με τον κανόνα 9 βλέπουμε συσχέτιση E10 με E78 και I10.

ICD10: I10 – Ιδιοπαθής (πρωτοπαθής) υπέρταση – Essential (primary) hypertension

Ασθενείς : 1762. Γυναίκες: 1028, άνδρες: 734

FP-Growth: min support 0.1

Create Association Rules: min confidence 0.7

1. [Substance_ATORVASTATIN CALCIUM TRIHYDRATE] --> [Diagnosis_E78]
(confidence: 0.964)
2. [Substance_SIMVASTATIN] --> [Diagnosis_E78] (confidence: 0.930)
3. [Substance_METFORMIN HYDROCHLORIDE] --> [Diagnosis_E10]
(confidence: 0.827)

Από τους κανόνες 1,2 βλέπουμε συσχέτιση E78 και I10, επειδή με I10 και τη χορήγηση δραστικής ουσίας για την χοληστερόλη, τότε έχουμε χοληστερίνη. Με τον

κανόνα 3 φαίνεται, ότι με I10 και τη χορήγηση δραστικής ουσίας για τον διαβήτη, τότε έχουμε E10.

Ασθενείς : 1762. Γυναίκες: 1028, άνδρες: 734

GSP: min support 0.1

1. 0.414: <Diagnosis_E78, Diagnosis_I10>
2. 0.242: <Diagnosis_I10, Diagnosis_K21>
3. 0.218: <Diagnosis_I10, Substance_VALSARTAN>
4. 0.173: <Diagnosis_I10, Substance_IRBESARTAN>
5. 0.158: <Diagnosis_E10, Diagnosis_I10>
6. 0.151: <Diagnosis_I10, Diagnosis_I25.5>
7. 0.144: <Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
8. 0.128: <Diagnosis_I10, Substance_OLMESARTAN MEDOXOMIL>
9. 0.128: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
10. 0.119: <Diagnosis_E10, Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
11. 0.106: <Diagnosis_F41.9, Diagnosis_I10>

Από τον κανόνα 1 βλέπουμε συσχέτιση E78 και I10. Από τον κανόνα 2 βλέπουμε συσχέτιση I10 και K21. Από τους κανόνες 3 και 4 βλέπουμε την δραστική ουσία για την υπέρταση. Από τους κανόνες 5, 7 και 10 βλέπουμε συσχέτιση E10 και I10 καθώς και την χορήγηση δραστικής ουσίας για τον διαβήτη. Από τον κανόνα 6 βλέπουμε συσχέτιση I10 και I25.5. Από τους κανόνες 8 και 9 βλέπουμε συσχέτιση I10 και K21 με την χορήγηση δραστικής ουσίας για την χοληστερόλη. Από τον κανόνα 11 βλέπουμε συσχέτιση I10 και F41.9 (Διάφορες αγχώδεις διαταραχές).

ICD10: I10 & ICD10: E78

Ασθενείς : 729. Γυναίκες: 448, άνδρες: 281

FP-Growth: min support 0.1

Create Association Rules: min confidence 0.7

1. [Substance_AMLODIPINE] --> [Substance_VALSARTAN] (confidence: 0.987)
2. [Substance_OMEPRAZOLE] --> [Diagnosis_K21] (confidence: 0.956)
3. [Substance_METFORMIN HYDROCHLORIDE] --> [Diagnosis_E10] (confidence: 0.827)

Από τον κανόνα 1 βλέπουμε ότι με I10 και E78, με τη χορήγηση δραστικής ουσίας για την υπέρταση AMLODIPINE, δίνουμε και τη δραστική ουσία για την υπέρταση VALSARTAN. Από τον κανόνα 2 βλέπουμε ότι με I10 και E78 και με τη χορήγηση δραστικής ουσίας για την γαστροοισοφαγική παλινδρόμηση, τότε έχουμε K21. Από τον κανόνα 3 βλέπουμε ότι με I10 και E78 και τη χορήγηση δραστικής ουσίας για τον διαβήτη, τότε έχουμε E10.

Ασθενείς : 729. Γυναίκες: 448, άνδρες: 281

GSP: min support 0.1

1. 0.309: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
2. 0.203: <Diagnosis_E78, Diagnosis_I10, Diagnosis_I25.5>
3. 0.196: <Diagnosis_E10, Diagnosis_E78, Diagnosis_I10>
4. 0.191: <Diagnosis_E78, Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
5. 0.158: <Diagnosis_E10, Diagnosis_E78, Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
6. 0.119: <Diagnosis_E78, Diagnosis_F41.9, Diagnosis_I10>
7. 0.118: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21, Substance_OMEPRAZOLE>
8. 0.115: <Diagnosis_E78, Diagnosis_F33, Diagnosis_I10>

- 9. 0.111: <Diagnosis_E78, Diagnosis_I10, Substance_CLOPIDOGREL>
- 10. 0.111: <Diagnosis_E78, Diagnosis_I10, Diagnosis_I25.5, Diagnosis_K21>
- 11. 0.110: <Diagnosis_E11, Diagnosis_E78, Diagnosis_I10>

Από τον κανόνα 1 βλέπουμε συσχέτιση E78, I10 και K21. Από τον κανόνα 2 και 9 βλέπουμε συσχέτιση E78, I10 και I25.5 καθώς και την χορήγηση δραστικής ουσίας για την στεφανιαία νόσο. Από τους κανόνες 3, 4, 5 και 11 βλέπουμε συσχέτιση E10, E11, E78 και I10 καθώς και την χορήγηση δραστικής ουσίας για τον διαβήτη. Με τον κανόνα 6 βλέπουμε συσχέτιση E78, I10 και F41.9 (Διάφορες αγχώδεις διαταραχές). Με τον κανόνα 7 βλέπουμε συσχέτιση I10 και K21 και την χορήγηση δραστικής ουσίας για την γαστροοισοφαγική παλινδρόμηση. Με τον κανόνα 8 βλέπουμε συσχέτιση E78, I10 και F33 (καταθλιπτική διαταραχή). Από τον κανόνα 10 βλέπουμε συσχέτιση E78, I10, I25.5 και K21.

Γενικά συμπεράσματα για τα αποτελέσματα των datasets ανά ασθενή

Βλέπουμε ότι υπάρχει θετική συσχέτιση υψηλής χοληστερόλης (E78), υπέρτασης (I10) και στεφανιαίας νόσου (I25.5). Εμφανίζονται οι διάφοροι παράγοντες κινδύνου υπέρτασης και υψηλής χοληστερόλης, όπως διάφορες αγχώδεις διαταραχές (F41.9), ινσουλινοεξαρτώμενος σακχαρώδης διαβήτης (E10), μη ινσουλινοεξαρτώμενος σακχαρώδης διαβήτης (E11) και καταθλιπτική διαταραχή (F33).

Για την γαστροοισοφαγική παλινδρόμηση με οισοφαγίτιδα (K21), ένας παράγοντας κινδύνου είναι τα φάρμακα, όπως οι ανταγωνιστές βήτα υποδοχέων

για την υπέρταση ή τις καρδιακές παθήσεις και οι ανταγωνιστές διαύλων ασβεστίου για την υπέρταση. Βλέπουμε σε πολλούς κανόνες συσχέτιση E78, I10, I25.5 και K21.

6.3.2. Αποτελέσματα των datasets ανά επίσκεψη

ICD10: E78 – Αμιγής υπερχοληστερολαιμία – Pure hypercholesterolaemia

Εγγραφές : 7455. Γυναίκες: 805, άνδρες: 490

FP-Growth: min support 0.1

Create Association Rules: min confidence 0.7

1. [Substance_HYDROCHLOROTHIAZIDE] --> [Diagnosis_I10]
(confidence: 0.932)
2. [Substance_VALSARTAN] --> [Diagnosis_I10]

Από τους κανόνες 1,2 βλέπουμε συσχέτιση E78 και I10, επειδή με E78 και τη χορήγηση δραστικής ουσίας για την υπέρταση, τότε έχουμε υπέρταση.

Εγγραφές : 7455. Γυναίκες: 805, άνδρες: 490

GSP: min support 0.1

1. 0.563: <Diagnosis_E78, Diagnosis_I10>
2. 0.174: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
3. 0.162: <Diagnosis_E78, Diagnosis_I25.5>
4. 0.150: <Diagnosis_E10, Diagnosis_E78>
5. 0.131: <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN
CALCIUM TRIHYDRATE> <Diagnosis_E78, Diagnosis_I10,
Substance_ATORVASTATIN CALCIUM TRIHYDRATE>

Από τον κανόνα 1 βλέπουμε συσχέτιση E78 και I10. Από τον κανόνα 2 βλέπουμε συσχέτιση E78, I10 και K21. Από τον κανόνα 3 βλέπουμε συσχέτιση E78

και I25.5. Από τον κανόνα 4 βλέπουμε συσχέτιση E10 και E78. Από τον κανόνα 5 βλέπουμε συσχέτιση E78 και I10 με την χορήγηση δραστικής ουσίας για την χοληστερόλη.

ICD10: I10 – Ιδιοπαθής (πρωτοπαθής) υπέρταση – Essential (primary) hypertension

Εγγραφές : 10136. Γυναίκες: 1028, άνδρες: 734

FP-Growth: min support 0.1

Create Association Rules: min confidence 0.7

1. [Substance_ATORVASTATIN CALCIUM TRIHYDRATE]
--> [Diagnosis_E78] (confidence: 0.955)
2. [Substance_SIMVASTATIN] --> [Diagnosis_E78] (confidence: 0.924)

Από τους κανόνες 1,2 βλέπουμε συσχέτιση E78 και I10, επειδή με I10 και την χορήγηση δραστικής ουσίας για την χοληστερόλη, τότε έχουμε χοληστερίνη.

Εγγραφές : 10136. Γυναίκες: 1.028, άνδρες: 734

GSP: min support 0.1

1. 0.161: <Diagnosis_I10> <Diagnosis_K21>
2. 0.158: <Diagnosis_E10, Diagnosis_I10>
3. 0.128: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
4. 0.125: <Diagnosis_K21> <Diagnosis_K21>
5. 0.119: <Diagnosis_E10, Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
6. 0.119: <Diagnosis_E78, Diagnosis_I10> <Diagnosis_E78, Diagnosis_I10> <Diagnosis_E78, Diagnosis_I10> <Diagnosis_E78, Diagnosis_I10>
7. 0.115: <Diagnosis_I25.5> <Diagnosis_I10>
8. 0.106: <Diagnosis_F41.9, Diagnosis_I10>

Από τον κανόνα 1 βλέπουμε συσχέτιση I10 και K21. Από τον κανόνα 2 βλέπουμε συσχέτιση E10 και I10. Από τον κανόνα 3 βλέπουμε συσχέτιση E78, I10 και K21. Από τον κανόνα 4 βλέπουμε διαδοχικά γεγονότα που εμφανίζουν K21. Από τον κανόνα 5 βλέπουμε συσχέτιση E10 με I10 καθώς και την χορήγηση δραστικής ουσίας για τον διαβήτη. Από τον κανόνα 6 βλέπουμε διαδοχικά γεγονότα που εμφανίζουν E78 και I10. Από τον κανόνα 7 βλέπουμε διαδοχικά γεγονότα που εμφανίζουν I25.5 και I10. Με τον κανόνα 8 βλέπουμε συσχέτιση I10 και F41.9 (Διάφορες αγχώδεις διαταραχές).

ICD10: I10 & ICD10: E78

Εγγραφές : 5370. Γυναίκες: 448, άνδρες: 281

FP-Growth: min support 0.1

Create Association Rules: min confidence 0.6

1. [Substance_CLOPIDOGREL] --> [Diagnosis_I25.5] (confidence: 0.903)
2. [Diagnosis_I10, Diagnosis_E78] --> [Diagnosis_K21] (confidence: 0.623)

Από τον κανόνα 1 βλέπουμε ότι με την χορήγηση δραστικής ουσίας για την στεφανιαία νόσο τότε έχουμε I25.5. Από τον κανόνα 2 βλέπουμε συσχέτιση E78, I10 και K21.

Εγγραφές : 5370. Γυναίκες: 448, άνδρες: 281

GSP: min support 0.22

1. 0.309: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
2. 0.228: <Diagnosis_I10> <Diagnosis_K21>

3. 0.224: <Diagnosis_E78> <Diagnosis_K21>

Από τους κανόνες 1,2 και 3 βλέπουμε συσχέτιση E78, I10 και K21.

Γενικά συμπεράσματα για τα αποτελέσματα των datasets ανά επίσκεψη

Όπως και προηγουμένως, βλέπουμε ότι υπάρχει θετική συσχέτιση υψηλής χοληστερόλης (E78), υπέρτασης (I10) και στεφανιαίας νόσου (I25.5). Εμφανίζονται οι διάφοροι παράγοντες κινδύνου υπέρτασης και υψηλής χοληστερόλης, όπως διάφορες αγχώδεις διαταραχές (F41.9), ινσουλινοεξαρτώμενος σακχαρώδης διαβήτης (E10) και καταθλιπτική διαταραχή (F33).

Για την γαστροοισοφαγική παλινδρόμηση με οισοφαγίτιδα (K21), ένας παράγοντας κινδύνου είναι τα φάρμακα, όπως οι ανταγωνιστές βήτα υποδοχέων για την υπέρταση ή τις καρδιακές παθήσεις και οι ανταγωνιστές διαύλων ασβεστίου για την υπέρταση. Βλέπουμε σε πολλούς κανόνες συσχέτιση E78, I10 και K21.

7. Συμπεράσματα

7.1. Σύνοψη της προτεινόμενης προσέγγισης

Στην παρούσα διπλωματική εργασία επιχειρήθηκε η αναζήτηση γνώσης από την παραγωγή κανόνων συσχέτισης και ακολουθιακών προτύπων, τόσο νόσων, όσο και δραστικών ουσιών, από ασθενείς που έχουν υψηλή χοληστερόλη (E78), υπέρταση (I10) αλλά και από I10 & E78 ταυτόχρονα. Το σύνολο των ιατρικών δεδομένων προέρχεται από την ηλεκτρονική συνταγογράφηση στην περιοχή της Αθήνας, από Γενικό Ιατρό, για το χρονικό διάστημα από 7/1/2012 έως 31/7/2015.

Δημιουργήθηκαν έξι σύνολα δεδομένων με το NeatBeans και ακολουθήθηκαν τα βήματα της διαδικασίας της εξόρυξης γνώσης. Τα 6 σύνολα δεδομένων ήταν : ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ημερομηνία επίσκεψης, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά ημερομηνία επίσκεψης, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10 & E78, ανά ημερομηνία επίσκεψης, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υπέρταση (I10), ανά ασθενή, στο σύνολο όλων των επισκέψεων του, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν υψηλή χοληστερόλη (E78), ανά ασθενή, στο σύνολο όλων των επισκέψεων του, ένα σύνολο δεδομένων με τις διαγνώσεις σε ICD10 και τις δραστικές ουσίες των ασθενών που έχουν I10 & E78, ανά ασθενή, στο σύνολο όλων των επισκέψεων του.

Επιλέξαμε τις μεθόδους που θα χρησιμοποιηθούν στα έξι σύνολα δεδομένων για την εξαγωγή χρησίμων πληροφοριών που θα οδηγήσουν σε γνώση. Ως μεθόδους επιλέξαμε τους αλγόριθμους FP-Growth για την εξαγωγή κανόνων συσχέτισης και GSP για την εξαγωγή ακολουθιακών προτύπων. Κάθε κανόνας συσχέτισης εκφράζεται με δύο μετρικές, την υποστήριξη (support) και την εμπιστοσύνη (confidence). Το βασικό πρόβλημα στην εξαγωγή κανόνων συσχέτισης είναι ότι εξάγεται μεγάλος αριθμός κανόνων. Για να μειωθεί ο αριθμός των κανόνων που εξάγονται, έχουμε ως περιορισμούς την ελάχιστη υποστήριξη (minimum support) και την ελάχιστη εμπιστοσύνη (minimum confidence). Στα σύνολα δεδομένων με ασθενείς που πάσχουν από υπέρταση (I10) και υψηλή χοληστερόλη (E78) ταυτόχρονα, ο αριθμός των κανόνων ήταν πάρα πολύ μεγάλος, όποτε εφαρμόσαμε σαν πρόσθετη μετρική το στατιστικό μέτρο Lift. Κάθε ακολουθιακό πρότυπο εκφράζεται με μια μετρική, την υποστήριξη (support). Το βασικό πρόβλημα στην εξαγωγή ακολουθιακών προτύπων είναι ότι εξάγεται μεγάλος αριθμός προτύπων. Για να μειωθεί ο αριθμός ακολουθιακών προτύπων που εξάγονται, έχουμε ως περιορισμό την ελάχιστη υποστήριξη (minimum support).

Τέλος, χρησιμοποιήθηκε το λογισμικό Rapid Miner, για την παραγωγή κανόνων συσχέτισης και ακολουθιακών προτύπων, το οποίο είναι ένα από τα πιο αναγνωρισμένα εργαλεία στο πεδίο της εξόρυξης γνώσης. Από τα αποτελέσματα έγινε καθαρισμός των κανόνων και επιλέχθηκαν κανόνες που περιείχαν διαφορετική νόσο και δραστικές ουσίες.

7.2. Μελλοντικές επεκτάσεις

Η ανακάλυψη γνώσης από ιατρικές βάσεις δεδομένων είναι μια περιοχή έρευνας που εξελίσσεται με ραγδαίους ρυθμούς. Στο πλαίσιο της παρούσας διπλωματικής μελετήθηκε η παραγωγή κανόνων συσχέτισης και ακολουθιακών προτύπων, τόσο νόσων, όσο και δραστικών ουσιών, από ασθενείς που πάσχουν είτε από υψηλή χοληστερόλη (E78), είτε από υπέρταση (I10), αλλά και από ασθενείς που πάσχουν ταυτόχρονα από υπέρταση (I10) και υψηλή χοληστερόλη (E78).

Η παρούσα διπλωματική εργασία θα μπορούσε να επεκταθεί, ερευνώντας και εφαρμόζοντας uplift modelling for personalised medicine. Το uplift modelling είναι μια προβλεπτική τεχνική εξόρυξης γνώσης, που τελευταία εφαρμόζεται στην εξατομικευμένη ιατρική (personalised medicine).

Οι παραδοσιακές μέθοδοι ταξινόμησης προβλέπουν την κατηγορία από ένα σύνολο μεταβλητών πρόβλεψης. Το uplift modelling προσπαθεί να προβλέψει τη διαφορά μεταξύ της ομάδας θεραπείας (treatment group), όπου έχει γίνει κάποια ενέργεια και της ομάδας ελέγχου (control group), όπου δεν έχει γίνει καμία ενέργεια. (Rzepakowski & Jaroszewicz, 2010; Rzepakowski & Jaroszewicz, 2012; Jaskowski & Jaroszewicz, 2012)

Η εξατομικευμένη ιατρική αποτελεί μία μοναδική προσέγγιση, κατά την οποία, εξειδικευμένα χαρακτηριστικά ενός ασθενούς, καθορίζουν τον σχεδιασμό της θεραπευτικής αγωγής που θα ακολουθηθεί και γίνεται προσδιορισμός της προδιάθεσης του ασθενούς σε μια συγκεκριμένη νόσο. Τέτοια εξειδικευμένα χαρακτηριστικά είναι το ιατρικό ιστορικό, το ιστορικό προηγούμενης θεραπείας, οι περιβαλλοντικοί παράγοντες κ.α.

Η εξατομικευμένη ιατρική επιχειρεί να καθορίσει από την αρχή την αποτελεσματικότητα της θεραπείας, διαχωρίζοντας ασθενείς και ασθένειες σε υποσύνολα, με βάση τα εξειδικευμένα χαρακτηριστικά τους. Έτσι, με την εξατομικευμένη θεραπεία, έχουμε μείωση των παρενεργειών και αύξηση της αποτελεσματικότητας της θεραπείας.

8. Παραρτήματα

8.1. Παραρτήματα I – Αποτελέσματα εκτέλεσης των datasets ανά ασθενή

8.1.1. FP-Growth

E78

1.295 ασθενείς. Γυναίκες: 805, άνδρες: 490

FP-Growth min support 0.1

Create Association Rules min confidence 0.7

1. [Substance_SIMVASTATIN, Substance_HYDROCHLOROTHIAZIDE] --> [Diagnosis_I10] (confidence: 0.985)
2. [Substance_VALSARTAN] --> [Diagnosis_I10] (confidence: 0.951)
3. [Substance_HYDROCHLOROTHIAZIDE] --> [Diagnosis_I10] (confidence: 0.943)
4. [Substance_METFORMIN HYDROCHLORIDE] --> [Diagnosis_E10] (confidence: 0.806)
5. [Diagnosis_E10] --> [Substance_METFORMIN HYDROCHLORIDE] (confidence: 0.794)
6. [Diagnosis_E10] --> [Diagnosis_I10] (confidence: 0.737)
7. [Substance_METFORMIN HYDROCHLORIDE] --> [Diagnosis_I10] (confidence: 0.728)
8. [Diagnosis_K21] --> [Diagnosis_I10] (confidence: 0.726)
9. [Diagnosis_I25.5] --> [Diagnosis_I10] (confidence: 0.705)

I10

1.762 ασθενείς. Γυναίκες: 1028, άνδρες: 734

FP-Growth min support 0.1

Create Association Rules min confidence 0.7

1. [Substance_ATORVASTATIN CALCIUM TRIHYDRATE] --> [Diagnosis_E78]
(confidence: 0.964)
2. [Substance_SIMVASTATIN] --> [Diagnosis_E78] (confidence: 0.930)
3. [Substance_METFORMIN HYDROCHLORIDE] --> [Diagnosis_E10]
(confidence: 0.827)
4. [Diagnosis_E10] --> [Substance_METFORMIN HYDROCHLORIDE]
(confidence: 0.753)

I10 & E78

729 ασθενείς. Γυναίκες: 448, άνδρες: 281 (κατάταξη κατά lift)

FP-Growth min support 0.1

Create Association Rules min confidence 0.7

1. [Substance_AMLODIPINE] --> [Substance_VALSARTAN] (confidence: 0.987)
2. [Substance_OMEPRAZOLE] --> [Diagnosis_K21] (confidence: 0.956)
3. [Substance_A/VICTORIA/361/2011 (H3N2)-DERIVED STRAIN USED IVR-165] --> [Substance_A/CALIFORNIA/7/2009 (H1N1) PDM09-DERIVED STRAIN USED NYMC X-179A] (confidence: 0.948)
4. [Substance_METFORMIN HYDROCHLORIDE] --> [Diagnosis_E10]
(confidence: 0.827)
5. [Diagnosis_E10] --> [Substance_METFORMIN HYDROCHLORIDE]
(confidence: 0.804)

8.1.2. GSP

E78

1.295 ασθενείς. Γυναίκες: 805, άνδρες: 490

GSP min support 0.1

1. 0.563: <Diagnosis_E78, Diagnosis_I10>
2. 0.471: <Diagnosis_E78, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
3. 0.347: <Diagnosis_E78, Substance_SIMVASTATIN>
4. 0.249: <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
5. 0.246: <Diagnosis_E78, Substance_HYDROCHLOROTHIAZIDE>
6. 0.239: <Diagnosis_E78, Diagnosis_K21>
7. 0.232: <Diagnosis_E78, Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE>
8. 0.215: <Diagnosis_E78, Diagnosis_I10, Substance_SIMVASTATIN>
9. 0.174: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
10. 0.162: <Diagnosis_E78, Diagnosis_I25.5>
11. 0.151: <Diagnosis_E78, Substance_ROSUVASTATIN CALCIUM>
12. 0.150: <Diagnosis_E10, Diagnosis_E78>
13. 0.147: <Diagnosis_E78, Substance_METFORMIN HYDROCHLORIDE>
14. 0.143: <Diagnosis_E78, Substance_VALSARTAN>
15. 0.136: <Diagnosis_E78, Diagnosis_I10, Substance_VALSARTAN>
16. 0.119: <Diagnosis_E10, Diagnosis_E78, Substance_METFORMIN HYDROCHLORIDE>
17. 0.114: <Diagnosis_E78, Diagnosis_I10, Diagnosis_I25.5>
18. 0.113: <Diagnosis_E78, Diagnosis_K21, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
19. 0.110: <Diagnosis_E10, Diagnosis_E78, Diagnosis_I10>

20. 0.107: <Diagnosis_E78, Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
21. 0.104: <Diagnosis_E78, Substance_HYDROCHLOROTHIAZIDE, Substance_SIMVASTATIN>
22. 0.103: <Diagnosis_E78, Substance_ATORVASTATIN CALCIUM TRIHYDRATE, Substance_HYDROCHLOROTHIAZIDE>
23. 0.103: <Diagnosis_E78, Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE, Substance_SIMVASTATIN>
24. 0.102: <Diagnosis_E78, Substance_AMLODIPINE BESYLATE>
25. 0.101: <Diagnosis_E78, Substance_IRBESARTAN>

I10

1.762 ασθενείς. Γυναίκες: 1.028, άνδρες: 734

GSP min support 0.1

1. 0.414: <Diagnosis_E78, Diagnosis_I10>
2. 0.400: <Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE>
3. 0.242: <Diagnosis_I10, Diagnosis_K21>
4. 0.218: <Diagnosis_I10, Substance_VALSARTAN>
5. 0.190: <Diagnosis_I10, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
6. 0.183: <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
7. 0.173: <Diagnosis_I10, Substance_IRBESARTAN>
8. 0.170: <Diagnosis_I10, Substance_SIMVASTATIN>
9. 0.170: <Diagnosis_E78, Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE>
10. 0.164: <Diagnosis_I10, Substance_AMLODIPINE BESYLATE>
11. 0.158: <Diagnosis_E10, Diagnosis_I10>
12. 0.158: <Diagnosis_E78, Diagnosis_I10, Substance_SIMVASTATIN>

13. 0.151: <Diagnosis_I10, Diagnosis_I25.5>
14. 0.144: <Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
15. 0.128: <Diagnosis_I10, Substance_OLMESARTAN MEDOXOMIL>
16. 0.128: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
17. 0.120: <Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE,
Substance_VALSARTAN>
18. 0.119: <Diagnosis_E10, Diagnosis_I10, Substance_METFORMIN
HYDROCHLORIDE>
19. 0.106: <Diagnosis_F41.9, Diagnosis_I10>

I10 & E78

729 ασθενείς. Γυναίκες: 448, άνδρες: 281

GSP min support 0.1

1. 0.443: <Diagnosis_E78, Diagnosis_I10, Substance_ATOMARVASTATIN
CALCIUM TRIHYDRATE>
2. 0.412: <Diagnosis_E78, Diagnosis_I10,
Substance_HYDROCHLOROTHIAZIDE>
3. 0.383: <Diagnosis_E78, Diagnosis_I10, Substance_SIMVASTATIN>
4. 0.309: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
5. 0.241: <Diagnosis_E78, Diagnosis_I10, Substance_VALSARTAN>
6. 0.203: <Diagnosis_E78, Diagnosis_I10, Diagnosis_I25.5>
7. 0.196: <Diagnosis_E10, Diagnosis_E78, Diagnosis_I10>
8. 0.191: <Diagnosis_E78, Diagnosis_I10, Substance_METFORMIN
HYDROCHLORIDE>
9. 0.182: <Diagnosis_E78, Diagnosis_I10,
Substance_HYDROCHLOROTHIAZIDE, Substance_SIMVASTATIN>
10. 0.174: <Diagnosis_E78, Diagnosis_I10, Substance_AMLODIPINE
BESYLATE>
11. 0.171: <Diagnosis_E78, Diagnosis_I10, Substance_IRBESARTAN>

12. 0.169: <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN
CALCIUM TRIHYDRATE, Substance_HYDROCHLOROTHIAZIDE>
13. 0.158: <Diagnosis_E78, Diagnosis_I10, Substance_ROSUVASTATIN
CALCIUM>
14. 0.158: <Diagnosis_E10, Diagnosis_E78, Diagnosis_I10,
Substance_METFORMIN HYDROCHLORIDE>
15. 0.152: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21,
Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
16. 0.144: <Diagnosis_E78, Diagnosis_I10,
Substance_ACETYLSALICYLIC ACID>
17. 0.132: <Diagnosis_E78, Diagnosis_I10,
Substance_A/VICTORIA/361/2011 (H3N2)-DERIVED STRAIN USED IVR-
165>
18. 0.130: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21,
Substance_HYDROCHLOROTHIAZIDE>
19. 0.130: <Diagnosis_E78, Diagnosis_I10,
Substance_HYDROCHLOROTHIAZIDE, Substance_VALSARTAN>
20. 0.125: <Diagnosis_E78, Diagnosis_I10,
Substance_A/CALIFORNIA/7/2009 (H1N1)PDM09-DERIVED STRAIN USED
NYMC X-179A>
21. 0.125: <Diagnosis_E78, Diagnosis_I10,
Substance_A/CALIFORNIA/7/2009 (H1N1)PDM09-DERIVED STRAIN USED
NYMC X-179A, Substance_A/VICTORIA/361/2011 (H3N2)-DERIVED
STRAIN USED IVR-165>
22. 0.123: <Diagnosis_E78, Diagnosis_I10, Substance_OMEPRAZOLE>
23. 0.122: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21,
Substance_SIMVASTATIN>
24. 0.121: <Diagnosis_E78, Diagnosis_I10, Substance_OLMESARTAN
MEDOXOMIL>
25. 0.119: <Diagnosis_E78, Diagnosis_F41.9, Diagnosis_I10>

- 26. 0.118: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21, Substance_OMEPRAZOLE>
- 27. 0.115: <Diagnosis_E78, Diagnosis_F33, Diagnosis_I10>
- 28. 0.114: <Diagnosis_E78, Diagnosis_I10, Substance_EZETIMIBE>
- 29. 0.111: <Diagnosis_E78, Diagnosis_I10, Substance_CLOPIDOGREL>
- 30. 0.111: <Diagnosis_E78, Diagnosis_I10, Diagnosis_I25.5, Diagnosis_K21>
- 31. 0.110: <Diagnosis_E11, Diagnosis_E78, Diagnosis_I10>
- 32. 0.110: <Diagnosis_E78, Diagnosis_I10, Diagnosis_M81>
- 33. 0.107: <Diagnosis_E78, Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE, Substance_IRBESARTAN>
- 34. 0.106: <Diagnosis_E78, Diagnosis_I10, Substance_AMLODIPINE>
- 35. 0.104: <Diagnosis_E78, Diagnosis_I10, Substance_AMLODIPINE, Substance_VALSARTAN>
- 36. 0.103: <Diagnosis_E78, Diagnosis_I10, Substance_ALLOPURINOL>
- 37. 0.103: <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN, CALCIUM TRIHYDRATE, Substance_VALSARTAN>

8.2. Παραρτήματα II - Αποτελέσματα εκτέλεσης των datasets ανά επίσκεψη

8.2.1. FP-Growth

E78

7.455 Εγγραφές. Γυναίκες: 805, άνδρες: 490

FP-Growth min support 0.1

Create Association Rules min confidence 0.7

- 1. [Substance_HYDROCHLOROTHIAZIDE] --> [Diagnosis_I10]
(confidence: 0.932)

2. [Substance_VALSARTAN] --> [Diagnosis_I10]

I10

10.136 Εγγραφές. Γυναίκες: 1028, άνδρες: 734

FP-Growth min support 0.1

Create Association Rules min confidence 0.7

1. [Substance_ATORVASTATIN CALCIUM TRIHYDRATE]
--> [Diagnosis_E78] (confidence: 0.955)
2. [Substance_ATORVASTATIN CALCIUM TRIHYDRATE] --> [Diagnosis_E78]
(confidence: 0.948)
3. [Substance_SIMVASTATIN] --> [Diagnosis_E78] (confidence: 0.924)

I10 & E78

5.370 Εγγραφές. Γυναίκες: 448, άνδρες: 281 (κατάταξη κατά lift)

FP-Growth min support 0.1

Create Association Rules min confidence 0.6

1. [Substance_CLOPIDOGREL] --> [Diagnosis_I25.5] (confidence: 0.903)
2. [Diagnosis_I10, Diagnosis_E78] --> [Substance_ATORVASTATIN CALCIUM TRIHYDRATE] (confidence: 0.844)
3. [Diagnosis_I10, Diagnosis_E78] -->
[Substance_HYDROCHLOROTHIAZIDE] (confidence: 0.742)
4. [Diagnosis_I10, Diagnosis_E78] --> [Substance_SIMVASTATIN]
(confidence: 0.727)
5. [Diagnosis_I10, Diagnosis_E78] --> [Substance_VALSARTAN]
(confidence: 0.707)

6. [Diagnosis_I10, Diagnosis_E78] --> [Diagnosis_K21] (confidence:
0.623)

8.2.2. GSP

E78

7.455 Εγγραφές. Γυναίκες: 805, άνδρες: 490

GSP min support 0.13

Έγινε επιλογή των κανόνων που περιείχαν διαφορετική νόσο και δραστικές ουσίες

1. 0.563: <Diagnosis_E78, Diagnosis_I10>
2. 0.331: <Diagnosis_E78, Diagnosis_I10> <Diagnosis_I10>
3. 0.322: <Diagnosis_E78> <Diagnosis_E78, Diagnosis_I10>
4. 0.286: <Diagnosis_I10> <Diagnosis_I10> <Diagnosis_I10>
5. 0.268: <Diagnosis_I10> <Diagnosis_E78> <Diagnosis_E78>
6. 0.256: <Diagnosis_I10> <Diagnosis_I10> <Diagnosis_E78,
Diagnosis_I10>
7. 0.232: <Diagnosis_E78, Diagnosis_I10,
Substance_HYDROCHLOROTHIAZIDE>
8. 0.215: <Diagnosis_E78, Diagnosis_I10, Substance_SIMVASTATIN>
9. 0.174: <Diagnosis_I10, Diagnosis_K21>
10. 0.174: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
11. 0.171: <Diagnosis_E78> <Diagnosis_I10> <Diagnosis_E78,
Diagnosis_I10> <Diagnosis_E78, Diagnosis_I10>
12. 0.167: <Diagnosis_I10> <Substance_ATORVASTATIN CALCIUM
TRIHYDRATE>
13. 0.162: <Diagnosis_E78, Diagnosis_I25.5>
14. 0.158: <Diagnosis_E78> <Diagnosis_K21>
15. 0.155: <Diagnosis_I10> <Substance_HYDROCHLOROTHIAZIDE>

16. 0.151: <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN
CALCIUM TRIHYDRATE> <Diagnosis_E78>
17. 0.150: <Diagnosis_E10, Diagnosis_E78>
18. 0.145: <Diagnosis_E78> <Diagnosis_E78, Diagnosis_K21>
19. 0.144: <Diagnosis_I10, Substance_ATORVASTATIN CALCIUM
TRIHYDRATE> <Diagnosis_E78, Substance_ATORVASTATIN CALCIUM
TRIHYDRATE>
20. 0.141: <Diagnosis_I10> <Substance_SIMVASTATIN>
21. 0.137: <Diagnosis_E78, Substance_HYDROCHLOROTHIAZIDE>
<Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE>
22. 0.136: <Diagnosis_E78, Diagnosis_I10, Substance_VALSARTAN>
23. 0.133: <Diagnosis_I10, Substance_ATORVASTATIN CALCIUM
TRIHYDRATE> <Diagnosis_E78, Diagnosis_I10,
Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
24. 0.133: <Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE>
<Diagnosis_E78, Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE>
25. 0.131: <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN
CALCIUM TRIHYDRATE> <Diagnosis_E78, Diagnosis_I10,
Substance_ATORVASTATIN CALCIUM TRIHYDRATE>

I10

10.136 Εγγραφές. Γυναίκες: 1.028, άνδρες: 734

GSP min support 0.1

Έγινε επιλογή των κανόνων που περιείχαν διαφορετική νόσο και δραστικές ουσίες

1. 0.414: <Diagnosis_E78, Diagnosis_I10>
2. 0.265: <Diagnosis_I10> <Diagnosis_E78>
3. 0.240: <Diagnosis_E78, Diagnosis_I10> <Diagnosis_E78>

4. 0.183: <Diagnosis_E78, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
5. 0.170: <Diagnosis_E78, Substance_HYDROCHLOROTHIAZIDE>
6. 0.161: <Diagnosis_I10> <Diagnosis_K21>
7. 0.158: <Diagnosis_E10, Diagnosis_I10>
8. 0.158: <Diagnosis_E78, Substance_SIMVASTATIN>
9. 0.149: <Diagnosis_I10> <Diagnosis_I10, Diagnosis_K21>
10. 0.143: <Diagnosis_I10, Diagnosis_K21> <Diagnosis_I10>
11. 0.128: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>
12. 0.125: <Diagnosis_K21> <Diagnosis_K21>
13. 0.123: <Diagnosis_E78, Substance_ATORVASTATIN CALCIUM TRIHYDRATE> <Diagnosis_I10>
14. 0.119: <Diagnosis_E10, Diagnosis_I10, Substance_METFORMIN HYDROCHLORIDE>
15. 0.119: <Diagnosis_K21> <Diagnosis_I10, Diagnosis_K21>
16. 0.119: <Diagnosis_E78, Diagnosis_I10> <Diagnosis_E78, Diagnosis_I10> <Diagnosis_E78, Diagnosis_I10> <Diagnosis_E78, Diagnosis_I10>
17. 0.117: <Diagnosis_E78> <Diagnosis_E78, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
18. 0.115: <Diagnosis_I25.5> <Diagnosis_I10>
19. 0.112: <Diagnosis_I10> <Diagnosis_I10, Diagnosis_K21> <Diagnosis_I10>
20. 0.110: <Diagnosis_E78> <Diagnosis_E78, Diagnosis_I10, Substance_ATORVASTATIN CALCIUM TRIHYDRATE>
21. 0.109: <Diagnosis_I10> <Diagnosis_E78, Substance_HYDROCHLOROTHIAZIDE>
22. 0.107: <Diagnosis_E78, Substance_HYDROCHLOROTHIAZIDE> <Diagnosis_I10>
23. 0.106: <Diagnosis_F41.9, Diagnosis_I10>
24. 0.106: <Diagnosis_I10> <Diagnosis_I25.5>

25. 0.106: <Diagnosis_I10, Diagnosis_K21> <Diagnosis_I10,
Diagnosis_K21>

26. 0.100: <Diagnosis_E78, Substance_HYDROCHLOROTHIAZIDE>
<Diagnosis_I10, Substance_HYDROCHLOROTHIAZIDE>

I10 & E78

5.370 Εγγραφές. Γυναίκες: 448, άνδρες: 281

GSP min support 0.22

Έγινε επιλογή των κανόνων που περιείχαν διαφορετική νόσο και δραστικές ουσίες

1. 0.309: <Diagnosis_E78, Diagnosis_I10, Diagnosis_K21>

2. 0.228: <Diagnosis_I10> <Diagnosis_K21>

3. 0.224: <Diagnosis_E78> <Diagnosis_K21>

9. Βιβλιογραφία

- Alwan, A. (2011). *Global status report on noncommunicable diseases 2010*. World Health Organization
- Akdag, B., Fenkci, S., Degirmencioglu, S., Rota, S., Sermez, Y., & Camdeviren, H. (2006). Determination of risk factors for hypertension through the classification tree method. *Advances in therapy*, 23(6), 885-892.
- Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.
- Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).
- Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on* (pp. 3-14). IEEE.
- Bhatnagar, D., Soran, H., & Durrington, P. N. (2008). Hypercholesterolaemia and its management. *BMJ*, 337.
- Boschi-Pinto, C., Lopez, A. D., Murray, C. J., Lozano, R., & Inoue, M. (2001). *Age standardization of rates: a new WHO standard* (pp. 1-12). Geneva: World Health Organization.
- Chang, C. D., Wang, C. C., & Jiang, B. C. (2011). Using data mining techniques for multi-diseases prediction modeling of hypertension and

hyperlipidemia by common risk factors. *Expert systems with applications*, 38(5), 5507-5513.

- Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine*, 26(1), 1-24.
- Cutler, J. A., Sorlie, P. D., Wolz, M., Thom, T., Fields, L. E., & Roccella, E. J. (2008). Trends in hypertension prevalence, awareness, treatment, and control rates in United States adults between 1988–1994 and 1999–2004. *Hypertension*, 52(5), 818-827.
- Dogan, S., & Turkoglu, I. (2008). Diagnosing hyperlipidemia using association rules. *Mathematical and Computational Applications*, 13(3), 193-202.
- Durrington, P. (2003). Dyslipidaemia. *The Lancet*, 362(9385), 717-731.
- Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., & Murray, C. J. (2002). Selected major risk factors and global and regional burden of disease. *The Lancet*, 360(9343), 1347-1360.
- Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3), 9.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).

- Jaskowski, M., & Jaroszewicz, S. (2012, June). Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*.
- Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), 65.
- Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71-82.
- Lewington, S., Clarke, R., Qizilbash, N., Peto, R., & Collins, R. (2002). Prospective studies collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*, 360(9349), 1903-1913.
- Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., ... & Aryee, M. (2013). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet*, 380(9859), 2224-2260.
- Mendis, S., Armstrong, T., Bettcher, D., Branca, F., Lauer, J., Mace, C., ... & Tang, K. C. (2014). Global status report on noncommunicable diseases 2014. *Switzerland: World Health Organization*, 14-15.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques: concepts and techniques*. Elsevier.
- Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In *ACM SIGMOD Record* (Vol. 29, No. 2, pp. 1-12). ACM.

- Hofmann, M., & Klinkenberg, R. (Eds.). (2013). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.
- Hooper, L., Summerbell, C. D., Thompson, R., Sills, D., Roberts, F. G., Moore, H., & Davey Smith, G. (2011). Reduced or modified dietary fat for preventing cardiovascular disease. *Cochrane Database Syst Rev*, 7(7).
- Roddick, J. F., & Spiliopoulou, M. (2002). A survey of temporal knowledge discovery paradigms and methods. *Knowledge and Data Engineering, IEEE Transactions on*, 14(4), 750-767.
- Rzepakowski, P., & Jaroszewicz, S. (2010, December). Decision trees for uplift modeling. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*(pp. 441-450). IEEE.
- Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2), 303-327.
- Simovici, D. A. Data Mining of Medical Data: Opportunities and Challenges in Mining Association Rules, *University of Massachusetts Boston*, 2011.
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson.
- Tang, J. L., Smith, G. D., Armitage, J. M., Lancaster, T., Silagy, C. A., Fowler, G. H., ... & Ebrahim, S. (1998). Systematic review of dietary intervention trials to lower blood total cholesterol in free-living subjectsCommentary: Dietary change, cholesterol reduction, and the public health—what does meta-analysis add?. *Bmj*, 316(7139), 1213-1220.

- Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, 29(3), 583-588.
- Williams, B., Poulter, N. R., Brown, M. J., Davis, M., McInnes, G. T., Potter, J. F., ... & Thom, S. M. (2004). British Hypertension Society guidelines for hypertension management 2004 (BHS-IV): summary. *BMJ: British Medical Journal*, 328(7440), 634.
- Wilson, A. M., Thabane, L., & Holbrook, A. (2004). Application of data mining techniques in pharmacovigilance. *British journal of clinical pharmacology*, 57(2), 127-134.
- World Health Organization. (2004). *International statistical classification of diseases and related health problems* (Vol. 1). World Health Organization.
- World Health Organization, & UNAIDS. (2007). *Prevention of cardiovascular disease*. World Health Organization.
- World Health Organization. (2009). *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization.
- World Health Organization. (2011). *Causes of death 2008: data sources and methods*. Geneva: World Health Organization.
- World Health Organization (WHO). (2015). *A global brief on hypertension: silent killer, global public health crisis*. Women.