

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση και Επεξεργασία Ιατρικών Δεδομένων και Σχεδίαση της Υποκείμενης Γνώσης

Μαρία Χριστοδούλου Α.Μ: Π2015001

Επιβλέπων Καθηγητής: Φοίβος Μυλωνάς

Επιτροπή: Κάτια Λήδα Κερμανίδου

Θεμιστοκλής Έξαρχος

Κέρκυρα

Περίληψη

Στον τομέα της ιατρικής και πιο συγκεκριμένα στα πλαίσια ενός οργανωμένου νοσοκομειακού περιβάλλοντος, η εξεύρεση της γνώσης και η εξαγωγή της είναι πολύ κρίσιμες διαδικασίες. Από μόνη της η ιατρική δεν είναι σε θέση να αναδείξει τη γνώση, η οποία στην ουσία είναι προϊόν της εμπειρίας, αλλά και της συχνότητας εμφάνισης της σε περιστατικά ασθενών. Η συγκεκριμένη γνώση εάν εξαχθεί από τις μεγάλες βάσεις δεδομένων, εντός των οποίων είναι κρυμμένη, είναι σε θέση να ισχυροποιήσει τα θεωρητικά ιατρικά συμπεράσματα και τα συμπεράσματα που περιγράφονται σε ιατρικά εγχειρίδια, αλλά και να αυξήσει αυτή τη θεωρητική γνώση. Οι ειδικοί του τομέα είναι σε θέση να εξάγουν αντικειμενικά αποτελέσματα στηριζόμενοι στην επιστήμη τους, κάνοντας χρήση υποκειμενικών και ουσιαστικών στοιχείων παραμέτρων, όπως είναι το ιατρικό ιστορικό, τα αποτελέσματα εξετάσεων και άλλα. Μέσα σε αυτά τα πλαίσια σημαντικό ρόλο έχουν και οι οντολογίες. Μία οντολογία είναι η αναπαράσταση μιας οντότητας ή των εννοιών μιας γνωστικής περιοχής, έτσι ώστε να μπορεί να την επεξεργαστεί ο Η/Υ. Μια οντολογία ορίζει τους όρους που χρησιμοποιούνται για να περιγράψουν και να αναπαραστήσουν. Οι οντολογίες ερμηνεύουν λοιπόν την πληροφορία και επικοινωνούν με άλλα εργαλεία λογικής. Με τη βοήθεια εργαλείων, εξάγουν λογικά συμπεράσματα για τις έννοιες που αναπτύσσονται μέσα σε αυτές και τις σχέσεις μεταξύ τους.

Ο σκοπός της παρούσας εργασίας είναι η πειραματική εφαρμογή μεθόδων αναζήτησης γνώσης σε πραγματικά ιατρικά δεδομένα. Στόχος είναι να αναζητηθεί και να αναδειχθεί η κρυμμένη γνώση, η οποία θα μπορέσει να αξιολογηθεί από ειδικούς στον τομέα της υγείας. Χρησιμοποιήθηκαν ιατρικά δεδομένα που σχετίζονται με καρδιαγγειακές παθήσεις. Στα πλαίσια της εργασίας πραγματοποιήθηκαν τρία πειράματα, από τα οποία προέκυψε πως τα ποσοστά της ακρίβειας κάθε αλγόριθμου για την μεμονωμένη εφαρμογή του φίλτρου Attribute Selection που έγινε στο πείραμα 1 κυμαίνονται στα ίδια ποσοστά ακριβείας που εξάχθηκαν στα πείραμα 2 και 3. Το πείραμα 3 έχει σχεδόν τα ίδια αποτελέσματα (accuracy, precision, recall, f-measure) με το πείραμα 2.

Η μείωση των χαρακτηριστικών επιφέρει πολλές φορές καλύτερα αποτελέσματα λόγω του ότι μπορούν να αγνοηθούν τα περιττά ή άσχετα με την τιμή εξόδου χαρακτηριστικά. Επίσης, διακρίνοντας ποια χαρακτηριστικά είναι πιο σημαντικά για το αποτέλεσμα μιας διαδικασίας, μπορεί να αποκτηθεί διαίσθηση για το πραγματικό πρόβλημα, επιτρέποντας στους ειδικούς του τομέα να το αντιμετωπίσουν αποτελεσματικότερα. Κατά την εφαρμογή του φίλτρου Class Balancer έχει παρατηρηθεί ότι με την εξισορρόπηση κλάσεων για τους περισσότερους αλγόριθμους χάνεται χρήσιμη πληροφορία σχετικά με τις συχνότητες εμφάνισης.

Abstract

In the field of medicine and more specifically in the context of an organized hospital environment, finding knowledge and extracting it is a very critical process. Medicine alone is not able to highlight knowledge, which is essentially a product of experience, but also of the frequency of its occurrence in patient cases. This knowledge, if extracted from the large databases in which it is hidden, is able to strengthen the theoretical medical conclusions and the conclusions described in medical textbooks, but also to increase this theoretical knowledge. Experts in the field are able to extract objective results based on their science, using subjective and substantive parameters, such as medical history, test results and more. In this context, ontologies also play an important role. An ontology is the representation of an entity or the concepts of a cognitive area so that it can be processed by the computer. An ontology defines the terms used to describe and represent. Ontologies therefore interpret information and communicate with other tools of logic. With the help of tools, they draw logical conclusions about the concepts that develop within them and the relationships between them.

The purpose of this research is the experimental application of knowledge seeking methods to real medical data. The aim is to seek out and reveal the hidden knowledge, which will be able to be evaluated by health experts. Medical data related to cardiovascular disease were used. As part of the work, three experiments were performed, which showed that the accuracy percentages of each algorithm for the individual application of the Attribute Selection filter performed in Experiment 1 are the same as the accuracy rates extracted in Experiment 2 and 3. Experiment 3 has almost the same results (accuracy, precision, recall, f-measure) as in Experiment 2.

Reducing features often results in better results because features that are unnecessary or unrelated to the output value can be ignored. Also, by distinguishing which features are most important for the outcome of a process, intuition can be gained about the real problem, allowing industry experts to deal with it more effectively. When applying the Class Balancer filter, it has been observed that by balancing classes for most algorithms, useful information about occurrence frequencies is lost.

Πίνακας Περιεχομένων

Κεφάλαιο 1: Εισαγωγή

- 1.1 Δομή διπλωματικής εργασίας
- 1.2 Στόχος διπλωματικής εργασίας

Κεφάλαιο 2: Θεωρητικό υπόβαθρο δεδομένων

- 2.1 Περιγραφή δεδομένων
- 2.2 Τύπος δεδομένων και ποιότητα δεδομένων
- 2.3 Προ-επεξεργασία δεδομένων
 - 2.3.1 Καθορισμός δεδομένων - Data cleaning
 - 2.3.2 Ενοποίηση δεδομένων - Integration
 - 2.3.3 Μετασχηματισμοί δεδομένων - Data transformation
- 2.4 Ανάλυση των δεδομένων βάση συσχετίσεων

Κεφάλαιο 3: Θεωρητικό υπόβαθρο εξόρυξης δεδομένων

- 3.1 Ρόλος της μηχανικής μάθησης
- 3.2 Διαδικασία εξόρυξης δεδομένων
- 3.3 Απαιτήσεις στην εξόρυξη δεδομένων
- 3.4 Είδη και τεχνικές εξόρυξης δεδομένων
 - 3.4.1 Κατηγοριοποίηση - Classification
 - 3.4.2 Ομαδοποίηση (συσταδοποίησης) - Clustering
 - 3.4.3 Κανόνες συσχέτισης - Association rules
 - 3.4.4 Πρότυπα ακολουθιών - Sequential Patterns
 - 3.4.5 Ομοιότητα χρονολογικών σειρών - Similarity of time series
 - 3.4.6 Παρουσίαση συνόψεων - Summarization
 - 3.4.7 Παλινδρόμηση - Regression
 - 3.4.8 Μέθοδος πρόβλεψης – Predictive method
- 3.5 Ρόλος της εξόρυξης δεδομένων στη ιατρική

Κεφάλαιο 4: Χρήση εργαλείου Weka για την εξόρυξη γνώσης από δεδομένα και την υλοποίηση αλγορίθμων

- 4.1 Αλγόριθμοι Πειραμάτων Μάθησης

4.2 Εργαλείο Weka

4.3 Περιγραφή του συνόλου δεδομένων Cardiovascular Disease

4.4 Εισαγωγή δεδομένων στο Weka

4.5 Υλοποίηση αλγορίθμων - Εξαγωγή αποτελεσμάτων και σύγκριση

Κεφάλαιο 5: Σχεδίαση της υποκείμενης γνώσης

5.1 Ορισμοί και κίνητρα χρήσης οντολογιών

5.2 Αρχές σχεδίασης και ανάπτυξης οντολογιών

5.3 Ανοιχτά ζητήματα στην ανάπτυξη οντολογιών

5.4 Διαχείριση ιατρικών δεδομένων μέσω οντολογιών

5.5 Οντολογίες ιατρικών δεδομένων

Κεφάλαιο 6: Συμπεράσματα και μελλοντικές κατευθύνσεις

6.1 Συμπεράσματα

6.2 Μελλοντικές κατευθύνσεις

Βιβλιογραφία

Κεφάλαιο 1: Εισαγωγή

1.1 Δομή διπλωματικής εργασίας

Η παρούσα εργασία αποτελείται από έξι κεφάλαια, στα πλαίσια των οποίων επιχειρείται η κάλυψη του θέματος, το οποίο πραγματεύεται.

Στο πρώτο κεφάλαιο περιλαμβάνεται η εισαγωγή, όπου γίνεται μία γενική περιγραφή του θέματος, και παρουσιάζονται ο σκοπός και η δομή της εργασίας.

Στο δεύτερο κεφάλαιο περιγράφονται βασικές έννοιες όπως τα δεδομένα, ο τύπος και η ποιότητα τους, η διαδικασία της προ-επεξεργασίας τους, αλλά και η ανάλυση τους στη βάση συσχετίσεων.

Στο τρίτο κεφάλαιο παρουσιάζεται η έννοια της εξόρυξης δεδομένων. Συγκεκριμένα περιγράφεται ο ρόλος της μηχανικής μάθησης, η διαδικασία, οι απαιτήσεις, τα είδη και οι τεχνικές της εξόρυξης δεδομένων, καθώς επίσης και ο ρόλος που διαδραματίζει η εξόρυξη δεδομένων στην ιατρική.

Στο τέταρτο κεφάλαιο παρουσιάζεται το εργαλείο Weka και πως αυτό συμβάλει στην εξόρυξη γνώσεις από δεδομένα και την υλοποίηση αλγορίθμων.

Στο πέμπτο κεφάλαιο παρουσιάζεται η έννοια της οντολογίας. Πιο συγκεκριμένα, περιγράφονται οι ορισμοί και τα κίνητρα για τη χρήση οντολογιών, οι αρχές σχεδίασης και ανάπτυξης των οντολογιών, αλλά και τα ανοιχτά ζητήματα που υπάρχουν γύρω από την ανάπτυξη τους. Επίσης περιγράφεται η διαχείριση των ιατρικών δεδομένων μέσω των οντολογιών, καθώς επίσης και οι οντολογίες των ιατρικών δεδομένων.

Τέλος, το έκτο κεφάλαιο της εργασίας αποτελεί τη συνολική επισκόπηση της εργασίας, όπου παρουσιάζονται τα συμπεράσματα, αλλά και ορισμένες προτάσεις για πιθανές μελλοντικές ερευνητικές κατευθύνσεις.

1.2 Στόχος διπλωματικής εργασίας

Είναι κοινά αποδεκτό πως το μέγεθος που έχουν πληροφορίες που καθημερινά αποθηκεύονται σε μία πληθώρα βάσεων δεδομένων, αυξάνεται ολοένα και περισσότερο και μάλιστα με ιδιαίτερα ταχείς ρυθμούς. Τα δεδομένα, τα οποία αποθηκεύονται ανήκουν σε πολλές και διαφορετικές κατηγορίες από απλά έγγραφα έως και πιο περίπλοκα δεδομένα. Ενδεικτικά μπορούν να αναφερθούν ορισμένες από τις κατηγορίες των δεδομένων τα οποία αποθηκεύονται, και αφορούν σε συναλλαγές εταιρειών, οι οποίες διατηρούνται ως αρχείο των συναλλαγών μεταξύ εταιρειών, αλλά και εσωτερικής λειτουργίας των εταιρειών, σε δεδομένα επιστημονικής φύσης, σε ιατρικά και προσωπικά δεδομένα που συγκεντρώνουν συστήματα υγείας και δημόσιοι φορείς με σκοπό τη βελτίωση της ποιότητας των παρεχόμενων υπηρεσιών, σε εικόνες και βίντεο και σε μηνύματα ηλεκτρονικού ταχυδρομείου.

Σε παλαιότερες δεκαετίες το σύνολο αυτών των δεδομένων αποθηκεύονταν κατά τρόπο ανεξέλεγκτο στα πλαίσια της μεγάλης δύναμης που μπορούσε να προσφέρει η ψηφιακή αποθήκευση, χωρίς ουσιαστικά να δίνεται προσοχή στις δυνατότητες που διέθεταν οι υπάρχουσες δομές αποθήκευσης δεδομένων για την επεξεργασία τους. Η πρακτική αυτή είχε ως αποτέλεσμα τη δημιουργία δομημένων συστημάτων διαχείρισης βάσεων δεδομένων, τα οποία είχαν τη δυνατότητα να προσφέρουν πιο αποδοτικές και αποτελεσματικές υπηρεσίες για την ανάκτηση πληροφοριών μέσα από μεγάλες σε όγκο συλλογές δεδομένων. Παρόλα αυτά τα συγκεκριμένα συστήματα διέθεταν λίγες δυνατότητες επεξεργασίας, οι οποίες σχετίζονταν με ερωτήματα σχεσιακού τύπου, έτσι ώστε να μπορεί να γίνει επιλογή υποσυνόλων δεδομένων από ένα γενικό σύνολο με βάση κριτήρια και διαδικασίες στατιστικής επεξεργασίας δεδομένων, έχοντας θέσει ο χρήστης ένα συγκεκριμένο στόχο προς αναζήτηση.

Καθώς οι πληροφορίες γνωρίζουν μία ραγδαία αύξηση, παρατηρείται συχνά ότι τα δεδομένα στα οποία σήμερα έχουμε πρόσβαση είναι κατά πολύ περισσότερα από όσα είμαστε σε θέση να διαχειριστούμε. Μάλιστα σε αρκετές περιπτώσεις τόσο το πλήθος των δεδομένων, όσο και το μέγεθος των βάσεων δεδομένων παρουσιάζουν αύξηση τόσο όσον αφορά το πλήθος των εγγράφων όσο και το πλήθος των πεδίων. Στο περιθώριο τέτοιων τεράστιων βάσεων δεδομένων, υπάρχει μία γνώση, η οποία δεν είναι ορατή,, προφανής και γνωστή από πριν, η οποία όμως εάν έρθει στο προσκήνιο μπορεί να είναι ιδιαιτέρως χρήσιμη. Τη συγκεκριμένη ανάγκη επιχειρεί να καλύψει ένα νέο επιστημονικό πεδίο, που ασχολείται με την ανακάλυψη της γνώσης σε βάσεις δεδομένων. Στα πλαίσια αυτού του επιστημονικού πεδίου λαμβάνουν χώρα τεχνικές και μέθοδοι για την εξόρυξη γνώσης, προκειμένου να βρεθούν πρότυπα και να κατασκευαστούν μοντέλα με βάση αυτά τα δεδομένα. Στη σημερινή εποχή, το συγκεκριμένο πεδίο, αλλά και οι εφαρμογές που σχετίζονται με αυτό συζητούνται ολοένα και περισσότερο όσον αφορά τα ιατρικά δεδομένα.

Στον τομέα της ιατρικής και πιο συγκεκριμένα στα πλαίσια ενός οργανωμένου νοσοκομειακού περιβάλλοντος, η εξεύρεση της γνώσης και η εξαγωγή της είναι πολύ κρίσιμες διαδικασίες. Η ιατρική από μόνη της δε μπορεί να αναδείξει τη γνώση, η οποία κατά βάση είναι αποτέλεσμα τόσο της εμπειρίας όσο και της συχνότητας με την οποία εμφανίζονται περιστατικά ασθενών. Η εν λόγω γνώση εάν εξαχθεί από τις ογκώδεις βάσεις δεδομένων, θα συμβάλει στην ενίσχυση των θεωρητικών ιατρικών συμπερασμάτων, αλλά και των αποτελεσμάτων που αναφέρονται σε ιατρικά άρθρα, καθώς επίσης και στον εμπλουτισμό της θεωρητικής γνώσης. Το επιστημονικό προσωπικό που ασχολείται με το συγκεκριμένο τομέα μπορεί να εξάγει αντικειμενικά αποτελέσματα χρησιμοποιώντας υποκειμενικά και ουσιαστικά στοιχεία, όπως τα αποτελέσματα των εξετάσεων και το ιατρικό ιστορικό.

Στα πλαίσια της παρούσας εργασίας, τα ιατρικά δεδομένα σχετίζονται με την καρδιαγγειακή νόσο. Λαμβάνοντας υπόψιν μας το γεγονός πως ο πληθυσμός γερνάει, η αστικοποίηση και η παγκοσμιοποίηση εντείνονται, οι καρδιαγγειακές παθήσεις αρχίζουν να αποτελούν ένα από τα πιο σοβαρά προβλήματα που καλείται να αντιμετωπίσει σήμερα η ανθρωπότητα (Galasso et al., 2015; De Rosa et al., 2018). Αν και γίνονται πολλές προσπάθειες για να βελτιωθεί η πρόληψη, η διάγνωση και η θεραπεία, οι καρδιαγγειακές παθήσεις αντιπροσωπεύουν την κύρια αιτία αναπηρίας και πρόωρου θανάτου σε όλο τον κόσμο. Από την έρευνα προκύπτει πως σε ετήσια βάση 17 εκατομμύρια περίπου θάνατοι είναι αποτέλεσμα καρδιαγγειακών αιτιών, ενώ αναμένεται πως μέχρι το 2030 αυτός ο αριθμός θα αγγίξει τα 23,3 εκατομμύρια. Η σοβαρή αναντιστοιχία που υπάρχει ανάμεσα στην ανάγκη για δεδομένα και τους πραγματικούς πόρους που είναι διαθέσιμοι, τονίζει την ανάγκη που υπάρχει να αναπτυχθούν νέες μορφές καρδιαγγειακής έρευνας, η οποία να μπορεί να αναλύει με οικονομικά αποδοτικό τρόπο μεγάλες ποσότητες πληροφοριών. Μεγάλος αριθμός δεδομένων είναι συχνά εύκολα διαθέσιμα, ωστόσο, επειδή συλλέγονται για άλλους λόγους, δεν επεξεργάζονται στα πλαίσια συμβατικής ερευνητικής πρακτικής. Σε άλλες περιπτώσεις τα δεδομένα είναι μέρος συνόλων δεδομένων τα οποία δε σχετίζονται μεταξύ τους. Η βελτίωση των πόρων είναι αποτέλεσμα της άρσης των εμποδίων ανάμεσα στις πηγές των δεδομένων και τη δημιουργία πλατφορμών που είναι ικανές να διαχειριστούν μεγάλο όγκο πληροφοριών (Silverio et al., 2019).

Μέσα σε αυτά τα πλαίσια ο σκοπός της εργασίας είναι η πειραματική εφαρμογή μεθόδων αναζήτησης γνώσης σε πραγματικά ιατρικά δεδομένα. Στόχος είναι να αναζητηθεί και να αναδειχθεί η κρυμμένη γνώση, η οποία θα μπορέσει να αξιολογηθεί από ειδικούς στον τομέα της υγείας. Χρησιμοποιήθηκαν ιατρικά δεδομένα που σχετίζονται με καρδιαγγειακές παθήσεις.

Κεφάλαιο 2: Θεωρητικό υπόβαθρο δεδομένων

2.1 Περιγραφή δεδομένων

Η έννοια των δεδομένων αναφέρεται συνήθως σε μία σειρά από ακατέργαστα δεδομένα, τα οποία μπορεί να είναι μια συλλογή από κείμενα, αριθμούς ή και σύμβολα, τα οποία δεν έχουν κανένα νόημα. Ως εκ τούτου, είναι σημαντικό τα δεδομένα να υποβάλλονται σε επεξεργασία ή να παρέχονται με ένα πλαίσιο, έτσι ώστε να αποκτούν νόημα, καθώς από μόνα τους τα δεδομένα δεν έχουν κάποιο νόημα. Τα δεδομένα παίρνουν νόημα και μετατρέπονται σε πληροφορίες μόνο όταν ερμηνεύονται. Τα δεδομένα αποτελούνται από πρωτογενή, ακατέργαστα γεγονότα και αριθμούς. Όταν αυτά τα δεδομένα υποβληθούν σε επεξεργασία με τη μορφή συνόλων, σύμφωνα με το ερευνητικό περιβάλλον εντός του οποίου πρόκειται να χρησιμοποιηθούν, τότε είναι σε θέση να παράσχουν πληροφορίες (Zins, 2007).

Τα δεδομένα αναφέρονται σε ακατέργαστες εισόδους πληροφοριών που όταν υποβάλλονται σε επεξεργασία ή οργανώνονται οδηγούν σε μία ερμηνεία που νόημα και ουσία. Οι πληροφορίες είναι συνήθως το επεξεργασμένο αποτέλεσμα των δεδομένων. Όταν τα δεδομένα υποβάλλονται σε επεξεργασία προκειμένου να προκύψουν οι πληροφορίες, τότε αυτά γίνονται κατανοητή και αποκτούν σημασία (Sanders, 2016).

Σε ένα σύστημα πληροφορικής τα σύμβολα, οι χαρακτήρες, οι εικόνες ή οι αριθμοί είναι δεδομένα. Αυτές είναι οι εισροές που πρέπει να επεξεργαστεί ένα σύστημα πληροφορικής προκειμένου να παράγει μια ουσιαστική ερμηνεία. Με άλλα λόγια, τα δεδομένα σε ουσιαστική μορφή γίνονται πληροφορίες. Οι πληροφορίες μπορούν να αφορούν γεγονότα, πράγματα, έννοιες ή οτιδήποτε σχετίζεται με το σχετικό θέμα. Μπορεί να παρέχει απαντήσεις σε ερωτήσεις όπως ποιος, τι, πότε, γιατί, και πώς (Sanders, 2016).

Η παρουσίαση δεδομένων παίζει καθοριστικό ρόλο στην έρευνα. Οι ερευνητές μπορούν να πείσουν την έρευνά τους στον αναγνώστη με την αποτελεσματική παρουσίαση δεδομένων. Βασικά, υπάρχουν δύο τύποι παρουσίασης δεδομένων: αριθμητικοί και γραφικοί (Slutsky, 2014; Praharaj, 2017).

Στα πλαίσια της αριθμητικής παρουσίασης υπάρχουν διάφοροι τύποι αριθμητικής παρουσίασης των δεδομένων, συμπεριλαμβανομένης της τακτοποίησής τους σε αύξουσα σειρά, φθίνουσα σειρά και ταξινόμηση των δεδομένων στη μορφή πίνακα. Ο δεύτερος τύπος παρουσίασης είναι η γραφική παρουσίαση. Τα γραφήματα είναι μια κοινή μέθοδος για την οπτική απεικόνιση των σχέσεων στα δεδομένα. Ένα γράφημα είναι μια γραφική αναπαράσταση των δεδομένων, στην οποία τα δεδομένα αντιπροσωπεύονται από σύμβολα, όπως γραμμές σε ένα γράφημα ράβδων, γραμμές σε ένα γράφημα γραμμών ή φέτες σε ένα γράφημα πίτας. Τα γραφήματα μάς επιτρέπουν να μελετήσουμε τη σχέση αιτίας-και-αποτελέσματος μεταξύ δύο μεταβλητών. Τα γραφήματα συμβάλλουν στη μέτρηση της έκτασης μιας μεταβλητής όταν μια άλλη μεταβλητή αλλάζει κατά ένα συγκεκριμένο ποσό (Franzblau & Chung, 2012; Slutsky, 2014).

Υπάρχουν διάφοροι τύποι γραφικής παρουσίασης (McDonald, 2001; Durbin, 2004):

- Ραβδόγραμμα. Ένα γράφημα ράβδων είναι η παρουσίαση δεδομένων χρησιμοποιώντας ορθογώνιες ράβδους, με ύψη ή μήκη ανάλογα με τις τιμές που αντιπροσωπεύουν. Ο αναγνώστης μπορεί εύκολα να συγκρίνει την ποσότητα παρατηρώντας το μήκος της ράβδου. Στο γράφημα ράβδων, οι ράβδοι μπορούν να απεικονιστούν οριζόντια ή κάθετα. Στον άξονα x, χρησιμοποιείται κατηγορηματική μεταβλητή, ενώ στον άξονα y, χρησιμοποιείται αριθμητική τιμή. Το γράφημα ράβδων έχει τρεις τύποι: απλό, παρακείμενο και αθροιστικό. Τα δύο τελευταία ονομάζονται επίσης γράφημα πολλαπλών ράβδων. Στο απλό γράφημα ράβδων, χρησιμοποιούνται το πολύ δύο

μεταβλητές (μία κατηγορική και μία ποσοτική), ενώ στο διάγραμμα πολλαπλών ράβδων, χρησιμοποιούνται το πολύ τρεις μεταβλητές (δύο κατηγορηματικές και μία ποσοτική). Τα πολλαπλά γραφήματα ράβδων είναι χρήσιμα όταν ένας ερευνητής θέλει να συγκρίνει σχήματα δύο ή περισσότερων διαφορετικών δεδομένων.

- Γράφημα γραμμής. Είναι εναλλακτικό γράφημα από το αυτό του ραβδογράμματος. Ένα γράφημα γραμμής είναι ένα είδος γραφήματος που αντιπροσωπεύει δεδομένα με τρόπο που μια σειρά σημείων πρόκειται να συνδεθεί με τμήματα ευθειών γραμμών. Η διαφορά μεταξύ γραφήματος γραμμής και ραβδογράμματος είναι πως το ραβδόγραμμα αντιπροσωπεύεται από ορθογώνιο ενώ το γράφημα γραμμής εμφανίζεται με γραμμή, αν και τα δύο χρησιμοποιούνται για τον ίδιο σκοπό.
- Διαγράμματα πίτας. Ένα γράφημα πίτας ορίζεται ως ένα γράφημα που περιέχει έναν κύκλο και χωρίζεται σε τομείς. Τα μήκη τόξου των τομέων είναι ανάλογα με την αριθμητική τιμή που αντιπροσωπεύουν. Χρησιμοποιείται μόνο για τα κατηγορηματικά δεδομένα.
- Ιστόγραμμα και πολύγωνο συχνότητας. Ένα ιστόγραμμα αντιπροσωπεύει την κατανομή συχνότητας μιας συνεχούς μεταβλητής της οποίας οι περιοχές είναι ανάλογες με τις αντίστοιχες συχνότητες. Ένα ιστόγραμμα είναι αρκετά παρόμοιο με το γράφημα ράβδων καθώς και τα δύο αποτελούνται από ορθογώνιες ράβδους. Η διαφορά είναι ότι δεν υπάρχει κενό μεταξύ δύο ράβδων στο ιστόγραμμα. Το ιστόγραμμα χρησιμοποιείται για τον έλεγχο της κανονικής κατανομής των συνεχών δεδομένων και έχει μόνο μία συνεχή μεταβλητή και δεν χρησιμοποιούνται κατηγορηματικές μεταβλητές για να το σχεδιάσουν, ενώ στο γράφημα ράβδων, απαιτούνται τουλάχιστον δύο μεταβλητές, συμπεριλαμβανομένης μιας ποσοτικής και μιας κατηγορικής μεταβλητής. Τα πολύγωνα συχνότητας εξυπηρετούν τον ίδιο σκοπό με τα ιστογράμματα, αλλά είναι ιδιαίτερα χρήσιμα για τη σύγκριση συνόλων δεδομένων. Τα πολύγωνα συχνότητας είναι επίσης μια καλή επιλογή για την εμφάνιση αθροιστικών κατανομών συχνότητας. Όταν τα μεσαία σημεία των κορυφών των ορθογώνιων ράβδων στο ιστόγραμμα ενώνονται, δημιουργείται το πολύγωνο συχνότητας.
- Γραμμή σφαλμάτων. Οι ράβδοι σφάλματος είναι γραφικές παραστάσεις της μεταβλητότητας των δεδομένων και χρησιμοποιούνται στο γράφημα για να υποδείξουν το σφάλμα ή την αβεβαιότητα (τυπική απόκλιση / τυπικό σφάλμα / διάστημα εμπιστοσύνης) σε μια αναφερόμενη μέτρηση (μέσος όρος). Δίνουν μια γενική ιδέα για το πόσο ακριβής είναι μια μέτρηση ή αντίστροφα, πόσο μακριά από την αναφερόμενη τιμή.
- Γραφικές παραστάσεις box. Οι εν λόγω γραφικές παραστάσεις χαρακτηρίζουν ένα δείγμα χρησιμοποιώντας το ελάχιστο.

- Διάγραμμα διασποράς. Ένα διάγραμμα διασποράς είναι ένα γράφημα στο οποίο οι τιμές δύο ποσοτικών μεταβλητών απεικονίζονται σε δύο άξονες. Το μοτίβο των σημείων που προκύπτουν αποκαλύπτουν οποιαδήποτε συσχέτιση που υπάρχει μεταξύ μεταβλητών για ένα σύνολο δεδομένων. Τα γραφήματα διασποράς δείχνουν πόσο μια μεταβλητή επηρεάζεται από μια άλλη. Η σχέση μεταξύ δύο μεταβλητών ονομάζεται συσχέτιση τους. Εάν τα σημεία δεδομένων κάνουν μια ευθεία γραμμή που πηγαίνει από την αρχή σε υψηλές τιμές x - και y -, τότε οι μεταβλητές λέγεται ότι έχουν θετική συσχέτιση. Εάν η γραμμή πηγαίνει από μια υψηλή τιμή στον άξονα y κάτω σε μια υψηλή τιμή στον άξονα x -, οι μεταβλητές έχουν αρνητικό συσχετισμό. Σε περίπτωση που δεν εμφανίστηκε τάση, καλείται μη συσχέτιση.
- Διάγραμμα Bland – Altman. Ένα διάγραμμα Bland – Altman (διαφορά διαγράμματος) είναι μια μέθοδος σχεδίασης δεδομένων που χρησιμοποιείται για την ανάλυση της συμφωνίας μεταξύ δύο διαφορετικών προσδιορισμών. Στο διάγραμμα Bland-Altman, οι διαφορές (μεταξύ των δύο μεθόδων) γράφονται σε σχέση με τους μέσους όρους των δύο μεθόδων. Εναλλακτικά, μπορούμε να επιλέξουμε να σχεδιάσουμε τις διαφορές (μεταξύ των δύο μεθόδων) έναντι μιας από τις δύο μεθόδους, εάν αυτή είναι μια μέθοδος αναφοράς και των δύο μεθόδων.
- Διάγραμμα “δάσος”. Ένα διάγραμμα “δάσος” είναι μια γραφική απεικόνιση των εκτιμώμενων αποτελεσμάτων από μια σειρά επιστημονικών μελετών που ασχολούνται με το ίδιο ερώτημα, μαζί με τα συνολικά αποτελέσματα. Είναι μια γραφική αναπαράσταση μιας μετα-ανάλυσης. Συνήθως συνοδεύεται από έναν πίνακα με αναφορές (συγγραφέας και ημερομηνία) των μελετών με το εκτιμώμενο αποτέλεσμα που περιλαμβάνεται στη μετα-ανάλυση.
- Άλλες γραφικές μέθοδοι. Εκτός από τα παραπάνω, υπάρχουν μερικές άλλες γραφικές μέθοδοι, που χρησιμοποιούνται στις ερευνητικές μελέτες, αν και είναι λιγότερο δημοφιλείς, συμπεριλαμβανομένων του διαγράμματος περιοχής, του πολικού διαγράμματος, της γραφικής παράστασης Youden, και του υψηλού-χαμηλού γραφήματος.

2.2 Τύπος δεδομένων και ποιότητα δεδομένων

Οι τύποι δεδομένων είναι μια σημαντική έννοια της στατιστικής, η οποία πρέπει να γίνει κατανοητή, για την ορθή εφαρμογή στατιστικών μετρήσεων στα δεδομένα και, ως εκ τούτου, για τη σωστή σύναψη ορισμένων υποθέσεων σχετικά με αυτά. Από την έρευνα εντοπίζουμε δύο διαφορετικούς τύπους δεδομένων, τα κατηγοριακά δεδομένα και τα αριθμητικά δεδομένα (Mishra, 2018).

Τα κατηγοριακά δεδομένα χρησιμοποιούνται κατά κύριο λόγο για να μετρήσουν τον τύπο και όχι το ποσό, και μπορούν να αναλυθούν σε δύο επιμέρους κατηγορίες, στα ονομαστικά και τα διατεταγμένα δεδομένα.

Τα ονομαστικά δεδομένα είναι δεδομένα που εκχωρούνται σε κατηγορίες ή ετικέτες ή μια μεγάλη σειρά δεδομένων όπου ο αριθμός εκχωρείται τυχαία. Οι κατηγορίες ή οι ετικέτες δεν μπορούν να ταξινομηθούν ή να τοποθετηθούν σε διάταξη και δεν σχετίζονται μεταξύ τους.

Τα διατεταγμένα δεδομένα είναι δεδομένα όπου οι αριθμοί έχουν χρησιμοποιηθεί για την τοποθέτηση αντικειμένων σε μια σειρά. Κατηγοριοποιείται σε μια σταθερή, συγκεκριμένη σειρά και μπορούμε μόνο να πούμε ότι ένα πράγμα είναι περισσότερο ή λιγότερο από κάτι άλλο στην κλίμακα. Αν και τα δεδομένα μπορούν να ταξινομηθούν κατά σειρά, δεν σημαίνει ότι βρίσκονται σε ίσα διαστήματα μεταξύ των κατηγοριών.

Τα αριθμητικά δεδομένα διακρίνονται σε μία σειρά από κατηγορίες που περιλαμβάνουν τα διακριτά δεδομένα και τα συνεχή δεδομένα.

Τα διακριτά δεδομένα περιγράφουν δεδομένα, των οποίων οι τιμές είναι διακριτές και ξεχωριστές. Με λίγα λόγια ως διακριτά δεδομένα περιγράφουμε τα δεδομένα που μπορούν να πάρουν μόνο συγκεκριμένες τιμές. Αυτός ο τύπος δεδομένων δεν μπορεί να υπολογιστεί αλλά μπορεί να καταμετρηθεί.

Τα συνεχή δεδομένα αντιπροσωπεύουν μετρήσεις και επομένως οι τιμές τους δεν μπορούν να καταμετρηθούν αλλά μπορούν να υπολογιστούν. Τα συνεχή δεδομένα μπορούν να αναλυθούν περαιτέρω σε δεδομένα διαστήματος και αναλογίας.

Στα δεδομένα διαστήματος οι τιμές του διαστήματος αντιπροσωπεύουν παραγγεληθείσες μονάδες που έχουν την ίδια διαφορά. Επομένως, μιλάμε για δεδομένα διαστήματος όταν έχουμε μια μεταβλητή που περιέχει αριθμητικές τιμές που ταξινομούνται και όπου γνωρίζουμε τις ακριβείς διαφορές μεταξύ των τιμών. Το πρόβλημα με τα δεδομένα τιμών διαστήματος είναι ότι δεν έχουν “πραγματικό μηδέν”. Με δεδομένα διαστήματος, μπορούμε να προσθέσουμε και να αφαιρέσουμε, αλλά δεν μπορούμε να πολλαπλασιάσουμε, να διαιρέσουμε ή να υπολογίσουμε τους λόγους. Επειδή δεν υπάρχει αληθινό μηδέν, δεν μπορούν να εφαρμοστούν πολλές περιγραφικές και συμπερασματικές στατιστικές.

Στα δεδομένα αναλογίας οι τιμές αναλογίας είναι επίσης ταξινομημένες μονάδες που έχουν την ίδια διαφορά. Οι τιμές αναλογίας είναι ίδιες με τις τιμές διαστήματος, με τη διαφορά ότι έχουν απόλυτο μηδέν.

Όταν οι άνθρωποι σκέφτονται την ποιότητα των δεδομένων, συνήθως αναφέρονται μόνο στην ακρίβεια. Ωστόσο, η ποιότητα των δεδομένων είναι κάτι παραπάνω από απλή ακρίβεια δεδομένων.

Άλλες σημαντικές διαστάσεις όπως η πληρότητα και η συνέπεια είναι απαραίτητες προκειμένου να χαρακτηριστεί πληρέστερα η ποιότητα των δεδομένων.

Ο παράγοντας της ακρίβειας μπορεί να αξιολογηθεί για διαφορετικά επίπεδα διακριτότητας ενός μοντέλου δεδομένων, που κυμαίνονται από μεμονωμένες τιμές έως ολόκληρες βάσεις δεδομένων. Για τιμές μεμονωμένων δεδομένων, η ακρίβεια μετρά την απόσταση μεταξύ μιας τιμής v και μιας τιμής v' που θεωρείται σωστή. Μπορούν να αναγνωριστούν δύο είδη ακρίβειας, δηλαδή μια συντακτική ακρίβεια και μια σημασιολογική ακρίβεια. Η συντακτική ακρίβεια μετράται μέσω συναρτήσεων σύγκρισης που αξιολογούν την απόσταση μεταξύ v και v' . Η απόσταση επεξεργασίας είναι ένα απλό παράδειγμα της λειτουργίας σύγκρισης, λαμβάνοντας υπόψιν το κόστος μετατροπής μιας συμβολοσειράς s σε συμβολοσειρά s' μέσω μιας ακολουθίας εισαγωγών χαρακτήρων, διαγραφών και αντικαταστάσεων. Υπάρχουν πιο σύνθετες λειτουργίες σύγκρισης που λαμβάνουν υπόψη παρόμοιο ήχο, μεταθέσεις κ.λπ. (Elfekey et al., 2002). Η σημασιολογική ακρίβεια καταγράφει τις περιπτώσεις στις οποίες το v είναι συντακτικά σωστή τιμή, αλλά διαφέρει από τον (Scannapieco et al., 2005).

Ο παράγοντας της πληρότητας μπορεί γενικά να οριστεί ως ο βαθμός στον οποίο τα δεδομένα έχουν επαρκές εύρος και βάθος για τη συγκεκριμένη εργασία (Wang & Madnick, 1989). Οι Pipino et al., (2002) προσδιορίζουν τρεις τύπους πληρότητας. Η πληρότητα σχήματος ορίζεται ως ο βαθμός στον οποίο οι οντότητες και τα χαρακτηριστικά δεν λείπουν από το σχήμα. Η πληρότητα στηλών είναι συνάρτηση των τιμών που λείπουν σε μια στήλη πίνακα. Η πληρότητα του πληθυσμού ισοδυναμεί με την αξιολόγηση των τιμών που λείπουν σε σχέση με έναν πληθυσμό αναφοράς.

Ο παράγοντας της συνοχής καταγράφει την παραβίαση των σημασιολογικών κανόνων που ορίζονται σε ένα σύνολο στοιχείων δεδομένων. Αναφορικά με τη σχεσιακή θεωρία, οι περιορισμοί της ακεραιότητας αποτελούν ένα παράδειγμα τέτοιων σημασιολογικών κανόνων. Οι περιορισμοί ακεραιότητας είναι ιδιότητες που πρέπει να ικανοποιούνται από όλες τις παρουσίες ενός σχήματος βάσης δεδομένων. Είναι δυνατόν να διακρίνουμε δύο κύριες κατηγορίες περιορισμών ακεραιότητας, συγκεκριμένα: περιορισμούς ενδο-σχέσεων και περιορισμούς μεταξύ σχέσεων. Οι περιορισμοί ακεραιότητας μεταξύ σχέσεων μπορούν να αφορούν μεμονωμένα χαρακτηριστικά (ονομάζονται επίσης και περιορισμοί τομέα) ή πολλαπλά χαρακτηριστικά μιας σχέσης (Scannapieco et al., 2005).

2.3 Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων είναι ένα από τα πιο σημαντικά στάδια που λαμβάνουν χώρα στα πλαίσια της ανακάλυψης γνώσης σε βάσεις δεδομένων. Η προεπεξεργασία συχνά απαιτεί το μεγαλύτερο ποσοστό της συνολικής προσπάθειας που καταβάλεται στη διαδικασία της ερμηνείας

και ανάλυσης των δεδομένων. Ο λόγος για αυτό είναι πως τα δεδομένα τα ίδια δεν είναι στην κατάλληλη μορφή, αλλά δεν είναι “καθαρά”, με αποτέλεσμα να μην υπάρχει λόγος να αναφέρεται κανείς για ποιότητα αποτελεσμάτων πριν λάβει χώρα η διαδικασία της προεπεξεργασίας δεδομένων. Στα πλαίσια της προεπεξεργασίας δεδομένων εντοπίζονται τρεις λειτουργίες που λαμβάνουν χώρα, ο καθαρισμός δεδομένων, η ενοποίηση δεδομένων, και οι μετασχηματισμοί των δεδομένων.

2.3.1 Καθαρισμός δεδομένων - Data cleaning

Ο καθαρισμός δεδομένων είναι μια διαδικασία αφαίρεσης θορύβου και ασυνεπών δεδομένων ή άσχετων δεδομένων. Σε γενικές γραμμές, τα δεδομένα που λαμβάνονται, τόσο από τη βάση δεδομένων μιας εταιρείας όσο και από τα πειραματικά αποτελέσματα, έχουν ατελή περιεχόμενα όπως χαμένα δεδομένα, μη έγκυρα δεδομένα ή απλά τυπογραφικά λάθη. Επιπλέον, υπάρχουν επίσης χαρακτηριστικά δεδομένων που δεν σχετίζονται με την υπόθεση εξόρυξης δεδομένων στην οποία συμμετέχουν. Τα άσχετα δεδομένα απορρίπτονται επίσης καλύτερα. Ο καθαρισμός δεδομένων θα επηρεάσει επίσης την απόδοση των τεχνικών εξόρυξης δεδομένων επειδή τα δεδομένα που διαχειρίζονται θα μειώσουν τον αριθμό και την πολυπλοκότητα (Han, 2006).

2.3.2 Ενοποίηση δεδομένων - Integration

Η ενοποίηση δεδομένων είναι η συγχώνευση δεδομένων από διάφορες βάσεις δεδομένων σε μια νέα βάση δεδομένων. Όχι συχνά, τα δεδομένα που απαιτούνται για την εξόρυξη δεδομένων δεν προέρχονται μόνο από μία βάση δεδομένων, αλλά επίσης προέρχονται από διάφορες βάσεις δεδομένων ή αρχεία κειμένου. Η ενοποίηση δεδομένων πραγματοποιείται στο χαρακτηριστικό φόρουμ που προσδιορίζει μοναδικές οντότητες όπως χαρακτηριστικά ονομάτων, τύπους προϊόντων, αριθμούς πελατών και άλλα. Η ενσωμάτωση δεδομένων πρέπει να γίνει προσεκτικά, επειδή τα σφάλματα στην ενοποίηση δεδομένων μπορούν να παράγουν αποκλίνοντα αποτελέσματα και ακόμη και να παραπλανήσουν τη δράση αργότερα. Για παράδειγμα, εάν η ενοποίηση δεδομένων με βάση τον τύπο προϊόντος αποδεικνύει ότι συνδυάζει προϊόντα από διαφορετικές κατηγορίες, τότε θα υπάρχει συσχέτιση μεταξύ προϊόντων που στην πραγματικότητα δεν υπάρχουν (Han, 2006).

2.3.3 Μετασχηματισμοί δεδομένων - Data transformation

Τα δεδομένα μετατρέπονται ή συνδυάζονται σε μορφή κατάλληλη για επεξεργασία στην εξόρυξη δεδομένων. Ορισμένες μέθοδοι εξόρυξης δεδομένων απαιτούν ειδικές μορφές δεδομένων προτού εφαρμοστούν. Για παράδειγμα, ορισμένες τυπικές μέθοδοι όπως η ανάλυση συσχέτισης και η ο-

μαδοποίηση μπορούν να δεχτούν μόνο κατηγορηματική εισαγωγή δεδομένων. Επομένως, τα δεδομένα με τη μορφή συνεχούς αριθμητικού αριθμού πρέπει να χωριστούν σε διάφορα διαστήματα. Αυτή η διαδικασία ονομάζεται συχνά μετασχηματισμός δεδομένων (Han, 2006).

2.4 Ανάλυση των δεδομένων βάση συσχετίσεων

Οι κανόνες σύνδεσης σχετίζονται με τη μελέτη του “τι είναι μαζί”. Οι κανόνες σύνδεσης θέλουν να παρέχουν αυτές τις πληροφορίες με τη μορφή σχέσεων “if-then” ή “if-then”. Αυτός ο κανόνας υπολογίζεται από πιθανά δεδομένα (Santoso, 2007). Η ανάλυση συσχέτισης είναι επίσης γνωστή ως μία από τις μεθόδους εξόρυξης δεδομένων που αποτελεί τη βάση διαφόρων άλλων μεθόδων εξόρυξης δεδομένων. Ιδιαίτερα ένα από τα στάδια της ανάλυσης συσχέτισης που ονομάζεται ανάλυση προτύπων υψηλής συχνότητας προσελκύει την προσοχή πολλών ερευνητών να παράγουν αποτελεσματικούς αλγόριθμους. Η σημασία ενός συσχετιστικού κανόνα μπορεί να γίνει γνωστή με δύο παραμέτρους, την υποστήριξη (τιμή υποστήριξης), που είναι το ποσοστό των στοιχείων συνδυασμού στη βάση δεδομένων και της εμπιστοσύνης (τιμή βεβαιότητας), δηλαδή την ισχυρή σχέση μεταξύ των στοιχείων στους συσχετιστικούς κανόνες ως μια διαδικασία εύρεσης όλων των συσχετιστικών κανόνων που πληρούν τις ελάχιστες απαιτήσεις για υποστήριξη (ελάχιστη υποστήριξη) και τις ελάχιστες απαιτήσεις για εμπιστοσύνη (ελάχιστη εμπιστοσύνη) (Pramudiono, 2007; Kumar & Chong, 2018).

Υπάρχουν διάφοροι αλγόριθμοι που έχουν αναπτυχθεί σχετικά με τους κανόνες συσχέτισης, αλλά υπάρχει ένας κλασικός αλγόριθμος που χρησιμοποιείται συχνά, δηλαδή ο εκ των προτέρων αλγόριθμος. Η βασική ιδέα αυτού του αλγορίθμου είναι η ανάπτυξη συχνών αντικειμένων. Με τη χρήση ενός αντικειμένου και την αναδρομική ανάπτυξη συχνών αντικειμένων με δύο στοιχεία, τρία στοιχεία και ούτω καθεξής σε συχνό σύνολο αντικειμένων με όλα τα μεγέθη. Για την ανάπτυξη συχνών συνόλων με δύο αντικείμενα, μπορούν να χρησιμοποιηθούν συχνές ομάδες. Ο λόγος είναι ότι εάν το σύνολο ενός στοιχείου δεν υπερβαίνει την ελάχιστη υποστήριξη, τότε οποιοδήποτε μεγαλύτερο μέγεθος αντικειμένου δεν θα υπερβαίνει αυτήν την ελάχιστη υποστήριξη. Σε γενικές γραμμές, η ανάπτυξη συνόλων με στοιχεία k χρησιμοποιεί συχνές σειρές με στοιχεία $k - 1$ που αναπτύχθηκαν στο προηγούμενο βήμα. Κάθε βήμα απαιτεί έναν μόνο έλεγχο ολόκληρης της βάσης δεδομένων. Στις συσχετίσεις υπάρχουν όροι προγενέστεροι και συνεπαγόμενοι, προηγούμενοι για την αναπαράσταση του τμήματος “if” και επακόλουθοι της εκπροσώπησης του τμήματος “then”. Σε αυτήν την ανάλυση, το προηγούμενο και το συνακόλουθο είναι μια ομάδα στοιχείων που δεν έχουν σχέση μεταξύ τους (Santoso, 2007; Kumar & Chong, 2018).

Από τον μεγάλο αριθμό κανόνων που θα μπορούσαν να αναπτυχθούν, είναι απαραίτητο να υπάρχουν κανόνες που να είναι αρκετά ισχυροί, το επίπεδο εξάρτησης μεταξύ αντικειμένων στο προηγούμενο και το συνακόλουθο. Για να μετρηθεί η ισχύς των κανόνων αυτής της ένωσης, χρησιμοποιήθηκαν μέτρα υποστήριξης και εμπιστοσύνης. Υποστήριξη είναι η αναλογία μεταξύ του αριθμού των συναλλαγών που περιέχουν προηγούμενες και επακόλουθες συναλλαγές. Η εμπιστοσύνη είναι η αναλογία μεταξύ του αριθμού των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία στο προηγούμενο και συνεπεία του αριθμού των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία στο προηγούμενο. Το πρώτο βήμα σε έναν *a priori* αλγόριθμο είναι πως η υποστήριξη κάθε στοιχείου υπολογίζεται με σάρωση της βάσης δεδομένων. Αφού ληφθεί υποστήριξη από κάθε στοιχείο, τα στοιχεία που έχουν υποστήριξη μεγαλύτερη από την ελάχιστη υποστήριξη επιλέγονται ως μοτίβα υψηλής συχνότητας με μήκος 1 ή συχνά συντομεύονται ως 1 σετ στοιχείων. Η συντομογραφία *k*-σετ στοιχείων σημαίνει ένα σύνολο που αποτελείται από *k* στοιχεία. Η δεύτερη επανάληψη παράγει 2 αντικείμενα που κάθε σετ έχει δύο αντικείμενα. Πρώτον, ένα υποψήφιο σετ 2 τεμαχίων δημιουργείται από ένα συνδυασμό όλων των σετ 1 τεμαχίου. Αυτό το υποψήφιο σετ 2 στοιχείων μετρά την υποστήριξή του με σάρωση της βάσης δεδομένων. Υποστήριξη σημαίνει τον αριθμό των συναλλαγών στη βάση δεδομένων που περιέχουν και τα δύο στοιχεία στον υποψήφιο σετ 2 στοιχείων. Αφού ληφθεί η υποστήριξη όλων των υποψήφιων σετ 2 στοιχείων, το υποψήφιο σετ 2 στοιχείων που πληροί τις ελάχιστες απαιτήσεις υποστήριξης μπορεί να προσδιοριστεί ως σετ 2 τεμαχίων που είναι επίσης μοτίβο υψηλής συχνότητας με μήκος 2 (Pramudiono, 2007; Kumar & Chong, 2018).

Κεφάλαιο 3: Θεωρητικό υπόβαθρο εξόρυξης δεδομένων

3.1 Ο ρόλος της μηχανικής μάθησης

Η μηχανική μάθηση αποτελεί κλάδο της τεχνητής νοημοσύνης που ως στόχο έχει να προσφέρει τη δυνατότητα στις μηχανές να εκτελούν εργασίες κατά τρόπο επιδέξιο μέσω της χρήσης έξυπνων λογισμικών. Οι στατιστικές μέθοδοι μάθησης είναι ο κύριος άξονας στον οποίο στηρίζεται το ευφυές λογισμικό που χρησιμοποιείται προκειμένου να αναπτυχθεί η μηχανική νοημοσύνη. Καθώς οι αλγόριθμοι της μηχανικής μάθησης έχουν ανάγκη δεδομένων, το πεδίο μελέτης είναι ανάγκη να σχετιστεί με τη βάση δεδομένων. Μέσα σε αυτά τα πλαίσια διαμορφώνονται και όροι όπως το *Discovery Knowledge from Data (KDD)*, η εξόρυξη δεδομένων και η αναγνώριση προτύπων (Ławrynowicz & Tresp, 2014). Για τον Mitchell (2006) το πεδίο μελέτης της μηχανικής μάθησης αποτελεί μια φυσική ανάπτυξη της τομής της Επιστήμης των Υπολογιστών και της Στατιστικής.

Ενώ η Επιστήμη των Υπολογιστών στρέφει το ενδιαφέρον της κατά κύριο λόγο στο μη αυτόματο τρόπο προγραμματισμού των υπολογιστών, η Μηχανική Εκμάθηση ενδιαφέρεται σχετικά με το πως να κάνει τους υπολογιστές να προγραμματίζουν οι ίδιοι. Από την άλλη, ενώ η Στατιστική στρέφει το ενδιαφέρον της στα συμπεράσματα που μπορούν να συναχθούν από τα δεδομένα, η Μηχανική Εκμάθηση ενδιαφέρεται για την ενσωμάτωση πρόσθετων ερωτήσεων όσον αφορά τις υπολογιστικές αρχιτεκτονικές και τους αλγόριθμους που μπορούν να χρησιμοποιηθούν έτσι ώστε να υπάρξει πιο αποτελεσματικότερη λήψη, αποθήκευση, εύρεση, ανάκτηση και συγχώνευση αυτών των δεδομένων, αλλά και την οργάνωση πολλών δευτερευόντων καθηκόντων μάθησης και ζητημάτων υπολογιστικής ιχνηλασιμότητας (Ławrynowicz & Tresp, 2014).

Υπάρχουν μία σειρά από εργασίες, τις οποίες ο άνθρωπος μπορεί να εκτελέσει καταβάλλοντας μικρή προσπάθεια και χωρίς να είναι σε θέση να εξηγήσει τον τρόπο που αυτές εκτελούνται. Η αναγνώριση της ομιλίας άλλων ανθρώπων για παράδειγμα αποτελεί ένα χαρακτηριστικό παράδειγμα, καθώς αν και ο άνθρωπος πραγματοποιεί την εργασία με ελάχιστη δυσκολία, δεν είναι σε θέση να εξηγήσει το μηχανισμό πίσω από την εκτέλεση. Λόγω της έλλειψης κατανόησης ενός τέτοιου φαινομένου (αναγνώριση ομιλίας σε αυτήν την περίπτωση), δεν είναι δυνατόν να δημιουργηθούν αλγόριθμοι για τέτοια σενάρια. Αυτό το κενό στην κατανόηση είναι που γεφυρώνουν οι αλγόριθμοι μηχανικής μάθησης (Mohri et al., 2012).

Δεν αποτελεί στόχο η κατανόηση των υποκείμενων διαδικασιών που συμβάλουν στη μάθηση του ανθρώπου. Σε αυτήν την περίπτωση τα προγράμματα των υπολογιστών μπορούν να κάνουν τις μηχανές να μάθουν και να εκτελέσουν εργασίες, όπως είναι η πρόβλεψη. Η μάθηση έχει ως στόχο. Το μοντέλο μπορεί να θεωρηθεί ως προσέγγιση της διαδικασίας που θέλουμε να μιμούνται τα μηχανήματα (Ławrynowicz & Tresp, 2014).

Στα πλαίσια αυτά υπάρχει η πιθανότητα σφάλματος σε κάποια εισαγωγή. Ωστόσο τις πιο πολλές φορές το μοντέλο είναι σε θέση να παράσχει ορθές απαντήσεις. Ως εκ τούτου, ένα άλλο μέτρο απόδοσης (πέραν της απόδοσης των μετρήσεων της ταχύτητας και της χρήσης μνήμης) για έναν αλγόριθμο μηχανικής μάθησης είναι η ακρίβεια που έχουν τα αποτελέσματα. Οι εργασίες μηχανικής μάθησης μπορούν να ταξινομηθούν στις ακόλουθες τέσσερις ομάδες (Ławrynowicz & Tresp, 2014):

- Στην επιτηρούμενη μάθηση
- Στη μη επιτηρούμενη μάθηση
- Στην ημι-επιτηρούμενη μάθηση
- Στη μάθηση με ενίσχυση

Επιτηρούμενη μάθηση

Στην επιτηρούμενη μάθηση, ο στόχος είναι να συναγάγουμε μια συνάρτηση ή μια χαρτογράφηση από δεδομένα που φέρουν ετικέτα. Τα δεδομένα αποτελούνται από το φορέα εισόδου X και το φορέα εξόδου Y από ετικέτες. Μια ετικέτα από τον φορέα Y είναι η εξήγηση του αντίστοιχου παραδείγματος εισόδου από τον φορέα εισόδου X . Μαζί σχηματίζουν ένα παράδειγμα. Με άλλα λόγια, τα δεδομένα περιλαμβάνουν παραδείγματα. Εάν η επισήμανση δεν υπάρχει για το φορέα εισόδου X , τότε το X είναι δεδομένα χωρίς ετικέτα. Ο φορέας εξόδου Y αποτελείται από ετικέτες για κάθε παράδειγμα που υπάρχει στα δεδομένα. Αυτές οι ετικέτες για τον φορέα εξόδου παρέχονται από τον επόπτη. Συχνά, αυτοί οι επόπτες είναι άνθρωποι, αλλά οι μηχανές μπορούν επίσης να χρησιμοποιηθούν για τέτοια σήμανση. Οι κρίσεις για τον άνθρωπο είναι ακριβότερες από τις μηχανές, αλλά τα υψηλότερα ποσοστά σφάλματος στα δεδομένα που επισημαίνονται από μηχανές υποδηλώνουν την υπεροχή της ανθρώπινης κρίσης. Τα δεδομένα με μη αυτόματη σήμανση είναι ένας πολύτιμος και αξιόπιστος πόρος για εποπτευόμενη μάθηση. Ωστόσο, σε ορισμένες περιπτώσεις, τα μηχανήματα μπορούν να χρησιμοποιηθούν για αξιόπιστη επισήμανση. Δύο ομάδες ή κατηγορίες αλγορίθμων εμπίπτουν στην ομπρέλα της εποπτευόμενης μάθησης, η παλινδρόμηση και η ταξινόμηση.

Μη επιτηρούμενη μάθηση

Στην μη εποπτευόμενη μάθηση, στερούμαστε εποπτικών ή εκπαιδευτικών δεδομένων. Με άλλα λόγια, όλα όσα έχουμε είναι δεδομένα χωρίς ετικέτα. Η ιδέα είναι να βρεθεί μια κρυφή δομή σε αυτά τα δεδομένα. Μπορεί να υπάρχουν διάφοροι λόγοι για τα δεδομένα να μην έχουν ετικέτα. Μπορεί να οφείλεται στη μη διαθεσιμότητα χρημάτων για την πληρωμή της μη αυτόματης επισήμανσης ή της εγγενούς φύσης των ίδιων των δεδομένων. Με πολλές συσκευές συλλογής δεδομένων, τώρα τα δεδομένα συλλέγονται με πρωτοφανή ρυθμό. Η ποικιλία, η ταχύτητα και ο όγκος είναι οι διαστάσεις στις οποίες φαίνεται και κρίνεται το Big Data. Είναι σημαντικό να ληφθεί κάτι από αυτά τα δεδομένα χωρίς τον επόπτη. Αυτή είναι και η πρόκληση για τον σημερινό επαγγελματία μηχανικής μάθησης.

Ημι-επιτηρούμενη μάθηση

Σε αυτόν τον τύπο μάθησης, τα δεδομένα που δίνονται είναι ένα μείγμα ταξινομημένων και μη ταξινομημένων δεδομένων. Αυτός ο συνδυασμός επισημασμένων και μη επισημασμένων δεδομένων χρησιμοποιείται για τη δημιουργία ενός κατάλληλου μοντέλου για την ταξινόμηση των δεδομένων. Στις περισσότερες περιπτώσεις, τα δεδομένα με επισήμανση είναι λιγοστά και τα δεδομένα

χωρίς επισήμανση είναι σε αφθονία. Ο στόχος της ημι-επιτηρούμενης ταξινόμησης είναι να υπάρξει ένα μοντέλο που θα προβλέπει τάξεις μελλοντικών δεδομένων δοκιμών καλύτερα από εκείνες από το μοντέλο που δημιουργείται χρησιμοποιώντας μόνο τα δεδομένα με ετικέτα.

Μάθηση με ενίσχυση

Η μέθοδος εκμάθησης ενίσχυσης στοχεύει στη χρήση παρατηρήσεων που συλλέγονται από την αλληλεπίδραση με το περιβάλλον για τη λήψη ενεργειών που θα μεγιστοποιούν την ανταμοιβή ή θα ελαχιστοποιούν τον κίνδυνο. Για την παραγωγή ευφυνών προγραμμάτων (που ονομάζονται επίσης πράκτορες), η ενίσχυση της μάθησης ακολουθεί τα ακόλουθα βήματα:

- Η κατάσταση εισαγωγής παρατηρείται από τον πράκτορα.
- Η λειτουργία λήψης αποφάσεων χρησιμοποιείται για να κάνει τον πράκτορα να εκτελέσει μια ενέργεια.
- Μετά την εκτέλεση της δράσης, ο πράκτορας λαμβάνει ανταμοιβή ή ενίσχυση από το περιβάλλον.
- Οι πληροφορίες ζεύγους κατάστασης δράσης για την ανταμοιβή αποθηκεύονται.

Χρησιμοποιώντας τις αποθηκευμένες πληροφορίες, η πολιτική για συγκεκριμένη κατάσταση όσον αφορά τη δράση μπορεί να ρυθμιστεί, βοηθώντας έτσι στη βέλτιστη λήψη αποφάσεων για τον πράκτορα.

3.2 Διαδικασία εξόρυξης δεδομένων

Συχνά ο όρος εξόρυξης δεδομένων ταυτίζεται με τον ορό μηχανική μάθηση αφού πολλές τεχνικές της εφαρμόζονται στη διαδικασία εξόρυξης δεδομένων. Επιπρόσθετα, εμπλέκονται και άλλα πεδία κατά την πραγματοποίηση διαδικασίας εξόρυξης δεδομένων όπως για παράδειγμα η στατιστική ανάλυση παρέχοντας παραστατικότητα σε μοντέλα, αναγνώριση προτύπων, εκτέλεση αλγόριθμων και οι βάσεις δεδομένων.

Η εξόρυξη δεδομένων είναι η διαδικασία κατά την οποία διεξάγεται η ανεύρεση γνώσης μέσα από την αξιοποίηση μεγάλων όγκων δεδομένων δημιουργώντας έτσι νέα πρότυπα, τα οποία αποσκοπούν στην επίλυση προβλημάτων. Για την πραγματοποίηση εξόρυξης δεδομένων γίνεται η εφαρμογή των παρακάτω στρατηγικών και τεχνικών διαδικασιών (Larose, 2005; Han et al., 2012):

- Καθορισμός προβλήματος. Προτού ξεκινήσει το πρακτικό μέρος της διαδικασίας εξόρυξης δεδομένων, ορίζεται το πρόβλημα προς επίλυση εξερευνώντας τις ανάγκες που πρέπει να καλυφθούν

θέτοντας στόχους. Αφού, κατανοηθεί η φύση του προβλήματος μεταβαίνει η διαδικασία στο επόμενο στάδιο, το οποίο αφορά τη συλλογή και αποθήκευση της κατάλληλης γνώσης για την διεκπεραίωση του εκάστοτε προβλήματος.

- Συλλογή – αποθήκευση δεδομένων. Επιλέγονται οι κατάλληλες πηγές συλλογής δεδομένων (μια συγκεκριμένη πηγή ή συνδυασμός πολλών πηγών), και καθορίζεται ο αποθηκευτικός χώρος (βάση δεδομένων, αρχεία). Μεγάλη έμφαση πρέπει να δοθεί και στον τύπο που θα έχουν τα δεδομένα και τα δείγματα. Αξιοσημείωτη είναι η αναφορά στην παρακάτω πηγή ορισμού που ορίζει με τον πιο ορθό τρόπο τη σχέση μιας βάσης δεδομένων με την διαδικασία εξόρυξης δεδομένων καθώς διεξάγεται το στάδιο ανακάλυψης γνώσης μέσα από τη μεθολογία αυτή.
- Προεπεξεργασία δεδομένων. Τα δεδομένα αποκτούν την κατάλληλη μορφή έτσι ώστε στην συνέχεια να μπορεί να διεξαχθεί ορθά το στάδιο της εφαρμογής αλγορίθμων. Η προεπεξεργασία είναι ένα από τα πιο καθορίστηκα και σημαντικά στάδια της διαδικασίας εξόρυξης δεδομένων. Αφορά τον καθαρισμό δεδομένων (data cleaning) από την ύπαρξη θορύβου και τον εντοπισμό ελλιπών πληροφοριών όπως (π.χ. κενά, μη έγκυρες τιμές, μη συμβατοί τύποι δεδομένων κλπ.). Ο θόρυβος επιφέρει σφάλματα τιμών (λανθασμένες τιμές) στα χαρακτηριστικά και διπλότυπες εγγραφές όπου αποκλείουν από τις πραγματικές τους τιμές. Ο θόρυβος μπορεί να προέρχεται από διάφορες αιτίες όπως λάθη που οφείλονται από τον άνθρωπο, ιδιορρυθμίες του λειτουργικού συστήματος στο οποίο βρίσκεται το περιβάλλον εκτέλεσης εξόρυξης δεδομένων κλπ. Τα δεδομένα στις πλείστες περιπτώσεις αφού περάσουν από το στάδιο της προ επεξεργασίας μετά αποθηκεύονται στις μόνιμες τους βάσης αποθήκευσης. Η διαδικασία της προ επεξεργασίας μπορεί να διεξαχθεί πολλές φορές μέχρι τα δεδομένα να φτάσουν στη επιθυμητή μορφή με όσο το δυνατό λιγότερες αποκλίσεις, στοχεύοντας στην μεγιστοποίηση της ακρίβειας.
- Μετασχηματισμός δεδομένων. Υπάρχουν περιπτώσεις όπου για λόγους πρακτικότητας τα δεδομένα θα ήταν πιο ορθό να αλλάξει ο αριθμητικός τους χαρακτήρας έτσι ώστε να μπορούν να εφαρμοστούν οι τεχνικές και οι αλγόριθμοι εξόρυξης. Μερικοί μέθοδοι μετασχηματισμού είναι η κανονικοποίηση δεδομένων χρησιμοποιώντας δεκαδική κλιμάκωση, εφαρμογή μαθηματικών τύπων για κανονικοποίηση ελάχιστου – μέγιστου, κανονικοποίηση με χρήση των τιμών Z-score και λογαριθμική κανονικοποίηση. Επιπρόσθετα, άλλοι μέθοδοι είναι η τροποποίηση τύπων δεδομένων, η ορθή επιλογή και η δημιουργία σχετικών χαρακτηριστικών και στιγμιότυπων.
- Εφαρμογή αλγορίθμων εξόρυξης δεδομένων. Παρέχεται ένα ευρύ φάσμα από διάφορα είδη τεχνικών και αλγορίθμων διαφόρων κατηγοριών όπου εφαρμόζονται στα προεπεξεργασμένα και μετασχηματισμένα σύνολα δεδομένων με σκοπό τη δημιουργία ενός μοντέλου. Το νέο αυτό μοντέλο δίνει στον άνθρωπο νέα εφόδια γνώσης, υποστηρίζει την λήψη της καλύτερης δυνατής

απόφασης, δίνει τιμές σε χαρακτηρίστηκα, παρέχει την δυνατότητα σύγκρισης αποτελεσμάτων και το ποιο σημαντικό από όλα παρέχει εξαγωγή νέας γνώσης και ανάπτυξη της ήδη υπάρχουσας. Οι βασικοί τύποι μοντέλων είναι τα περιγραφικά και τα μοντέλα πρόβλεψης.

- Ερμηνεία και αξιολόγηση εγκυρότητας αποτελεσμάτων. Στο στάδιο της ερμηνείας και της αξιολόγησης εγκυρότητας λαμβάνεται η απόφαση για το εάν το μοντέλο που έχει δημιουργηθεί είναι δεκτό και κατάλληλο να προσδώσει λύση στο εκάστοτε πρόβλημα για το οποίο αποσκοπούσε η υλοποίησή του. Η γνώση που έχει εξαχθεί παίρνει την κατάλληλη μορφή χρησιμοποιώντας τους απαραίτητους όρους για να γίνει κατανοητή προς του τελικούς χρήστες. Η ερμηνεία και η αξιολόγηση μπορεί να αναπαρασταθεί με πολλές μορφές (π.χ. στατιστική ανάλυση, ευρετική ανάλυση, πειραματική ανάλυση κλπ.). Εάν το μοντέλο που έχει δημιουργηθεί δεν αντιπροσωπεύει την επιθυμητή λύση για το πρόβλημα και δεν καλύπτει τις απαιτήσεις που έχουν τεθεί στο πρωταρχικό στάδιο η διαδικασία θα πρέπει να επαναληφθεί.

3.3 Απαιτήσεις στην εξόρυξη δεδομένων

Στην εξόρυξη δεδομένων, για να θεωρηθεί ότι ένα μοντέλο έχει δημιουργηθεί επιτυχώς και παρέχει την απαιτούμενη αποτελεσματικότητα, βασικό θεμέλιο είναι η πραγματοποίηση διερευνητικής διαδικασίας πάνω στο είδος των χαρακτηριστικών καθώς και η αναμέτρηση των απαιτήσεων που πρέπει να καλυφθούν μέσα από την διαδικασία ανάπτυξης τεχνικών εξόρυξης δεδομένων. Μέσα σε αυτά τα πλαίσια οι απαιτήσεις που εντοπίζονται είναι οι ακόλουθες (Chen et al., 1996):

- Διαχείριση διαφορετικών τύπων δεδομένων. Σε πρωταρχικά στάδια, απαιτείται η διερεύνηση και ο κατάλληλος χειρισμός μεταξύ των διαφοροποιημένων τύπων δεδομένων και διαφορετικών βάσεων δεδομένων έτσι ώστε το σύστημα εξόρυξης να εφαρμόζεται δυναμικά προσδίδοντας την προ απαιτούμενη αποδοτικότητα και συνάμα αποτελεσματικότητα ανεξάρτητα από την ύπαρξη διαφορετικών τύπων δεδομένων και βάσεων δεδομένων. Η ύπαρξη συγκεκριμένων διαφορετικών τύπων δεδομένων και στόχων μπορεί να οδηγήσει σε δυσκολίες με επακόλουθο το σύστημα να αδυνατεί να χειρίζεται όλες τις διαφοροποιημένες περιπτώσεις. Κλασικά παραδείγματα οι περιπτώσεις όπου το σύστημα καλείται να εξάγει γνώση από σχεσιακές βάσεις δεδομένων, χωροχρονικές βάσεις δεδομένων, βάσεις δεδομένων πολυμέσων κλπ. όπου απαιτούνται συγκεκριμένοι τύποι δεδομένων και εφαρμογών.
- Απαιτήσεις στη απόδοση και πρόοδο αλγορίθμων εξόρυξης γνώσης. Μια άλλη σημαντική απαίτηση, αφορά το πόσο αποδοτικοί και προοδευτικοί μπορεί να είναι οι αλγόριθμοι εξόρυξης δεδομένων κατά την εφαρμογή τους στα σύνολα δεδομένων. Ένας τρόπος να εξεταστεί η αποδοτικότητα και η προοδευτικότητα των αλγορίθμων είναι να τεθεί ως μέτρο αξιολόγησης ο χρόνος

εκτέλεσης τους, θέτοντας ως στόχο να τείνει όσο το δυνατό πιο πολύ προς τα επιθυμητά χρονικά πλαίσια, ιδικά στις περιπτώσεις διαχείρισης μεγάλων συνόλων δεδομένων και μεγάλων βάσεων δεδομένων.

- Η αξία της ακρίβειας στην απαίτηση εξαγωγής – αναπαράστασης ποιοτικής εξορυγμένης γνώσης. Απαιτήσεις παρουσιάζονται και μετά την ολοκλήρωση της διαδικασίας εξόρυξης δεδομένων αφού η αξία της ακρίβειας αποτελεί μέτρο ύψιστης σημασίας σχετικά με το πως παρουσιάζονται τα εξορυγμένα δεδομένων σε μια βάση δεδομένων. Χρησιμοποιώντας τον όρο ακρίβεια εκφράζεται η απαίτηση που πρέπει να καλυφθεί σχετικά με την ποιότητα των εξορυγμένων δεδομένων, δηλαδή αν είναι απαλλαγμένα από σφάλματα όπως για παράδειγμα ο θόρυβος και οι outliers, όπου σε τέτοιες περιπτώσεις το σύστημα καλείται να εφαρμόσει τους κατάλληλους μηχανισμούς καταπολέμησης σφαλμάτων. Η εξορυγμένη γνώση για να παραμένει ποιοτική πρέπει να εφαρμόζεται συχνή επισκόπηση πάνω σε αυτή με σκοπό τη εξέλιξη και τη βελτίωση της.
- Διαφορετικοί τύποι υποβολής ερωτημάτων και εξαγωγής – αναπαράστασης της εξορυγμένης γνώσης. Η ανάγκη για προσβασιμότητα στην εξορυγμένη γνώση από τους μη εξειδικευμένους χρήστες, έχει επιφέρει την απαίτηση για παρουσίαση των εξορυγμένων δεδομένων μέσα από μια μέθοδο αναπαράστασης κατάλληλα διαμορφωμένη έτσι ώστε τα εκάστοτε δεδομένα να είναι κατανοητά προς αυτούς. Οι χρήστες μπορούν να υποβάλλουν στο σύστημα ερωτήματα διαφορετικού τύπου με αποτέλεσμα το σύστημα να καλείται να διαχειρίζεται αναλόγως τα εισερχόμενα ερωτήματα. Παράλληλα, οι χρήστες μπορούν να εξάγουν διαφορετικούς τύπους γνώσεων μέσα από μεγάλα σύνολα δεδομένων, έτσι για την αποδοτικότερη εκμείευση αποτελεσμάτων πρέπει να καλύπτεται η ανάγκη έκφρασης των ερωτημάτων και η εξαγωγή της εξορυγμένης γνώσης μέσα από τη χρήση γλώσσας υψηλού επιπέδου ή μέσα από κατάλληλα διαμορφωμένα γραφικά διεπαφών.
- Διαλογική ανακάλυψη γνώσης σε πολλαπλά επίπεδα. Μέσα από την διαλογική ανακάλυψη, ο χρήστης ερχόμενος σε επαφή με το σύστημα του παρέχεται η δυνατότητα να υποβάλλει ερωτήματα τα οποία αλλάζουν τον προσδιορισμό των δεδομένων με σκοπό να πραγματοποιηθεί μια διαδικασία εξόρυξης δεδομένων πιο εξονυχιστικά. Έτσι επιτυγχάνεται, η μελέτη των δεδομένων και των αποτελεσμάτων εκτέλεσης μιας διαδικασίας εξόρυξης γνώσης μέσα από πολλά επίπεδα, καλύπτοντας την επιθυμία των χρηστών για πιο λεπτομερή έρευνα των εξορυγμένων δεδομένων.
- Εξαγωγή γνώσης μέσα από διαφορετικές πηγές δεδομένων. Ο χρήστης μπορεί να αποκτήσει πρόσβαση σε διάφορες πηγές δεδομένων μέσα από την πραγματοποίηση μιας σύνδεσης είτε τοπικού χαρακτήρα είτε παγκόσμιου, έτσι προκύπτει η κατασκευή μεγάλου όγκου κατανεμημένων

και ετερογενών βάσεων δεδομένων. Για να μπορεί να επιτευχθεί η σωστή διαχείριση σε βάσεις δεδομένων μεγάλου όγκου χρειάζεται να εφαρμοστούν παράλληλοι και καταναεμημένοι αλγόριθμοι εξόρυξης δεδομένων.

3.4 Είδη και τεχνικές εξόρυξης δεδομένων

Υπάρχουν διάφορες μέθοδοι που χρησιμοποιούνται για την εξόρυξη γνώσης από δεδομένα. Οι τεχνικές αυτές όχι μόνο απαιτούν συγκεκριμένους τύπους δομών δεδομένων, αλλά υλοποιούνται και με πολλούς διαφορετικούς αλγόριθμους. Στις τεχνικές αυτές εντοπίζουμε την κατηγοριοποίηση, την ομαδοποίηση, τους κανόνες συσχέτισης, τα πρότυπα ακολουθιών, την ομοιότητα χρονολογικών σειρών, την παρουσίαση συνόψεων, την παλινδρόμηση, και τη μέθοδο πρόβλεψης.

3.4.1 Κατηγοριοποίηση - Classification

Οι τεχνικές ταξινόμησης στην εξόρυξη δεδομένων είναι σε θέση να επεξεργάζονται μεγάλο αριθμό δεδομένων. Μπορεί να χρησιμοποιηθεί για την πρόβλεψη κατηγοριοποιημένων ετικετών κλάσης και ταξινόμησης δεδομένων με βάση το σύνολο εκπαίδευσης και τις ετικέτες τάξης και μπορεί να χρησιμοποιηθεί για την ταξινόμηση των πρόσφατα διαθέσιμων δεδομένων. Ο όρος θα μπορούσε να καλύπτει οποιοδήποτε πλαίσιο στο οποίο λαμβάνεται κάποια απόφαση ή πρόβλεψη βάσει των διαθέσιμων προς το παρόν πληροφοριών. Διαδικασίες ταξινόμησης αποτελούν αναγνωρισμένη μέθοδος για τη λήψη επανειλημμένων αποφάσεων σε νέες καταστάσεις (Porkizhi, 2017).

Η ταξινόμηση είναι η πιο συχνά εφαρμοζόμενη τεχνική εξόρυξης δεδομένων, η οποία χρησιμοποιεί ένα σύνολο προ-ταξινομημένων παραδειγμάτων για την ανάπτυξη ενός μοντέλου που μπορεί να ταξινομήσει τον πληθυσμό των αρχείων γενικά. Αυτή η προσέγγιση χρησιμοποιεί συχνά αλγόριθμους ταξινόμησης βάσει δένδρων αποφάσεων ή νευρωνικών δικτύων. Η διαδικασία ταξινόμησης δεδομένων περιλαμβάνει μάθηση και ταξινόμηση. Στην εκμάθηση τα δεδομένα εκπαίδευσης αναλύονται με αλγόριθμο ταξινόμησης. Στην ταξινόμηση τα δεδομένα δοκιμής χρησιμοποιούνται για την εκτίμηση της ακρίβειας των κανόνων ταξινόμησης. Εάν η ακρίβεια είναι αποδεκτή, οι κανόνες μπορούν να εφαρμοστούν στις νέες πλειάδες δεδομένων (Porkizhi, 2017). Στην κατηγοριοποίηση εντοπίζουμε τα εξής μοντέλα:

- Ταξινόμηση με επαγωγή δέντρου αποφάσεων
- Ταξινόμηση του Bayes
- Νευρωνικά δίκτυα
- Υποστήριξη διανυσματικών μηχανών (SVM)
- Ταξινόμηση με βάση τις ενώσεις

3.4.2 Ομαδοποίηση (συσταδοποίηση) - Clustering

Ομαδοποίηση είναι η ομαδοποίηση ενός συγκεκριμένου συνόλου αντικειμένων με βάση τα χαρακτηριστικά τους, συγκεντρώνοντας τα σύμφωνα με τις ομοιότητές τους. Όσον αφορά την εξόρυξη δεδομένων, αυτή η μεθοδολογία χωρίζει τα δεδομένα που εφαρμόζουν έναν ειδικό αλγόριθμο ένταξης, πιο κατάλληλο για την επιθυμητή ανάλυση πληροφοριών. Αυτή η ανάλυση ομαδοποίησης επιτρέπει σε ένα αντικείμενο να μην ανήκει σε ένα σύμπλεγμα ή να ανήκει αυστηρά σε αυτό, κάνοντας αυτό το είδος ομαδοποίησης να αποτελεί ένα σκληρό διαχωρισμό. Από την άλλη πλευρά, ο χαλαρός διαχωρισμός δηλώνει ότι κάθε αντικείμενο ανήκει σε ένα σύμπλεγμα σε καθορισμένο βαθμό. Πιο συγκεκριμένα τμήματα μπορούν να δημιουργηθούν σαν αντικείμενα που ανήκουν σε πολλαπλές συστάδες, να αναγκάσει ένα αντικείμενο να συμμετάσχει σε ένα μόνο σύμπλεγμα ή ακόμη και να δημιουργήσει ιεραρχικά δέντρα σε ομαδικές σχέσεις (Porkizhi, 2017). Η ομαδοποίηση μπορεί να θεωρηθεί ως αναγνώριση παρόμοιων κατηγοριών αντικειμένων. Χρησιμοποιώντας τεχνικές ομαδοποίησης μπορούμε να εντοπίσουμε περαιτέρω πυκνές και αραιές περιοχές στο χώρο αντικειμένων και να ανακαλύψουμε το συνολικό μοτίβο κατανομής και συσχετίσεις μεταξύ των χαρακτηριστικών δεδομένων. Η προσέγγιση ταξινόμησης μπορεί επίσης να χρησιμοποιηθεί για αποτελεσματικά μέσα διάκρισης ομάδων ή κατηγοριών αντικειμένων, αλλά καθίσταται δαπανηρή, ώστε η ομαδοποίηση μπορεί να χρησιμοποιηθεί ως προσέγγιση προεπεξεργασίας για επιλογή και ταξινόμηση υποομάδων χαρακτηριστικών (Porkizhi, 2017). Στην ομαδοποίηση εντοπίζουμε τους εξής τύπους μεθόδων:

- Μέθοδοι με βάση το πλέγμα
- Μέθοδοι βάσει μοντέλου
- Μέθοδοι που βασίζονται σε κεντροειδή
- Μέθοδοι που βασίζονται στη συνδεσιμότητα
- Μέθοδοι με βάση την πυκνότητα

3.4.3 Κανόνες συσχέτισης - Association rules

Η εξόρυξη των κανόνων συσχέτισης αποτελεί μία διαδικασία, η οποία έχει ως στόχο την εύρεση συχνών προτύπων, συσχέτισεων, ή αιτιωδών δομών από σύνολα δεδομένων που βρίσκονται σε διάφορα είδη βάσεων δεδομένων όπως σχεσιακές βάσεις δεδομένων, βάσεις δεδομένων συναλλαγών και άλλες μορφές αποθετηρίων δεδομένων. Ο συγκεκριμένος τύπος εύρεσης συμβάλει στη λήψη συγκεκριμένων αποφάσεων. Οι αλγόριθμοι κανόνων συσχέτισης πρέπει να είναι σε θέση να δημιουργούν κανόνες με τιμές εμπιστοσύνης μικρότερες από μία. Παρόλα αυτά οι πιθανοί κανόνες

σύνδεσης ενός δεδομένου συνόλου δεδομένων σε γενικές γραμμές είναι μεγάλος και ένα υψηλό ποσοστό κανόνων είναι συχνά μικρής αξίας (εάν όχι μηδαμινής) (Porkizhi, 2017). Στους κανόνες συσχέτισης εντοπίζουμε τους εξής τύπους:

- Πολυεπίπεδος κανόνας συσχέτισης
- Πολυδιάστατος κανόνας συσχέτισης
- Σταθμισμένοι κανόνες σύνδεσης
- Ποσοτικός κανόνας συσχέτισης
- Πολυσχεσιακοί κανόνες σύνδεσης
- Ασαφείς κανόνες σύνδεσης

3.4.4 Πρότυπα ακολουθιών - Sequential patterns

Ο όρος ακολουθιακά πρότυπα μπορεί περιγράφει είτε σύνολα αντικειμένων (itemsets), είτε ακολουθιών (sequences), είτε δομών (structures), οι οποίες κάνουν την εμφάνιση τους μέσα σε μια βάση δεδομένων με συχνότητα μεγαλύτερη από μια προκαθορισμένη. Ο στόχος είναι να εξορυχθούν τα σχετικά με το χρόνο ή άλλα ακολουθιακά πρότυπα (Raval, 2012). Οι βασικοί αλγόριθμοι που χρησιμοποιούνται προκειμένου να προσδιοριστούν τα ακολουθιακά πρότυπα είναι τρεις, οι οποίοι εξετάζουν, άμεσα ή έμμεσα, την *a priori* ιδιότητα στα πλαίσια της οποίας κάθε μη κενή υπό-συμβολοσειρά από ένα ακολουθιακό πρότυπο αποτελεί ακολουθιακό πρότυπο. Οι τρεις αλγόριθμοι είναι ο GSP, ο SPADE, και ο PrefixSpan. Ο GSP αναζητά και ελέγχει υποψήφια ακολουθιακά πρότυπα οριζόντιας μορφής αποθήκευσης. Ο αλγόριθμος SPADE είναι σχεδόν ίδιος με τον GSP, με τη μόνη διαφορά πως ο SPADE ελέγχει υποψήφια ακολουθιακά πρότυπα κατακόρυφης αποθήκευσης. Ο PrefixSpan, ο οποίος προέρχεται από τον αλγόριθμο FP-Growth, δε διαθέτει σύνολα υποψήφια υπό-συμβολοσειρών (Ramakrishnan & Rakesh, 1996; Zaki, 2001; PrexSpan, 2001).

3.4.5 Ομοιότητα χρονολογικών σειρών - Similarity of time series

Υπάρχουν ορισμένα μεγέθη, στα οποία παρατηρείται μία χρονική εξέλιξη. Η εξέλιξη αυτή αναπαριστάται με τη χρήση χρονοσειρών (time series), οι οποίες αποτελούν ακολουθίες σημείων. Οι χρονοσειρές είναι μετρήσεις μεγέθους στα πλαίσια της διάρκειας του χρόνου. Οι μέθοδοι που χρησιμοποιούνται για την ανάλυση των χρονοσειρών, στην ουσία αναλύουν δεδομένων διαφορετικών χρονικών περιόδων και εξάγουν συμπεράσματα, τα οποία χρησιμεύουν μόνο στα πλαίσια του φαινομένου με το οποίο σχετίζονται τα δεδομένα. Έτσι εάν οι τιμές παρουσιάζουν κανονικότητες στις διακυμάνσεις τους στη διάρκεια του χρόνου, τότε αυτές οι διακυμάνσεις μπορούν να χρησιμοποιηθούν για να διατυπωθούν προβλέψεις. Μέσα από την ανάλυση των χρονοσειρών γίνεται η

μελέτη της τιμής που έχει ένα γνώρισμα, όπως αυτό μεταβάλλεται μέσα στην πορεία του χρόνου. Οι τιμές κατά κύριο λόγο λαμβάνονται σε ίσα χρονικά διαστήματα. Προκειμένου μία χρονοσειρά να παρασταθεί οπτικά γίνεται χρήση ενός διαγράμματος χρονοσειρών (Raval, 2012).

3.4.6 Παρουσίαση συνόψεων - Summarization

Η παρουσίαση συνόψεων παρουσιάζει τα δεδομένα σε υποσύνολα τα οποία συνοδεύονται από απλές περιγραφές. Η σύνοψη των δεδομένων περιγράφεται επίσης και ως χαρακτηρισμός ή γενίκευση. Στα πλαίσια της παρουσίασης συνόψεων εξάγονται ή παράγονται αντιπροσωπευτικές πληροφορίες σχετικά με τις βάσεις δεδομένων. Αυτό επιτυγχάνεται μέσα από την ανάκτηση τμημάτων από τα δεδομένα. Εναλλακτικά υπάρχει η δυνατότητα να εξαχθούν συνοπτικές πληροφορίες από τα δεδομένα. Με λίγα λόγια η παρουσίαση των συνόψεων χαρακτηρίζει στην ουσία τα περιεχόμενα που διαθέτει μία βάση δεδομένων (Raval, 2012).

3.4.7 Παλινδρόμηση - Regression

Μια εργασία παλινδρόμησης αρχίζει με ένα σύνολο δεδομένων για το οποίο οι τιμές-στόχοι είναι γνωστές. Η τεχνική της παλινδρόμησης μπορεί να προσαρμοστεί και για τις ανάγκες της πρόβλεψης. Η ανάλυση της παλινδρόμησης μπορεί να χρησιμοποιηθεί και στη μοντελοποίηση των σχέσεων ανάμεσα σε μία ή και περισσότερες ανεξάρτητες και εξαρτημένες μεταβλητές. Στη διαδικασία της εξόρυξης δεδομένων οι μεταβλητές απόκρισης σχετίζονται με αυτό που πρέπει να προβλεφθεί και οι ανεξάρτητες μεταβλητές αποτελούν τα γνωστά χαρακτηριστικά. Ωστόσο στην πραγματικότητα τα περισσότερα προβλήματα δεν αποτελούν απλά ζήτημα πρόβλεψης. Σε περιπτώσεις όπως οι τιμές των μετοχών, ο όγκος των πωλήσεων ή η πιθανότητα να αποτύχει ένα προϊόν δεν είναι εύκολο να προβλεφθούν, καθώς η πρόβλεψη εξαρτάται από μία σειρά πολύπλοκων αλληλεπιδράσεων ανάμεσα σε πολλαπλές μεταβλητές πρόβλεψης. Ως εκ τούτου, τεχνικές που είναι πιο περιπλοκές, όπως τα νευρωνικά δίκτυα, τα δέντρα αποφάσεων και η λογιστική παλινδρόμηση, ίσως είναι πιο αναγκαίες για να προβλεφθούν μελλοντικές τιμές. Οι ίδιοι αυτοί τύποι μοντέλων συχνά είναι σε θέση να χρησιμοποιηθούν τόσο στην παλινδρόμηση όσο και στην ταξινόμηση. Για παράδειγμα, ο αλγόριθμος δέντρου αποφάσεων CART (Δέντρα ταξινόμησης και παλινδρόμησης) μπορεί να χρησιμοποιηθεί για να δημιουργηθούν δέντρα ταξινόμησης και δέντρα παλινδρόμησης. Τα νευρωνικά δίκτυα επίσης μπορούν να δημιουργήσουν μοντέλα ταξινόμησης και παλινδρόμησης (Porkizhi, 2017). Οι τύποι μεθόδων παλινδρόμησης που εντοπίζονται είναι οι εξής:

- Γραμμική παλινδρόμηση
- Μη γραμμική παλινδρόμηση

- Γραμμική παλινδρόμηση πολλαπλών παραλλαγών
- Πολυπαραγοντική μη γραμμική παλινδρόμηση

3.4.8 Μέθοδος πρόβλεψης - Predictive method

Αρκετές από τις πρακτικές εφαρμογές, οι οποίες χρησιμοποιούνται προκειμένου εξορυχθεί η γνώση μπορεί να θεωρηθεί ότι αποτελούν προβλέψεις μελλοντικών καταστάσεων με τη βοήθεια γνώσης από προηγούμενα και σημερινά δεδομένα. Η πρόβλεψη μπορεί να θεωρηθεί ως ένας τύπος κατηγοριοποίησης, με τη διαφορά όμως πως η πρόβλεψη σχετίζεται περισσότερο με μία τιμή μελλοντικής κατάστασης παρά μίας τρέχουσας. Ως εκ τούτου η συζήτηση αφορά για ένα είδος εφαρμογής και όχι σε μία προσέγγιση μοντελοποίησης. Παρά το γεγονός πως μπορεί να γίνει πρόβλεψη των μελλοντικών τιμών με τεχνικές ανάλυσης χρονοσειρών ή παλινδρόμησης, είναι δυνατή και η χρήση άλλων προσεγγίσεων (Raval, 2012).

3.5 Ο ρόλος της εξόρυξης δεδομένων στην ιατρική

Με την είσοδο της μηχανοργάνωσης στις μονάδες υγείας, τη σταθερή ανάπτυξη που παρατηρείτε στην ανάπτυξη των ηλεκτρονικών αρχείων για τους ασθενείς, αλλά και τα συστήματα για την κλινική ιατρική πληροφορική, η συλλογή των δεδομένων στα πλαίσια της συνήθους κλινικής χρήσης είναι πλέον κοινή πρακτική. Γενικά, το περιβάλλον της υγειονομικής περίθαλψης χαρακτηρίζεται από πληθώρα πληροφοριών, αλλά δεν έχει προχωρήσει στο ζήτημα της εξόρυξης γνώσης που προκύπτει από αυτές τις πληροφορίες. Παρόλο που τα συστήματα υγειονομικής περίθαλψης διαθέτουν πολλά δεδομένα, ωστόσο δεν υπάρχουν τα κατάλληλα εργαλεία να αναλύσουν συνθετικά αυτά τα δεδομένα έτσι ώστε να εντοπιστούν τάσεις και σχέσεις που πιθανόν υπάρχουν σε αυτά. Η εξόρυξη δεδομένων αποτελεί ένα πολύ σημαντικό βήμα στην ανακάλυψη της γνώσης και αυτός είναι ο λόγος που τα τελευταία χρόνια η βιομηχανία πληροφοριών έχει δείξει ιδιαίτερο ενδιαφέρον προς αυτήν την κατεύθυνση. Η διαδικασία για να εντοπιστεί η γνώση περιλαμβάνει τον καθορισμό των δεδομένων, την ενσωμάτωση των δεδομένων, την επιλογή των δεδομένων, την αναγνώριση προτύπων εξόρυξης δεδομένων και την παρουσίαση της γνώσης. Η διαδικασία αυτή είναι επαναλαμβανόμενη. Πιο συγκεκριμένα η διαδικασία της εξόρυξης δεδομένων μπορεί να φέρει εις πέρας την περιγραφή, τη συσχέτιση, την ταξινόμηση, την ομαδοποίηση, την πρόβλεψη, αλλά και την ανάλυση των χρονοσειρών. Επιπρόσθετα, μπορεί να προσφέρει αυτόματη αναγνώριση προτύπων, ενώ επιχειρεί να βρει και πρότυπα σε δεδομένα, όπου οι παραδοσιακές στατιστικές μέθοδοι δε λειτουργούν.

Προκειμένου να οργανωθεί η υγειονομική περίθαλψη με επιτυχία είναι απαραίτητο να υπάρχει η δυνατότητα να συλλέγονται, να αποθηκεύονται και να αναλύονται δεδομένα. Χωρίς τη χρήση της εξόρυξης δεδομένων δεν είναι εύκολο να γίνει πλήρης αξιοποίηση του δυναμικού των δεδομένων που συλλέγονται στα πλαίσια της οργάνωσης της υγειονομικής περίθαλψης, λαμβάνοντας υπόψιν πως τα δεδομένα που είναι υπό ανάλυση είναι πολλά. Μέσα από την εξόρυξη ιατρικών δεδομένων παρέχεται η δυνατότητα να εξερευνηθούν κρυμμένα μοτίβα στα σύνολα των δεδομένων του ιατρικού τομέα, τα οποία είναι δυνατόν να χρησιμοποιηθούν στα πλαίσια της κλινικής διάγνωσης. Την ίδια στιγμή, αυτά τα δεδομένα μπορούν να λειτουργήσουν ως πληροφορία και γνώση συμβάλλοντας στον έλεγχο, το κόστος, αλλά και τη διατήρηση υψηλού επιπέδου ποιότητας στη φροντίδα των ασθενών. Τα αρχεία των θεραπειών μεγάλου αριθμού ασθενών θα αποθηκεύονται, θα μηχανογραφούνται και με τη βοήθεια των τεχνικών εξόρυξης δεδομένων θα είναι σε θέση να δώσουν απαντήσεις σε μία σειρά σημαντικών και κρίσιμων ζητημάτων που έχουν σχέση με την υγειονομική περίθαλψη.

Κεφάλαιο 4: Χρήση εργαλείου Weka για την εξόρυξη γνώσης από δεδομένα και την υλοποίηση αλγορίθμων

4.1 Αλγόριθμοι Πειραμάτων Μάθησης

Το πειραματικό μέρος βασίζεται στη κατηγοριοποίηση καθώς όπως έχει αναφερθεί και προηγουμένως αποτελεί την πιο συχνά εφαρμοσμένη τεχνική εξόρυξης δεδομένων. Αυτός είναι και ο λόγος της επιλογής των αλγορίθμων JRip, IBk, J48 και Random Forest.

Ο αλγόριθμος Rip εφαρμόζει την μέθοδο Ripper και είναι υπεύθυνος για την κατασκευή ενός συνόλου κανόνων μέσα από την επαναληπτική πρόσθεση κανόνων στα πλαίσια ενός κενού συνόλου από αυτούς έτσι ώστε να υπάρξει κάλυψη κάθε θετικού παραδείγματος. Η δημιουργία των κανόνων γίνεται μέσα της προσθήκης συνθηκών στην υπόθεση ενός κανόνα (κάνοντας αρχή από μία κενή υπόθεση) και ολοκληρώνεται όταν πλέον τα αρνητικά παραδείγματα δεν καλύπτονται. Επίσης γίνεται χρήση συνδυασμού των τεχνικών μήκους ελάχιστης περιγραφής και cross-validation με σκοπό να αποτραπεί η υπερπροσαρμογή. Ο Rip μπορεί να χρησιμοποιηθεί μεταξύ άλλων και στα συνεχή δεδομένα, ενώ έχει υψηλή απόδοση και στο χρόνο εκτέλεσης ανεξαρτήτως μεγέθους βάσεων δεδομένων. Ο Rip αποτελείται από τέσσερα βασικά στάδια, την φάση της κατασκευής, την φάση της επέκτασης, την φάση του κλαδέματος και την φάση της βελτιστοποίησης.

Ο αλγόριθμος IBk (k-nearest neighbor) καθορίζει μια λίστα πλησιέστερων γειτόνων υπολογίζοντας την απόσταση ή την ομοιότητα που ενδεχόμενος να υπάρχει μεταξύ κάθε δείγματος ελέγχου και σε όλα τα δείγματα εκπαίδευσης. Δηλαδή, τα αποτελέσματα που εξάγει κάθε φορά τα κρατάει

μέσα στη μνήμη και όταν δοθεί μια άγνωστη πλειάδα ο ταξινομητής KNN ψάχνει να βρει το πρότυπο χώρους όσον αφορά τις k πλειάδες εκπαίδευσης, οι οποίες βρίσκονται πιο κοντά στην πλειάδα που είναι άγνωστη. Οι k πλειάδες εκπαίδευσης είναι και οι “πιο κοντινοί γείτονες” k στην άγνωστη πλειάδα (Pang-Ning et al., 2010; 2017).

Ο αλγόριθμος J48 (C4.5) βασίζεται σε δέντρα απόφασης και σε μια στατιστική ιδιότητα, που καλείται κέρδος πληροφορίας (information gain), προκειμένου να επιλέξει το γνώρισμα που θα ελέγξει σε κάθε κόμβο του δέντρου. Δηλαδή, όποιο έχει το μεγαλύτερο information gain μπαίνει στην ρίζα του δέντρου. Τα δεδομένα χωρίζονται σε αυτά που μας βοηθούν περισσότερο, που μας δίνουν περισσότερη πληροφορία σχετικά με την έξοδο που θέλουμε να βγάλουμε και σε αυτά που μας βοηθούν λιγότερο. Αν κάθε παρατήρηση που αφορά στο δείγμα ανήκει στην ίδια κλάση, τότε έχουμε μετατροπή του κόσμου σε φύλλο. Σε διαφορετική περίπτωση γίνεται επιλογή του γνωρίσματος που διαχωρίζει κατά τον καλύτερο δυνατό τρόπο τις παρατηρήσεις που αφορούν το δείγμα αναλόγως και της κλάσης στην οποία ανήκουν. Η εν λόγω διαδικασία επαναλαμβάνεται σε όλα τα υποσύνολα, τα οποία προέκυψαν από τους κλάδους στο προηγούμενο βήμα, από τα οποία αποτελείται το δείγμα. Η ολοκλήρωση της επανάληψης προκύπτει όταν μία τουλάχιστον από τις ακόλουθες συνθήκες εξόδου ικανοποιηθεί (Τεχνικές Εξόρυξης Δεδομένων, 2015-2016):

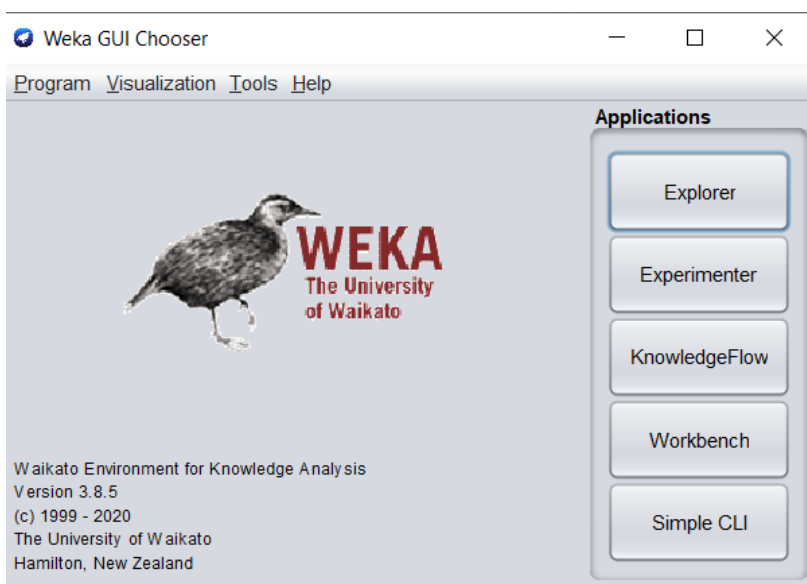
- Κάθε παρατήρηση που αφορά ένα κόμβο ανήκει στην ίδια κλάση.
- Όσον αφορά τη διαδικασία διαχωρισμού του δείγματος δεν εντοπίζονται άλλα γνωρίσματα. Στην περίπτωση αυτή υπάρχει μετατροπή του κόμβου σε φύλλο.
- Δεν εντοπίζονται περαιτέρω παρατηρήσεις όσον αφορά το υποσύνολο του δείγματος, το οποίο να ορίζεται από τον κλάδο.

Ο Random Forest αποτελεί έναν αλγόριθμο μηχανικής εκμάθησης, ο οποίος είναι εύχρηστος και ευέλικτος, και παράλληλα είναι υπεύθυνος για την παραγωγή, έστω και χωρίς να υπάρχει συντονισμός υπερ-παραμέτρων, τις πιο πολλές φορές ενός μεγάλου αποτελέσματος. Επιπρόσθετα είναι ίσως και ο αλγόριθμος που χρησιμοποιείται πιο πλού. Ο λόγος για αυτό είναι δύο χαρακτηριστικά που έχει και σχετίζονται με την ποικιλομορφία και την απλότητα του (μπορεί να χρησιμοποιηθεί για εργασίες ταξινόμησης και παλινδρόμησης).

4.2 Εργαλείο Weka

Το Weka είναι ένα σύστημα μηχανικής μάθησης (συλλογή αλγορίθμων μηχανικής μάθησης) το οποίο αναπτύχθηκε από το πανεπιστήμιο Waikato στη Νέα Ζηλανδία γραμμένο στη γλώσσα προγραμματισμού java καθιστώντας το κατά αυτό τον τρόπο ικανό να είναι προσβάσιμο από τον κάθε ένα ανεξάρτητα από το τι λειτουργικό σύστημα χρησιμοποιεί.

Παρέχει μια πληθώρα αλγορίθμων μηχανικής μάθησης κατηγοριοποίησης, συσταδοποίησης και συσχέτισης δεδομένων. Ουσιαστικά, εμπεριέχει τεχνικές μηχανικής μάθησης οι οποίες εφαρμόζονται στα διάφορα προβλήματα του πραγματικού κόσμου αποσκοπώντας στην επίτευξη εξόρυξης δεδομένων (13 Οδηγός WEKA).



Εικόνα 1: Περιβάλλον εκκίνησης Weka – Chooser

Στο περιβάλλον εκκίνησης προσφέρονται στον χρήστη οι εφαρμογές Explorer, Experimenter, KnowledgeFlow, Workbench και Simple CLI (13 Οδηγός WEKA).

Explorer: Η εφαρμογή αυτή είναι η πιο πολυχρησιμοποιημένη, καθώς δίνει στο χρήστη τα πιο σημαντικά εργαλεία όπως είναι η προ επεξεργασία δεδομένων, οι αλγόριθμοι μηχανικής μάθησης (Classify - κατηγοριοποίησης, Cluster - ομαδοποίησης, Associate – κανόνες συσχέτισης), επιλογή σχετικών χαρακτηριστικών και η αναπαράσταση δεδομένων μέσω της χρήσης γραφημάτων (13 Οδηγός WEKA)

Experimenter: Εφαρμογή πραγματοποίησης πειραμάτων όπου προσδίδει την δυνατότητα σύγκρισης διαφορετικών μοντέλων και παρουσιάζει τα αποτελέσματα χρησιμοποιώντας πίνακα (13 Οδηγός WEKA)

KnowledgeFlow: Παρέχει τα ίδια εργαλεία με τον Explorer μέσω διαφορετικής διεπαφής.

Στα πειράματα που ακολουθούν θα χρησιμοποιήσουμε το περιβάλλον Explorer (13 Οδηγός WEKA)

4.3 Περιγραφή του συνόλου δεδομένων Cardiovascular Disease

Τα ιατρικά δεδομένα που χρησιμοποιήθηκαν για την πειραματική εκτέλεση των αλγορίθμων αφορούν τη Καρδιαγγειακή νόσο. Το σύνολο δεδομένων περιέχει 70000 εγγραφές από δεδομένα ασθενών και αποτελείται από τα εξής 13 κελιά χαρακτηριστικών:

Χαρακτηριστικά	Περιγραφή
id	Αριθμός id για κάθε ασθενή
age	Ηλικία ασθενή σε μέρες
gender	Φύλο ασθενή (1: woman, 2: man)
height	Ύψος ασθενή cm
weight	Βάρος ασθενή kg
ap_hi	Συστολή της πίεσης αίματος – αριθμητική τιμή
ap_lo	Διαστολή της πίεσης αίματος – αριθμητική τιμή
cholesterol	Μέτρηση της χοληστερόλης (1: normal, 2: above normal, 3: well above normal)
gluc	Μέτρηση της γλυκόζης (1: normal, 2: above normal, 3: well above normal)
smoke	Εάν καπνίζει ο ασθενής ή όχι (1: yes, 0: no) binary
alco	Εάν καταναλώνει αλκοόλ ο ασθενής ή όχι (1: yes, 0: no) binary
active	Εάν είναι φυσικά δραστήριος ο ασθενής ή όχι (1: yes, 0: no) binary
cardio	Εάν ο ασθενής παρουσιάζει καρδιαγγειακές παθήσεις ή όχι (1: yes, 0: no) binary, (Είναι μεταβλητή στόχος)

Οι εγγραφές από τα δεδομένα ασθενών του συνόλου δεδομένου είναι αριθμητικές (numeric) δεδομένα σε κατηγορίες με τεκμηριωμένες πληροφορίες, αποτελέσματα ιατρικών εξετάσεων και πληροφορίες που δόθηκαν από τους ασθενείς.

4.4 Εισαγωγή δεδομένων στο Weka

Κατά την διαδικασία εισαγωγής του αρχείου που περιέχει το σύνολο δεδομένων θα πρέπει πρώτα να μετατραπεί στη κατάλληλη μορφή, δηλαδή από αρχείο .csv σε αρχείο .arff όπως παρουσιάζεται παρακάτω. Η μετατροπή μορφής του αρχείου, είναι απαραίτητο βήμα αφού το Weka αναμένει, το συγκεκριμένο format (ARFF) αρχείου.

Σχέση

@relation cardio_train

Πεδία

@attribute id numeric

@attribute age numeric

@attribute gender numeric

@attribute height numeric

@attribute weight numeric

@attribute ap_hi numeric

@attribute ap_lo numeric

@attribute cholesterol numeric

@attribute gluc numeric

@attribute smoke numeric

@attribute alco numeric

@attribute active numeric

@attribute cardio numeric

Δεδομένα

@data

0,18393,2,168,62,110,80,1,1,0,0,1,0

1,20228,1,156,85,140,90,3,1,0,0,1,1
 2,18857,1,165,64,130,70,3,1,0,0,0,1
 3,17623,2,169,82,150,100,1,1,0,0,1,1
 4,17474,1,156,56,100,60,1,1,0,0,0,0
 8,21914,1,151,67,120,80,2,2,0,0,0,0
 9,22113,1,157,93,130,80,3,1,0,0,1,0
 12,22584,2,178,95,130,90,3,3,0,0,1,1
 13,17668,1,158,71,110,70,1,1,0,0,1,0
 ...

@relation<όνομα>: Δεσμευμένη έκφραση η οποία ορίζεται στη πρώτη γραμμή του αρχείου RTFF και ονομάζει το σύνολο των δεδομένων που πρόκειται να εισαχθούν.

@attribute <όνομα χαρακτηριστικού><τύπος δεδομένων>: Δεσμευμένη έκφραση όπου προσδίδει περιγραφή του χαρακτηριστικού και του τύπου του.

@data: Δεσμευμένη έκφραση που αναφέρει τα δεδομένα που ακολουθούν, διαχωρίζοντας τα με κόμμα.

4.5 Υλοποίηση αλγορίθμων - Εξαγωγή αποτελεσμάτων και σύγκριση

Για όλους τους αλγόριθμους έχει επιλεγεί η μέθοδος δοκιμής 10-fold cross validation, η οποία παίρνει το σύνολο δεδομένων και το χωρίζει σε 10 υποσύνολα. Κάθε φορά το k - οστό υποσύνολο σχηματίζει ένα σύνολο ελέγχου, ενώ τα υπόλοιπα αποτελούν το σύνολο εκπαίδευσης.

Μέτρα αξιολόγησης:

- Accuracy αναφέρεται στο πόσο κοντά είναι το αποτέλεσμα με την πραγματικότητα και ορίζεται ως (Τεχνικές Εξόρυξης Δεδομένων 2015-2016)

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- Positive predictive value (PPV) αναφέρεται και ως Precision είναι η δυνατότητα του διαγνωστικού ελέγχου να ανιχνεύσει την πραγματικότητα και ορίζεται ως (Τεχνικές Εξόρυξης Δεδομένων 2015-2016)

$$\text{Precision} = \frac{tp}{tp + fp}$$

- Sensitivity αναφέρεται και ως Recall περιγράφει το ποσοστό των πραγματικά θετικών αποτελεσμάτων και ορίζεται ως (Τεχνικές Εξόρυξης Δεδομένων 2015-2016)

$$\text{Recall} = \frac{tp}{tp + fn}$$

- F-Measure είναι η δυνατότητα του διαγνωστικού ελέγχου να μην περιλαμβάνει τις αρνητικές κλάσεις (Τεχνικές Εξόρυξης Δεδομένων)

$$\text{F-measure} = \frac{\text{precision} * \text{recall}}{\beta \text{precision} + (1 - \beta) \text{recall}}$$

- TP Αληθώς Θετικά, τα δεδομένα που ταξινομήθηκαν ορθά ως θετικά π.χ. ανήκουν στη κλάση 1 (θετικό αποτέλεσμα). (Τεχνικές Εξόρυξης Δεδομένων 2015-2016)
- TN Αληθώς Αρνητικά, τα δεδομένα που ταξινομήθηκαν ορθά ως αρνητικά π.χ. ανήκουν στη κλάση 0 (αρνητικό αποτέλεσμα). (Τεχνικές Εξόρυξης Δεδομένων 2015-2016)
- FP Ψευδώς Θετικά, τα δεδομένα που ταξινομήθηκαν λάθος ως θετικά ενώ είναι αρνητικά π.χ. στη κλάση 1 (θετικό αποτέλεσμα) ενώ θα έπρεπε να ήταν στη κλάση 0 (αρνητικό αποτέλεσμα). (Τεχνικές Εξόρυξης Δεδομένων 2015-2016)
- FN Ψευδώς Αρνητικά, τα δεδομένα που ταξινομήθηκαν λάθος ως αρνητικά ενώ είναι θετικά π.χ. στη κλάση 0 (αρνητικό αποτέλεσμα) ενώ θα έπρεπε να ήταν στη κλάση 1 (θετικό αποτέλεσμα). (Τεχνικές Εξόρυξης Δεδομένων 2015-2016)

Η μεταβλητή στόχος (class cardio) αρχικά ήταν numeric έτσι έπρεπε να μετατραπεί σε nominal (ονομαστική) για την πραγματοποίηση των παρακάτω πειραμάτων μάθησης. Οι υπόλοιπες κλάσεις έχουν παραμείνει numeric για να μην υπάρχει απώλεια πληροφορίας.

Αλγόριθμοι	Μέτρα Αξιολόγησης			
	Accuracy	Precision	Recall	F-Measure
JRip	72.73%	0.727	0.727	0.727
IBk	56.65%	0.567	0.567	0.566
J48	72.90%	0.730	0.729	0.729
RandomForest	72.45%	0.725	0.725	0.724

Πίνακας 1.1: Αρχικά αποτελέσματα με φίλτρο nominal στην κλάση cardio

Τα παραπάνω αποτελέσματα έχουν εξαχθεί με σκοπό να χρησιμοποιηθούν στα πειράματα 1, 2 και 3 ως συγκριτικές αρχικές τιμές. Έχει εφαρμοστεί μόνο το φίλτρο nominal στη κλάση cardio (μεταβλητή στόχο) καθώς οι αλγόριθμοι JRip και J48 για να μπορούν να τρέξουν αναμένουν την κλάση που θα επιλεγεί για πρόβλεψη να έχει ονομαστικό χαρακτήρα. Στο πίνακα 1.1 όπου είναι συγκεντρωμένα όλα τα αρχικά αποτελέσματα εστιάζοντας στη στήλη Accuracy φαίνεται ότι ο αλγόριθμος J48 έχει το υψηλότερο ποσοστό ακρίβειας 72.90% και ο αλγόριθμος IBk το χαμηλότερο 56.65% ποσοστό ακρίβειας.

Πείραμα 1:

Στο πείραμα 1 επιλέχθηκαν 3 φίλτρα τα οποία είχαν εφαρμοστεί μεμονωμένα για κάθε αλγόριθμο. Πρώτα το φίλτρο Attribute Selection μετά το φίλτρο SpreadSubsample και το φίλτρο Class Balancer για κάθε ένα από τους 4 αλγορίθμους ξεχωριστά.

Αλγόριθμοι	Εφαρμογή Φίλτρων		
	Attribute Selection	SpreadSubsample	Class Balancer
JRip 72.73% 0.727 0.727 0.727	Accuracy: 72.78% Precision: 0.728 Recall: 0.728 F-Measure: 0.728	Accuracy: 72.75% Precision: 0.728 Recall: 0.728 F-Measure: 0.727	Accuracy: 72.61% Precision: 0.726 Recall: 0.726 F-Measure: 0.726
IBk 56.65% 0.567 0.567 0.566	Accuracy: 64.77% Precision: 0.649 Recall: 0.648 F-Measure: 0.647	Accuracy: 56.53% Precision: 0.565 Recall: 0.565 F-Measure: 0.565	Accuracy: 56.64% Precision: 0.567 Recall: 0.566 F-Measure: 0.566
J48 72.90% 0.730 0.729 0.729	Accuracy: 73.26% Precision: 0.735 Recall: 0.733 F-Measure: 0.732	Accuracy: 72.99% Precision: 0.731 Recall: 0.730 F-Measure: 0.730	Accuracy: 72.98% Precision: 0.731 Recall: 0.730 F-Measure: 0.730
RandomForest 72.45% 0.725 0.725 0.724	Accuracy: 65.88 % Precision: 0.659 Recall: 0.659 F-Measure: 0.659	Accuracy: 72.55% Precision: 0.726 Recall: 0.726 F-Measure: 0.725	Accuracy: 72.49% Precision: 0.726 Recall: 0.725 F-Measure: 0.725

Πίνακας 1.2: Συγκριτικά αποτελέσματα εφαρμογής φίλτρων

Παρατηρώντας τον πίνακα 1.2 δείχνει ότι κάτω από το όνομα κάθε αλγορίθμου είναι καταγεγραμμένα τα αρχικά αποτελέσματα από τον πίνακα 1.1 έτσι ώστε να μπορεί να γίνει μια ολοκληρωμένη σύγκριση μεταξύ των αρχικών αποτελεσμάτων και των αποτελεσμάτων από την εφαρμογή φίλτρων. Στην περίπτωση του Attribute Selection όπου επιλέγονται τα πιο σημαντικά attributes ο αλγόριθμος J48 είχε το μεγαλύτερο ποσοστό ακρίβειας 73.26%, ο αλγόριθμος IBk δείχνει και αυτός βελτίωση στην ακρίβεια από το 56.65% στο 64.77% όπως και ο JRip με ελάχιστη άνοδο από 72.73% στο 72.78%. Ο αλγόριθμος RandomForest έχει παρουσιάσει πτώση στη ακρίβεια από 72.45% στο 65.88% και αυτό μπορεί να συμβαίνει λόγω του ότι χάθηκε κάποια πληροφορία με την αφαίρεση των ακόλουθων χαρακτηριστικών: id, gender, height, weight, gluc, smoke και alco.

Το SpreadSubsample παράγει ένα τυχαίο υποσύνολο ενός συνόλου δεδομένων. Αυτό το φίλτρο επιτρέπει στο χρήστη να καθορίσει το μέγιστο "spread" μεταξύ της σπανιότερης και της πιο κοινής κλάσης. Όλοι οι αλγόριθμοι εκτός ο IBk έχουν μια μικρή αύξηση στη ακρίβεια τους, ο J48 από 72.90% στο 72.99%, ο JRip από 72.73% στο 72.75% και ο RandomForest από 72.45% στο 72.55%.

Το Class Balancer μοιράζει τα instances σε ίσο πλήθος για κάθε κλάση έτσι ώστε να έχουν το ίδιο συνολικό βάρος. Σκοπός της εφαρμογής του φίλτρου αυτού ήταν για να αποφευχθεί η υπερεκπαίδευση (overfitting) έτσι ώστε ένας αλγόριθμος να μπορέσει να δημιουργήσει μια πιο γενικευμένη εικόνα για το τι χαρακτηρίζεται ως 0 (αρνητικό αποτέλεσμα) ή 1 (θετικό αποτέλεσμα). Ο J48 και ο Random Forest έχουν δείξει ελάχιστη αύξηση ακρίβειας από 72.90% στο 72.98% και από 72.45% στο 72.49% αντίστοιχα. Οι υπόλοιποι αλγόριθμοι είχαν ελάχιστη μείωση στην ακρίβεια δηλαδή, ο JRip από 72.73% στο 72.61% και IBk από 56.65% στο 56.64%. Η εφαρμογή του Class Balancer δεν είχε επιφέρει σημαντική αύξηση ή μείωση στην τιμή της ακρίβειας των αλγορίθμων. Οι αλγόριθμοι JRip και IBk είχαν ελάχιστη μείωση στα αποτελέσματα τους συγκριτικά με τα αρχικά αποτελέσματα και αυτό μπορεί να οφείλεται στο ότι το φίλτρο μοίρασε τυχαία τα δεδομένα που είχαν έξοδο 0 (αρνητικό αποτέλεσμα) ή 1 (θετικό αποτέλεσμα) για να τα εξισορροπήσει, με αποτέλεσμα να χαθεί χρήσιμη πληροφορία σχετικά με τις συχνότητες εμφάνισης των features που χαρακτηρίζουν τις κλάσεις.

Ακολουθούν με λεπτομέρεια οι μετρήσεις και των τριών τεχνικών:

Για Attribute Selection:

- Filter → Unsupervised → attribute → Numeric to Nominal (attributeIndices = last) προς εφαρμογή για την κλάση (μεταβλητή στόχο) cardio.
- Filter → Supervised → attribute → Attribute Selection (CSF) με μέθοδο αναζήτησης BestFirst προς εφαρμογή.

Αλγόριθμοι	Class 0	Class 1
JRip	TP Rate: 0.743 FP Rate: 0.287	TP Rate: 0.713 FP Rate: 0.257
IBk	TP Rate: 0.688 FP Rate: 0.392	TP Rate: 0.608 FP Rate: 0.312
J48	TP Rate: 0.778 FP Rate: 0.313	TP Rate: 0.687 FP Rate: 0.222
RandomForest	TP Rate: 0.665 FP Rate: 0.348	TP Rate: 0.652 FP Rate: 0.335

Πίνακας 1.2.1: Συγκριτικά αποτελέσματα κλάσεων

Βάση του παραπάνω πίνακα το FP Rate της κλάσης 0 είναι σχετικά υψηλό για όλους τους αλγόριθμους κατά την εφαρμογή του φίλτρου Attribute Selection έχοντας το χαμηλότερο FP Rate ο αλγόριθμος JRip με 0.287. Επίσης ο αλγόριθμος J48 έχει το υψηλότερο TP Rate με 0.778 για την κλάση 0 και για την ίδια κλάση χαμηλότερο TP Rate έχει ο RandomForest με 0.665. Για την κλάση 1 μεγαλύτερο TP Rate έχει ο JRip με 0.713 και μεγαλύτερο FP Rate ο αλγόριθμος RandomForest με 0.335.

Για SpreadSubsample:

- Filter → Unsupervised → attribute → Numeric to Nominal προς εφαρμογή για την κλάση (μεταβλητή στόχο) cardio.
- Filter → Supervised → instances → SpreadSubsample

Αλγόριθμοι	Class 0	Class 1
JRip	TP Rate: 0.747 FP Rate: 0.292	TP Rate: 0.708 FP Rate: 0.253
IBk	TP Rate: 0.567 FP Rate: 0.436	TP Rate: 0.564 FP Rate: 0.433
J48	TP Rate: 0.757 FP Rate: 0.297	TP Rate: 0.703 FP Rate: 0.243
RandomForest	TP Rate: 0.753 FP Rate: 0.302	TP Rate: 0.698 FP Rate: 0.247

Πίνακας 1.2.2: Συγκριτικά αποτελέσματα κλάσεων

Σύμφωνα με τον ποιο πάνω πίνακα μέσω της εφαρμογής του SpreadSubsample το TP Rate και για τις δύο κλάσεις είναι υψηλό. Για την κλάση 0 μεγαλύτερο TP Rate έχει ο αλγόριθμος J48 με 0.757 και για την ίδια κλάση μεγαλύτερο FP Rate έχει ο IBk με 0.436. Επίσης, για την κλάση 1 υψηλότερο FP Rate έχει ο IBk με 0.433 και το μεγαλύτερο TP Rate έχει ο αλγόριθμος JRip με 0.708.

Για Class Balancer:

- Filter → Unsupervised → attribute → Numeric to Nominal προς εφαρμογή για την κλάση (μεταβλητή στόχο) cardio.
- Filter → Supervised → instances → Class Balancer.

Αλγόριθμοι	Class 0	Class 1
JRip	TP Rate: 0.746 FP Rate: 0.293	TP Rate: 0.707 FP Rate: 0.254
IBk	TP Rate: 0.569 FP Rate: 0.436	TP Rate: 0.564 FP Rate: 0.431

J48	TP Rate: 0.758 FP Rate: 0.298	TP Rate: 0.702 FP Rate: 0.242
RandomForest	TP Rate: 0.752 FP Rate: 0.302	TP Rate: 0.698 FP Rate: 0.248

Πίνακας 1.2.3: Συγκριτικά αποτελέσματα κλάσεων

Παρατηρώντας τον ποιο πάνω πίνακα υψηλότερο FP Rate για την κλάση 0 έχει ο αλγόριθμος IBk με 0.436 και το μικρότερο FP Rate για την ίδια κλάση το έχει ο JRip με 0.293. Μεγαλύτερο FP Rate για την κλάση 1 το έχει ο IBk με 0.431 και το μικρότερο FP Rate για την ίδια κλάση το έχει ο J48 με 0.242. Το ποιο υψηλό TP Rate για την κλάση 0 το έχει ο J48 με 0.758 και το μικρότερο TP Rate για την ίδια κλάση ο IBk με 0.569. Για τη κλάση 1 το μεγαλύτερο TP Rate το έχει ο JRip με 0.707 και το μικρότερο ο IBk με 0.564.

Πείραμα 2:

Αλγόριθμοι	Εφαρμογή Φίλτρων	
	Attribute Selection + Class Balancer	
JRip	72.73% 0.727 0.727 0.727	Accuracy: 72.61% Precision: 0.727 Recall: 0.726 F-Measure: 0.726
IBk	56.65% 0.567 0.567 0.566	Accuracy: 63.56% Precision: 0.636 Recall: 0.636 F-Measure: 0.635
J48	72.90% 0.730 0.729 0.729	Accuracy: 73.27% Precision: 0.735 Recall: 0.733 F-Measure: 0.732

RandomForest	
72.45%	Accuracy: 65.88%
0.725	Precision: 0.659
0.725	Recall: 0.659
0.724	F-Measure: 0.659

Πίνακας 2.1: Συγκριτικά αποτελέσματα εφαρμογής φίλτρων

Στο πείραμα 2 εφαρμόστηκαν μαζί τα φίλτρα Attribute Selection (CSF) (επιλογή χαρακτηριστικών) και Class Balancer (εξισορρόπηση κλάσεων) για κάθε ένα αλγόριθμο. Στο πίνακα 5.5 είναι καταγεγραμμένα τα αρχικά αποτελέσματα κάτω από κάθε αλγόριθμο και δίπλα είναι τα αποτελέσματα από την εφαρμογή των φίλτρων. Ο αλγόριθμος J48 έχει το μεγαλύτερο ποσοστό ακρίβειας 73.27% παρατηρώντας ότι είχε μια μικρή βελτίωση συγκριτικά με τα αρχικά αποτελέσματα από το 72.90% στο 73.27%. Επίσης, βελτίωση έχει παρουσιάσει και ο αλγόριθμος IBk από το 56.65% στο 63.56%. Οι αλγόριθμοι JRip και RandomForest είχαν μείωση του ποσοστού ακρίβειας από το 72.73% στο 72.61% και από το 72.45% στο 65.88% αντίστοιχα.

Ακολουθούν με λεπτομέρεια οι μετρήσεις:

- Filter → Unsupervised → attribute → Numeric to Nominal (attributeIndices = last) προς εφαρμογή για την κλάση (μεταβλητή στόχο) cardio.
- Filter → Supervised → attribute → Attribute Selection (CSF) με μέθοδο αναζήτησης BestFirst προς εφαρμογή.
- Filter → Supervised → instances → Class Balancer.

Αλγόριθμοι	Class 0	Class 1
JRip	TP Rate: 0.752 FP Rate: 0.300	TP Rate: 0.700 FP Rate: 0.248
IBk	TP Rate: 0.613 FP Rate: 0.341	TP Rate: 0.659 FP Rate: 0.387
J48	TP Rate: 0.778 FP Rate: 0.313	TP Rate: 0.687 FP Rate: 0.222

RandomForest	TP Rate: 0.666	TP Rate: 0.652
	FP Rate: 0.348	FP Rate: 0.334

Πίνακας 2.1.1: Συγκριτικά αποτελέσματα κλάσεων

Βάση των παραπάνω αποτελεσμάτων το μεγαλύτερο TP Rate για την κλάση 0 το έχει ο J48 με 0.778 και το μικρότερο TP Rate ο IBk με 0.613. Για τη κλάση 1 υψηλότερο TP Rate έχει ο JRip με 0.700 και χαμηλότερο ο RandomForest με 0.652. Μεγαλύτερο FP Rate για την κλάση 0 έχει ο Random-Forest με 0.348 και μικρότερο ο JRip με 0.300. Επίσης, για την κλάση 1 υψηλότερο FP Rate έχει ο IBk με 0.387 και χαμηλότερο ο J48 με 0.222.

Πείραμα 3:

Αλγόριθμοι	Εφαρμογή Φίλτρων	
	Attribute Selection + SpreadSubsample	
JRip		
72.73%	Accuracy: 72.79%	
0.727	Precision: 0.728	
0.727	Recall: 0.728	
0.727	F-Measure: 0.728	
IBk		
56.65%	Accuracy: 64.63%	
0.567	Precision: 0.658	
0.567	Recall: 0.658	
0.566	F-Measure: 0.658	
J48		
72.90%	Accuracy: 73.13%	
0.730	Precision: 0.733	
0.729	Recall: 0.731	
0.729	F-Measure: 0.731	
RandomForest		
72.45%	Accuracy: 65.81%	
0.725	Precision: 0.658	
0.725	Recall: 0.658	
0.724	F-Measure: 0.658	

Πίνακας 3.1: Συγκριτικά αποτελέσματα εφαρμογής φίλτρων

Στο πείραμα 3 εφαρμόστηκαν μαζί τα φίλτρα Attribute Selection (CSF) (επιλογή χαρακτηριστικών) και SpreadSubsample για κάθε ένα αλγόριθμο. Στο πίνακα 3.1 είναι καταγεγραμμένα τα αρχικά αποτελέσματα κάτω από κάθε αλγόριθμο και δίπλα είναι τα αποτελέσματα από την εφαρμογή των φίλτρων. Ο αλγόριθμος J48 έχει το μεγαλύτερο ποσοστό ακρίβειας 73.13% παρατηρώντας ότι είχε μια μικρή βελτίωση συγκριτικά με τα αρχικά αποτελέσματα από το 72.90% στο 73.13%. Επίσης, βελτίωση έχει παρουσιάσει ο αλγόριθμος IBk από το 56.65% στο 64.63% και ο JRip από το 72.73% στο 72.79%. Ο αλγόριθμος RandomForest είχε μείωση του ποσοστού ακρίβειας από το 72.45% στο 65.81%.

Ακολουθούν με λεπτομέρεια οι μετρήσεις:

- Filter → Unsupervised → attribute → Numeric to Nominal (attributeIndices = last) προς εφαρμογή για την κλάση (μεταβλητή στόχο) cardio.
- Filter → Supervised → attribute → Attribute Selection (CSF) με μέθοδο αναζήτησης BestFirst προς εφαρμογή.
- Filter → Supervised → instances → SpreadSubsample

Αλγόριθμοι	Class 0	Class 1
JRip	TP Rate: 0.744 FP Rate: 0.288	TP Rate: 0.712 FP Rate: 0.256
IBK	TP Rate: 0.688 FP Rate: 0.395	TP Rate: 0.605 FP Rate: 0.312
J48	TP Rate: 0.776 FP Rate: 0.313	TP Rate: 0.687 FP Rate: 0.224
RandomForest	TP Rate: 0.665 FP Rate: 0.349	TP Rate: 0.651 FP Rate: 0.335

Πίνακας 3.1.1: Συγκριτικά αποτελέσματα κλάσεων

Βάση των παραπάνω αποτελεσμάτων το μεγαλύτερο TP Rate για την κλάση 0 το έχει ο J48 με 0.776 και το μικρότερο TP Rate ο RandomForest με 0.665. Για τη κλάση 1 υψηλότερο TP Rate έχει ο JRip με 0.712 και χαμηλότερο ο IBK με 0.605. Μεγαλύτερο FP Rate για την κλάση 0 έχει ο IBK με 0.395

και μικρότερο ο JRip με 0.288. Επίσης, για την κλάση 1 υψηλότερο FP Rate έχει ο RandomForest με 0.335 και χαμηλότερο ο J48 με 0.224.

Κεφάλαιο 5: Σχεδίαση της υποκείμενης γνώσης

5.1 Ορισμοί και κίνητρα της χρήσης οντολογιών

Μία οντολογία αποτελεί την αναπαράσταση είτε των εννοιών που υπάρχουν σε μία γνωστική περιοχή είτε μίας οντότητας, προκειμένου να είναι σε θέση ο Η/Υ να την επεξεργαστεί. Η οντολογία περιλαμβάνει έναν κατάλογο όρων που είναι πεπερασμένος, αλλά και τη σχέση που υπάρχει ανάμεσα σε αυτούς τους όρους. Οι συγκεκριμένοι όροι είναι σημαντικές έννοιες με τη μεταξύ τους σχέση να έχει χαρακτήρα ιεραρχικό στα πλαίσια ομάδων και υποομάδων. Μία οντολογία ως εκ τούτου ορίζει τους όρους στα πλαίσια των οποίων περιγράφεται και αναπαριστάται μία περιοχή γνώσης και με τη βοήθεια της οντολογίας μπορούμε να προσδώσουμε στις εκφράσεις σημασία, δηλαδή περιεχόμενο που έχει νόημα. Ως εκ τούτου οι οντολογίες είναι σε θέση να ερμηνεύσουν την πληροφορία, αλλά και να επικοινωνήσουν και με άλλα εργαλεία λογικής. Μέσα από τη χρήση εργαλείων, καθίσταται δυνατή η εξαγωγή λογικών συμπερασμάτων γύρω από τις έννοιες που δημιουργούνται εντός αυτών, αλλά και των μεταξύ τους σχέσεων (Keet, 2020).

Εξαιτίας της δυναμικότητας των οντολογιών, οι χρήσεις τους μπορούν να υποδιαιρεθούν στις παρακάτω κατηγορίες:

- Επικοινωνία (communication): Οι οντολογίες επιτρέπουν την κοινή αντίληψη και διευκολύνουν την επικοινωνία ανάμεσα σε ανθρώπους που έχουν διαφορετικές απόψεις και ανάγκες, αλλά προέρχονται και διαφορετικά γνωστικά αντικείμενα.
- Διαλειτουργικότητα (interoperability): Με τις οντολογίες μπορεί να επιτευχθεί η διαλειτουργικότητα ανάμεσα σε συστήματα. Μέσα σε αυτά τα πλαίσια, καθίσταται δυνατή η χρήση των οντολογιών ως μία κοινή γλώσσα αναφοράς, η οποία συμβάλει να μεταφραστούν όλα τα διαφορετικά λεξιλόγια που πιθανόν χρησιμοποιούν τα διαφορετικά μέρη που έχει ένα σύστημα, λαμβάνοντας υπόψιν ότι είναι αρκετά δύσκολο να υιοθετηθεί ένα και μόνο λεξιλόγιο.
- Μηχανική συστημάτων (engineering systems): Η χρήση των οντολογιών μπορεί να αποφέρει οφέλη σε επίπεδο, προδιαγραφών, αξιοπιστίας και επαναχρησιμοποίησης πόρων. Η κοινή κατανόηση ενός προβλήματος μπορεί να συμβάλει να καθοριστούν οι απαιτήσεις και οι προδιαγραφές ενός συστήματος λογισμικού και να κατανοηθούν οι σχέσεις μεταξύ των συνιστωσών του συστήματος, ιδιαίτερα εάν στα πλαίσια αυτής της διαδικασίας εμπλέκονται και οι ομάδες, οι οποίες κάνουν χρήση διαφορετικής οντολογίας. Μία αναπαράσταση που είναι τυπική συμβάλει να ελεγχθεί ο έλεγχος της συνέπειας μέσα από την αύξηση της αξιοπιστίας του λογισμικού. Η

κοινή συμφωνία αποτελεί το θεμέλιο της τυπικής κωδικοποίησης των οντοτήτων, των διεργασιών, των σημαντικών οντοτήτων και του συστήματος, αλλά και των σχέσεων που αναπτύσσονται ανάμεσα τους στο πεδίο ενδιαφέροντος. Αυτή η τυπική κωδικοποίηση είναι δυνατόν να αποτελέσει ένα συστατικό που μπορεί να επαναχρησιμοποιηθεί εντός ενός συστήματος λογισμικού.

- Στο Σημασιολογικό Ιστό και στον Ιστό γενικότερα (Semantic Web): Ιδιαίτερως σημαντική είναι η ανάπτυξη ενός μηχανισμού, στα πλαίσια του οποίου θα δίνεται η δυνατότητα στις μηχανές να συγκεντρώνουν περιεχόμενο από διάφορες πηγές, να προχωρούν σε επεξεργασία των πληροφοριών, αλλά και να κάνουν ανταλλαγή των αποτελεσμάτων τους με άλλα προγράμματα και χρήστες. Η ιδέα πίσω από την ανάπτυξη Σημασιολογικού Ιστού (SemanticWeb) κινείται προς αυτή την κατεύθυνση, ενώ παράλληλα πολύ σημαντική θεωρείται και η συμβολή των οντολογιών σε αυτήν την κατεύθυνση. Μέσω των οντολογιών διευκολύνεται ο διαμοιρασμός και η επαναχρησιμοποίηση της γνώσης προκειμένου να επιτυγχάνεται η επικοινωνία ανάμεσα σε ετερογενή και διανεμημένα συστήματα εφαρμογών (Γαϊτάνου & Γεγρατσούλης, 2006). Σχετικά με την αναζήτηση στον Παγκόσμιο Ιστό, οι οντολογίες συμβάλουν στη βελτίωση της ακρίβειας των αναζητήσεων.
- Συστήματα διαχείρισης γνώσης (Knowledge management): Χρησιμοποιούνται με σκοπό να μοντελοποιήσουν τη γνώση και ως εργαλείο διερμηνείας των μεταδεδομένων που χρησιμοποιούνται από διάφορες εφαρμογές και υπηρεσίες πληροφόρησης.
- Ηλεκτρονικό εμπόριο (e-Business): Χρησιμοποιούνται από πράκτορες λογισμικού για την απόκτηση γνώσης και πληροφορίας. Συγκεκριμένα στο ηλεκτρονικό εμπόριο οι οντολογίες εκτελούν λειτουργίες όπως η σύγκριση τιμών του ίδιου προϊόντος σε διαφορετικά ηλεκτρονικά καταστήματα.
- Ψηφιακές Βιβλιοθήκες (Digital Libraries): Χρησιμοποιούνται κυρίως για καλύτερη ταξινόμηση του υλικού τους.
- Ανάκτηση Πληροφοριών (Information Retrieval): Χρησιμοποιούνται για να οδηγήσουν την αναζήτηση, έτσι ώστε το σύστημα να επιστρέψει περισσότερα σχετικά αποτελέσματα.

5.2 Αρχές σχεδίασης και ανάπτυξης οντολογιών

Ο σχεδιασμός και η ανάπτυξη μιας οντολογίας πρέπει να διέπεται από κάποια κριτήρια σχεδίασης και ένα σύνολο αρχών. Αυτές οι βασικές αρχές παρουσιάζονται παρακάτω (Slimani, 2014):

- Σαφήνεια (clarity): μια οντολογία πρέπει να εκφράζει το νόημα των όρων που καθορίζει κατά τρόπο αποτελεσματικό. Το θεματικό πεδίο που περιγράφει μια οντολογία πρέπει να αποδίδεται

με σαφήνεια και οι ορισμοί πρέπει να είναι αντικειμενικοί. Αν και το κίνητρο για να καθοριστεί μία έννοια είναι δυνατόν να προκύψει είτε από υπολογιστικές απαιτήσεις είτε από κοινωνικές καταστάσεις, είναι σημαντικό ο ορισμός να είναι ανεξάρτητος από το υπολογιστικό και κοινωνικό πλαίσιο.

- Συνοχή ή Συνέπεια (coherence): μια οντολογία είναι απαραίτητο να έχει εσωτερική συνέπεια και ως εκ τούτου οι έννοιες και τα αξιώματα είναι σημαντικό να χαρακτηρίζονται από συνέπεια και λογική συνέχεια. Μία οντολογία είναι σημαντικό να δίνει τη δυνατότητα συμπερασμών που να είναι συνεπής με τους ορισμούς. Αν βρεθεί μία πρόταση, η οποία αντικρούει κάποιο παράδειγμα ή έναν ορισμό, ο οποίος έχει δοθεί άτυπα, τότε η οντολογία είναι ασυνεπής.
- Επεκτασιμότητα (extendibility): μια οντολογία πρέπει να σχεδιάζεται κατά τρόπο που να της προσφέρεται η δυνατότητα να επεκτείνεται και να εξειδικεύεται και ως προς τα στιγμιότυπα αλλά και ως προς τις έννοιες και τις σχέσεις που εμπεριέχονται σε αυτή. Πρέπει κάποιος να μπορεί να ορίσει νέους όρους στα πλαίσια ειδικών χρήσεων, επί τη βάση του υπάρχοντος λεξικού, κατά τρόπο που να μην απαιτείται να αναθεωρηθούν οι υπάρχοντες ορισμοί. Η επεκτασιμότητα είναι απαραίτητη καθώς οι ανάγκες χρήσης της οντολογίας είναι δύσκολο να καθοριστούν από την αρχή επακριβώς.
- Ελάχιστη οντολογική δέσμευση (minimal ontological commitment): μια οντολογία είναι σημαντικό να απαιτεί την ύπαρξη της ελάχιστης οντολογικής δέσμευσης, η οποία θα είναι ικανή να υποστηρίξει τις προβλεπόμενες και επιθυμητές λειτουργίες στα πλαίσια των οποίων διαμοιράζεται η γνώση. Είναι σημαντικό να υπάρχουν ορισμένες αξιώσεις αναφορικά με το χώρο ο οποίος μοντελοποιείται και όσο δυνατόν λιγότερες αιτήσεις για τον κόσμο που διαμορφώνει, επιτρέποντας στα δεσμευμένα μέρη οντολογική ελευθερία για να εξειδικευθούν και να συγκεκριμενοποιήσουν την οντολογία όπως απαιτείται. Τα αξιώματα που εμπεριέχονται σε μια οντολογία πρέπει να είναι γενικευμένα. Πρέπει να αποφεύγεται η σαφής αναφορά στο πεδίο ενδιαφέροντος που μοντελοποιείται, έτσι ώστε να υπάρχει ελευθερία για εξειδίκευση και εφαρμογή σε συγκεκριμένες περιοχές καθώς και για επαναχρησιμοποίηση της οντολογίας σε άλλα πεδία ενδιαφέροντος.
- Ελάχιστη προκατάληψη κωδικοποίησης (minimal encoding bias): η εννοιολόγηση πρέπει να καθορίζει το επίπεδο γνώσης χωρίς να εξαρτάται από ένα συγκεκριμένο επίπεδο συμβολισμού που κωδικοποιεί. Για το λόγο αυτό, η κωδικοποίηση της οντολογίας είναι σημαντικό να γίνεται με τον πιο απλό τρόπο, ζητώντας κατά το δυνατόν πιο απλούς και μικρούς κώδικες για να αναπαρασταθεί, οι οποίοι να μην επηρεάζουν την πληροφορία. Πιο συγκεκριμένα σημαίνει ότι το μοντέλο

που επιδιώκουμε να αναπτύξουμε, στα πλαίσια της μοντελοποίησης ενός φαινομένου, είναι σημαντικό να σχετίζεται άρρηκτα από το επίπεδο της γνώσης και μόνο. Για να επιτευχθεί αυτό σημαντικό ρόλο διαδραματίζει ο διαμοιρασμός της γνώσης, καθώς όλοι όσοι εμπλέκονται σε αυτή τη διαδικασία είναι δυνατόν να υλοποιηθούν με διάφορους τρόπους.

- Τυποποίηση των ονομάτων (standardization of names): προτείνεται η χρήση ορισμένων συμβάσεων αναφορικά με όρους που έχουν σχέση μεταξύ τους, προκειμένου να κατανοηθεί ευκολότερα η οντολογία.

Τα κύρια μοντέλα που υπάρχουν για να αναπτύσσονται οι οντολογίες είναι αυτό της από πάνω προς τα κάτω ανάπτυξης και της από κάτω από προς τα πάνω ανάπτυξης. Και τα δύο μοντέλα έχουν τόσο πλεονεκτήματα όσο και μειονεκτήματα. Η επιλογή ενός από τα δύο από ένα μηχανικό οντολογιών σχετίζεται κατά κύριο λόγο από την εφαρμογή, την οποία αφορούν οι αναπτυσσόμενες οντολογίες.

Με βάση το μοντέλο ανάπτυξης από πάνω προς τα κάτω, στην αρχή γίνεται η ανάπτυξη μίας οντολογίας που περιγράφει το θεματικό πεδίο στα πλαίσια του οποίου εντάσσεται η εφαρμογή, κάνοντας χρήση θεμελιωδών εννοιών του πεδίου. Ακολούθως συνεχίζουμε με την κύρια θεματική οντολογία προκειμένου να αποτυπωθούν συγκεκριμένες έννοιες, έως τη στιγμή που η οντολογία θα είναι σε θέση να παράσχει μία ολοκληρωμένη περιγραφή του συνόλου των δεδομένων, τα οποία οφείλει να καλύπτει. Για μία τέτοια προσέγγιση οι μηχανικοί και οι ειδικοί πρέπει να προχωρήσουν στο χαρακτηρισμό του θεματικού πεδίου και στον προσδιορισμό των βασικών εννοιών και συσχετίσεων. Ωστόσο, στην πράξη είναι αρκετά δύσκολο τις πιο πολλές φορές να υπάρξει μία τέτοιου είδους συμφωνία ανάμεσα στους ειδικούς. Όταν όμως ξεπεραστεί η εν λόγω δυσκολία, η οντολογία που παράγεται είναι ένα θεματικό μοντέλο πλούσιο σημασιολογικά, το οποίο μπορεί να παράσχει πλήρη και σαφή αναπαράσταση ενός θεματικού πεδίου, το οποίο στη συνέχεια θα μπορέσει να χρησιμοποιηθεί προκειμένου να συσχετιστούν έννοιες, οι οποίες εντάσσονται εντός διαφορετικών θεματικών περιοχών.

Στα πλαίσια της παραγωγής οντολογιών, όπου ακολουθείται το μοντέλο ανάπτυξης από πάνω προς τα κάτω, είναι χρήσιμο να γίνεται εκ νέου χρήση οντολογιών που έχουν κατασκευάσει άλλοι. Οι οντολογίες αυτές ονομάζονται θεμελιώδεις. Κάνοντας χρήση οντολογιών που υπάρχουν ήδη, επιτυγχάνεται, εκτός της επαναχρησιμοποίησης πόρων, και η κατασκευή ενός οντολογικού μοντέλου πιο συνεπούς και πλήρους. Κάνοντας χρήση των εμπειριών και της δουλειάς που έχουν κάνει άλλοι, οι μηχανικοί μπορούν να επεκτείνουν τις έννοιες που βρίσκονται σε θεμελιώδεις οντολογίες προκειμένου να καλύπτονται οι ανάγκες των αναπτυσσόμενων εφαρμογών. Τέλος μέσα από τη

χρήση θεμελιωδών οντολογιών μπορεί να επιτευχθεί η διασύνδεση οντολογιών διαφορετικών θεματικών πεδίων.

Μία ουσιαστική πρόκληση που τίθεται από το μοντέλο ανάπτυξης από πάνω προς τα κάτω αφορά στην επέκταση της θεματικής οντολογίας, προκειμένου οι έννοιες που είναι χαμηλού επιπέδου να αναπαριστούν και να περιγράφουν τα δεδομένα της θεματικής περιοχής με ακρίβεια, ενώ παράλληλα διατηρούν τη συνέπεια που χαρακτηρίζει το μοντέλο. Μέσα σε αυτά τα πλαίσια είναι σημαντικό κατά την ανάπτυξη της θεματικής οντολογίας να δίνεται η δέουσα σημασία στα δεδομένα. Ως εκ τούτου είναι σημαντική, για τη διαδικασία της ανάπτυξης, η εμπλοκή ατόμων που είναι γνώστες της θεματικής περιοχής.

Το μοντέλο ανάπτυξης από πάνω προς τα κάτω μπορεί να λειτουργήσει πιο καλά εάν ο σκοπός του είναι η αναπαράσταση μιας θεματικής περιοχής, αλλά και η χρήση αυτής της αναπαράστασης ως πεδίο σύγκλισης των απόψεων που έχουν οι ειδικοί.

Ένα εναλλακτικός τρόπος για να προσεγγίσουμε την ανάπτυξη οντολογιών είναι μέσα από τη χρήση δεδομένων ως αφετηρία, αναπτύσσοντας για αρχή οντολογίες που εστιάζονται στην περιγραφή πηγών δεδομένων που είναι συγκεκριμένες, και εν συνεχεία η επέκταση και η διασύνδεση των οντολογιών αυτών θα οδηγήσει σε μεγαλύτερα επίπεδα αφαίρεσης. Η από κάτω προς τα πάνω προσέγγιση μπορεί να διασφαλίσει το γεγονός πως θα υπάρξει περιγραφή των δεδομένων με ακρίβεια, αφού η οντολογία στήνεται επί αυτών. Στην περίπτωση που τα δεδομένα μπορούν να παράσχουν μία καλά αναπαράσταση του θεματικού πεδίου, η εν λόγω προσέγγιση μπορεί να αναπτύξει ενός πλήρες και ακριβές θεματικό μοντέλο. Λαμβάνοντας ωστόσο υπόψιν πως οι λεπτομέρειες της εφαρμογής και οι πιθανές ασυνέπειες μπορούν να γίνουν αντιληπτές μόνο μέσα στα πλαίσια μίας υψηλού επιπέδου αφαίρεσης θεματικής οντολογίας, υπάρχει η περίπτωση της μη παροχής πλήρους και ακριβούς περιγραφής του θεματικού πεδίου, αλλά και των αναπτυσσόμενων εντός αυτού σχέσεων, καθώς εστιάζει στις ιδιαιτερότητες που έχουν τα δεδομένα.

Αφού δημιουργηθούν οι οντολογίες χαμηλού επιπέδου αφαίρεσης, οι οποίες έχουν τη βάση τους στα δεδομένα και ως εκ τούτου σε πιο αφηρημένα μοντέλα, συχνά φαίνεται πως οι οντολογίες που έχουν κατασκευάσει άλλοι είναι αρκετά χρήσιμες. Οι οντολογίες αυτές είναι γνωστές και ως βοηθητικές, στα πλαίσια των οποίων αναπαρίστανται κοινές έννοιες, οι οποίες βρίσκονται σε διαφορετικά θεματικά πεδία και είναι δυνατόν να γίνει χρήση τους ως είναι στα πλαίσια επέκτασης ή αναπαράσταση αντίστοιχων εννοιών από το θεματικό πεδίο.

Υπάρχουν διάφορες μεθοδολογίες ανάπτυξης οντολογιών. Σύμφωνα με τους Noy & McGuinness (2001), η μεθοδολογία που ακολουθείται συνήθως για την ανάπτυξη μιας οντολογίας, αποτελείται από τα παρακάτω βήματα:

- Καθορισμός του τομέα και του πεδίου της οντολογίας.
- Επαναχρησιμοποίηση υπαρχουσών οντολογιών.
- Απαρίθμηση των σημαντικών όρων της οντολογίας.
- Ορισμός και ιεράρχηση κλάσεων.
- Ορισμός των κλάσεων και ιδιοτήτων (slots).
- Ορισμός των απόψεων (facets) των ιδιοτήτων.
- Δημιουργία οντοτήτων.

Ωστόσο μέχρι σήμερα, πολλές ερευνητικές ομάδες έχουν προτείνει διάφορες μεθοδολογίες και περιβάλλοντα ανάπτυξης οντολογιών. Παρατηρώντας διάφορες μεθοδολογίες συμπεραίνουμε ότι, παρόλο που έχουν κοινά σημεία, δεν υπάρχει κάποια που να ταιριάζει και να ανταποκρίνεται σε όλες τις περιπτώσεις. Επειδή κάθε περίπτωση είναι διαφορετική, η επιλογή της κατάλληλης μεθοδολογίας εξαρτάται από διάφορους παράγοντες (π.χ. ο αριθμός των εμπλεκόμενων στην ανάπτυξη της οντολογίας, το είδος της οντολογίας, το πεδίο ενδιαφέροντος, η τελική εφαρμογή που θα χρησιμοποιήσει την οντολογία, η εμπειρία, το επίπεδο της γνώσης των εμπλεκόμενων, η διαθεσιμότητα των απαραίτητων πόρων).

Παρακάτω παρουσιάζονται οι σημαντικότερες μεθοδολογίες ανάπτυξης οντολογιών (Corcho et al., 2002):

- Μεθοδολογίες που σχεδιάζουν μια οντολογία «εκ του μηδενός» (from scratch).
- Μεθοδολογίες που σχεδιάζουν μια οντολογία χρησιμοποιώντας υπάρχουσες οντολογίες (reengineering ontologies).
- Μεθοδολογίες Συνεργατικής Ανάπτυξης Οντολογιών (cooperative ontologies).
- Μεθοδολογίες Συγχώνευσης Οντολογιών (merging ontologies)

5.3 Ανοιχτά ζητήματα στην ανάπτυξη οντολογιών

Οι γλώσσες του Σημασιολογικού Ιστού, και ειδικά η OWL, έχουν μια απότομη καμπύλη μάθησης (Warren et al., 2014; Vigo et al., 2014) και απαιτούν αλλαγή προοπτικής, ειδικά για άτομα που προέρχονται από μηχανική λογισμικού, αντικειμενοστραφή προγραμματισμό ή σχεσιακά υπόβαθρα βάσης δεδομένων. Οι νεοεισερχόμενοι αντιμετωπίζουν όχι μόνο το τρομακτικό καθήκον της δημιουργίας ενός νέου τύπου μοντέλου για το πρόβλημα ή τον τομέα τους, αλλά επίσης προσπαθούν να βρουν το σωστό εργαλείο - οντολογικός επεξεργαστής, οπτικοποίηση, αιτιολογητής - και διαδικασία ροής εργασίας / ανάπτυξης, ενώ οι πόροι για τη λήψη μίας τεκμηριωμένης απόφασης είναι σπάνια.

Μία βασική πτυχή της οντολογικής μηχανικής είναι η επαναχρησιμοποίηση, τόσο των οντολογιών όσο και των τμημάτων των οντολογιών. Η απόδειξη αυτής της πρόκλησης είναι οι αμέτρητες δημοσιεύσεις στις λίστες αλληλογραφίας του Σηματολογικού Ιστού από ενδιαφερόμενους χρήστες που προσπαθούν να βρουν μια οντολογία για έναν συγκεκριμένο τομέα ή εργασία. Η επαναχρησιμοποίηση τμημάτων οντολογικών μεμονωμένων οντοτήτων ή σετ αξιωμάτων είναι συχνά απαραίτητη κατά τη διάρκεια της οντολογικής ανάπτυξης, ωστόσο η εφαρμογή της επαναχρησιμοποίησης στην πράξη είναι δύσκολη. Αρκετές δημοσιευμένες μελέτες έχουν δείξει ότι το επίπεδο της οντολογικής επαναχρησιμοποίησης είναι χαμηλό (Ghazvinian et al., 2011; Poveda Villalon et al., 2012; Kamdar et al., 2017).

5.4 Διαχείριση ιατρικών δεδομένων μέσω οντολογιών

Στα πλαίσια των μονάδων υγείας αλλά και της καθημερινής εξέλιξης της έρευνας της βιοϊατρικής δημιουργούνται μεγάλες ποσότητες δεδομένων διαφορετικής φύσης και είδους. Η κατασκευή οντολογιών στο πεδίο εφαρμογής σε διάφορους τομείς που έχουν στο επίκεντρο τους περιορισμούς και ταξινόμηση έχει αποδειχθεί πως είναι αρκετά χρήσιμη για τη βιοϊατρική. Αυτό συμβαίνει για τους παρακάτω λόγους:

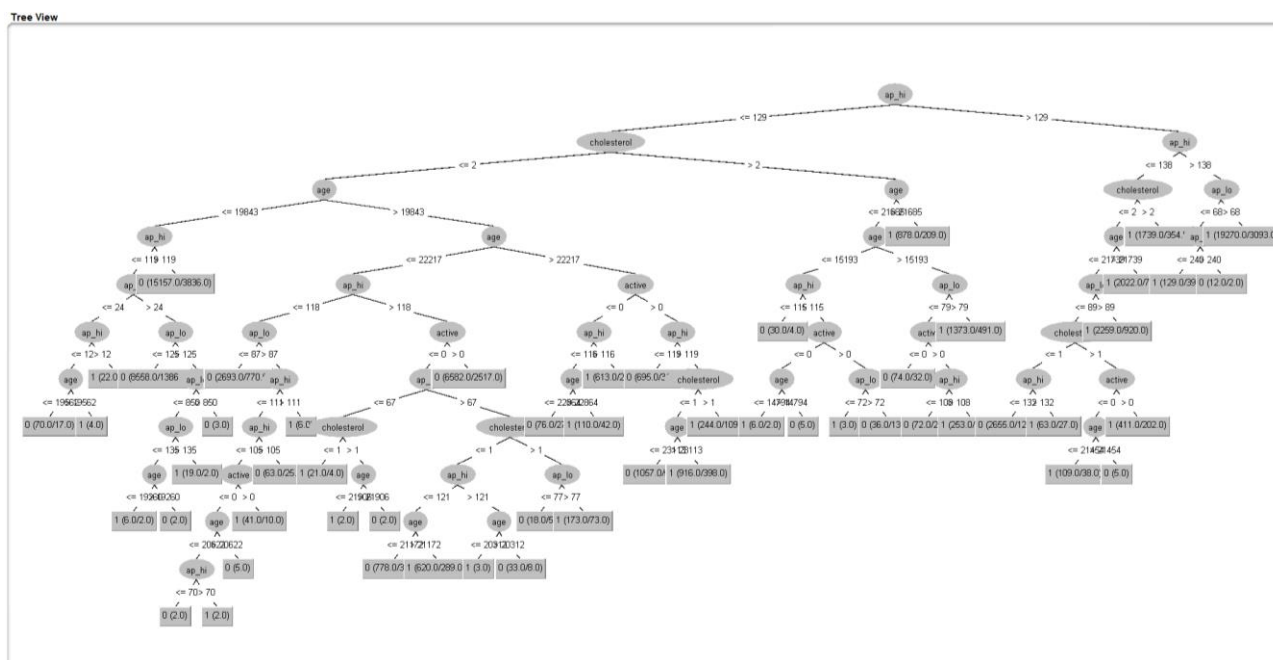
- Για να υπάρχει καλύτερη διαχείριση της σχετικής με την ιατρική ορολογία. Σε γενικές γραμμές τα παραδοσιακά συστήματα ορολογίας θεωρείται πως δεν είναι σε θέση να αντιμετωπίσουν τις απαιτήσεις που έχουν τα πληροφοριακά συστήματα της ιατρικής, τα οποία βασίζονται στη σαφή επικοινωνία ανάμεσα στις πολύπλοκες ιατρικές και βιολογικές πληροφορίες σε μορφή που να είναι επεξεργάσιμη και κατανοητή από τους Η/Υ (Chen et al., 2007). Οι οντολογίες είναι σε θέση να χρησιμοποιηθούν προκειμένου να επιτευχθεί αυτός ο στόχος, κατά κύριο λόγο επειδή δεν περιγράφουν απλά ιατρικούς όρους, αλλά αναπαριστούν και αναλύουν τη δομή που έχουν οι έννοιες που σχετίζονται με τους ιατρικούς όρους (Chen et al., 2007). Ως εκ τούτου, υιοθετώντας μία οντολογική προσέγγιση για να διαχειριστούμε την ιατρική ορολογία, διευκολύνονται ορισμένοι από τους στόχους που σχετίζονται με τις δραστηριότητες στους τομείς της βιολογίας (Smith et al., 2005) και της κλινικής ιατρικής (Rector, 2003).
- Διανομή, διαλειτουργικότητα και ενοποίηση δεδομένων ιατρικής φύσης. Η διανομή δεδομένων ιατρικής φύσης, αλλά και η διαλειτουργικότητα ανάμεσα στα διαφορετικά πληροφοριακά συστήματα είναι από ένα από τα πιο σημαντικά ζητήματα στο πεδίο της ιατρικής πληροφορικής. Η ύπαρξη μίας κοινής σηματολογίας, όπως το Μοντέλο Αναφοράς Πληροφοριών (HL7 Reference Information Model), το οποίο έχει ως στόχο την αναπαράσταση των σηματολογικών και λεξικο-

λογικών συνδέσεων ανάμεσα στις πληροφορίες που ανταλλάσσονται, είναι ένα καλό παράδειγμα για τη χρήση των οντολογιών σε αυτό το πλαίσιο. Ωστόσο το πρόβλημα της σαφούς αναπαράστασης των πληροφοριών και των αντικειμένων που ανταλλάσσονται παραμένει ακόμα ανοιχτό (Vizenor et al., 2004). Η υλοποίηση μίας ενοποίησης όπως η παραπάνω περιέχει την υποστήριξη μία σειράς από συστήματα στήριξης αποφάσεων, τα οποία είναι σε θέση να κάνουν συλλογισμούς και να προχωρήσουν στην εξαγωγή συμπερασμάτων σύμφωνα με τις ιατρικές πληροφορίες που ανταλλάσσονται.

- Συστήματα υποστήριξης απόφασης και εκ νέου χρήση της γνώσης. Τα συστήματα που βασίζονται στη γνώση και τα οποία υποστηρίζουν διάφορες εφαρμογές αποφάσεων στον τομέα της ιατρικής περίθαλψης, έχουν τη βάση τους στο μεγάλο όγκο που χαρακτηρίζει την υπάρχουσα γνώση της ιατρικής. Προκειμένου να γίνει κατανοητή η γνώση των οντολογιών, οι τελευταίες συμβάλουν στο πεδίο της αναπαράστασης, αλλά και της εκ νέου χρήσης από εφαρμογές που είναι ανεξάρτητες. Αυτό έχει ως αποτέλεσμα να μην απαιτείται πρόσθετο κόστος και προσπάθεια για να αναπτυχθούν νέα συστήματα.

5.5 Οντολογίες ιατρικών δεδομένων

Στο παρόν σημείο ακολουθεί η μετατροπή του δέντρου J48 όπου εξάχθηκε μέσω του εργαλείου Weka σε οντολογία. Το δέντρο του J48 έχει παρθεί από το πείραμα 3 το οποίο συνδυάζει τα φίλτρα Attribute Selection + SpreadSubsample.



Εικόνα 2: Οπτικοποίηση του δέντρου J48

Κεφάλαιο 6: Συμπεράσματα και μελλοντικές κατευθύνσεις

6.1 Συμπεράσματα

Συγκρίνοντας τα 3 πειράματα που πραγματοποιήθηκαν συμπεραίνουμε ότι:

Στο πείραμα 2 και στο πείραμα 3 σχετικά με τις μετρήσεις των TP Rate και FP Rate των κλάσεων οι τιμές των αλγόριθμων ήταν σχεδόν οι ίδιες με την μεμονωμένη εφαρμογή του φίλτρου Attribute Selection που έγινε στο πείραμα 1. Τα ποσοστά της ακρίβειας κάθε αλγορίθμου για την μεμονωμένη εφαρμογή του φίλτρου Attribute Selection που έγινε στο πείραμα 1 κυμαίνονται στα ίδια ποσοστά ακρίβειας που εξάχθηκαν στα πείραμα 2 και 3.

Το πείραμα 3 έχει σχεδόν τα ίδια αποτελέσματα (accuracy, precision, recall, f-measure) με το πείραμα 2 οπότε το συμπέρασμα που εξάγεται είναι ότι οι συνδυασμοί των φίλτρων Attribute Selection + Class Balancer για πείραμα 2 και Attribute Selection + SpreadSubsample για πείραμα 3 δεν έχουν επιφέρει ιδιαίτερη αύξηση ή μείωση στα ποσοστά της ακρίβειας των αλγορίθμων, έχουν παρόμοια επίδραση στο σετ δεδομένων.

Η μείωση των χαρακτηριστικών επιφέρει πολλές φορές καλύτερα αποτελέσματα λόγω του ότι μπορούν να αγνοηθούν τα περιττά ή άσχετα με την τιμή εξόδου χαρακτηριστικά. Επίσης, διακρίνοντας ποια χαρακτηριστικά είναι πιο σημαντικά για το αποτέλεσμα μιας διαδικασίας, μπορεί να αποκτηθεί διαίσθηση για το πραγματικό πρόβλημα, επιτρέποντας στους ειδικούς του τομέα να το αντιμετωπίσουν αποτελεσματικότερα. Αυτή η πτυχή της επιλογής χαρακτηριστικών είναι ιδιαιτέρως σημαντική για προβλήματα βιοπληροφορικής. Τέλος, χρησιμοποιώντας λιγότερα χαρακτηριστικά, η εκπαίδευση ταξινομητών είναι γρηγορότερη ενώ μειώνονται και οι απαιτήσεις σε αποθηκευτικό χώρο.

Κατά την εφαρμογή του φίλτρου Class Balancer έχει παρατηρηθεί ότι με την εξισορρόπηση κλάσεων για τους περισσότερους αλγόριθμους χάνεται χρήσιμη πληροφορία σχετικά με τις συχνότητες εμφάνισης. Καλύτερο θα ήταν λοιπόν, να γίνεται εμπλουτισμός του σετ δεδομένων με πραγματικά δεδομένα, από ειδικούς έτσι ώστε να έχουμε όσο το δυνατόν λιγότερη απόκλιση μεταξύ κλάσεων. Η παρέμβαση στα δεδομένα αλλοιώνει τον «χαρακτήρα» τους με αποτέλεσμα να μπερδεύονται οι αλγόριθμοι.

6.2 Μελλοντικές κατευθύνσεις

Η παρούσα μελέτη θα μπορούσε να θεωρηθεί ως το πρώτο βήμα μίας προσπάθειας να προσεγγίσουμε το βέλτιστο μέτρο απόστασης σε ένα πρόβλημα που σχετίζεται με την αναζήτηση ομοιοτήτων σε ιατρικά δεδομένα υψηλών διαστάσεων. Επίσης αποτελεί μία πρώτη προσπάθεια να εξαχθεί κρυμμένη πληροφορία μέσα από την ανακάλυψη συχνών ακολουθιακών προτύπων.

Παρόλο που τα αποτελέσματα και τα συμπεράσματα που προέκυψαν από την παρούσα έρευνα είναι πολύ σημαντικά, για να είναι σε θέση να ισχυροποιηθούν, είναι απαραίτητο σε μελλοντικές έρευνες να υπάρξει εμπλουτισμός των αρχείων δεδομένων με πιο πολλές εγγραφές, καθώς επίσης και με επιπλέον πεδία, παρέχοντας έτσι τη δυνατότητα να οδηγηθούμε και σε επιπλέον μοντέλα, συμπεράσματα, αλλά και κανόνες που θα είναι πιο ισχυροί και ποιοτικοί. Αυτό θα έχει ως αποτέλεσμα να βοηθηθεί η ιατρική επιστήμη στην προσπάθεια της για βελτίωση, αλλά και διασφάλιση της ποιότητας ζωής των ασθενών.

Βιβλιογραφία

Γαϊτάνου, Π. & Γεργατσούλης, Μ. (2006). *Διαχείριση οντολογιών: μελέτη και εμβάθυνση στα βασικά προβλήματα που την αφορούν και παρουσίαση υπαρχόντων βιβλιοθηκών οντολογιών*. 15ο Πανελλήνιο Συνέδριο Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Πάτρα, Ελλάδα, 1-3 Νοεμβρίου.

Chen, M. et al. (1996). Data Mining: An Overview from Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8, 866-883.

Chen, Y. et al. (2007). Analysis of a Study of the Users, Uses, and Future Agenda of the UMLS. *J Am Med Inform Assoc*, 14, 221-231.

Corcho, O. et al. (2002). Methodologies, tools and languages for building ontologies. Where is the meeting point?. *Data & Knowledge Engineering*, 46, 41-64.

De Rosa, R. et al. (2018). High on-treatment platelet reactivity and outcome in elderly with non ST-segment elevation acute coronary syndrome - Insight from the GEPRESS study. *Int J Cardiol.*, 259, 20-25.

Durbin, C.G. (2004). Effective use of tables and figures in abstracts, presentations, and papers. *Respir Care*, 49, 1233-1237.

Elfekey, M. et al. (2002). TAILOR: A Record Linkage Toolbox. *IEEE International Conference on Data Engineering*.

Franzblau, L.E. & Chung, K.C. (2012). Graphs, tables, and figures in scientific publications: the good, the bad, and how not to be the latter. *J Hand Surg Am.*, 37, 591-596.

Galasso, G. et al. (2015). Effect of an invasive strategy on outcome in patients $\geq$ 75 years of age with non-ST-elevation acute coronary syndrome. *Am J Cardiol.*, 115, 576-580.

Ghazvinian, A. et al. (2011). How orthogonal are the OBO Foundry ontologies?. *J. Biomedical Semantics*, 2.

Han, J. & Kamber, M. (2006). *Data Mining Concepts and Techniques*. London: Morgan Kauffman publishers.

Han, J. et al. (2012). *Data Mining: Concepts and Techniques*. New York: Elsevier.

Kamdar, M.R. et al. (2017). A systematic analysis of term reuse and term overlap across biomedical ontologies. *Semantic Web*, 8, 853-871.

Keet, M. (2020). *An Introduction to Ontology Engineering*. Cape Town: University of Cape Town.

Kumar, S. & Chong, I. (2018). Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States. *International Journal of Environmental Research and Public Health*, 15.

Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. London: John Willey & Sons, Inc.

Ławrynowicz, A. & Tresp, V. (2014). Introducing Machine Learning. In J. Lehmann & J. Voelker (eds.). *Perspectives on Ontology Learning* (pp.35-50) Berlin: AKA Heidelberg /IOS Press.

Lozano, R. et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, 380, 2095-2128.

McDonald, J.C. (2001). Charts, graphs and tables – Reporting the data. *Radiat Prot Dosimetry*, 95, 291-293.

Mitchell, T.M. (2006). *The Discipline of Machine Learning*, Machine Learning Department technical report CMU-ML-06-108. Pittsburgh, PA: Carnegie Mellon University.

Mishra, P. (2018). Scales of Measurement and Presentation of Statistical Data. *Annals of Cardiac Anaesthesia*, 21, 419-422.

Mohri, M. et al. (2012). *Foundations of Machine Learning*. Cambridge: The MIT Press.

PrexSpan (2001). PrexSpan: Mining Sequential Patterns Efficiently by Prex-Projected Pattern Growth. In *Proceedings of the 17th International Conference on Data Engineering. ICDE '01*. Washington, DC, USA: IEEE Computer Society.

Noy, N.F. & McGuinness, D.L. (2001). *Ontology Development 101: a guide to creating your first ontology*. Διαθέσιμο από: <http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf>.

Peleg, M. et al. (2003). Comparing Computer - interpretable Guideline Models: A Case-study Approach. *J Am Med Inform Assoc*, 10, 52-68.

Pipino, L. et al. (2002). Data Quality Assessment. *Communications of the ACM*, 45.

Porkizhi, M. (2017). A Study of Data Mining Techniques And Its Applications. *IJSART*, 3, 1402-1406.

Poveda Villalon, M. et al. (2012). The landscape of ontology reuse in linked data. In *Proceedings of Ontology Engineering in a Data-driven World (OEDW 2012), Galway, Ireland, Informatica*.

Praharaj, S.K. (2017). Illustrate well to get noticed: Graphs and figures in research papers. *Kerala J Psychiatry*, 30, 115-119.

Ramakrishnan, S. & Rakesh A. (1996). Mining sequential patterns: Generalizations and performance improvements. In P. Apers et al. (eds.). *Advances in Database Technology. EDBT '96: 5th International Conference on Extending Database Technology Avignon, France, March 25-29, 1996 Proceedings*. Berlin: Springer Berlin Heidelberg.

Raval, K.M. (2012). Data Mining Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2, 439-442.

Rector, A. (2003). Medical informatics. In *The description logic handbook: theory, implementation, and applications* (406-426). Cambridge: Cambridge University Press.

Sanders, J.D. (2016). Defining Terms: Data, Information and Knowledge. *SAI Computing Conference*, 1-6.

Scannapieco, M. et al. (2005). Data Quality at a Glance. *Datenbank-Spektrum*, 14, 6-14.

Silverio, A. et al. (2019). Big Health Data and Cardiovascular Diseases: A Challenge for Research, an Opportunity for Clinical Care. *Frontiers in medicine*, 6.

Slimani, T. (2014). A Study Investigating Typical Concepts and Guidelines for Ontology Building. *Journal of Emerging Trends in Computing and Information Sciences*, 5, 886-893.

Slutsky D. J. (2014). The effective use of graphs. *Journal of wrist surgery*, 3, 67-68.

Smith, B. et al. (2005). Relations in biomedical ontologies. *Genome Biology*, 6.

Wang, P.Y. & Madnick, S.E. (1989). The Inter-Database Instance Identification Problem in Integrating Autonomous Systems. *Proceedings of the Fifth International Conference on Data Engineering*, 46–55.

Warren, P. et al. (2014). The Usability of Description Logics - Understanding the Cognitive Difficulties Presented by Description Logics, in: *The Semantic Web: Trends and Challenges - 11th International*

Conference, ESWC 2014, Anissaras, Crete, Greece, May 25- 29, 2014. Proceedings. *Lecture Notes in Computer Science*, 8465, 550–564.

Vigo, M. et al. (2014). Overcoming the pitfalls of ontology authoring: Strategies and implications for tool design. *Int. J. Hum.-Comput. Stud.*, 72, 835-845.

Vizenor, L. et al. (2004). Foundation for the electronic health record: an ontological analysis of the HL7's reference information model. *IFOMIS Reports*, 1-14.

Zaki, M.J. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, 42, 31-60.

Zins, C. (2007). Conceptions of information science. *Journal of the American Society for Information Science and Technology*, 58, 335-350.

13 Οδηγός WEKA, 313-314.

Διαθέσιμο από: https://repository.kallipos.gr/bitstream/11419/1239/2/Kef._13.pdf

Pang-Ning Tan, Michael Steinbach, Vipin Kumar (2017, 2010). Εισαγωγή στην εξόρυξη δεδομένων, 249-250.

Τεχνικές Εξόρυξης Δεδομένων, 78-79, 85-86, .

Διαθέσιμο από: <https://docplayer.gr/49055074-Tehnikes-exoryxis-dedomenon.html>