ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«**ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ**»

ε.δε.μ^2

Βελτίωση της Διαλειτουργικότητας των Δομημένων ή Ημι-Δομημένων Δεδομένων με χρήση Τεχνικών Μηχανικής Μάθησης και Μοντέλων Γλώσσας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ευθυμίου Κ. Χονδρογιάννη

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Αθανάσιος Βουλόδημος, Επίκουρος Καθηγητής ΕΜΠ

ΟΚΤΩΒΡΙΟΣ 2023

Η σελίδα αυτή είναι σκόπιμα λευκή

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«**ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ**»

ε.δε.μ^2

Βελτίωση της Διαλειτουργικότητας των Δομημένων ή Ημι-Δομημένων Δεδομένων με χρήση Τεχνικών Μηχανικής Μάθησης και Μοντέλων Γλώσσας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευθυμίου Κ. Χονδρογιάννη

Εξεταστική Επιτροπή: Αθανάσιος Βουλόδημος

Ανδρέας-Γεώργιος Σταφυλοπάτης

Θεοδώρα Βαρβαρίγου

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Οκτωβρίου 2023.

Αθήνα, Οκτώβριος 2023

.....	.....	.....
Αθανάσιος Βουλόδημος	Ανδρέας-Γεώργιος Σταφυλοπάτης	Θεοδώρα Βαρβαρίγου
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγήτρια Ε.Μ.Π.

.....

Ευθύμιος Κ. Χονδρογιάννης

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Πρόλογος

Στα πλαίσια της Διπλωματικής μου εργασίας για το Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών (ΔΠΜΣ) «Επιστήμη Δεδομένων και Μηχανική Μάθηση» ασχολήθηκα με τη μελέτη των τεχνικών μηχανικής μάθησης (συμπεριλαμβανομένων των κλασικών τεχνικών μηχανικής μάθησης και των βαθιών νευρωνικών δικτύων) και πως μπορούν αυτές να συμβάλουν στη βελτίωση της διαλειτουργικότητας των δεδομένων των οργανισμών. Για τον σκοπό αυτό, αναπτύχθηκε ένα σύστημα που διευκολύνει την εναρμόνιση των δεδομένων που προέρχονται από διαφορετικούς οργανισμούς, μέσω της χρήσης τεχνικών μηχανικής μάθησης και προ-εκπαιδευμένων μοντέλων γλώσσας.

Η εργασία που παρουσιάζεται στις επόμενες σελίδες εκπονήθηκε κατά την διάρκεια του έτους 2023 υπό την επίβλεψη του κ. Αθανάσιου Βουλόδημου, τον οποίο και θα ήθελα να ευχαριστήσω, για την υποστήριξη που μου παρείχε στην προσπάθεια αυτή.

Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου, για τη στήριξή τους όλα αυτά τα χρόνια.

Ευθύμιος Κ. Χονδρογιάννης

Οκτώβριος 2023

Η σελίδα αυτή είναι σκόπιμα λευκή

Πίνακας Περιεχομένων

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

Περίληψη

Τα δεδομένα που συλλέγονται από διάφορους οργανισμούς, που δραστηριοποιούνται σε ένα συγκεκριμένο πεδίο γνώσης, μπορούν να βοηθήσουν στη λήψη σημαντικών αποφάσεων. Ειδικά στον χώρο της κλινικής έρευνας, η από κοινού εξέταση των δεδομένων που προέρχονται από διαφορετικά κέντρα έρευνας και αφορούν μία συγκεκριμένη διαταραχή μπορεί να συμβάλει στην καλύτερη μελέτη του φαινομένου και την εξαγωγή αξιόπιστων αποτελεσμάτων. Ωστόσο, οι διαφορές που υπάρχουν στον *τρόπο αναπαράστασης των δεδομένων* αυτών αποτελεί συχνά εμπόδιο στην περαιτέρω επεξεργασία και χρήση τους. Στα πλαίσια της εργασίας αυτής μελετήθηκαν διάφορες *τεχνικές μηχανικής μάθησης* καθώς και ο τρόπος που μπορούν αυτές να συμβάλουν στην επίλυση του παραπάνω προβλήματος. Επιπρόσθετα, αναπτύχθηκαν εργαλεία και μηχανισμοί που συμβάλλουν στη *βελτίωση της δια-λειτουργικότητας των δεδομένων* μέσω της χρήσης των τεχνικών αυτών. Τα εργαλεία και οι μηχανισμοί που αναπτύχθηκαν, χρησιμοποιήθηκαν για την εναρμόνιση *πραγματικών δεδομένων* που προέρχονται από διαφορετικά κέντρα έρευνας και αφορούν ασθενείς που πάσχουν από ένα συγκεκριμένο σύνδρομο. Τα αποτελέσματα της εργασίας αυτής έδειξαν ότι τα εργαλεία που αναπτύχθηκαν και ειδικότερα οι μηχανισμοί μηχανικής μάθησης που χρησιμοποιήθηκαν μπορούν να συμβάλουν σημαντικά στην εναρμόνιση των δεδομένων αυτών.

Λέξεις Κλειδιά

Δεδομένα, Εναρμόνιση Δεδομένων, Μοντέλα Γλώσσας, Τεχνικές Μηχανικής Μάθησης, Διαλειτουργικότητα Συστημάτων

Abstract

Data collected by different organizations operatijng in a particular field can help them to take important decisions. Especially in the healthcare and clinical research domains, the collaborative analysis of patient data coming from different research institutes can help the clinical experts in the study of the under investigation disorder and the publication of more accurate results. However, the heterogeneity of data poses an important barrier in the computer-based analysis of patient data collected so far. In this work we have studied several machine learning techniques that could potentially mediate the aforementioned problem. Also, several tools and mechanisms were developed that facilitate the harmonization of data, using machine learning techniques. The developed tools and mechanisms were accordingly applied for the harmonization of patient data coming from two different research organisations. The results verified that the tolls developed and the machine learning techniques used can really help in the harmonization of real data and hence improve their interoperability.

Keywords

Data, Data Harmonization, Language Models, Machine Learning Techniques, System Interoperability

1. Εισαγωγή

Η ανάλυση των δεδομένων που προέρχονται από διάφορους οργανισμούς, μπορεί να συμβάλει στη βελτίωση των εσωτερικών διεργασιών του οργανισμού και στην εξαγωγή χρήσιμων συμπερασμάτων, τα οποία θα μπορούσαν να προσδώσουν στον οργανισμό ένα ανταγωνιστικό πλεονέκτημα, έναντι των υπολοίπων οργανισμών που δραστηριοποιούνται σε ένα συγκεκριμένο πεδίο. Ειδικά στον χώρο της κλινικής έρευνας, η ανάλυση δεδομένων που προέρχονται από διαφορετικά κέντρα έρευνας, είναι ιδιαίτερα χρήσιμη, καθώς δίνει τη δυνατότητα στους ερευνητές να έχουν πρόσβαση σε ένα σημαντικό όγκο διαφορετικών δεδομένων, το οποίο, με τη σειρά του, επιτρέπει την εις βάθος ανάλυση των δεδομένων αυτών (π.χ., μελέτη υποκατηγοριών) και την καλύτερη γενίκευση των αποτελεσμάτων [1]. Αυτό, με τη σειρά του, μπορεί να συμβάλει στην ισχυροποίηση των συνεργαζόμενων κέντρων έρευνας στον συγκεκριμένο χώρο και τη δημιουργία κέρδους, στην περίπτωση των επιχειρήσεων (π.χ., φαρμακευτικών εταιριών).

Ωστόσο, η από κοινού χρησιμοποίηση των δεδομένων αυτών προϋποθέτει τη δυνατότητα εντοπισμού τους και ακολούθως την ένταξη και χρησιμοποίησή τους από τα σχετικά λογισμικά πακέτα των οργανισμών. Συνεπώς, έννοιες που σχετίζονται με τη διαλειτουργικότητα (interoperability) των δεδομένων και γενικότερα των συστημάτων, αποκτούν ιδιαίτερο ενδιαφέρον. Προς αυτήν την κατεύθυνση, η καταγραφή των απαραίτητων πληροφοριών για το σύνολο των δεδομένων αυτών (metadata) καθώς επίσης και η έκφραση των δεδομένων αυτών με βάση ένα κοινό από τους συνεργαζόμενους οργανισμούς μοντέλο, είναι ιδιαίτερα σημαντική. Η διαδικασία αυτή είναι, εν γένει, δύσκολη, εξαιτίας των σημαντικών διαφορών που υπάρχουν στον τρόπο αναπαράστασης των δεδομένων αυτών, γεγονός που δυσκολεύει την από κοινού χρήση τους. Ειδικά στον χώρο της κλινικής έρευνας και περίθαλψης υπάρχουν σημαντικές δομικές και σημασιολογικές διαφορές, που καθιστούν την εναρμόνιση των δεδομένων εξαιρετικά δύσκολη, η οποία γίνεται ακόμη πιο δύσκολη, εάν αναλογιστούμε ότι θα πρέπει να διαχειριστούμε προσωπικά δεδομένα και κατ' επέκταση να λάβουμε υπόψη το ευρύτερο πλαίσιο που τα περιβάλλει.

Οι τεχνικές μηχανικής μάθησης έχουν σημειώσει ιδιαίτερη πρόοδο τα τελευταία χρόνια, ειδικά στον χώρο επεξεργασίας της φυσικής γλώσσας και εικόνας, εκμεταλλευόμενες τις τεχνολογικές προόδους που έχουν γίνει από την πλευρά του υλικού (π.χ., κάρτες επεξεργασίας γραφικών γενικού σκοπού). Ειδικά στον χώρο επεξεργασίας κειμένου, έχουν αναπτυχθεί αρκετά διαφορετικά μοντέλα, τα οποία επιτρέπουν τη διανυσματική αναπαράσταση των λέξεων, λαμβάνοντας υπόψη τη σημασία που αυτές έχουν στο κείμενο και συνεπώς διευκολύνουν την εφαρμογή αλγορίθμων επεξεργασίας κειμένου που ανήκουν στο ευρύτερο

πεδίο της τεχνικής νοημοσύνης. Λαμβάνοντας υπόψη ότι η χρήση της φυσικής γλώσσας είναι εκτεταμένη ακόμη και στη δομημένη ή ημι-δομημένη αναπαράσταση των δεδομένων ενός οργανισμού, είναι προφανές ότι τα μοντέλα γλώσσας και οι τεχνικές μηχανικής μάθησης αναμένεται να έχουν έναν σημαντικό ρόλο στη διαδικασία εναρμόνισης των δεδομένων που προέρχονται από διαφορετικούς παρόχους.

Στην εργασία αυτή παρουσιάζουμε την προσέγγιση που ακολουθήσαμε και τα εργαλεία που αναπτύχθηκαν για την εναρμόνιση των δεδομένων των ασθενών που πάσχουν από ένα συγκεκριμένο σύνδρομο καθώς και τη χρήση τεχνικών μηχανικής μάθησης για τη διευκόλυνση της διαδικασίας αυτής. Ειδικότερα, οι τεχνικές μηχανικής μάθησης χρησιμοποιήθηκαν τόσο κατά τη δημιουργία της κοινής αναπαράστασης των δεδομένων όσο και κατά τη μετατροπή – έκφραση των δεδομένων αυτών με τους όρους του μοντέλου που αναπτύχθηκε μέσω της χρήσης κανόνων συσχέτισης. Τα αποτελέσματα έδειξαν ότι τα εργαλεία αυτά και ειδικότερα οι τεχνικές μηχανικής μάθησης μπορούν να συμβάλουν στην καλύτερη αναπαράσταση των εναρμονισμένων δεδομένων και να επιταχύνουν τη διαδικασία αυτή μέσω του αυτόματου εντοπισμού πιθανών συσχετίσεων μεταξύ των όρων διαφορετικών μοντέλων. Ωστόσο, ο ρόλος του χρήστη στην παραπάνω διαδικασία είναι καθοριστικός, καθώς έχει τη δυνατότητα να εξετάσει περαιτέρω τα ενδιάμεσα αποτελέσματα που προκύπτουν από την εφαρμογή των αλγορίθμων και να καλύψει πιθανά κενά (π.χ., μη αυτομάτως εντοπισμένες συσχετίσεις μεταξύ των όρων) μέσω των εργαλείων που αναπτύχθηκαν.

1.1. Δομή Εγγράφου

Η εργασία αυτή είναι οργανωμένη ως εξής. Αρχικά, στην Ενότητα 2 παρουσιάζουμε σχετική εργασία και γνώση σχετικά με την εναρμόνιση των δεδομένων και τις τεχνικές μηχανικής μάθησης. Στη συνέχεια, στην Ενότητα 3 παρουσιάζουμε την προσέγγιση που ακολουθήσαμε για την εναρμόνιση πραγματικών δεδομένων ασθενών που προέρχονται από διαφορετικά κέντρα έρευνας, τα εργαλεία που αναπτύχθηκαν για τον σκοπό αυτό και τις τεχνικές μηχανικής μάθησης και εξόρυξης γνώσης που χρησιμοποιήθηκαν Τα αποτελέσματα της εργασίας αυτής παρουσιάζονται στην Ενότητα 4 καθώς και σχετικά θέματα προς συζήτηση. Τέλος, στην Ενότητα 5 συνοψίζουμε τα κύρια σημεία της εργασίας αυτής. Στο παράρτημα που υπάρχει στην Ενότητα 6 υπάρχει πληροφορία για θέματα που σχετίζονται έμμεσα ή άμεσα με αυτά που αναφέρονται στην εργασία αυτή.

2. Σχετική Εργασία και Γνώση

2.1. Διαλειτουργικότητα και Ετερογένεια

2.1.1. Διαλειτουργικότητα Συστημάτων

Ο όρος διαλειτουργικότητα (interoperability) αναφέρεται στην ικανότητα των συστημάτων να ανταλλάσουν πληροφορία και να μπορούν να τη χρησιμοποιήσουν για την επίτευξη των κοινών σκοπών τους [2]. Ειδικά στον χώρο των επιχειρήσεων, η διαλειτουργικότητα αναφέρεται σε διάφορα επίπεδα αλληλεπίδρασης (π.χ., επίπεδο δεδομένων) και συστήματα (π.χ., σύστημα διαχείρισης δεδομένων). Μη επαρκής διαλειτουργικότητα μεταξύ των συστημάτων μπορεί να δημιουργήσει σοβαρά προβλήματα σε έναν οργανισμό. Συνεπώς, είναι απαραίτητη η κατανόηση των δυνατών σημείων και των αδυναμιών των επιχειρήσεων σε θέματα που σχετίζονται με τη διαλειτουργικότητα και η συνεχής παρακολούθηση της κατάστασης (Interoperability Assessment - INAS), έτσι ώστε να μπορούν να εντοπιστούν εγκαίρως πιθανά προβλήματα και να τα διαχειριστούν κατάλληλα.

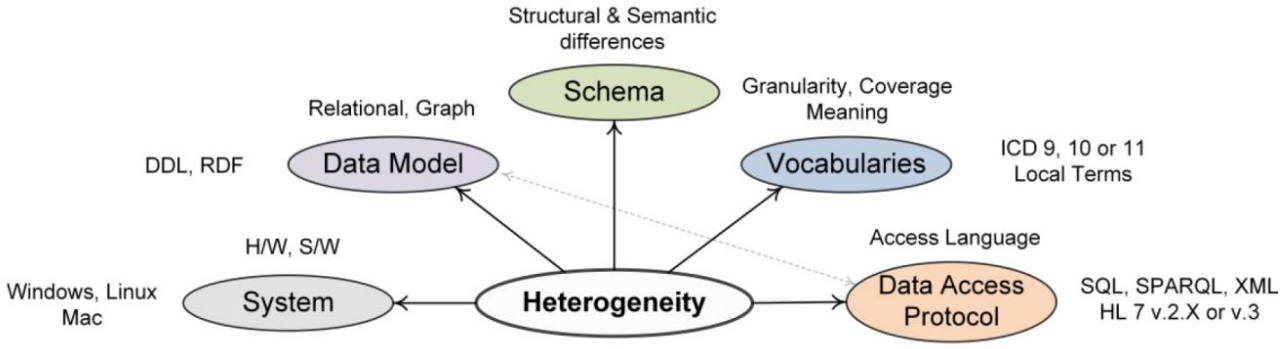
Το Ευρωπαϊκό Πλαίσιο για τη Διαλειτουργικότητα (European Interoperability Framework – EIF[1]) ορίζει τέσσερα επίπεδα διαλειτουργικότητας. Το τεχνικό επίπεδο (technical layer) εστιάζει στη δυνατότητα των IT συστημάτων / υπηρεσιών να επικοινωνούν μεταξύ τους. Το σημασιολογικό επίπεδο (semantic layer) επικεντρώνεται στην ικανότητα των συστημάτων να κατανοούν τα δεδομένα που ανταλλάσσονται. Το επίπεδο οργάνωσης (organizational layer) εστιάζει στη δυνατότητα «συγχρονισμού» / «ευθυγράμμισης» των εσωτερικών διεργασιών των οργανισμών. Τέλος το νομικό επίπεδο (legal layer) ασχολείται με τα νομικά θέματα και τις πολιτικές των οργανισμών.

Προβλήματα διαλειτουργικότητας μπορεί να υπάρχουν σε ένα ή περισσότερα από τα παραπάνω επίπεδα. Για τον σκοπό αυτό, θα πρέπει κάθε οργανισμός να παρακολουθεί την τρέχουσα κατάσταση και, κατά πόσο είναι εναρμονισμένος με το περιβάλλον στο οποίο δραστηριοποιείται (potentiality assessment), να εξετάζει τη δυνατότητα επικοινωνίας / συμβατότητας (compatibility assessment) των συστημάτων του με κάποιο άλλο σύστημα στον χώρο αυτό και τέλος να λαμβάνει υπόψη το κόστος που χρειάζεται για τη δημιουργία δυσλειτουργικών εφαρμογών (performance assessment).

[1] European Interoperability Framework (EIF), https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vkcw9q2gvczu

2.1.2. Ετερογένεια Πηγών Δεδομένων

Ο τρόπος αναπαράστασης και αποθήκευσης των δεδομένων (και των μετα-δεδομένων) που ανήκουν σε ένα συγκεκριμένο πεδίο γνώσης είναι διαφορετικός από οργανισμό σε οργανισμό (ή ακόμη και τμήματα του ίδιου του οργανισμού) εξαιτίας του ανεξάρτητου τρόπου σχεδιασμού των σχετικών συστημάτων και των αντίστοιχων βάσεων δεδομένων των οργανισμών αυτών και της περιορισμένης χρήσης ανοιχτών, ευρέως χρησιμοποιούμενων προτύπων που υπάρχουν στον χώρο τους. Κατά συνέπεια, υπάρχουν σημαντικές διαφορές στον τρόπο οργάνωσης των δεδομένων (data model), στα μοντέλα που έχουν αναπτυχθεί (schema) και στους όρους / κωδικοποιήσεις που χρησιμοποιούνται (vocabularies), καθώς και στον τρόπο που μπορούμε να έχουμε πρόσβαση σε αυτά (data access protocol) –Εικόνα 1. Τα παραπάνω είναι γνωστά ως ετερογένεια των δεδομένων (data heterogeneity) και δυσκολεύουν την πρόσβαση και ανάλυση του συνόλου των δεδομένων αυτών με έναν ενιαίο τρόπο.



Εικόνα 1*:* Ετερογένεια Πηγών Δεδομένων

Ειδικά στον χώρο της κλινικής έρευνας, ο τρόπος αναπαράστασης των δεδομένων διαφέρει σημαντικά από οργανισμό σε οργανισμό λόγω της ιδιαιτερότητας του πεδίου αυτού. Ειδικότερα, υπάρχουν πολλές διαφορετικές ανθρώπινες εκφράσεις για την περιγραφή των βιοϊατρικών όρων [3], η σημασιολογία των οποίων είναι, εν γένει, αρκετά περίπλοκη. Για τη διαλειτουργική αναπαράσταση και ανταλλαγή των δεδομένων που σχετίζονται με την υγεία των ασθενών στον χώρο της κλινικής έρευνας και περίθαλψης, έχουν αναπτυχθεί διάφορα πρότυπα στον χώρο αυτό [4], με τα πιο γνωστά να είναι αυτά που έχουν δημοσιευτεί από τους οργανισμούς HL7[2] (Health Level Seven) και CDISC[3] (Clinical Data Interchange Standards Consortium), τα οποία μπορούν επίσης να συνδυαστούν με άλλα πρότυπα, που έχουν να κάνουν με την καταγραφή και οργάνωση των όρων ενός συγκεκριμένου πεδίου γνώσης, όπως είναι η Διεθνής

[2] Health Level 7 (HL7), https://www.hl7.org/
[3] Clinical Data Interchange Standards Consortium (CDISC), https://www.cdisc.org/

Κωδικοποίηση των Ασθενειών (International Classification of Diseases – ICD[4]) και η Κατηγοριοποίηση των Φαρμακευτικών Χημικών Ουσιών (Anatomical Therapeutic Chemical – ATC[5] – classification). Ωστόσο, η χρήση των προτύπων αυτών από τους οργανισμούς είναι, εν γένει, περιορισμένη και, σε συνδυασμό με τις δυσκολίες ευθυγράμμισης που υπάρχουν ακόμη και μεταξύ διαφορετικών εκδόσεων των ίδιων προτύπων, καθιστούν την ομαλή επικοινωνία μεταξύ των συστημάτων εξαιρετικά δύσκολη.

Στην περίπτωση που πρόκειται να διαχειριστούμε προσωπικά δεδομένα (ή εν γένει άλλα ιδιωτικά δεδομένα), υπάρχουν επιπρόσθετα θέματα ασφάλειας και ιδιωτικότητας που θα πρέπει να ληφθούν υπόψη, καθώς τα δεδομένα αυτά υπόκεινται σε αυστηρούς εθνικούς νόμους και κανόνες. Για παράδειγμα, από το 2018 είναι σε ισχύ στην Ευρώπη ο Γενικός Κανόνας Προστασίας Δεδομένων (General Data Privacy Regulation - GDPR) [5], ο οποίος καθορίζει λεπτομερώς τις συνθήκες που θα πρέπει να πληρούνται σχετικά με την συλλογή, αποθήκευση και διαχείριση των προσωπικών δεδομένων των επιχειρήσεων και των οργανισμών.

2.2. Εναρμόνιση και Ενοποίηση Δεδομένων

2.2.1. Ορισμός Εναρμόνισης και Ενοποίησης Δεδομένων

Η Ενοποίηση Δεδομένων (Data Integration)[6] είναι η διαδικασία που ακολουθείται για τη συλλογή δεδομένων από αρκετές διάσπαρτες πηγές δεδομένων (π.χ., Σχεσιακές Βάσεις Δεδομένων) και την τοποθέτησή τους σε ένα κοινό σύστημα διαχείρισης δεδομένων (όπως είναι για παράδειγμα μία Αποθήκη Δεδομένων – Data Warehouse). Η διαδικασία αυτή, γνωστή και ως διαδικασία εξαγωγής-μετατροπής-φόρτωσης (Extract-Transform-Load – ETL – process) περιλαμβάνει τα ακόλουθα τρία βήματα. Αρχικά γίνεται εξαγωγή (extract) των απαραίτητων δεδομένων από τις σχετικές πηγές δεδομένων. Στη συνέχεια γίνεται καθαρισμός και μετατροπή των δεδομένων αυτών (transformation), έτσι ώστε αυτά να είναι εκφρασμένα με την απαραίτητη μορφή / σχήμα. Τέλος τα «καθαρά» δεδομένα εισέρχονται / «φορτώνονται» (load) στο κοινό σύστημα διαχείρισης δεδομένων.

Η Εναρμόνιση Δεδομένων (Data Harmonization)[7] είναι μια διαδικασία παρόμοια με αυτή που ακολουθείται για την Ενοποίηση των Δεδομένων (Data Integration). Ωστόσο, αποσκοπεί στην αναπαράσταση των δεδομένων που προέρχονται από διαφορετικές πηγές / βάσεις δεδομένων με ένα κοινό

[4] International Classification of Diseases (ICD), https://www.who.int/standards/classifications/classification-of-diseases
[5] Anatomical Therapeutic Chemical (ATC) Classification, https://www.who.int/tools/atc-ddd-toolkit/atc-classification
[6] What is Data Integration? https://www.integrate.io/glossary/what-is-data-integration/
[7] What is Data Harmonization? https://www.integrate.io/glossary/what-is-data-harmonization/

μοντέλο / σχήμα. Συνεπώς, η προσέγγιση που συνήθως ακολουθείται είναι παρόμοια με τη διαδικασία εξαγωγής-μετατροπής-φόρτωσης (ETL process), που ακολουθείται και στην περίπτωση της ενοποίησης των δεδομένων. Ωστόσο, στην περίπτωση αυτή δίνεται ιδιαίτερη έμφαση στη διαδικασία της μετατροπής των δεδομένων και ιδιαίτερα στην αποτελεσματική διαχείριση των συντακτικών και σημασιολογικών διαφορών που υπάρχουν μεταξύ των διαφορετικών μοντέλων / σχημάτων αναπαράστασης των δεδομένων.

Επίσης, κατά την παραπάνω διαδικασία εξαγωγής-μετατροπής-φόρτωσης, θα πρέπει να λάβουμε υπόψη διάφορους παράγοντες που έχουν να κάνουν (α) με το μέγεθος και τον τρόπο αναπαράστασης των δεδομένων των πηγών, (β) τον διαφορετικό τρόπο αναπαράστασης των δεδομένων αυτών, (γ) την ακρίβεια και την αξιοπιστία των δεδομένων που έχουν καταγραφεί, (δ) την ταχύτητα με την οποία παράγονται νέα δεδομένα ή γίνονται αλλαγές στα ήδη υπάρχοντα και τέλος (ε) τις διαδικασίες ή σκοπούς που σκοπεύουμε να υποστηρίξουμε. Στα παραπάνω, θα πρέπει να λάβουμε υπόψη και άλλους παράγοντες που έχουν να κάνουν με την ασφάλεια και την ιδιωτικότητα των δεδομένων, καθώς και πιθανούς νομικούς και ηθικούς περιορισμούς που μπορεί να τα συνοδεύουν.

2.2.2. Τεχνικές Εναρμόνισης και Ενοποίησης Δεδομένων

Η εναρμόνιση των δεδομένων μπορεί να γίνει είτε κατά τη διαδικασία της συλλογής των δεδομένων από τους εκάστοτε οργανισμούς (prospective data harmonization), μέσω της χρήσης κοινών πρωτοκόλλων για τον σκοπό αυτό, είτε μετά τη συλλογή των δεδομένων (retrospective data harmonization), ειδικά όταν τα δεδομένα έχουν ήδη συλλεχθεί από τους οργανισμούς αυτούς (secondary use of data). Πολλές φορές η δεύτερη προσέγγιση είναι και η μόνη επιλογή.

Για την εναρμόνιση των δεδομένων που προέρχονται από διαφορετικά κέντρα και μελέτες, από τεχνικής άποψης μπορούμε είτε να μεταφέρουμε τα δεδομένα σε «έναν» κεντρικό υπολογιστή και ακολούθως να τα μετατρέψουμε στην επιθυμητή μορφή είτε να τα διατηρήσουμε στους υπολογιστές στους οποίους αυτά βρίσκονται [6]. Υπάρχουν, ωστόσο, αρκετοί περιορισμοί σχετικά με την ασφάλεια και ιδιωτικότητα των δεδομένων, οι οποίοι μπορούν να μετριαστούν, μεταφέροντας την επεξεργασία (και κατ' επέκταση μετατροπή/εναρμόνιση και ανάλυση) των δεδομένων αυτών κοντά στον χώρο δημιουργίας τους. Στην περίπτωση αυτή, η διαδικασία που συνήθως ακολουθείται περιλαμβάνει την αποστολή δεσμίδων εντολών (scripts), οι οποίες διαβάζουν μόνο τα δεδομένα (read-only) και ακολούθως τα μετατρέπουν στην επιθυμητή / εναρμονισμένη μορφή. Η περιγραφή της διαδικασίας εναρμόνισης που ακολουθήθηκε (documentation) και η επικύρωση των εναρμονισμένων δεδομένων (validation) είναι επίσης σημαντικές, αν και συχνά παραλείπονται.

Για την εναρμόνιση των δεδομένων, η προσέγγιση που περιγράφεται στο έργο [7] βασίζεται στην ύπαρξη των σχετικών δεδομένων στους εκάστοτε οργανισμούς (γνωστή ως διαθεσιμότητα δείγματος – Sample Availability - SAIL) και επιτρέπει στους ερευνητές να ξεπεράσουν προσωρινά τα προβλήματα που υπάρχουν σχετικά με την ιδιωτικότητα των δεδομένων. Εφόσον οι ερευνητές εντοπίσουν τα σχετικά σύνολα που τους ενδιαφέρουν, μπορούν να τεκμηριώσουν την ανάγκη πρόσβασης και να προχωρήσουν στις σχετικές ενέργειες, έτσι ώστε να λάβουν την άδεια πρόσβασης στα δεδομένα αυτά.

Οι συγγραφείς της δημοσίευσης [8] περιγράφουν έναν γενικό τρόπο που θα μπορούσαμε να ακολουθήσουμε για την εναρμόνιση των δεδομένων, σύμφωνα με τον οποίο θα πρέπει αρχικά να καθοριστούν τα «κοινά» πεδία (common data elements) που περιγράφουν επαρκώς τις σημαντικές οντότητες ενός πεδίου γνώσης (στην εργασία [7] τα πεδία αυτά αναφέρονται ως μεταβλητές ενδιαφέροντος – Variables of Interest - VOI) και ακολούθως να εκφράσουν τα δεδομένα που υπάρχουν από κάθε πάροχο με βάση αυτά, ακολουθώντας μια προκαθορισμένη διαδικασία. Αρχικά (φάση 1), θα πρέπει να εντοπίσουμε τα αντίστοιχα πεδία σε κάθε πάροχο (schema matching), λαμβάνοντας υπόψη τον τρόπο που αυτά έχουν αναπαρασταθεί σε καθέναν από αυτούς. Έπειτα (φάση 2), θα πρέπει να καθορίσουμε τον τρόπο συσχέτισης των πεδίων αυτών με τα κοινά πεδία του μοντέλου που έχει αναπτυχθεί (semantic mapping), έτσι ώστε οι συσχετίσεις αυτές να μπορούν να χρησιμοποιηθούν αυτόματα από κάποιο λογισμικό πακέτο, για να καλύψουν τις ανάγκες μας. Τέλος (φάση 3), θα πρέπει να εκφραστούν τα δεδομένα με βάση τα στοιχεία του «κοινού» μοντέλου που έχει καθοριστεί μέσω μιας διαδικασίας, η οποία αναλαμβάνει να φορτώσει τα δεδομένα, να τα μετατρέψει στην κατάλληλη μορφή και να τα αποθηκεύσει ακολούθως στη βάση (ETL – process).

Στο έγγραφο αυτό [9], οι συγγραφείς εστιάζουν στη σχέση που υπάρχει μεταξύ της ενοποίησης δεδομένων και της μηχανικής μάθησης. Ειδικότερα, η μηχανική μάθηση θα μπορούσε να χρησιμοποιηθεί κατά την ενοποίηση δεδομένων σε αρκετές διαφορετικές περιπτώσεις. Η επίλυση οντοτήτων (entity resolution) αποσκοπεί στον εντοπισμό εγγραφών που αναφέρονται στην ίδια οντότητα του πραγματικού κόσμου και θα μπορούσε να βοηθηθεί από τη χρήση τεχνικών μηχανικής μάθησης (επιβλεπόμενων ή μη επιβλεπόμενων). Οι τεχνικές αυτές θα μπορούσαν, επίσης, να συμβάλουν στην επίλυση διαφορών / συγκρούσεων (conflicts) που εντοπίζονται σε διαφορετικές πηγές δεδομένων (ως μέρος της διαδικασίας συγχώνευσης – data fusion). Η εξαγωγή δεδομένων (data extraction) έχει ως στόχο την εξαγωγή δομημένων δεδομένων από μη-δομημένα δεδομένα (όπως είναι το κείμενο) ή ημι-δομημένα δεδομένα (όπως είναι τα δεδομένα των διαδικτυακών σελίδων, που εσωτερικά αναπαριστώνται ως δένδρα – DOM-tree) και μπορεί, επίσης, να βελτιωθεί με τη χρήση τεχνικών που ανήκουν στο πεδίο της μηχανικής μάθησης. Τέλος, ο

εντοπισμός συσχετίσεων μεταξύ διαφορετικών μοντέλων (schema alignment) μπορεί να βελτιωθεί μέσω της χρήσης τεχνικών μηχανικής μάθησης. Λόγω της ιδιαίτερης σημασίας που έχουν οι τεχνικές μηχανικής μάθησης στην εναρμόνιση / ενοποίηση των δεδομένων, θα τις δούμε αναλυτικά στην ενότητα 2.3.

Σημειώνουμε ότι στο έγγραφο [10] υπάρχει μια συστηματική ανάλυση της βιβλιογραφίας (Systematic Literature Review) [11] σχετικά με την εναρμόνιση ετερογενών δεδομένων. Όπως φαίνεται στα αποτελέσματα της ανάλυσης αυτής, τα αρχικά δεδομένα μπορεί να είναι δομημένα, ήμι-δομημένα ή μη-δομημένα και να έχουν αναπαρασταθεί με διαφορετική μορφή, ακόμη και όταν ανήκουν σε κάποια από τις παραπάνω κατηγορίες (π.χ., μέσω της χρήσης SQL, RDF, XML, JSON, CSV, κτλ.). Επίσης, οι τεχνολογίες που χρησιμοποιούνται για την αποθήκευση και επεξεργασία των δεδομένων αυτών σχετίζονται άμεσα με τις ιδιαιτερότητες της κάθε πηγής (π.χ., μέγεθος δεδομένων) και μπορεί να διαφέρουν (π.χ., HDFS [12] και MapReduce [13], Spark [14], κτλ.). Για τα παραπάνω υπάρχουν περισσότερες πληροφορίες στο παράρτημα και ειδικότερα στην ενότητα 6.2. Σχετικά με την εναρμόνιση των δεδομένων δίνεται ιδιαίτερη έμφαση στα μη-δομημένα δεδομένα και ειδικότερα στα δεδομένα που είναι σε κείμενο (text processing/mining) με τη χρήση κλασικών τεχνικών επεξεργασίας φυσικής γλώσσας και αλγορίθμων μηχανικής μάθησης (συμπεριλαμβανομένων των κλασικών τεχνικών μηχανικής μάθησης και των βαθιών νευρωνικών δικτύων) να έχουν ξεχωριστή θέση στα αντίστοιχα συστήματα. Επιπρόσθετα, αρκετά διαφορετικά μοντέλα γλώσσας (language models) έχουν χρησιμοποιηθεί για τη διανυσματική αναπαράσταση των λέξεων / φράσεων / προτάσεων / εγγράφων, τα οποία θα παρουσιάσουμε αναλυτικά στην ενότητα 2.4.

2.3. Τεχνικές Μηχανικής Μάθησης και Εξόρυξης Γνώσης

2.3.1. Κλασικές Τεχνικές Μηχανικής Μάθησης και Εξόρυξης Γνώσης

Οι τεχνικές μηχανικής μάθησης (Machine Learning - ML) μπορούν να οργανωθούν σε δύο ευρύτερες κατηγορίες, τις τεχνικές επιβλεπόμενης μάθησης (Supervised ML) και τις τεχνικές μη-επιβλεπόμενης μάθησης (Unsupervised ML), με βάση το κατά πόσο χρειάζονται επισημασμένα δεδομένα (labeled data) για να εκτελεστούν. Περισσότερες πληροφορίες για τα δεδομένα που χρησιμοποιούνται στα πλαίσια της μηχανικής μάθησης υπάρχουν στο παράρτημα και ειδικότερα στην ενότητα 6.1.

2.3.1.1. Κλασικές Τεχνικές Επιβλεπόμενης Μάθησης

Υπάρχουν αρκετές τεχνικές επιβλεπόμενης μάθησης (Supervised ML) συμπεριλαμβανομένων των τεχνικών Γραμμικής / Λογιστικής Παλινδρόμησης (Linear/Logistic Regression), οι οποίες αποτελούν ειδικές περιπτώσεις του Γενικευμένου Γραμμικού Μοντέλου (Generalized Linear Model) και της Γκαουσιανης

Διαχωριστικής Ανάλυσης (Gaussian Discrimination Analysis), η οποία αποτελεί έναν συγκεκριμένο τύπο Παραγωγικού Αλγορίθμου Μάθησης (Generative Learning Algorithm). Τα Μπεϋζιανά Δίκτυα (Bayesian Networks) [15] είναι γραφήματα που αναπαριστούν πιθανολογικά δίκτυα, στα οποία οι κόμβοι αντιπροσωπεύουν μεταβλητές και οι ακμές τις μεταξύ τους αλληλεπιδράσεις, με το απλούστερο από αυτά τα δίκτυα να είναι το Αφελές Μπεϋζιανό Δίκτυο (Naïve Bayes Network), στο οποίο τα χαρακτηριστικά των δεδομένων είναι ανεξάρτητα μεταξύ τους και εξαρτώνται μόνο από την κλάση των δεδομένων. Τα Δένδρα Αποφάσεων (Decision Trees) [16] είναι δομές που μοιάζουν με δένδρα και χρησιμοποιούνται συχνά για την κατηγοριοποίηση των δεδομένων, με βάση τις αποφάσεις που λαμβάνονται σε κάθε κόμβο, όπου εξετάζονται διάφορα χαρακτηριστικά των δεδομένων. Η απόδοσή τους μπορεί να βελτιωθεί αισθητά με τον «συνδυασμό» αρκετών τέτοιων Δένδρων Απόφασης. Η τεχνική αυτή ανήκει στην κατηγορία των τεχνικών Εκμάθησης Συνόλου (Ensemble Learning) και είναι γνωστή ως Τυχαίο Δάσος (Random Forest).

Ο αλγόριθμος των Κ Κοντινότερων Γειτόνων (K nearest neighbor - kNN) βασίζεται στον εντοπισμό των Κ κοντινότερων σημείων για την πρόβλεψη της επικρατέστερης κλάσης (στην περίπτωση των προβλημάτων κατηγοριοποίησης) και κατ' επέκταση προσπελαύνει όλα τα διαθέσιμα δεδομένα κάθε φορά, ενώ εξαρτάται από τον τρόπο μέτρησης της απόστασης μεταξύ δύο σημείων. Η Τεχνική Υποστήριξης Διανύσματος (Support Vector Machine - SVM) [17] είναι μια τεχνική που αποσκοπεί στην εύρεση του υπέρ-πλάνου (hyperplane) που διαχωρίζει καλύτερα τα δεδομένα, υπό την έννοια ότι η απόσταση των δεδομένων από αυτό μεγιστοποιείται και κατ' επέκταση το ρίσκο εσφαλμένης κατηγοριοποίησης ελαχιστοποιείται. Συνεπώς, η τεχνική αυτή είναι ιδανική για την περίπτωση των προβλημάτων δυαδικής ταξινόμησης και, σε συνδυασμό με την μέθοδο των πυρήνων (kernel methods), μπορεί να διαχειριστεί ικανοποιητικά αρκετές από τις περιπτώσεις εκείνες, στις οποίες τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα.

Η απόδοση ενός μοντέλου μηχανικής μάθησης εξαρτάται συχνά από πολλές άλλες παραμέτρους (γνωστές ως υπερπαράμετροι - hyperparameters), τις οποίες δεν μπορούμε να καθορίσουμε αυτόματα μέσω των τεχνικών της μηχανικής μάθησης. Ο αυτόματος εντοπισμός των κατάλληλων τιμών των υπερπαραμέτρων αυτών δεν είναι μία απλή διαδικασία, ειδικά όταν θα πρέπει να εξετάσουμε ταυτόχρονα αρκετούς συνδυασμούς αυτών. Η Μπεϋζιανή Βελτιστοποίηση (Bayesian Optimization) [18] διευκολύνει την επίλυση του προβλήματος βελτιστοποίησης των υπερπαραμέτρων (Hyperparameters Optimization Problem) και, σε συνδυασμό με τεχνικές μετα-μάθησης (Meta-Learning Techniques) και Αυτόματου Συνδυασμό Μεθόδων (Automatic Ensemble Construction), μπορεί να περιορίσει στο ελάχιστο τον ρόλο ενός ειδικού σε θέματα μηχανικής μάθησης στη διαδικασία δημιουργίας του μοντέλου. Ειδικότερα, η επιλογή του κατάλληλου αλγορίθμου και των βέλτιστων υπερπαραμέτρων αυτού, μπορεί να περιγραφεί ως ένα ενιαίο

πρόβλημα βελτιστοποίησης (Combined Algorithm Selection and Hyperparameters Optimization – CASH problem) [19] και κατ' επέκταση να λυθεί αυτόματα.

2.3.1.2. Κλασικές Τεχνικές Μη-επιβλεπόμενης Μάθησης

Οι τεχνικές ομαδοποίησης (clustering methods) [20] ανήκουν στην ευρύτερη κατηγορία των τεχνικών μη επιβλεπόμενης μάθησης (Unsupervised ML) και αποσκοπούν στον εντοπισμό μοτίβων σε δεδομένα υψηλών διαστάσεων (δηλαδή δεδομένα για καθένα από τα οποία έχουν καταγραφεί αρκετά χαρακτηριστικά), με τις πιο γνωστές από αυτές τις τεχνικές να είναι οι τεχνικές ιεραρχικής ομαδοποίησης (hierarchical clustering) και οι τεχνικές που βασίζονται στην ύπαρξη κάποιου κέντρου στις ομάδες των δεδομένων (centroid models), με την πιο ευρέως γνωστή από αυτές τις τεχνικές να είναι ο αλγόριθμος ο K-means. Οι αλγόριθμοι που βασίζονται στην πυκνότητα των δεδομένων (density models), όπως είναι ο αλγόριθμος DBSCAN [21] τοποθετούν τα δεδομένα/σημεία σε ομάδας κατά τέτοιο τρόπο, ώστε οι ομάδες που δημιουργούνται να είναι «πυκνές» και τα μέλη τους συνδεδεμένα.

Οι τεχνικές εντοπισμού συχνών μοτίβων και κανόνων (Frequent Pattern and Association Rules Mining) αποσκοπούν στον εντοπισμό προηγουμένων άγνωστων μοτίβων και κανόνων με βάση τη συχνότητα εμφάνισης αυτών στα δεδομένα. Ο FP-Growth αλγόριθμος [22] δημιουργεί μια συμπαγή αναπαράσταση των δεδομένων των συναλλαγών που έχουν καταγραφεί σε μια σχεσιακή βάση δεδομένων με την μορφή ενός δέντρου, το οποίο και, ακολούθως, προσπελαύνει για τον εντοπισμό των συχνών μοτίβων (Frequent Patterns - FPs) και είναι πολύ πιο αποδοτικός από τον Apriori αλγόριθμο [23], ο οποίος σε κάθε επανάληψη προσπελαύνει τα δεδομένα της βάσης. Τα συχνά μοτίβα που έχουν εντοπιστεί (δηλαδή αντικείμενα τα οποία συνυπάρχουν αρκετές φορές στις συναλλαγές που έχουν καταγραφεί) μπορούν ακολούθως να χρησιμοποιηθούν για την εξαγωγή των κανόνων συσχέτισης (association rules).

Στο σημείο αυτό να αναφέρουμε ότι το υπολογιστικό κόστος των αλγορίθμων αυτών (συμπεριλαμβανομένων των τεχνικών επιβλεπόμενης και μη επιβλεπόμενης μάθησης) αυξάνεται εκθετικά με τον αριθμό των χαρακτηριστικών των δεδομένων (the curse of dimensionality). Για τον λόγο αυτό επιχειρείται συχνά να μειώσουμε το πλήθος των χαρακτηριστικών των δεδομένων, μια τεχνική που είναι γνωστή ως μείωση της διασταλτικότητας (dimensionality reduction). Για τον σκοπό αυτό μπορούμε, για παράδειγμα, να υπολογίσουμε τη διακριτική ικανότητα (discrimination power) του κάθε χαρακτηριστικού (είτε μεμονωμένα είτε συνδυασμός αυτών), έτσι ώστε ακολούθως να εστιάσουμε σε αυτά που έχουν υψηλή διακριτική ικανότητα.

Μια άλλη προσέγγιση είναι η μετάβαση σε μια άλλη διάσταση, όπου εκεί μπορεί να είναι πιο εύκολη η επιλογή των χαρακτηριστικών αυτών. Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis - PCA) [24] είναι ένας γραμμικός μετασχηματισμός, στον οποίο τα δεδομένα αναπαριστώνται σε ένα νέο σύστημα συντεταγμένων, χωρίς να υπάρχει απώλεια πληροφορίας. Στο νέο αυτό σύστημα συντεταγμένων, ορισμένες από τις διαστάσεις μπορούν να αναπαραστήσουν μεγάλο μέρος της διασποράς των δεδομένων και κατ' επέκταση, επιλέγοντας έναν μικρότερο αριθμό «νέων» χαρακτηριστικών των δεδομένων, μπορούμε να διατηρήσουμε το μεγαλύτερο μέρος της διασποράς τους. Η τεχνική αυτή είναι ιδιαίτερα χρήσιμη για την οπτική αναπαράσταση των δεδομένων. Ωστόσο, δεν θα πρέπει να παραλείψουμε να αναφέρουμε ότι η παραπάνω μέθοδος εστιάζει στην καλύτερη περιγραφή του συνόλου των δεδομένων και όχι τόσο στον διαχωρισμό τους. Η Ανάλυση Γραμμικού Διαχωρισμού (Linear Discrimination Analysis – LDA) [25] αποσκοπεί στην εύρεση των διανυσμάτων, τα οποία επιτρέπουν τη μετάβαση σε ένα νέο (με μικρότερες διαστάσεις) σύστημα συντεταγμένων, στο οποίο τα δεδομένα των δύο κλάσεων διαχωρίζονται καλύτερα.

Μια άλλη ευρέως χρησιμοποιούμενη μέθοδος για τη γραφική απεικόνιση των δεδομένων με πάρα πολλές διαστάσεις στον δισδιάστατο ή τρισδιάστατο χώρο είναι η t-SNE (t-distributed Stochastic Neighbor Embedding) [26]. Η μέθοδος αυτή είναι μια μη-γραμμική μέθοδος για τη μείωση των διαστάσεων των δεδομένων και βασίζεται στην κατανομή του Student (t-Student)[8] για τον υπολογισμό της ομοιότητας μεταξύ των δεδομένων. Ειδικότερα, τα δεδομένα τοποθετούνται στον χώρο κατά τέτοιο τρόπο, ώστε σημεία που είναι πάρα πολύ πιθανόν να είναι γειτονικά να τοποθετούνται κοντά στον δισδιάστατο/τρισδιάστατο χώρο, αλλιώς μακριά.

2.3.1.3. Άλλες Κλασικές Τεχνικές

Εκ των προτέρων αναφέρουμε ότι στις δύο προηγούμενες ενότητες δεν έχουμε συμπεριλάβει σκοπίμως τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks – ANN ή απλά NN), τα οποία και θα παρουσιάσουμε σε ξεχωριστή υποενότητα πιο κάτω (ενότητα 2.3.2)

Η ημι-επιβλεπόμενη μάθηση (semi-supervised learning)[9], η οποία είναι επίσης γνωστή και ως ασθενής επίβλεψη (weak supervision), βασίζεται τόσο σε δεδομένα με ετικέτα (labeled data) όσο και σε δεδομένα χωρίς ετικέτα (unlabeled data) για την εκπαίδευση του αλγορίθμου. Γενικότερα μιλώντας, τα δεδομένα με ετικέτα χρησιμοποιούνται για την «αρχική» εκπαίδευση του μοντέλου, ενώ τα δεδομένα χωρίς ετικέτα για την «περαιτέρω» βελτίωση του μοντέλου με βάση άλλα παρόμοια δεδομένα χωρίς ετικέτα. Για την εφαρμογή της ημι-επιβλεπόμενης μάθησης, θα πρέπει να πληρούνται ορισμένες συνθήκες. Ειδικότερα, τα

[8] T-Distribution, https://www.scribbr.com/statistics/t-distribution/
[9] The Ultimate Guide to Semi-Supervised Learning, https://www.v7labs.com/blog/semi-supervised-learning-guide

δεδομένα που βρίσκονται κοντά θα πρέπει να ανήκουν στην ίδια κατηγορία (smoothness assumption). Επίσης, το υπερ-επίπεδο που διαχωρίζει τα δεδομένα (στην περίπτωση ενός προβλήματος κατηγοριοποίησης) δεν θα πρέπει να «περνά» μέσα από τις πυκνές περιοχές (low density assumption). Επιπλέον, τα δεδομένα που ανήκουν στην ίδια περιοχή χαμηλότερων διαστάσεων θα πρέπει να έχοχυν την ίδια ετικέτα (manifold assumption). Σύμφωνα με του συγγραφείς του έργου [27], μπορούμε να οργανώσουμε τις τεχνικές ημι-επιβλεπόμενης μάθησης σε δύο κατηγορίες: τις επαγωγικές τεχνικές (inductive) και τις μεταδοτικές τεχνικες (transductive). Στην πρώτη περίπτωση, οι τεχνικές στοχεύουν στο να δημιουργήσουν και διαφέρουν ως προς τον τρόπο χρησιμοποίησης των δεδομένων χωρίς ετικέα, ενώ στη δεύτερη περίπτωση, οι τεχνικές εστιάζουν στην οργάνωση των δεδομένων.

Η ενισχυτική μάθηση (reinforcement learning) αποσκοπεί στην εκπαίδευση ενός συστήματος, έτσι ώστε να «παίρνει» τις σωστές αποφάσεις, ανάλογα με την κατάσταση στην οποία βρίσκεται [28]. Κατ' επέκταση ένα τέτοιο σύστημα θα πρέπει να λαμβάνει υπόψη το περιβάλλον στο οποίο βρίσκεται και ακολούθως να κάνει τις απαραίτητες ενέργειες (actions), έτσι ώστε να μεγιστοποιείται το όφελος (reward) που προκύπτει από τις ενέργειες αυτές. Για τον σκοπό αυτό, θα πρέπει να εξερευνήσει το περιβάλλον αλλά και να λάβει υπόψη την γνώση που έχει ήδη αποκτηθεί από προηγούμενη εμπειρία (exploration versus exploitation). Ένας τέτοιος αλγόριθμος συχνά υποθέτει ότι υπάρχει ένα μοντέλο για το περιβάλλον (model-based), όπως είναι για παράδειγμα οι αλγόριθμοι value-iteration και policy-iteration[10]. Ωστόσο, υπάρχουν και αλγόριθμοι, όπως είναι για παράδειγμα ο Q-learning [29], οι οποίοι δεν βασίζονται στην ύπαρξη ενός μοντέλου για το περιβάλλον (model-free).

Τέλος αναφέρουμε ότι στην παραπάνω περιγραφή των τεχνικών μηχανικής μάθησης δεν έχουμε σκοπίμως αναφέρει τις τεχνικές που βασίζονται στην «εξ' αποστάσεως» μάθηση (Federated Learning) καθώς και τις τεχνικές «συνεχούς» μάθησης (Online Learning). Επίσης, δεν έχουμε ασχοληθεί ιδιαίτερα με τη μεταφορά γνώσης (Transfer Learning) [30]. Το θέμα αυτό είναι αρκετά σημαντικό, καθώς συμβάλει στην επαναχρησιμοποίηση δεδομένων και κυρίως μοντέλων (ειδικά βαθιών νευρωνικών δικτύων) που έχουν είδη δημιουργηθεί και καλύπτεται μερικώς στην επόμενη ενότητα που παρουσιάζουμε τις τεχνικές επεξεργασίας κειμένου και τα μοντέλα γλώσσας. Για περισσοτερες λεπτομέρειες για το θέμα αυτό, παραπέμπουμε τους αναγνώστες στο προαναφερθέν έγγραφο καθώς επίσης και στη δημοσίευση [31], που εξειδικεύεται στον χώρο της βαθιάς μάθησης.

[10] Policy and Value Iteration, https://towardsdatascience.com/policy-and-value-iteration-78501afb41d2

2.3.2. Βαθιά Νευρωνικά Δίκτυα

Τα τεχνικά νευρωνικά Δίκτυα (Artificial Neural Networks - NN) αποτελούνται από αρκετούς συνδεδεμένους τεχνικούς νευρώνες, που χρησιμοποιούν την πληροφορία που έχει αποκτηθεί από τους κόμβους δικτύου για την κατηγοριοποίηση των δεδομένων εισόδου. Οι κόμβοι ενός τέτοιου νευρωνικού δικτύου είναι συχνά οργανωμένοι σε επίπεδα (layers), στα οποία οι κόμβοι δεν επικοινωνούν μεταξύ τους παρά μόνο με τους κόμβους του προηγούμενου συνήθως επιπέδου. Κάθε κόμβος / νευρώνας ενός τεχνητού νευρωνικού δικτύου «συνδυάζει» την πληροφορία που του δίνεται ως είσοδο (είτε από τον χρήστη είτε από τους κόμβους του προηγούμενου συνήθως επιπέδου) μέσω της χρήσης βαρών και ακολούθως «περνάει» την τιμή αυτή μέσα από μια συνάρτηση ενεργοποίησης (activation function)[11], όπως για παράδειγμα είναι η Softmax και η ReLU (Rectified Linear Unit), για να προκύψει η έξοδος του κόμβου αυτού. Το πρώτο επίπεδο ενός νευρωνικού δικτύου αποτελείται από τους κόμβους εισόδου (input layer) ενώ το τελευταίο επίπεδο από τους κόμβους εξόδου (output layer). Ένα νευρωνικό δίκτυο μπορεί, επίσης, να έχει μηδέν, ένα ή περισσότερα ενδιάμεσα επίπεδα, τα οποία είναι γνωστά και ως κρυφά επίπεδα (hidden layers). Στην περίπτωση που έχει παραπάνω από ένα κρυφό επίπεδο, το νευρωνικό δίκτυο ονομάζεται βαθύ νευρωνικό δίκτυο (deep neural network), αν και ο όρος αυτός συνήθως χρησιμοποιείται, όταν το νευρωνικό δίκτυο αποτελείται από πολύ περισσότερα από 2 ή 3 κρυμμένα επίπεδα.

2.3.2.1. Τύποι Βαθιών Νευρωνικών Δικτύων

Ένα νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης (Feed Forward Neural Network - FFNN), όπως για παράδειγμα είναι ένας πολλαπλών επιπέδων νευρώνας (Multi-Layer Perceptron - MLP), αποτελείται από ένα η περισσότερα επίπεδα (κρυφά επίπεδα), στα οποία οι κόμβοι του κάθε επιπέδου βασίζονται στις εξόδους των κόμβων του (συνήθως αμέσως) προηγούμενου επιπέδου, για να υπολογίσουν την έξοδό τους. Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional NN - CNN) [32] είναι Βαθιά Νευρωνικά Δίκτυα Προώθησης-Τροφοδοσίας (Feed-Forward), τα οποία έχουν αραιές συνδέσεις μεταξύ των κόμβων των προηγούμενων επιπέδων και κατ' επέκταση οι τεχνικοί νευρώνες που ανήκουν στα βαθύτερα στρώματα επικοινωνούν έμμεσα με τους κόμβους που υπάρχουν στα προηγούμενα επίπεδα. Σε κάθε ένα τέτοιο νευρωνικό δίκτυο μπορούμε να διακρίνουμε τρία χαρακτηριστικά επίπεδα: το επίπεδο συσπείρωσης (convolution layer), στο οποίο συνδυάζεται πληροφορία από αρκετούς γειτονικούς κόμβους μέσω της χρήσης ενός πυρήνα (kernel), το επίπεδο υποδειγματοληψίας (pooling layer), στο οποίο συνοψίζεται η πληροφορία ορισμένων γειτονικών κόμβων μέσω μιας προκαθορισμένης συνάρτησης (π.χ., μέγιστης ή μέσης τιμής) και τέλος ένα πλήρως συνδεδεμένο επίπεδο (fully-connected layer), το οποίο λαμβάνει υπόψη όλες τις εξόδους του

[11] Activation Functions in Neural Networks, https://www.v7labs.com/blog/neural-networks-activation-functions

προηγούμενου επιπέδου για να υπολογίσει την έξοδό του. Σημειώνουμε ότι ένα τέτοιο νευρωνικό επίπεδο μπορεί να περιλαμβάνει ένα η περισσότερα convolution layers, τα οποία ακολουθούνται από ένα pooling layer και ο συνδυασμός αυτός μπορεί να επαναληφτεί αρκετές φορές, πριν το τελικό fully-connected layer.

Τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent NN - RNN) είναι ένας ειδικός τύπος νευρωνικών δικτύων, τα οποία έχουν σχεδιαστεί, έτσι ώστε να έχουν μνήμη (δηλαδή η συμπεριφορά τους εξαρτάται από την προηγούμενή τους εμπειρία). Ειδικοί τύποι των νευρωνικών αυτών δικτύων είναι τα δίκτυα που αποτελούνται από κόμβους με σύντομη και μακροπρόθεσμη μνήμη (Long-Short-Term Memory - LSTM) και Περιφραγμένες Επαναλαμβανόμενες Οντότητες (Gated Recurrent Units - GRU). Το LSTM μοντέλο [33] εκμεταλλεύεται την ιδέα της εισαγωγής αυτο-βρόχων (self-loops) για τη δημιουργία μονοπατιών και διευκολύνει τη ροή της πληροφορίας μεταξύ των κόμβων χωρίς αυτή να μεταβάλλεται, ενώ ένα ειδικό στοιχείο/πύλη (forget gate) καθορίζει εάν θα πρέπει η πληροφορία αυτή να προστεθεί στην κατάσταση του κόμβου (cell state) ή όχι. Το GRU μοντέλο [34] βασίζεται στη δημιουργία μονοπατιών, στα οποία η παράγωγος σε βάθος χρόνου παραμένει σταθερή. Συνεπώς, επιτρέπει στα δίκτυα να συσσώρευαν πληροφορία με το πέρασμα του χρόνου και την αξιοποιούν για την υποστήριξη ενός συγκεκριμένου σκοπού. Τα LSTM και GRU υπερτερούν των RNN, καθώς μπορούν να διαχειριστούν καλύτερα τις αλληλεξαρτήσεις μεταξύ των λέξεων, ειδικά όταν αυτές απέχουν αρκετά μεταξύ τους. Ωστόσο, η διαδικασία εκπαίδευσης των δικτύων αυτών (τόσο των LSTM και GRU όσο και των RNN) είναι, εν γένει, αργή (ακόμη και μέσω της χρήσης ενός επιταχυντή) και δύσκολη, εξαιτίας της εξαφάνισης ή έκρηξης της κλίσης (gradient vanish / explosion).

2.3.2.2. Εκπαίδευση Βαθιών Νευρωνικών Δικτύων

Η εκπαίδευση των νευρωνικών δικτύων είναι μία επαναληπτική διαδικασία, σύμφωνα με την οποία αλλάζουμε τις τιμές των παραμέτρων (βαρών του δικτύου) προς την αντίθετη κατεύθυνση της κλίσης της αντικειμενικής συνάρτησης (gradient descent method). Συνεπώς, δύο βασικοί παράγοντες που επηρεάζουν την μεταβολή των τιμών των βαρών του νευρωνικού δικτύου είναι η κατεύθυνση και κυρίως η τιμή της κλίσης (gradient) καθώς και ο βαθμός μάθησης (learning rate). Σε κάθε επανάληψη, η τιμή της κλίσης μπορεί να υπολογιστεί χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα (batch gradient descent), μέρος αυτών (mini-batch gradient descent) ή ακόμη και ένα μόνο δεδομένο (Stochastic Gradient Descent - SGD) [35]. Επίσης, η τιμή του βαθμού μάθησης επηρεάζει σημαντικά τη διαδικασία της εκπαίδευσης και μπορεί να είναι σταθερή ή να μειώνεται συνεχώς (ειδικά στην περίπτωση του SGD, ώστε να υπάρξει σύγκλιση σε κάποιο τοπικό ελάχιστο). Παραλλαγές της παραπάνω μεθόδου όπως είναι το Momentum [36] και Adagrad [37] λαμβάνουν υπόψη την κατεύθυνση προς την οποία κινούμαστε και τις μεταβολές που γίνονται στις παραμέτρους αντίστοιχα, έτσι ώστε να επιταχύνουμε τη διαδικασία της μάθησης. Μια από τις πιο γνωστές

παραλλαγές του SGD είναι η μέθοδος ADAM (Adaptive Moment Estimation) [38], η οποία πρακτικά συνδυάζει τις δύο παραπάνω προσεγγίσεις, λαμβάνοντας υπόψη το ιστορικό των αλλαγών και τις παραμέτρους που αλλάζουν συχνά τιμή.

Η χρήση ενός μικρού τμήματος/δέσμης δεδομένων (mini-batch) για την ανανέωση των τιμών των βαρών του νευρωνικού δικτύου συνήθως προτιμάται, καθώς προσφέρει μια υπολογιστικά αποτελεσματική λύση για την εκπαίδευσή του δικτύου. Ωστόσο, το γεγονός ότι χρησιμοποιείται κάθε φορά ένα μικρό μέρος των διαθέσιμων δεδομένων προκαλεί αλλαγές στην κατανομή των δεδομένων εισόδου του κάθε επιπέδου του νευρωνικου δικτύου (το φαινόμενο αυτό είναι γνωστό ως covariance shift), το οποίο δυσχεραινει τη διαδικασία της μάθησης. Χρησιμοποιώντας την τεχνική της Κανονικοποίησης των Δεδομένων Εισόδου (Batch Normalization) [39] σε κάθε επίπεδο του νευρωνικούς δικτύου, μπορούμε να περιορίσουμε το παραπάνω πρόβλημα και να επιταχύνουμε τη διαδικασία της μάθησης. Επίσης, πολλές φορές, «ανακατεύουμε» τα δεδομένα εισόδου, πριν την εκπαίδευση του μοντέλου, καθώς η σειρά με την οποία αυτά εμφανίζονται επηρεάζει τη διαδικασία της μάθησης (curriculum learning) [40].

2.3.2.3. Αρχιτεκτονικές Νευρωνικών Δικτύων

Ένας κωδικοποιητής-αποκωδικοποιητής (encoder-decoder) [41] αποτελείται από δύο οντότητες, τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder) και μας διευκολύνει στη μετάφραση μιας μεταβλητού μήκους πρότασης σε μια άλλη (sequence to sequence problem). Ο κωδικοποιητής δέχεται ως είσοδο την μεταβλητού μήκους πρόταση και παράγει μια προκαθορισμένου μήκους διανυσματική αναπαράσταση αυτής. Ο αποκωδικοποιητής ακολούθως χρησιμοποιεί τη διανυσματική αυτή αναπαράσταση της πρότασης, για να παράγει μια νέα μεταβλητού μήκους πρόταση. Ένας αυτο-κωδικοποιητής (Autoencoder) [42] βασίζεται στο παραπάνω μοντέλο ενός κωδικοποιητή-αποκωδικοποιητή και αποσκοπεί στη δημιουργία μιας πιο συνοπτικής αλλά σύγχρονως περιεκτικής αναπαράστασης των δεδομένων εισόδου.

Ένας μετατροπέας (Transformer) [43] ακολουθεί την αρχιτεκτονική ενός κωδικοποιητή-αποκωδικοποιητή (encoder-decoder) και, σε συνδυασμό με τον μηχανισμό της προσοχής (Attention Mechanism)[12], καταφέρνει να εντοπίσει τις αλληλεπιδράσεις μεταξύ των επιμέρους στοιχείων της πρότασης (κωδικοποιητής) και ακολούθως να χρησιμοποιήσει την πληροφορία αυτή για τη δημιουργία μιας νέας πρότασης (αποκωδικοποιητής). Για να το πετύχει αυτό, λαμβάνει υπόψη τόσο την εκάστοτε λέξη (κωδικοποίηση λέξης) όσο και τη θέση στην οποία βρίσκεται και έπειτα περνά την πληροφορία αυτή μέσα

[12] All you need to know about 'Attention' and 'Transformers' — In-depth Understanding — Part 1, https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-1-552f0b41d021

από αρκετά (για την ακρίβεια 6) πανομοιότυπα επίπεδα κωδικοποίησης και αποκωδικοποίησης. Κάθε επίπεδο κωδικοποίησης περιλαμβάνει έναν μηχανισμό προσοχής πολλαπλής κεφαλής (Multi-head Attention Mechanism)[13] ακολουθούμενο από ένα νευρωνικό δίκτυο εμπρόσθιας φόρτωσης (feed-forward neural network), αφού πρώτα γίνει κάποιου είδους κανονικοποίηση (layer normalization) [44]. Κάθε επίπεδο αποκωδικοποίησης περιλαμβάνει επιπρόσθετα ένα μηχανισμό προσοχής (masked multi-head attention), που επιδρά πάνω στην έξοδο του κωδικοποιητή και ακολουθείται από ένα επίπεδο κανονικοποίησης.

Τα παραγωγικά-αντιπαλότητας δίκτυα (Generative Adversarial Networks - GANs) [45] αποτελούνται από δύο μέρη, τον παραγωγό (generator) και τον διευκρινιστή (discriminator). Ο παραγωγός προσπαθεί να «κοροϊδέψει» τον διευκρινιστή, δημιουργώντας ψεύτικα δεδομένα. Ο διευκρινιστής ακολούθως προσπαθεί να ξεχωρίζει τα πραγματικά από τα ψεύτικα δεδομένα. Μέσω της διαδικασίας της μάθησης, ο παραγωγός μαθαίνει να δημιουργεί δείγματα που είναι πολύ κοντά στα πραγματικά δεδομένα, ενώ ο διευκρινιστής μαθαίνει να μπορεί να τα ξεχωρίζει. Ωστόσο, υπάρχει κίνδυνος να σταματήσει η εκπαίδευση των δύο δικτύων, όταν το ένα δίκτυο υπερτερεί του άλλου. Γι' αυτό, έχουν προταθεί διάφορες παραλλαγές της αρχιτεκτονικής αυτής, όπως είναι για παράδειγμα τα Wasserstein GAN (WGAN) [46]. Στα διπλής κατεύθυνσης παραγωγικά-αντιπαλότητας δίκτυα (Bidirectional Generative Adversarial Networks - BiGANs) [47] χρησιμοποιείται επιπρόσθετα ένας κωδικοποιητής για την κωδικοποίηση των δεδομένων, ενώ ο διευκρινιστής μαθαίνει να τα ξεχωρίζει, λαμβάνοντας υπόψη τόσο τα δεδομένα όσο και την κωδικοποιημένη τους μορφή.

Άλλες γνωστές αρχιτεκτονικές νευρωνικών δικτύων (ειδικά στον χώρο της επεξεργασίας εικόνας) είναι τα V-Net και U-Net νευρωνικά δίκτυα[14], τα οποία βασίζονται σε μια παραλλαγή ενός CNN δικτύου, γνωστή ως Πλήρες Συνελικτικό Δίκτυο (Fully Convolutional Network - FCN), η οποία βασίζεται στην άνω-δειγματοληψία (up sample) των επιπέδων μείωσης διαστάσεων (pooling layers). Τα Βαθιά Δίκτυα Πίστης (Deep Belief Networks - DBNs)[15] είναι επίσης μια μέθοδος μη επιβλεπόμενης μάθησης. Είναι πιθανολογικά παραγωγικά μοντέλα, που προσφέρουν μια διαφορετική προσέγγιση στη διακριτική/διαχωριστική φύση των παραδοσιακών τεχνικών νευρωνικών δικτύων. Απαρτίζονται από πολλά επίπεδα που αποτελούνται από δίκτυα, που ονομάζονται περιορισμένες μηχανές Μπολτζμαν (Restricted Boltzmann Machines) και αποσκοπούν να βελτιώσουν την κατηγοριοποίηση των δεδομένων.

[13] All you need to know about 'Attention' and 'Transformers' — In-depth Understanding — Part 2, https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-2-bf2403804ada
[14] The Hourglass Network, https://medium.com/@calleris.enrico/hourglass-network-6e74cdb9ce2f
[15] An Overview of Deep Belief Network (DBN) in Deep Learning, https://www.analyticsvidhya.com/blog/2022/03/an-overview-of-deep-belief-network-dbn-in-deep-learning/

2.4. Επεξεργασία Κειμένου – Μοντέλα Γλώσσας

2.4.1. Επεξεργασία Κειμένου

Η επεξεργασία κειμένου είναι ένας ευρύς όρος που περιλαμβάνει ένα μεγάλο σύνολο αλγορίθμων, τεχνικών και εφαρμογών, που έχουν ως στόχο την ανάλυση, επεξεργασία και κατανόηση του φυσικού κειμένου με απώτερο στόχο τη διευκόλυνση υλοποίησης των σχετικών εφαρμογών.

Κατά την επεξεργασία του κειμένου καλούμαστε να διαχειριστούμε αρκετά διαφορετικά θέματα, τα οποία μπορούν να οργανωθούν σε δύο ευρύτερες κατηγορίες [48]. Στην πρώτη κατηγορία ανήκουν οι εργασίες χαμηλού επιπέδου (low-level tasks), οι οποίες έχουν να κάνουν με τον εντοπισμό των ορίων των προτάσεων (sentence boundary detection), τον εντοπισμό των επιμέρους συστατικών, όπως είναι οι λέξεις και τα σημεία στίξης (tokenization), καθώς και το μέρος του λόγου που αυτά ανήκουν (part of speech tagging), τη διαχείριση της μορφολογίας των λέξεων (morphology) καθώς και την οργάνωσή τους σε μεγαλύτερες φράσεις (swallow chunking). Στη δεύτερη κατηγορία ανήκουν οι εργασίες υψηλού επιπέδου (high-level tasks) και περιλαμβάνουν την αναγνώριση και διόρθωση λεκτικών και γραμματικών λαθών (spelling/grammatical error identification and correlation), την αναγνώριση οντοτήτων (Named Entity Recognition) και των μεταξύ τους σχέσεων (Relation Extraction), τη διαχείριση των διαφορετικών εννοιών των λέξεων/φράσεων (word sense disambiguation) ή των συντομεύσεων αυτών (abbreviations sense disambiguation), την αναγνώριση της ύπαρξης άρνησης ή αβεβαιότητας στο κείμενο (negation and uncertainty identification), καθώς και την αναγνώριση του συναισθήματος αυτού που τα έχει γράψει (sentiment analysis) ή, εν γένει, την εξαγωγή χρήσιμης πληροφορίας από το κείμενο (information extraction).

Η ύπαρξη δημοσίως διαθέσιμων δεδομένων (π.χ., stop words, λέξεις ή φράσεις που συνήθως έχουν αρνητική ή θετική σημασία) και οντολογιών [49] (π.χ. λεξιλόγια γενικού σκοπού όπως το WordNet [50] και θησαυροί όρων ενός συγκεκριμένου πεδίου, όπως το MeSH[16]), σε συνδυασμό με την ύπαρξη γενικών κανόνων που διέπουν την οργάνωση του κειμένου και των όρων αυτού (π.χ. γραμματική της αγγλικής γλώσσας και ονοματολογία όρων ενός συγκεκριμένου επιστημονικού πεδίου, όπως είναι τα φάρμακα), έχει συμβάλει σημαντικά στη βαθύτερη κατανόηση του κειμένου (text mining) και την υλοποίηση των παραπάνω εφαρμογών υψηλού επιπέδου. Ωστόσο, τα τελευταία χρόνια η συνεισφορά της μηχανικής μάθησης στον χώρο αυτό είναι ιδιαίτερα σημαντική στην υλοποίηση των παραπάνω εργασιών (τόσο του

[16] Medical Subject Headings (MeSH) thesaurus, https://www.nlm.nih.gov/mesh/meshhome.html

χαμηλού όσο και του υψηλού επιπέδου), το οποίο μπορεί να γίνει κατανοητό, αν αναλογιστούμε ότι οι κανόνες πολλές φορές δεν ακολουθούνται από τους χρήστες (ειδικά κατά την ανεπίσημη επικοινωνία των χρηστών στα κοινωνικά δίκτυα). καθώς και ότι η γλώσσα που χρησιμοποιούμε συνεχώς εξελίσσεται.

2.4.2. Διανυσματική Αναπαράσταση Λέξεων

Η διανυσματική αναπαράσταση των λέξεων (aka word embeddings) δεν είναι κάτι νέο. Από τις αρχές του 20ου αιώνα η συσχέτιση μιας λέξης με ένα διάνυσμα πραγματικών αριθμών, όπου ο κάθε αριθμός σχετίζεται με μία διαφορετική πτυχή της λέξης, έχει λάβει ιδιαίτερη προσοχή και η δημοσίευση του Bengio κ.α., 2003 [51] επηρέασε σημαντικά την έρευνα στον χώρο αυτό. Ωστόσο, ήταν η εργασία του Mikolov κ.α., 2013 [52][53], που έδωσε σημαντική ώθηση στη διανυσματική αναπαράσταση των λέξεων, μέσω της χρήσης δύο διαφορετικών μοντέλων, γνωστά ως Continuous Bag of Words (CBOW) και Skip-Gram Model. Στην πρώτη περίπτωση, εκπαιδεύεται ένα Νευρωνικό Δίκτυο (με ένα μόνο κρυφό επίπεδο), έτσι ώστε να μπορεί να προβλέψει μία λέξη με βάση το σύνολο των λέξεων που προηγούνται και έπονται αυτής στο κείμενο. Στη δεύτερη περίπτωση, η κάθε λέξη χρησιμοποιείται για να προβλέψει τις άλλες λέξεις που πιθανώς να πλαισιώνουν τη λέξη αυτή (συμφραζόμενα). Αξίζει να αναφέρουμε ότι αυτά τα δύο μοντέλα μοιάζουν, ωστόσο κανένα από αυτά δε θεωρείται ανώτερο του άλλου. Για περισσότερες λεπτομέρειες για τα παραπάνω μοντέλα και τον τρόπο εκπαίδευσης των νευρωνικών δικτύων, οι αναγνώστες παραπέμπονται στην ανάγνωση του εγγράφου του Xin Rong, 2014 [54].

Η μάθηση της διανυσματικής αναπαράστασης κάθε λέξης προκύπτει από την εκπαίδευση ενός δικτύου με βάση ένα μεγάλο σύνολο από έγγραφα. Αν και τα έγγραφα αυτά παίζουν πρωταρχικό ρόλο στην εκπαίδευση του δικτύου, η διανυσματική αναπαράσταση των λέξεων δεν εξαρτάται σε μεγάλο βαθμό από αυτά, όπως φάνηκε στην εργασία του Yanshan Wand κ.α., 2018 [55]. Επομένως, για την υλοποίηση ενός συστήματος, όπως αυτό της εργασία μας, μπορούμε να χρησιμοποιήσουμε τη διανυσματική αναπαράσταση των λέξεων που έχει δημιουργηθεί με βάση κάποιο άλλο σύνολο δεδομένων, όπως αυτά που είναι διαθέσιμα στον ακόλουθο σύνδεσμο [56] και έχουν δημιουργηθεί από την επεξεργασία ειδησεογραφικών άρθρων. Η διανυσματική αναπαράσταση των φράσεων μπορεί να προκύψει είτε με την εκπαίδευση του νευρωνικού δικτύου, ώστε να αναγνωρίζει φράσεις αντί για λέξεις είτε με τον συνδυασμό της διανυσματικής αναπαράστασης των λέξεων που τις απαρτίζουν.

Μια διαφορετική προσέγγιση παρουσιάστηκε από τον Bojanowski κ.α., 2016 [57], στην οποία δίνεται έμφαση στις επιμέρους συμβολοακολουθίες από τις οποίες αποτελείται κάθε λέξη, και κατ' επέκταση, μας δίνει τη δυνατότητα να εξάγουμε χρήσιμη πληροφορία για κάθε μία από αυτές, με βάση τα επιμέρους

χαρακτηριστικά της (χαρακτήρες, συλλαβές, κτλ). Συνεπώς, μπορούμε να πετύχουμε καλύτερη διανυσματική αναπαράσταση των λέξεων, αξιοποιώντας τόσο τις λέξεις που τις πλαισιώνουν όσο και τις συμβολοακολουθίες από τις οποίες αυτές απαρτίζονται, όπως φάνηκε στην εργασία του Yijia Zhang κ.α., 2019 [58]. Ο Jeffrey Pennington κ.α., 2014 [59] ακολούθησε μια αρκετά διαφορετική προσέγγιση για τη δημιουργία των διανυσματικών αναπαραστάσεων των λέξεων, στην οποία δίνεται έμφαση σε όλο το κείμενο στο οποίο βρίσκεται η λέξη (GloVe). Κατ' επέκταση, με την προσέγγιση αυτή μπορούμε να εστιάσουμε στα ευρύτερα χαρακτηριστικά των λέξεων.

Τα word2vec και Glove μοντέλα μάς επιτρέπουν να μάθουμε μια, ανεξάρτητα του πλαισίου στο οποίο αυτή βρίσκεται (context-independent), διανυσματική αναπαράσταση κάθε λέξης. Το ELMo (Embeddings from Language Models) [60] έχει σχεδιαστεί κατά τέτοιο τρόπο, ώστε να μπορεί να διαχειριστεί την πολυσημία των λέξεων. Για να το πετύχουν αυτό, οι συγγραφείς του παραπάνω έργου βασίστηκαν στους χαρακτήρες των λέξεων και χρησιμοποίησαν ένα «2-επίπεδων» διπλής κατεύθυνσης (bi-directional) LSTM δίκτυο, το οποίο εκπαιδεύτηκε, έτσι ώστε να μπορεί να προβλέψει την επόμενη λέξη με βάση την προηγούμενη (forward LM) και το αντίστροφο (backward LM). Η διανυσματική αναπαράσταση των λέξεων προκύπτει από τον συνδυασμό των τιμών των εσωτερικών καταστάσεων των LSTM κόμβων των δύο δικτύων.

Το διπλής κατεύθυνσης μοντέλο κωδικοποίησης ενός μετατροπέα (Bidirectional Encoder Representation from Transformer - BERT) [61] δημιουργεί μια διανυσματική αναπαράσταση των λέξεων, λαμβάνοντας υπόψη τα συμφραζόμενα που προηγούνται ή έπονται της εκάστοτε λέξης στο κείμενο, μέσω της χρήσης ενός κωδικοποιητή (ένα από τα δύο βασικά τμήματα ενός μετατροπέα), ο οποίος εστιάζει στα σχετικά στοιχεία της εκάστοτε λέξης (attention mechanism). Η δημιουργία του BERT μοντέλου βασίζεται στα ακόλουθα δύο βήματα. Στο πρώτο βήμα, που ονομάζεται προ-εκπαίδευση (pre-training), χρησιμοποιούνται μη-επισημασμένα δεδομένα για τη δημιουργία ενός μοντέλου το οποίο σε ένα δεύτερο βήμα βελτιστοποιείται (fine-tuning), έτσι ώστε να μπορεί να καλύψει καλύτερα τις ανάγκες ενός συγκεκριμένου σκοπού. Ειδικότερα, στο πρώτο βήμα (μη επιβλεπόμενη μάθηση) το μοντέλο εκπαιδεύεται, έτσι ώστε (α) να μπορεί να προβλέψει ορισμένες λέξεις των προτάσεων (της τάξεως του 15%), οι οποίες έχουν τυχαία αποκρυφτεί (masked language model) και (β) να μπορεί να προβλέψει εάν μία πρόταση έπεται μιας άλλης (next sentence prediction). Για τον σκοπό αυτό, οι λέξεις της πρότασης αρχικά αντικαθιστώνται από διανύσματα που προκύπτουν από τον συνδυασμό των προ-εκπαιδευμένων διανυσματικών αναπαραστάσεων (token embeddings όπως για παράδειγμα τα WordPiece embeddings [62]) και των διανυσματικών αναπαραστάσεων της πρότασης στην οποία ανήκει η λέξη (sentence embeddings) και της

θέσης που έχει η λέξη σε αυτήν (position embeddings). Η έξοδος του μοντέλου ακολούθως συγκρίνεται με την αναμενόμενη και ακολούθως γίνονται οι σχετικές αλλαγές στις παραμέτρους του μοντέλου. Στο δεύτερο βήμα (επιβλεπόμενη μάθηση), προστίθεται ένα ακόμη επίπεδο και ακολούθως το μοντέλο εκπαιδεύεται για τις ανάγκες ενός συγκεκριμένου προβλήματος επεξεργασίας φυσικής γλώσσας. Κατά τη διαδικασία αυτή, οι παράμετροι του BERT μοντέλου μεταβάλλονται, έτσι ώστε να πετυχαίνεται το καλύτερο αποτέλεσμα.

Η εξαγωγή γνώσης από βιοϊατρικά κείμενα (biomedical text mining) είναι μια αρκετά δύσκολη διαδικασία και οι υπάρχουσες γενικού σκοπού διανυσματικές αναπαραστάσεις των λέξεων/φράσεων δυσκολεύονται να καλύψουν τις ανάγκες που υπάρχουν σε αυτό το πεδίο γνώσης, εξαιτίας της ιδιαιτερότητας που παρουσιάζουν τα κείμενα που ανήκουν στο πεδίο αυτό (π.χ., υπάρχουν αρκετοί όροι που τους συναντάμε στο πεδίο αυτό και έχουν ιδιαίτερη σημασία). Για τον σκοπό αυτό, δημιουργήθηκε το BioBERT μοντέλο [63]. Το μοντέλο αυτό έχει ακριβώς την ίδια αρχιτεκτονική με το BERT μοντέλο και οι τιμές των παραμέτρων του αρχικοποιήθηκαν με αυτές που είχαν προκύψει από την εκπαίδευση του BERT μοντέλου, με βάση τα άρθρα από το Wikipedia. Ωστόσο, στην περίπτωση αυτή, η προ-εκπαίδευση του BioBERT μοντέλου συνεχίστηκε με βάση τα άρθρα που έχουν δημοσιευτεί στο PubMed[17]. Ακολούθως, το μοντέλο βελτιστοποιήθηκε, έτσι ώστε να μπορεί να υποστηρίξει την Αναγνώριση Ονομαστικών Οντοτήτων (Named Entity Recognition), την Εξαγωγή Σχέσεων (Relation Extraction) και την Απάντηση Ερωτημάτων (Question Answering). Όπως έδειξαν οι μετρήσεις που έγιναν, το νέο αυτό μοντέλο πετυχαίνει αρκετά καλύτερα αποτελέσματα τόσο από το BERT μοντέλο όσο και από άλλα πρόσφατα γνωστά μοντέλα γλώσσας.

[17] PubMed, https://pubmed.ncbi.nlm.nih.gov/

3. Μεθοδολογία

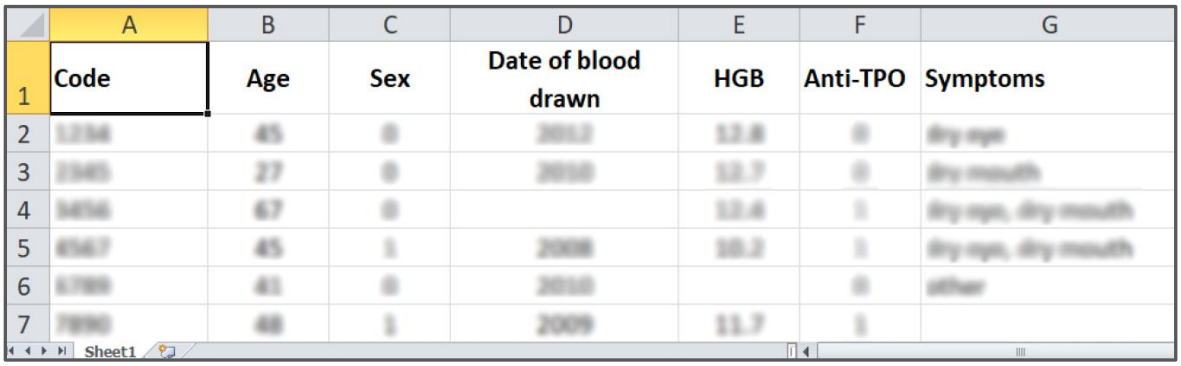
Στα πλαίσια της εργασίας αυτής ασχοληθήκαμε με την εναρμόνιση ιατρικών δεδομένων που προέρχονται από 2 διαφορετικά κέντρα έρευνας και αφορούν ασθενείς που πάσχουν από ένα συγκεκριμένο σύνδρομο, το πρωτοπαθές Σύνδρομο Σιόγκρεν (primary Sjögren Syndrome - pSS). Για τον σκοπό αυτό, αρχικά αναπτύχθηκε ένα μοντέλο για την κοινή αναπαράσταση των δεδομένων των ασθενών που προέρχονται από τους οργανισμούς αυτούς και ακολούθως υλοποιήθηκαν οι μηχανισμοί εκείνοι που επιτρέπουν στους χρήστες του εκάστοτε οργανισμού να εκφράσουν τα δεδομένα που έχουν στην κατοχή τους, με βάση τους όρους του μοντέλου αυτού. Ιδιαίτερο ενδιαφέρον στην παραπάνω διαδικασία έχουν οι τεχνικές μηχανικής μάθησης (συμπεριλαμβανομένων των τεχνικών επιβλεπόμενης και μη επιβλεπόμενης μάθησης), που χρησιμοποιήθηκαν για τη βαθύτερη κατανόηση των δεδομένων των ασθενών αλλά κυρίως για να επιταχύνουν τη διαδικασία της εναρμόνισης των δεδομένων και την αποφυγή πιθανών λαθών ή παραλήψεων.

3.1. Δεδομένα Οργανισμών

Τα δεδομένα των ασθενών προέρχονται από δύο διαφορετικά κέντρα έρευνας στον χώρο της υγείας και ήταν διαθέσιμα με τη μορφή MS Excel αρχείων. Πιο συγκεκριμένα, σε καθέναν από τους δύο αυτούς οργανισμούς υπήρχε ένας υπεύθυνος, ο οποίος είχε τοποθετήσει τα δεδομένα που είχε συγκεντρώσει για τους ασθενείς που «είχε» υπό την επίβλεψή του σε ένα MS Excel αρχείο, το οποίο είχε προκαθορισμένη και συνάμα αρκετά ευέλικτη μορφή. Ειδικότερα, κάθε MS Excel αρχείο είχε ένα μόνο φύλλο, στο οποίο τοποθετήθηκαν όλα τα δεδομένα των ασθενών, τα οποία ήταν σχετικά με το σύνδρομο που μελετήθηκε και εξυπηρετούσαν τους σκοπούς της έρευνας που ήθελαν να πραγματοποιήσουν. Οι γραμμές του φύλλου αυτού αντιστοιχούσαν στους ασθενείς και οι στήλες στα χαρακτηριστικά που είχαν καταγραφεί για κάθε ασθενή. Στο αρχείο αυτό τα ονόματα των χαρακτηριστικών είχαν τοποθετηθεί στην πρώτη γραμμή, ενώ στις υπόλοιπες γραμμές υπήρχαν τα δεδομένα των ασθενών.

Ένα παράδειγμα από ένα τέτοιο αρχείο φαίνεται στην Εικόνα 2. Στην εικόνα αυτή φαίνονται ξεκάθαρα (ένα υποσύνολο από) τα πεδία που έχουν καταγραφεί για κάθε ασθενή στον συγκεκριμένο οργανισμό. Ωστόσο, τα δεδομένα των ασθενών δεν φαίνονται σκοπίμως, καθώς σε αυτά επιτρέπεται η πρόσβαση μόνο στους υπεύθυνους του κάθε οργανισμού και η παράμετρος αυτή λήφθηκε σοβαρά υπόψη στην προσέγγιση που ακολουθήθηκε για την εναρμόνιση των δεδομένων των ασθενών. Σημειώνουμε ότι η μορφή αυτή του

αρχείου που ακολουθήσαμε επιτρέπει στους οργανισμούς να τοποθετήσουν τα δεδομένα των ασθενών που έχουν υπό την επίβλεψή τους και, ειδικότερα, τα χαρακτηριστικά εκείνα των ασθενών που θα ήθελαν να εξετάσουν σε συνεργασία και με τους υπόλοιπους οργανισμούς, με τον τρόπο που έχουν αυτά «εσωτερικά» αναπαρασταθεί στο σύστημα αποθήκευσης δεδομένων που ήδη χρησιμοποιούσαν, συμπεριλαμβανομένων των ονομάτων των παραμέτρων και των τιμών τους.



Εικόνα 2*:* Μορφή ενός MS Excel φύλλου με τα δεδομένα ενός οργανισμού

Στον Πίνακας 1 μπορούμε να δούμε τον αριθμό των ασθενών που υπάρχουν στο αντίστοιχο MS Excel αρχείο του κάθε οργανισμού, καθώς επίσης και το πλήθος των πεδίων που έχουν καταγραφεί για κάθε ασθενή σε καθένα από αυτά. Τα δεδομένα που έχουν καταγραφεί αφορούν βασικά δημογραφικά χαρακτηριστικά των ασθενών, συγκεκριμένες διαταραχές με τις οποίες έχουν διαγνωστεί, θεραπείες ή φαρμακευτικές ουσίες που έχουν λάβει, εργαστηριακές μετρήσεις που έχουν γίνει, ερωτηματολόγια που έχουν συμπληρωθεί και σχετίζονται άμεσα ή έμμεσα με το υπό-εξέταση σύνδρομο.

Κέντρο Έρευνας	**Πλήθος Ασθενών**	**Πλήθος Πεδίων**
Κέντρο Έρευνας 1	586	186
Κέντρο Έρευνας 2	286	163

Πίνακας 1*:* Δεδομένα των δύο Κέντρων Έρευνας

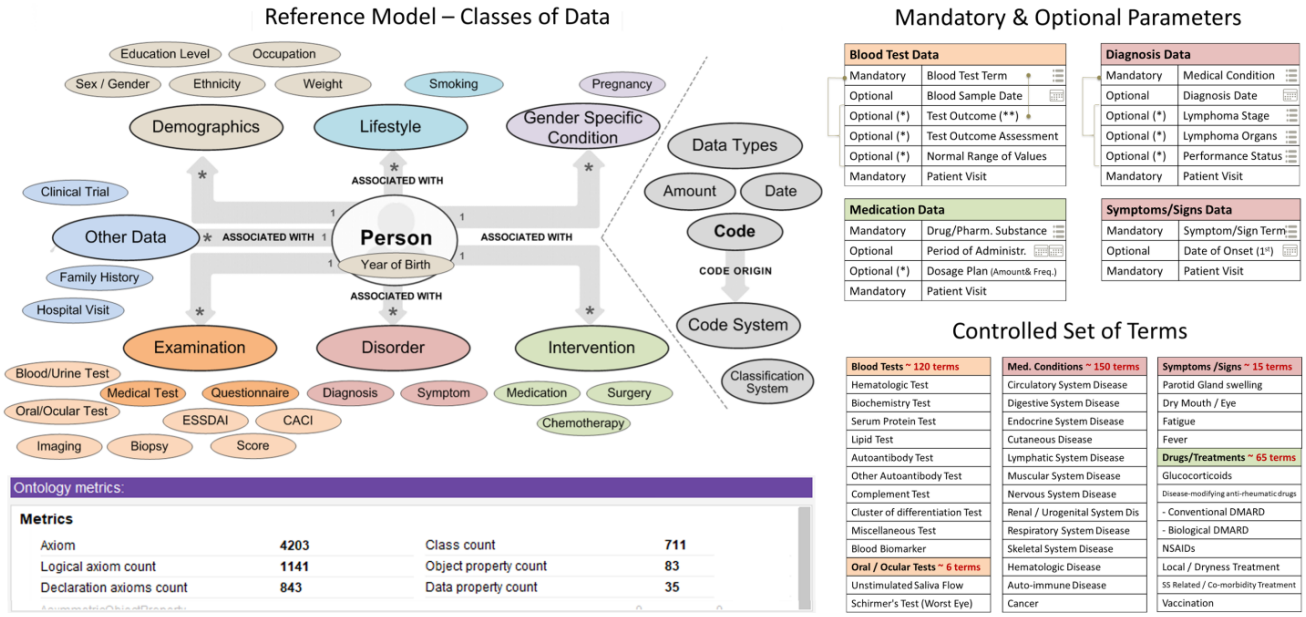
Στο σημείο αυτό να αναφέρουμε ότι οι ασθενείς που υπάρχουν σε καθένα από αυτά τα δύο κέντρα έρευνας και έχουν συμπεριληφθεί στα αρχεία αυτά είναι διαφορετικοί. Επίσης, τα ονόματα (και πολλές φορές οι τιμές) των πεδίων των δύο αυτών αρχείων είναι, εν γένει, διαφορετικά μεταξύ τους. Ενδεικτικά, αναφέρουμε ότι για την ίδια εργαστηριακή μέτρηση που έγινε μια δεδομένη χρονική στιγμή σε ένα υποσύνολο των ασθενών του ενός κέντρου έρευνας καταγράφηκε μόνο η τιμή που μετρήθηκε (χωρίς κάποια επιπρόσθετη πληροφορία για τον τρόπο που έγινε η εξέταση ή τις φυσιολογικές τιμές του εργαστηρίου

τους), ενώ για τους ασθενείς που συμμετείχαν στο άλλο κέντρο έρευνας υπήρχε πληροφορία μόνο για το εάν η τιμή αυτή της εργαστηριακής μέτρησης ήταν φυσιολογική ή όχι. Κατ' επέκταση, για τη σωστή ερμηνεία των τιμών αυτών, ήταν απαραίτητη η επικοινωνία με τους υπεύθυνους σε καθένα από τα δύο αυτά κέντρα έρευνας, λαμβάνοντας υπόψη τους περιορισμούς που υπάρχουν σχετικά με την ασφάλεια και την ιδιωτικότητα των δεδομένων των ασθενών.

3.2. Εναρμόνιση Δεδομένων

3.2.1. Μοντέλο Αναπαράστασης Δεδομένων

Για την Αναπαράσταση των Δεδομένων των Ασθενών που έχουν διαγνωστεί με το πρωτοπαθές Σύνδρομο Σιόγκρεν (pSS) δημιουργήθηκε ένα μοντέλο (Εικόνα 3), σύμφωνα με το οποίο τα δεδομένα των ασθενών μπορούν να οργανωθούν σε ευρύτερες κατηγορίες, όπως για παράδειγμα είναι οι Διαταραχές (όπως είναι οι Ασθένειες και τα Συμπτώματα), οι Παρεμβάσεις (όπως είναι οι Συνταγογραφήσεις Φαρμακευτικών ουσιών) και οι Εξετάσεις που έχουν γίνει, συμπεριλαμβανομένων των Εργαστηριακών Μετρήσεων και των Ερωτηματολογίων (σχετικών με το συγκεκριμένο Σύνδρομο) που συμπληρώθηκαν. Ο σχεδιασμός του μοντέλου αυτού βασιστήκαμε στην ανάλυση της σχετικής βιβλιογραφίας στον χώρο της κλινικής έρευνας και τη μελέτη ευρέως χρησιμοποιούμενων προτύπων στον χώρο αυτό [64].



Εικόνα 3: Μοντέλο Αναπαράστασης Εναρμονισμένων Δεδομένων

Για κάθε μία από τις κατηγορίες αυτές (συμπεριλαμβανομένων των υποκατηγοριών) καθορίστηκαν οι παράμετροι που θα θέλαμε να γνωρίζουμε και κατά πόσο οι παράμετροι αυτές είναι υποχρεωτικές ή όχι. Για παράδειγμα, για μία φαρμακευτική αγωγή θα θέλαμε να γνωρίζουμε τη δραστική ουσία που συνταγογραφήθηκε και προαιρετικά την χρονική περίοδο που λάμβανε ο ασθενής την ουσία αυτή και την δοσολογία αυτής. Επίσης, για κάθε μία από τις παραμέτρους καθορίσαμε το πεδίο τιμών τους, καθώς, επίσης, και τους όρους που μπορούν να χρησιμοποιηθούν. Για παράδειγμα, συμπεριλάβαμε στο μοντέλο τις δραστικές ουσίες που μας ενδιαφέρουν για τους ασθενείς με pSS και / ή τις ευρύτερες κατηγορίες στις οποίες αυτές ανήκουν (π.χ., μη στεροειδή αντιφλεγμονώδη φάρμακα – NSAIDs[18]), σε συνεργασία με ειδικούς στον χώρο της κλινικής έρευνας.

3.2.1.1. Συμβολή Τεχνικών Μηχανικής Μάθησης στον Εντοπισμό Όρων

Για τη δημιουργία του παραπάνω μοντέλου και ειδικότερα για τον προσδιορισμό των όρων των παραμέτρων, εξετάσαμε τη σημασία τόσο των ονομάτων των πεδίων όσο και των τιμών τους που χρησιμοποιούνται και στα δύο κέντρα έρευνας, με τη βοήθεια τεχνικών μηχανικής μάθησης και επεξεργασίας κειμένου. Πιο συγκεκριμένα, αρχικά εντοπίσαμε τα πεδία που υπάρχουν σε καθένα από τα δύο κέντρα έρευνας καθώς και τις τιμές τους (ειδικά στην περίπτωση που η τιμή τους προερχόταν από κάποιο προκαθορισμένο σύνολο τιμών), με βάση τα δεδομένα που υπήρχαν στα δύο MS Excel αρχεία, μέσω των εργαλείων που αναπτύχθηκαν και περιγράφονται αναλυτικά στην ενότητα 3.3. Στη συνέχεια, εξετάσαμε τα παραπάνω δεδομένα εφαρμόζοντας τεχνικές μη-επιβλεπόμενης μάθησης, με βάση τη διανυσματική αναπαράσταση των όρων (τόσο των πεδίων όσο και των δυνατών τιμών τους) για την οργάνωσή τους σε ευρύτερες κατηγορίες, τις οποίες ακολούθως παρουσιάσαμε σε ειδικούς στον χώρο της κλινικής έρευνας, οι οποίοι με τη σειρά τους τις εξέτασαν, προκειμένου να αποφασίσουν εάν θα έπρεπε να συμπεριλάβουν τις σχετικές κατηγορίες και τους συγκεκριμένους όρους στο μοντέλο.

Για τη διανυσματική αναπαράσταση των λέξεων/φράσεων βασιστήκαμε στο Word2Vec μοντέλο. Το μοντέλο αυτό (το οποίο περιγράφηκε συνοπτικά στην ενότητα 2.4.2) λαμβάνει υπόψη τους όρους που περιβάλλουν την κάθε λέξη / φράση (attention mechanism) για τη δημιουργία των διανυσματικών τους αναπαραστάσεων και κατ' επέκταση, μπορεί να διατηρήσει σημαντικό μέρος της σημασιολογίας του κάθε όρου. Επίσης, για την οργάνωση των όρων αυτών σε κατηγορίες (clusters) χρησιμοποιήθηκε ο αλγόριθμος K-means. Ο αλγόριθμος αυτός δέχεται ως είσοδο το πλήθος των clusters (K) τα οποία θέλουμε να δημιουργήσουμε και ακολούθως εντάσσει τους όρους σε αυτά (ανανεώνοντας παράλληλα την τιμή του κέντρου), μέσω μιας

[18] NHS - Non-steroidal anti-inflammatory drugs (NSAIDs), https://www.nhs.uk/conditions/nsaids/

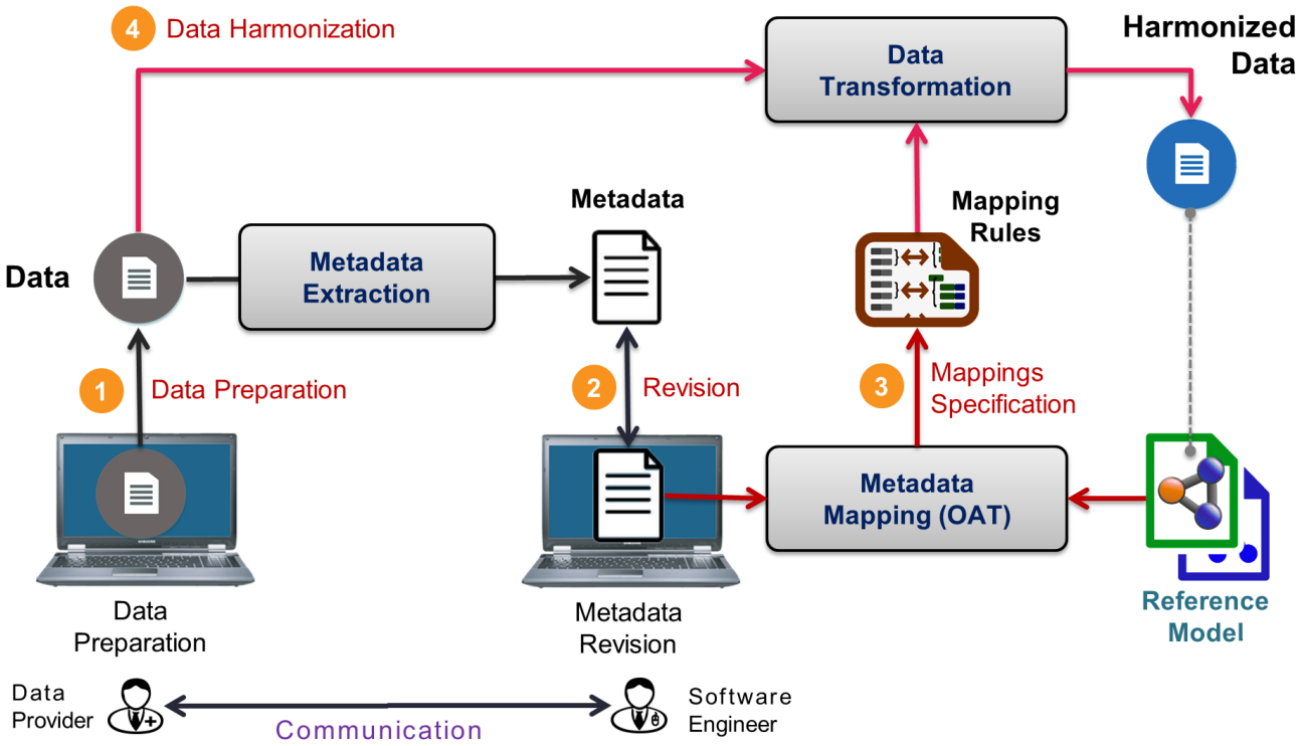
επαναληπτικής διαδικασίας, έτσι ώστε η απόσταση των όρων από το κέντρο της κάθε κατηγορίας να είναι η μικρότερη δυνατή. Για την αναπαράσταση της κάθε κατηγορίας επιλέξαμε τους 5 πιο ευρέως χρησιμοποιούμενους όρους (ή αυτούς που ήταν πιο κοντά στο κέντρο), με βάση τη συχνότητα εμφάνισής τους στην κατηγορία αυτή.

Στο σημείο αυτό αξίζει να αναφέρουμε ότι ο παραπάνω αλγόριθμος εντοπισμού μοτίβων χρησιμοποιήθηκε αρκετές φορές με διαφορετικό αριθμό συνόλων Κ κάθε φορά, δίνοντας έτσι τη δυνατότητα στους χρήστες να εξετάσουν τα διαθέσιμα δεδομένα, «πηγαίνοντας» από το γενικότερο σύνολο (για μικρές τιμές του Κ) προς το ειδικότερο (για μεγάλες τιμές του Κ). Επίσης, η παρουσίαση των συνόλων στον χρήστη βασίστηκε στο πλήθος των οντοτήτων που αυτά περιείχαν (με τα πιο μεγάλα σύνολα να εμφανίζονται στην αρχή της λίστας), ενώ από την παρουσίαση των συνόλων εξαιρέσαμε τα σύνολα εκείνα που περιλάμβαναν έναν και μόνο όρο.

Μέσω της ανάλυσης αυτής, οι ειδικοί στον χώρο της έρευνας είχαν τη δυνατότητα να εξετάσουν εύκολα και γρήγορα τους όρους και γενικότερα την πληροφορία που αποθηκεύεται στο σύνολο των αρχείων των οργανισμών αυτών και ακολούθως να αποφασίσουν τους όρους τους οποίους θα έπρεπε να συμπεριλάβουμε στο μοντέλο, έτσι ώστε να μπορέσουν στη συνέχεια να εκφράσουν τα αντίστοιχα δεδομένα με βάση το «κοινό» μοντέλο. Λαμβάνοντας υπόψη το γεγονός ότι το μοντέλο αυτό αναπτύχθηκε, για να μπορεί να υποστηρίξει την ενοποίηση των δεδομένων των ασθενών που πάσχουν από το συγκεκριμένο σύνδρομο και προέρχονται από διαφορετικά κέντρα έρευνας (πέραν των δύο κέντρων έρευνας που εξετάσαμε), συμπεριλάβαμε στο μοντέλο αρκετούς διαφορετικούς όρους/κατηγορίες, πέραν από αυτούς που συναντήσαμε με βάση υπάρχουσες ευρέως χρησιμοποιούμενες κατηγοριοποιήσεις των όρων του κάθε πεδίου και κυρίως με βάση τις υποδείξεις των ειδικών στον χώρο αυτό.

3.2.2. Διαδικασίας Εναρμόνισης Δεδομένων

Για τις ανάγκες ενοποίησης των δεδομένων αναπτύχθηκαν τρία διαφορετικά εργαλεία (παρουσιάζονται στην Εικόνα 4 και αναλύονται περαιτέρω στις επόμενες υπό-ενότητες), τα οποία επιτρέπουν στους υπεύθυνους του κάθενός από τα δύο κέντρα έρευνας να εκφράσουν αυτόματα τα δεδομένα των ασθενών τους με τα στοιχεία του μοντέλου αναφοράς (Reference Model), που έχει δημιουργηθεί με βάση τους κανόνες συσχέτισης (mapping rules), που έχουν καθοριστεί μεταξύ των παραμέτρων (και των τιμών) των δεδομένων τους και αυτών που έχουν οριστεί στο μοντέλο αναφοράς [65].



Εικόνα 4: Διαδικασία Εναρμόνισης Δεδομένων για ένα Κέντρο Έρευνας

Οι χρήστες έχουν πολύ σημαντικό ρόλο στην παραπάνω διαδικασία εναρμόνισης των δεδομένων και μπορούμε να τους χωρίσουμε σε δύο κατηγορίες με βάση τις γνώσεις που έχουν (π.χ., ιατρικές ή τεχνικές γνώσεις) και κυρίως, τη δυνατότητα να έχουν πρόσβαση στα δεδομένα των ασθενών.

Οι Πάροχοι Δεδομένων (Data Providers), στο πρόβλημα που μελετάται, είναι ειδικοί ερευνητές στον χώρο της ιατρικής, οι οποίοι έχουν πρόσβαση στα δεδομένα ασθενών του κέντρου έρευνας και μπορούν να τα χρησιμοποιήσουν για την επίτευξη συγκεκριμένων σκοπών (π.χ., εις βάθος μελέτη του συνδρόμου με το οποίο έχουν διαγνωστεί οι ασθενείς). Οι χρήστες αυτοί αρχικά ετοίμασαν ένα έγγραφο (στην περίπτωσή μας, ένα MS Excel φύλλο) με τα δεδομένα των ασθενών που θα ήθελαν να εκφράσουν με βάση το νέο μοντέλο που έχει αναπτυχθεί (μοντέλο αναφοράς). Στη συνέχεια, χρησιμοποιώντας το εργαλείο Εξαγωγής Μεταδεδομένων (Metadata Extraction), μπόρεσαν να εξάγουν αυτόματα τα απαραίτητα (για τις ανάγκες της ενοποίησης των δεδομένων) μετα-δεδομένα, όπως είναι τα πεδία (παράμετροι) που έχουν καταγραφεί για κάθε ασθενή και οι πιθανές τιμές τους, τα οποία αποθηκεύτηκαν σε ένα άλλο αρχείο (και αυτό είναι ένα MS Excel φύλλο), το οποίο έχει μια προκαθορισμένη μορφή (βήμα 1). Το αρχείο που προέκυψε, δεν περιέχει προσωπικά δεδομένα και κατ' επέκταση ήταν διαθέσιμο τόσο στους παρόχους δεδομένων όσο και σε ειδικούς στον χώρο της πληροφορικής (Software Engineers), οι οποίοι είχαν τη δυνατότητα να εξετάσουν τα

μεταδεδομένα που είχαν προκύψει και να ζητήσουν τις απαραίτητες διευκρινήσεις από τους παρόχους των δεδομένων, όπως είναι για παράδειγμα η διαδικασία που ακολουθήθηκε για τη συλλογή των δεδομένων και ορισμένες διευκρινίσεις για τη σημασία των παραμέτρων και των τιμών αυτών (βήμα-2).

Οι ειδικοί στον χώρο της πληροφορικής, οι οποίοι έχουν την απαραίτητη τεχνογνωσία και εξοικείωση με τα προγράμματα λογισμικού, συμπεριλαμβανομένου του εργαλείου καθορισμού συσχέτισης μεταξύ μοντέλων (Metadata Mapping Tool), μπόρεσαν ακολούθως να καθορίσουν επακριβώς τον τρόπο σύνδεσης των παραμέτρων αυτών με τις οντότητες του μοντέλου αναφοράς (δηλαδή κατά τέτοιον τρόπο, ώστε οι συσχετίσεις που έχουν καθοριστεί να μπορούν να χρησιμοποιηθούν αυτόματα για την εναρμόνιση των σχετικών δεδομένων), μέσω μιας ημι-αυτόματης διαδικασίας (βήμα-3). Οι κανόνες συσχέτισης που ορίστηκαν χρησιμοποιήθηκαν ακολούθως από τους παρόχους δεδομένων για την έκφραση – μετατροπή των δεδομένων τους (Data Transformation) με βάση το μοντέλο αναφοράς, μέσω μιας εντελώς αυτόματης διαδικασίας (βήμα-4). Τα εναρμονισμένα δεδομένα που προέκυψαν από την παραπάνω διαδικασία αποθηκεύτηκαν σε ένα OWL αρχείο.

3.2.2.1. Συμβολή Τεχνικών Μηχανικής Μάθησης στη Διαδικασία Εναρμόνισης Δεδομένων

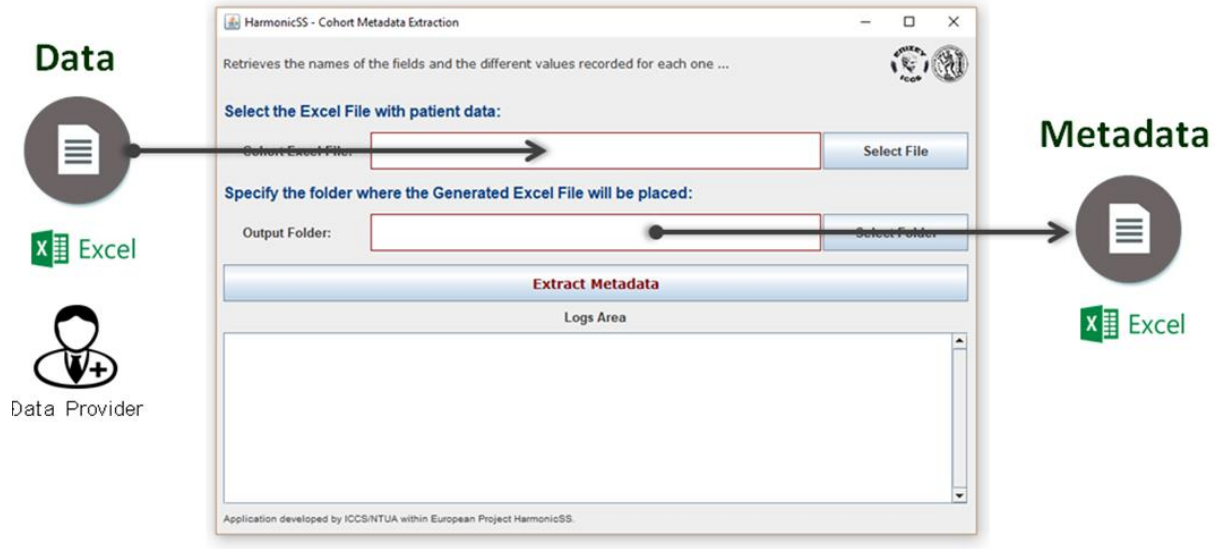
Οι αλγόριθμοι μηχανικής μάθησης και εξόρυξης γνώσης, μαζί με μεθόδους που ανήκουν στο ευρύτερο πεδίο της στατιστικής, χρησιμοποιήθηκαν κατά την παραπάνω διαδικασία, για να παρέχουν μια συνοπτική εικόνα των δεδομένων του καθενός κέντρου έρευνας και τον εντοπισμό εσφαλμένων ή ακραίων τιμών κατά τα βήματα 1 και 2 καθώς επίσης και για να βοηθήσουν τον χρήστη στον καθορισμό συσχετίσεων μεταξύ των στοιχείων των δύο μοντέλων με το μοντέλο αναφοράς που δημιουργήθηκε και κατ' επέκταση την εναρμόνιση των δεδομένων στα βήματα 3 και 4 και θα παρουσιαστούν αναλυτικά στις ακόλουθες υποενότητες κατά την περιγραφή των σχετικών εργαλείων.

3.3. Εργαλεία και Μηχανισμοί

3.3.1. Εργαλείο Εξαγωγής Μεταδεδομένων

Το Εργαλείο Εξαγωγής Μεταδεδομένων (Metadata Extraction Tool) είναι μια Desktop Εφαρμογή, η οποία διαθέτει ένα απλό γραφικό περιβάλλον (Εικόνα 5), μέσω του οποίου οι Πάροχοι Δεδομένων (Data Providers) μπορούν να εντοπίσουν αυτόματα τις παραμέτρους (πεδία) που έχουν καταγραφεί για κάθε οντότητα (στη δική μας περίπτωση, τις παραμέτρους των ασθενών) καθώς και επιπρόσθετη πληροφορία για κάθε μία από τις παραμέτρους αυτές, όπως είναι για παράδειγμα τον τύπο των δεδομένων, το εύρος τιμών (στην

περίπτωση μιας συνεχούς μεταβλητής) ή τις διαφορετικές τιμές που έχουν εντοπιστεί (στην περίπτωση μιας διακριτής μεταβλητής), κ.α.



Εικόνα 5: Γραφικό Περιβάλλον του Εργαλείου Εξαγωγής Μεταδεδομένων

Για την καλύτερη κατανόηση / περιγραφή των παραμέτρων, χρησιμοποιήθηκαν τεχνικές από το ευρύτερο πεδίο της στατιστικής για τον εντοπισμό του κέντρου των δεδομένων (centrality) καθώς και της λοξότητας (skewness), λαμβάνοντας υπόψη και το πεδίο τιμών τους. Γενικά μιλώντας, όταν η τιμή ενός πεδίου ήταν κάποιος ακέραιος ή πραγματικός αριθμός, υπολογίσαμε τη μέση τιμή και τη διασπορά των δεδομένων καθώς επίσης και το εύρος τιμών τους, ενώ, όταν η τιμή του ήταν κάποια άλλη συμβολοακολουθία, υπολογίσαμε το πλήθος των διαφορετικών συμβολοακολουθιών που χρησιμοποιήθηκαν καθώς και τις πέντε πιο ευρέως χρησιμοποιούμενες συμβολοακολουθίες.

Επίσης, χρησιμοποιήθηκε ο αλγόριθμος K-means για τον εντοπισμό πιθανών ακραίων τιμών. Ειδικότερα, οργανώσαμε τις τιμές κάθε παραμέτρου, των οποίων η τιμή ήταν κάποιος πραγματικός αριθμός σε δύο ευρύτερες κατηγορίες και ακολούθως εντοπίσαμε εκείνες οι οποίες περιείχαν έναν πολύ περιορισμένο αριθμό στοιχείων. Επιπρόσθετα, χρησιμοποιήσαμε τον αλγόριθμο FP-growth για τον εντοπισμό συχνών μοτίβων στο σύνολο των δεδομένων των ασθενών. Ο αλγόριθμός αυτός εφαρμόστηκε, λαμβάνοντας υπόψη τον τύπο των παραμέτρων και ειδικότερα την τιμή τους. Έμφαση δόθηκε στις παραμέτρους εκείνες, των οποίων η τιμή προερχόταν από κάποιο σύνολο τιμών (π.χ. NSAIDs presence: YES and drug: NSAIDs). Πιο συγκεκριμένα, για κάθε ασθενή δημιουργήθηκε ένα σύνολο από δεδομένα (π.χ., φύλο, ασθένειες, φάρμακα, κτλ.), τα οποία αποτέλεσαν την είσοδο στον αλγόριθμο εντοπισμού μοτίβων FP-growth. Ο

αλγόριθμος αυτός, αρχικά, αναπαριστά τα δεδομένα με τη μορφή ενός δένδρου (FP-tree), το οποίο, στη συνέχεια, χρησιμοποιείται για τον εντοπισμό των μοτίβων, που παρουσιάζονται παραπάνω από έναν προκαθορισμένο αριθμό φορών (threshold).

Τα μεταδεδομένα, που εξήχθησαν αυτόματα αποθηκεύτηκαν σε ένα άλλο MS Excel αρχείο, το οποίο περιλαμβάνει τρία φύλλα. Στο πρώτο φύλλο υπάρχει πληροφορία για το σύνολο των δεδομένων που έχουν καταγραφεί, όπως είναι για παράδειγμα το μέγεθος του αρχείου με τα αρχικά δεδομένα, η ημερομηνία που εκτελέστηκε η διαδικασία αυτή, το πλήθος των ασθενών που υπήρχαν στο MS Excel αρχείο με τα δεδομένα και το πλήθος των πεδίων που έχουν καταγραφεί για καθέναν από αυτούς. Στο φύλλο αυτό αποθηκεύτηκαν επίσης και τα συχνά εμφανιζόμενα μοτίβα που εντοπίστηκαν. Στο δεύτερο φύλλο (Εικόνα 6) υπάρχουν οι παράμετροι που έχουν καταγραφεί για κάθε μία από τις οντότητες καθώς και επιπρόσθετη πληροφορία για κάθε μία από αυτές (π.χ., τύπος δεδομένων, εύρος τιμών, κτλ.), ενώ στο τρίτο φύλλο υπάρχουν οι διαφορετικές τιμές που έχουν εντοπιστεί για ορισμένες από τις παραμέτρους αυτές.

A	B	C	D	E	F	G	H	I
ID	CATEGORY	SUBCATEGORY	FIELD NAME	DESCRIPTION	DATA TYPE	VALUES (NOTES/CONSTRAINTS)	CAN BE EMPTY	HAS MANY
A			SubjectID		STRING			
B			Gender		STRING	One of: [Female, Male]		
F			Symptoms		STRING	Examples: [dry mouth, ..]	YES	YES
K			Sample Date		DATE	Format: YEAR		
P			IgG > 15 g/L		STRING	One of: [NO, YES]	YES	
R			C3 value g/L		NUMBER	In Range: [0.43 , 1.5]	YES	
S			C4 value g/L		NUMBER	In Range: [0.05 , 0.4]	YES	

Fields will be filled in during the Metadata Analysis

Overview | Fields-Name | Fields-Values

Εικόνα 6: Ένα από τα τρία φύλλα του MS Excel αρχείου των Μεταδεδομένων

Στο σημείο αυτό να αναφέρουμε ότι η πληροφορία που υπάρχει και στα 3 φύλλα μπορεί να τροποποιηθεί κατά την εξέταση των μετα-δεδομένων που έχουν προκύψει από τους ειδικούς στον χώρο της πληροφορικής, σε συνεργασία με τους υπεύθυνους του αντίστοιχου κέντρου έρευνας, με απώτερο στόχο την καλύτερη και πληρέστερη περιγραφή του συνόλου των δεδομένων των ασθενών και των παραμέτρων που έχουν καταγραφεί. Ειδικότερα, στο πρώτο φύλλο μπορούν να προστεθούν επιπρόσθετα πεδία, όπως

είναι για παράδειγμα ο χρήστης που είναι υπεύθυνος για το σύνολο των δεδομένων των ασθενών. Στο δεύτερο φύλλο μπορούν να οργανωθούν οι παράμετροι σε ευρύτερες κατηγορίες και να προστεθεί επιπρόσθετη πληροφορία για κάθε μία από τις παραμέτρους των ασθενών (είτε από τους υπεύθυνους του κάθε κέντρου είτε από τους ειδικούς στον χώρο της πληροφορικής, κατόπιν επικοινωνίας με τους υπεύθυνους του αντίστοιχου κέντρου έρευνας) ή ακόμη και να τροποποιηθούν τα ήδη υπάρχοντα μετα-δεδομένα, εάν αυτό είναι απαραίτητο. Στο τρίτο φύλλο, οι διαφορετικοί όροι που χρησιμοποιούνται σε μία ή περισσότερες παραμέτρους, μπορούν να οργανωθούν καλύτερα και να προστεθεί επιπρόσθετη πληροφορία για καθέναν από αυτούς (π.χ., σύντομη περιγραφή του κάθε όρου).

3.3.2. Εργαλείο Καθορισμού Συσχέτισης

Το Εργαλείο Καθορισμού Συσχέτισης (Metadata Mapping Tool) αποτελεί μια τροποποιημένη/βελτιωμένη έκδοση του Εργαλείου Ευθυγράμμισης Οντολογιών (Ontology Alignment Tool) [66] και επιτρέπει στους χρήστες να γεφυρώσουν το χάσμα που υπάρχει μεταξύ των όρων δύο διαφορετικών μοντέλων. Ειδικότερα, το εργαλείο αυτό επιτρέπει στου χρήστες να καθορίσουν τον τρόπο συσχέτισης των παραμέτρων (και των τιμών τους) του αρχικού μοντέλου (source/institute model) με αυτούς που έχουμε ορίσει στο μοντέλο αναφοράς (target/reference model) με τη μορφή κανόνων συσχέτισης (mapping rules), έτσι ώστε αυτοί να μπορούν ακολούθως να χρησιμοποιηθούν για την εναρμόνιση των δεδομένων των ασθενών.

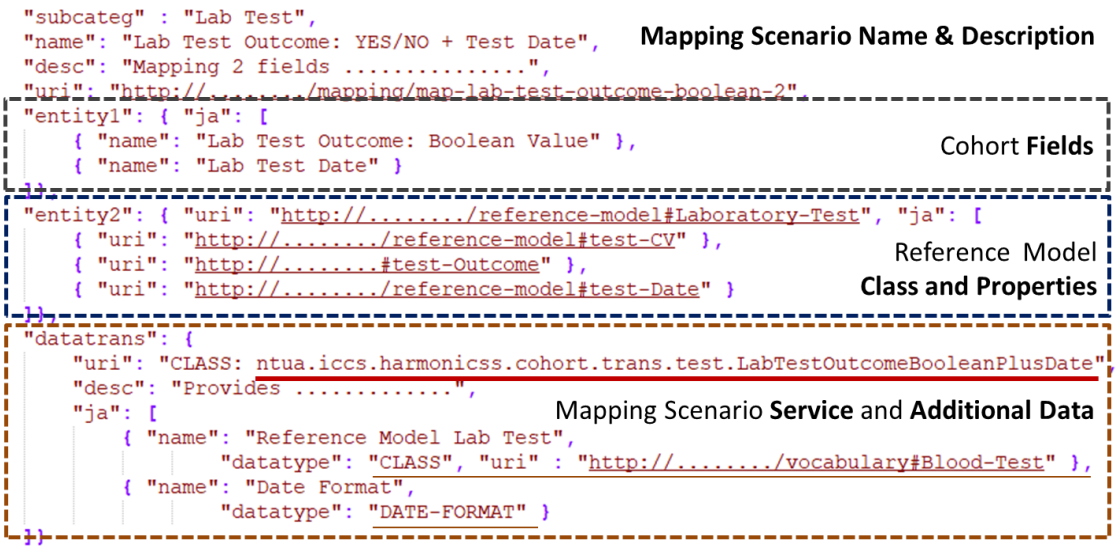
Κάθε κανόνας συσχέτισης καθορίζει τα στοιχεία που συμμετέχουν από τις δύο πλευρές (source & target elements), καθώς επίσης και τη διαδικασία που θα πρέπει να ακολουθήσουμε (transformation) για την έκφραση των αρχικών δεδομένων (source elements), με βάση τα στοιχεία του μοντέλου αναφοράς (target elements). Ένας κανόνας συσχέτισης μπορεί να έχει και επιπρόσθετη πληροφορία για τη σωστή μετατροπή των δεδομένων καθώς επίσης και επιπλέον πληροφορία για τον κανόνα, αυτό καθ' αυτό, όπως είναι για παράδειγμα η προέλευσή του (π.χ., προέρχεται από την αποδοχή ενός προτεινόμενου κανόνα συσχέτισης ή έχει οριστεί εκ του μηδενός, με βάση το γραφικό περιβάλλον της εφαρμογής).

Το εργαλείο αυτό (Εικόνα 7) υποστηρίζει την όλη διαδικασία καθορισμού συσχέτισης (mapping process), επιτρέποντας στους χρήστες να καθορίσουν τα δύο μοντέλα, να εξετάσουν ταυτόχρονα τα στοιχεία των δύο μοντέλων στην ίδια οθόνη (Tab 1), να διαχειριστούν τους πιθανούς τρόπους συσχέτισης μεταξύ των παραμέτρων και των τιμών τους που προτείνονται αυτόματα από το σύστημα (Tab 2), να ορίσουν εκ του μηδενός νέους κανόνες που δεν μπόρεσαν να εντοπιστούν αυτόματα από το εργαλείο αυτό (Tab 3) και τέλος (Tab 4) να εξάγουν τους κανόνες που έχουν καθοριστεί στην επιθυμητή (από τις υποστηριζόμενες) μορφή (π.χ., JSON, XML ή HTML για λόγους παρουσίασης).

Εικόνα 7: Ένα στιγμιότυπο του γραφικού περιβάλλοντος του Εργαλείου Καθορισμού Συσχετίσεων

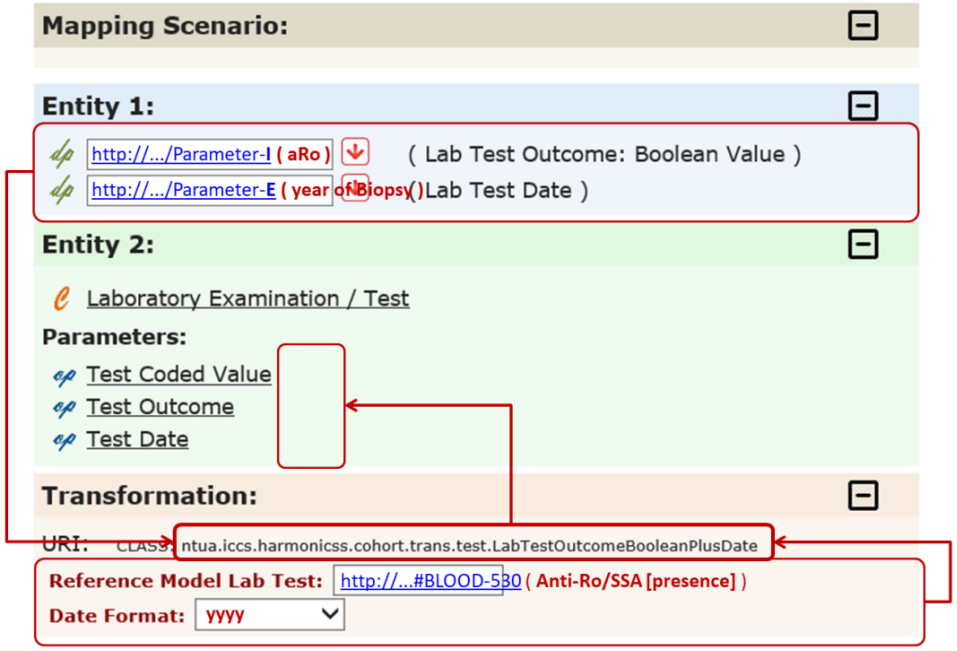
Το εργαλείο αυτό επιτρέπει στους χρήστες να καθορίσουν τόσο απλούς κανόνες, όπως είναι για παράδειγμα, όταν ένας όρος από το ένα μοντέλο έχει ακριβώς την ίδια σημασία με έναν άλλον όρο από το άλλο μοντέλο όσο και πιο περίπλοκους κανόνες, οι οποίοι προϋποθέτουν τον συνδυασμό της πληροφορίας που υπάρχει σε παραπάνω από μια παραμέτρους και την εφαρμογή του κατάλληλου μετασχηματισμού, έτσι ώστε να προκύψουν οι σχετικές τιμές των παραμέτρων του μοντέλου αναφοράς. Για παράδειγμα, τα δεδομένα μιας αιματολογικής εξέτασης μπορεί να είναι διάσπαρτα σε περισσότερα από ένα πεδία (π.χ., ημερομηνία και αποτέλεσμα συγκεκριμένης εξέτασης/μέτρησης) τα οποία θα πρέπει να λάβουμε υπόψη για τη δημιουργία μιας οντότητας του μοντέλου αναφοράς. Για τον σκοπό αυτό, αναπτύχθηκαν διάφορα σενάρια καθορισμού οντοτήτων του μοντέλου αναφοράς (mapping scenarios) και συνδέθηκαν με το εργαλείο αυτό [67]. Τα σενάρια αυτά είναι πρακτικά πρότυπα (templates) για τον καθορισμό μιας οντότητας του μοντέλου αναφοράς, με βάση ένα ή περισσότερα πεδία του μοντέλου της πηγής δεδομένων, τα οποία με τη σειρά τους μπορούν να χρησιμοποιηθούν (instantiated) πολλές φορές κατά τη διαδικασία του καθορισμού της συσχέτισης μεταξύ των οντοτήτων των δύο μοντέλων μέσω του γραφικού περιβάλλοντος.

Στην Εικόνα 8 μπορούμε να δούμε ένα σενάριο συσχέτισης που δημιουργήθηκε, το οποίο καθορίζει το πλήθος και τη σημασία των παραμέτρων που χρειάζεται να γνωρίζουμε, προκειμένου να δημιουργήσουμε την αντίστοιχη οντότητα του μοντέλου αναφοράς καθώς και τη διαδικασία που πρέπει να ακολουθήσουμε και κάποιες επιπρόσθετες παραμέτρους που χρειάζεται να γνωρίζουμε για τη σωστή επεξεργασία των δεδομένων των παραμέτρων του αρχικού μοντέλου.



Εικόνα 8: Ένα Σενάριο Καθορισμό Συσχέτισης για μια Αιματολογική Εξέταση

Το εργαλείο που αναπτύχθηκε επιτρέπει στον χρήστη να επιλέξει ποιο σενάριο συσχέτισης θα ήθελε να χρησιμοποιήσει και ακολούθως ενημερώνει αυτόματα τα πεδία που υπάρχουν στη σελίδα του για τον καθορισμό της συσχέτισης μεταξύ των όρων του μοντέλου δεδομένων του χρήστη και του μοντέλου αναφοράς που αναπτύχθηκε. Όπως φαίνεται και στην Εικόνα 9 ο χρήστης μπορεί να ορίσει μόνο το αριστερό μέρος του κανόνα συσχέτισης (entity 1), καθώς και να δώσει την απαραίτητη πληροφορία για τη σωστή ερμηνεία και επεξεργασία των τιμών των παραμέτρων αυτών.



Εικόνα 9: Καθορισμός Πεδίων ενός Σεναρίου Καθορισμού Συσχέτισης

Στην παραπάνω εικόνα δίνεται ιδιαίτερη έμφαση στον τρόπο επεξεργασίας των δεδομένων των παραμέτρων του αρχικού μοντέλου που ορίζει ο χρήστης, έτσι ώστε να δημιουργηθεί μια οντότητα του μοντέλου αναφοράς και να προσδιοριστούν οι τιμές των παραμέτρων της. Ειδικότερα, οι τιμές των πεδίων που ορίζει ο χρήστης «διαβάζονται» από την κλάση που έχει οριστεί ως συμβολοακολουθίες, τις οποίες επεξεργάζεται περαιτέρω, λαμβάνοντας υπόψη τις τιμές των επιπρόσθετων παραμέτρων που έχουν οριστεί και αφορούν τις τιμές των πεδίων αυτών. Στην κλάση αυτή είναι γνωστός εκ των προτέρων ο τύπος των δεδομένων που έχει ορίσει ο χρήστης και κατ' επέκταση η διαδικασία που πρέπει να ακολουθηθεί για τη δημιουργία της οντότητας που αναφέρεται στο δεξιό μέρος του κανόνα και τον προσδιορισμό των παραμέτρων που επίσης αναφέρονται.

Στο σημείο αυτό να αναφέρουμε ότι το εργαλείο αυτό επιτρέπει στους χρήστες να καθορίσουν τον τρόπο συσχέτισης μεταξύ των οντοτήτων των δύο μοντέλων σε διάφορα στάδια. Αρχικά, μπορούν να εστιάσουν στη βαθύτερη κατανόηση της πληροφορίας που έχει καταγραφεί στα διάφορα πεδία και στο κατά πόσο η πληροφορία αυτή αφορά π.χ. κάποια διαταραχή ή φαρμακευτική ουσία καθώς επίσης και εάν αναφέρεται σε κάποιο κωδικοποιημένο όρο και / ή χρονική στιγμή. Στη συνέχεια, έχοντας εντοπίσει τα μοτίβα που χρησιμοποιούνται και καθορίσει τα σχετικά σενάρια συσχέτισης, ο χρήστης μπορεί να τα χρησιμοποιήσει, για να καθορίσει επ' ακριβώς τον τρόπο σύνδεσης των πεδίων των δύο μοντέλων για τις ανάγκες της εναρμόνισης των δεδομένων. Τέλος, μπορεί να προχωρήσει και στην υλοποίηση των σχετικών μετασχηματισμών που αναφέρονται στα σενάρια (σε κάποια γλώσσα περιγραφής διαδικασιών – procedural language), έτσι ώστε, εν τέλει, να μπορούν οι κανόνες συσχέτισης που έχουν καθοριστεί να χρησιμοποιηθούν από τους παρόχους δεδομένων για την αυτόματη εναρμόνιση των δεδομένων του αντίστοιχου κέντρου έρευνας.

3.3.2.1. Δημιουργία Σεναρίων και Εντοπισμός Πιθανών Συσχετίσεων

Για τον καθορισμό του τρόπου συσχέτισης των παραμέτρων του καθενός από τους δύο οργανισμούς με τις αντίστοιχες του μοντέλου αναφοράς, αναπτύχθηκαν διάφορα σενάρια καθορισμού συσχέτισης, τα οποία ακολούθως χρησιμοποιήθηκαν καμία, μία ή περισσότερες φορές από κάθε οργανισμό. Λαμβάνοντας υπόψη το γεγονός ότι ο καθορισμός το σεναρίων είναι μια, εν γένει, χρονοβόρα διαδικασία, που απαιτεί (α) τον ορισμό των σεναρίων χρησιμοποιώντας την γλώσσα JSON αλλά και (β) την υλοποίηση των μηχανισμών που αναλαμβάνουν την οντολογική αναπαράσταση των τιμών των συγκριμένων παραμέτρων σε κάποια διαδικαστική γλώσσα, όπως είναι η JAVA και η Python, έγινε κάποια προεπεξεργασία των μετα-δεδομένων που εξήχθησαν από κάθε ένα από τα δύο κέντρα έρευνας και έγινε επιλογή των σεναρίων που αναπτύχθηκαν, έτσι ώστε να καλύψουν όλους τους δυνατούς συνδυασμούς που παρατηρήθηκαν, ενώ

προτεραιότητα (στην υλοποίηση) δόθηκε στα μοτίβα εκείνα που χρησιμοποιούνται συχνότερα έναντι των υπολοίπων.

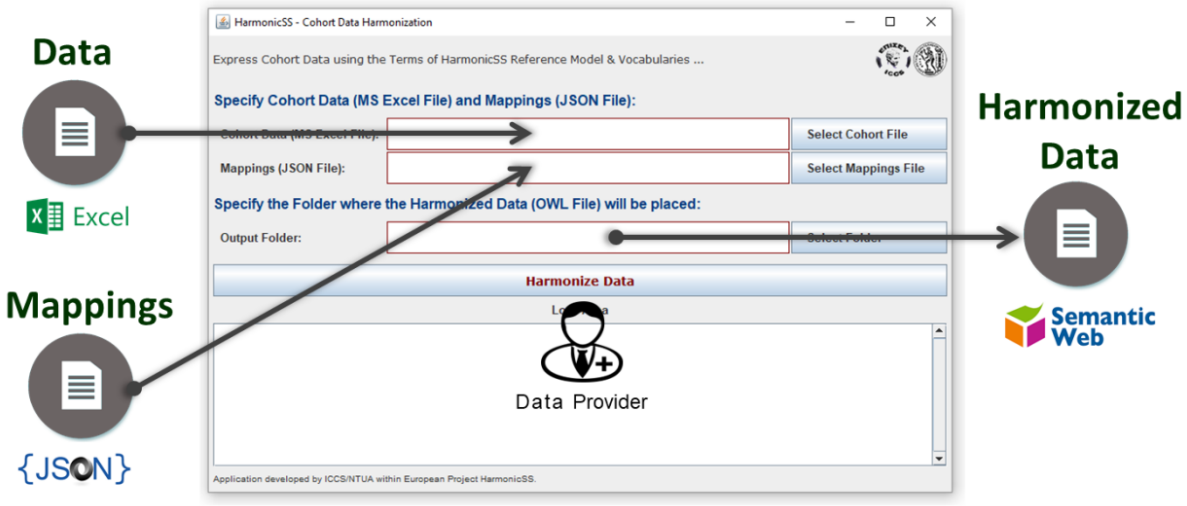
Οι τεχνικές μηχανικής μάθησης χρησιμοποιήθηκαν για τον εντοπισμό πιθανών συσχετίσεων μεταξύ τόσο των παραμέτρων ενός μοντέλου όσο και των τιμών του με τις οντότητες που ορίστηκαν στο μοντέλο αναφοράς. Ειδικότερα, οι τεχνικές μηχανικής μάθησης χρησιμοποιήθηκαν στις ακόλουθες δύο περιπτώσεις: (α) ευθυγράμμιση όρων και (β) δημιουργία σεναρίων συσχέτισης. Στην πρώτη περίπτωση, οι τεχνικές μηχανικής μάθησης χρησιμοποιήθηκαν για την αναζήτηση όρων του μοντέλου αναφοράς που έχουν πιθανώς την ίδια σημασία με αυτή που έχουν στο μοντέλο του καθενός από τα δύο κέντρα έρευνας. Στη δεύτερη περίπτωση, οι τεχνικές μηχανικής μάθησης χρησιμοποιήθηκαν για την αναζήτηση πιθανών σεναρίων συσχέτισης, λαμβάνοντας υπόψη τόσο τον τύπο των δεδομένων που απαιτούνται για τη χρησιμοποίηση καθενός από αυτά όσο και τη σημασία των παραμέτρων και των όρων των δύο μοντέλων. Και στις δύο παραπάνω περιπτώσεις βασιστήκαμε στη διανυσματική αναπαράσταση των ονομάτων των παραμέτρων και των όρων αυτών που συναντήσαμε σε καθένα από τα δύο κέντρα έρευνας. Για τον σκοπό αυτό, χρησιμοποιήθηκε το Word2Vec μοντέλο. Επίσης, λάβαμε υπόψη και τους κανόνες συσχέτισης που έχουν ήδη καθοριστεί.

Οι πιθανές συσχετίσεις, που εντοπίστηκαν ακολουθώντας την παραπάνω διαδικασία, παρουσιάστηκαν στον χρήστη μέσω του γραφικού περιβάλλοντος του εργαλείου που αναπτύχθηκε, έτσι ώστε να μπορέσει ο χρήστης ακολούθως να εξετάσει και μόνος του τα αντίστοιχα στοιχεία και να ορίσει εύκολα και γρήγορα τους κατάλληλους κανόνες συσχέτισης, αποδεχόμενος τις προτάσεις του εργαλείου αυτού. Στο σημείο αυτό, να αναφέρουμε ότι ο χρήστης έχει τη δυνατότητα να σταματήσει τον καθορισμό συσχετίσεων μεταξύ των δύο μοντέλων, αποθηκεύοντας την τρέχουσα κατάσταση και να επανέλθει και να συνεχίσει από το ίδιο σημείο που σταμάτησε. Αυτό είναι ιδιαίτερα σημαντικό, εάν λάβουμε υπόψη τα παραπάνω και το γεγονός ότι νέα σενάρια μπορούν να προστεθούν κατά τη διάρκεια καθορισμού των συσχετίσεων. Τέλος, να αναφέρουμε ότι η εργασία που απαιτείται για τα παραπάνω μπορεί να εκτελεστεί από διαφορετικά μέλη ενός οργανισμού. Ειδικότερα, άλλος να αναλάβει τον καθορισμό των σεναρίων συσχέτισης, άλλος την υλοποίηση του μετασχηματισμού σε μία διαδικαστική γλώσσα και άλλος, εν τέλει, να χρησιμοποιήσει τα σενάρια αυτά, για να ορίσει τους κανόνες συσχέτισης.

3.3.3. Εργαλείο Μετατροπής Δεδομένων

Το Εργαλείο Μετατροπής Δεδομένων (Data Transformation Tool) έχει ένα απλό γραφικό περιβάλλον (Εικόνα 10), που επιτρέπει στους Παρόχους Δεδομένων να εκφράσουν αυτόματα τα δεδομένα τους,

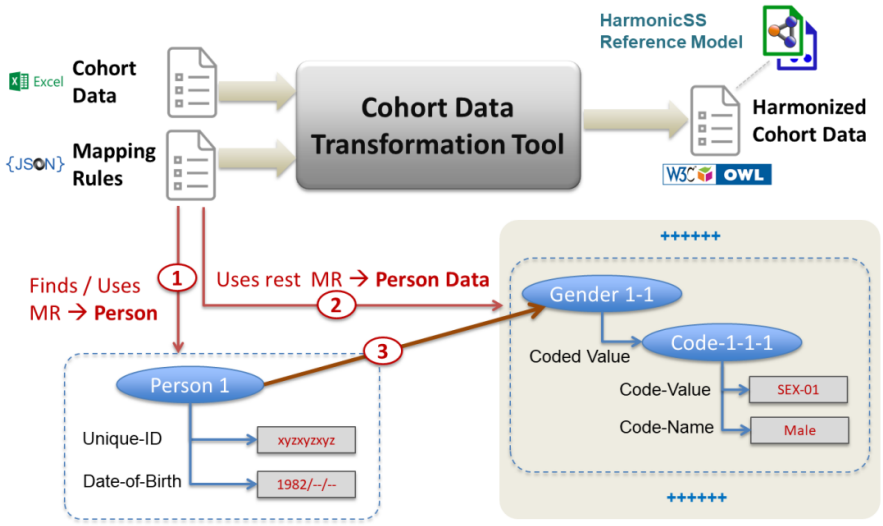
χρησιμοποιώντας τους όρους του Μοντέλου Αναφοράς, με βάση τους κανόνες συσχέτισης (mapping rules) που έχουν ήδη καθοριστεί. Συνεπώς, οι χρήστες θα πρέπει να δώσουν ως είσοδο το MS Excel έγγραφο που περιέχει τα αρχικά δεδομένα καθώς και το JSON έγγραφο με τους κανόνες συσχέτισης που έχουν καθοριστεί. Το αποτέλεσμα της διαδικασίας αυτής είναι ένα OWL[19] έγγραφο, που περιέχει τα εναρμονισμένα δεδομένα (δηλαδή τα αρχικά δεδομένα εκφρασμένα με βάση το μοντέλο αναφοράς).



Εικόνα 10: Γραφικό Περιβάλλον του Εργαλείου Μετατροπής Δεδομένων

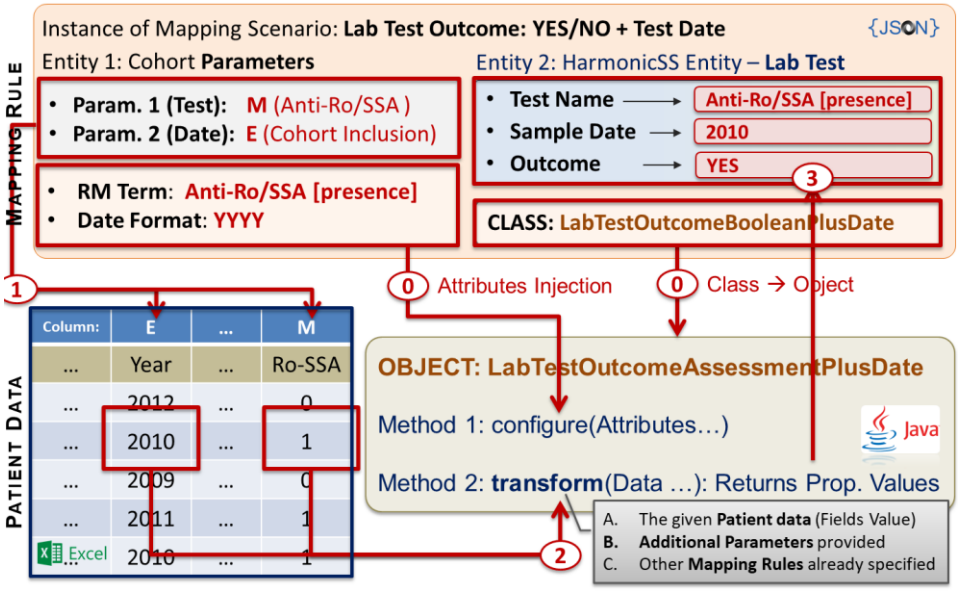
Το εργαλείο αυτό διαβάζει ένα προς ένα τα δεδομένα που έχουν καταγραφεί για κάθε οντότητα (στην προκειμένη περίπτωση, για κάθε ασθενή) και ακολούθως τα εκφράζει με βάση τις κλάσεις που έχουν καθοριστεί στο μοντέλο αναφοράς και τις παραμέτρους αυτών. Η όλη διαδικασία (Εικόνα 11) καθοδηγείται από τους κανόνες που έχουν καθοριστεί. Αρχικά, εντοπίζει και χρησιμοποιεί τον κανόνα συσχέτισης για τη δημιουργία της βασικής οντότητας του μοντέλου αναφοράς (στην περίπτωσή μας, Person), με βάση και τις παραμέτρους που έχουν καταγραφεί στα αντίστοιχα πεδία (βήμα 1). Ακολούθως, χρησιμοποιεί τους υπόλοιπους κανόνες συσχέτισης για την εναρμόνιση των υπολοίπων δεδομένων που ανήκουν στην κύρια αυτή οντότητα (π.χ., δεδομένα για ασθένειες, φάρμακα, εργαστηριακές μετρήσεις, κτλ) και τη δημιουργία των σχετικών οντοτήτων του μοντέλου αναφοράς, με βάση την πληροφορία που υπάρχει στα αντίστοιχα πεδία (βήμα 2). Οι οντότητες αυτές που δημιουργούνται (π.χ., Δεδομένα Διάγνωσης) συνδέονται, εν τέλει, με την κύρια οντότητα (στην προκειμένη περίπτωση, αυτή που αναπαριστά έναν Ασθενή) (βήμα 3).

[19] Web Ontology Language (OWL), https://www.w3.org/OWL/



Εικόνα 11: Εφαρμογή Κανόνων Συσχέτισης για την Εναρμόνιση των Δεδομένων

Κάθε φορά που θα πρέπει να χρησιμοποιηθεί ένας κανόνας (είτε για τη δημιουργία της κύριας οντότητας ή άλλων οντοτήτων που ακολούθως θα συνδεθούν με αυτή), ακολουθείται πάντα η ίδια διαδικασία (Εικόνα 12). Ειδικότερα, με βάση το αριστερό μέρος του κανόνα συσχέτισης, εντοπίζει τα δεδομένα του εκάστοτε ασθενή στα σχετικά πεδία (βήμα 1). Στη συνέχεια, εντοπίζει και χρησιμοποιεί τον μηχανισμό εκείνο που αναλαμβάνει να εντοπίσει / υπολογίσει / βρει τις τιμές των σχετικών παραμέτρων του δεξιού μέρους του κανόνα (βήμα 2). Τέλος, δημιουργεί μια οντότητα, που να ανήκει στην κλάση που αναφέρεται στο δεξιό μέρος, με τις παραμέτρους που αναφέρονται και τις τιμές τους που έχουν υπολογιστεί (βήμα 3).



Εικόνα 12: Εφαρμογή ενός Κανόνα Συσχέτισης για την Εναρμόνιση των σχετικών Δεδομένων

Στο σημείο αυτό να τονίσουμε ότι όλες οι υπηρεσίες που χρησιμοποιήθηκαν στο βήμα 2 έχουν την ίδια «υπογραφή» και κατ' επέκταση έχουν πάντα τις δύο μεθόδους που φαίνονται και στο παραπάνω σχήμα (configure και transform). Επίσης, η καθεμιά από αυτές τις υπηρεσίες είναι συνδεδεμένη με ένα συγκεκριμένο σενάριο συσχέτισης και κατ' επέκταση γνωρίζει εκ των προτέρων το πλήθος και τη σημασία των δεδομένων που θα δίνονται ως είσοδο, κάθε φορά που καλείται για να γίνει μια μετατροπή ορισμένων δεδομένων των ασθενών (π.χ., το πρώτο πεδίο θα περιέχει την τιμή της εργαστηριακής μέτρησης, ενώ το δεύτερο πεδίο την ημερομηνία που έγινε), καθώς επίσης και ορισμένες παράμετροι που έχει ορίσει ο χρήστης εκ των προτέρων για τη σωστή επεξεργασία των δεδομένων αυτών (π.χ., η συγκεκριμένη εργαστηριακή εξέταση που έγινε και η μορφή της ημερομηνίας).

Οι κανόνες μηχανικής μάθησης και εξόρυξης γνώσης μπορούν ακολούθως να εφαρμοστούν για τη δημιουργία μοντέλων και τον εντοπισμό συχνά εμφανιζόμενων μοτίβων και κανόνων, λαμβάνοντας υπόψη είτε τα εναρμονισμένα δεδομένα του κέντρου έρευνας (κυρίως για λόγους επιβεβαίωσης της ορθότητας της διαδικασίας εναρμόνισης των δεδομένων) είτε το σύνολο των εναρμονισμένων δεδομένων, για την δημιουργία πιο αξιόπιστων μοντέλων και την εξαγωγή ορθών αποτελεσμάτων. Ωστόσο, στη δεύτερη περίπτωση απαιτείται η έγκριση από τους υπεύθυνους των δύο κέντρων για την από κοινού χρήση – επεξεργασία των δεδομένων μέσω ενός υπολογιστή, για την επίτευξη ενός συγκεκριμένου στόχου.

Η σελίδα αυτή είναι σκόπιμα λευκή

4. Αποτελέσματα και Σχολιασμός

4.1. Αποτελέσματα Εναρμόνισης Πραγματικών Δεδομένων

4.1.1. Μοντέλο Αναφοράς

Για την έκφραση του μοντέλου (Εικόνα 3) βασιστήκαμε στις τεχνολογίες του σημασιολογικού ιστού. Ειδικότερα, το μοντέλο που αναπτύχθηκε έγινε διαθέσιμο με τη μορφή μιας OWL [68] οντολογίας, στην οποία συμπεριλάβαμε τόσο τις κατηγορίες των δεδομένων που καταγράφηκαν για κάθε ασθενή (π.χ., δημογραφικά στοιχεία, διαγνώσεις, συνταγογραφήσεις φαρμάκων, εργαστηριακές μετρήσεις, κτλ.) όσο και τις κατηγορίες των όρων που χρησιμοποιήθηκαν, για να καλύψουν τις ανάγκες του κάθε πεδίου (π.χ., διαταραχές, φαρμακευτικές ουσίες, εργαστηριακές εξετάσεις, κτλ.). Επίσης, εκμεταλλευόμενοι την εκφραστικότητα που μας παρέχει η γλώσσα αυτή, οργανώσαμε τις παραμέτρους σε ευρύτερες κατηγορίες και εκφράσαμε την αναγκαιότητα ύπαρξής τους στα δεδομένα με τη μορφή αξιωμάτων. Ο συνολικός αριθμός των κλάσεων και των παραμέτρων του μοντέλου αυτού καθώς επίσης και τα σχετικά αξιώματα που ορίστηκαν παρουσιάζονται στην Εικόνα 13.

Ontology metrics

Metrics		Object property axioms	
Axiom	4201	SubObjectPropertyOf	80
Logical axiom count	1139	ObjectPropertyDomain	73
Declaration axioms count	843	ObjectPropertyRange	75
Class count	711	**Data property axioms**	
Object property count	81	SubDataPropertyOf	33
Data property count	37	DataPropertyDomain	30
Individual count	1	DataPropertyRange	34
Annotation Property count	19	**Individual axioms**	
Class axioms		ClassAssertion	1
SubClassOf	806	**Annotation axioms**	
EquivalentClasses	4	AnnotationAssertion	2188
DisjointClasses	3		
Hidden GCI Count	6		

Εικόνα 13: Συνοπτική παρουσίαση των οντοτήτων του μοντέλου αναφοράς

Σημειώνουμε ότι το μοντέλο αυτό επιτρέπει την έκφραση των δεδομένων των ασθενών ενός κέντρου έρευνας με τη μορφή συγκεκριμένων οντοτήτων των αντίστοιχων κλάσεων (instance).

4.1.2. Κανόνες και Σενάρια Συσχέτισης

Για τον καθορισμό των συσχετίσεων μεταξύ των όρων των δύο μοντέλων με αυτούς του μοντέλου αναφοράς (mapping rules), αναπτύχθηκαν 22 διαφορετικά σενάρια συσχέτισης (mapping scenarios), μέσω των οποίων ορίστηκαν συνολικά 157 κανόνες συσχέτισης (mapping rules). Το πλήθος των διαφορετικών σεναρίων που χρησιμοποιήθηκαν σε καθένα από τα δύο κέντρα έρευνας, ο αριθμός των κανόνων συσχέτισης που ορίστηκαν σε αυτά και το πλήθος των διαφορετικών πεδίων που συμμετείχαν παρουσιάζονται στον Πίνακας 2.

Κέντρα	**Αριθμός Πεδίων**	**Αριθμός Κανόνων Συσχέτισης**	**Αριθμός Σεναρίων Συσχέτισης**	**Αριθμός Πεδίων που Χρησιμοποιήθηκαν**
Κέντρο Ερευνάς 1	186	80	22	117
Κέντρο Ερευνάς 2	163	77	22	108

Πίνακας 2: Σενάρια και Κανόνες Συσχέτισης ανά Κέντρο Έρευνας

Μέσω των κανόνων συσχέτισης, μπορέσαμε να εκφράσουμε τον τρόπο σύνδεσης των περισσότερων παραμέτρων (και των τιμών τους) που έχουν καταγραφεί σε κάθε ένα από τα 2 κέντρα έρευνας με τα στοιχεία που συμπεριλάβαμε στο μοντέλο αναφοράς (Πίνακας 3). Ωστόσο, ορισμένες παράμετροι των αρχικών δεδομένων αγνοήθηκαν κατά την παραπάνω διαδικασία, είτε γιατί η πληροφορία που είχε καταγραφεί σε αυτές τις παραμέτρους ήταν επανάληψη κάποιας πληροφορίας που θα μπορούσε να εξαχθεί από κάποιο άλλο πεδίο, είτε γιατί κρίθηκε ότι δεν ήταν σχετική με τους όρους του μοντέλου, κατόπιν συνεννόησης και με τους ειδικούς στον χώρο αυτό.

No.	**Σύντομη Περιγραφή**	**Πλήθος**
1	Disorder: Confirmation Term	42
2	Disorder: Confirmation Term + Date: String	9
3	Disorder: Current/Past/Never	3
4	Drug: Confirmation Term	11
5	Drug: Confirmation Term + Start Date: String	5

No.	Σύντομη Περιγραφή	Πλήθος
6	Drug: Current/Past/Never	3
7	Lab Test: Numeric Value	24
8	Lab Test: Numeric Value + Test Date: String	4
9	Lab Test: Confirmation Term	20
10	Lab Test: Confirmation Term + Test Date: String	3

Πίνακας 3: Ευρέως Χρησιμοποιούμενα Μοτίβα

Αξίζει να αναφέρουμε ότι δεν χρησιμοποιούνται όλα τα σενάρια αυτά με την ίδια συχνότητα. Ειδικότερα, οι περισσότεροι κανόνες συσχέτισης ορίστηκαν μέσω της χρήσης ενός περιορισμένου αριθμού σεναρίων, γεγονός που δείχνει ότι τα σενάρια αυτά μπορούν να χρησιμοποιηθούν για την εναρμόνιση των δεδομένων ενός μεγάλου μέρους ενός νέου οργανισμού που έχει δεδομένα που ανήκουν στον χώρο αυτό. Ωστόσο, μπορούμε εύκολα να προσθέσουμε νέα σενάρια συσχέτισης στα ήδη υπάρχοντα και ακολούθως να τα χρησιμοποιήσουμε, για να γεφυρώσουμε το χάσμα μεταξύ των διαφορετικών μοντέλων, ακολουθώντας τη διαδικασία που περιγράφηκε στην ενότητα 3.3.2.

4.1.3. Εναρμονισμένα Δεδομένα

Μέσω της χρήσης των κανόνων, ο υπεύθυνος καθενός από τα δύο κέντρα έρευνας μπόρεσε να εκφράσει αυτόματα τα δεδομένα των ασθενών με τους όρους του μοντέλου αναφοράς (harmonized data). Ειδικότερα στο Κέντρο Έρευνας 1, τα εναρμονισμένα δεδομένα αφορούν 586 ασθενείς και προέρχονται από την εναρμόνιση των 117 πεδίων που είχαν καταγραφεί γι αυτούς στο αρχικό MS Excel αρχείο. Αντίστοιχα, στο Κέντρο Έρευνας 2, τα εναρμονισμένα δεδομένα αφορούν 286 ασθενείς και προέρχονται από 108 διαφορετικά πεδία του αρχικού MS Excel αρχείου. Τα δεδομένα αυτά (λόγω του μικρού αριθμού δεδομένων των ασθενών) αποθηκεύτηκαν σε ένα OWL αρχείο σε καθένα από τα δύο αυτά κέντρα έρευνας, το οποίο περιείχε τόσο τον τρόπο αναπαράστασης των δεδομένων (δηλαδή τον ορισμό των οντοτήτων του μοντέλου) όσο και τα δεδομένα των ασθενών.

4.1.4. Τεχνολογίες Προγραμματισμού

Το σύστημα που αναπτύχθηκε και ειδικότερα, τα Εργαλεία που δημιουργήθηκαν, υλοποιήθηκαν σε Java, μέσω της χρήσης των σχετικών βιβλιοθηκών και πλαισίων. Πιο συγκεκριμένα, το Εργαλείο Εντοπισμού Μεταδεδομένων και το Εργαλείο Μετατροπής Δεδομένων είναι desktop εφαρμογές, οι οποίες

χρησιμοποιούν τις βιβλιοθήκες Apache POI[20] και Apache Jena[21] για την επεξεργασία των MS Excel εγγράφων και των OWL οντολογιών αντίστοιχα. Το Εργαλείο Καθορισμού Συσχετίσεων είναι μια Web Εφαρμογή, η οποία βασίστηκε στην JavaScript βιβλιοθήκη jQuery[22] για τη δημιουργία του γραφικού περιβάλλοντος και στις Java βιβλιοθήκες, που αναφέρθηκαν και πιο πριν για την επεξεργασία των δεδομένων.

Σχετικά την ανάλυση και περαιτέρω επεξεργασία των δεδομένων υλοποιήθηκαν τα σχετικά Script σε Python, που αναλαμβάνουν τόσο να διαβάσουν το περιεχόμενο των αρχείων με τα μεταδεδομένα και το μοντέλο, όσο και να εκτελέσουν τους σχετικούς αλγορίθμους. Για τον σκοπό αυτό, βασιστήκαμε σε υπάρχουσες βιβλιοθήκες της γλώσσας αυτής (ενδεικτικά αναφέρουμε: Numpy, Pandas, MatPlotLib, Scikit-Learn, Mlxtend, Gensim, κτλ.) αλλά και του προ-εκπαιδευμένου Word2Vec μοντέλου, για τη διανυσματική αναπαράσταση των λέξεων. Σημειώνουμε ότι οι τεχνικές εντοπισμού συχνών μοτίβων που χρησιμοποιήθηκαν χρειάζεται να προσπελάσουν το σύνολο των δεδομένων των ασθενών και κατ' επέκταση, για αυτόν τον σκοπό, δημιουργήθηκαν ξεχωριστά python scripts, έτσι ώστε να μπορέσουν ακολούθως να εκτελεστούν από τους παρόχους δεδομένων.

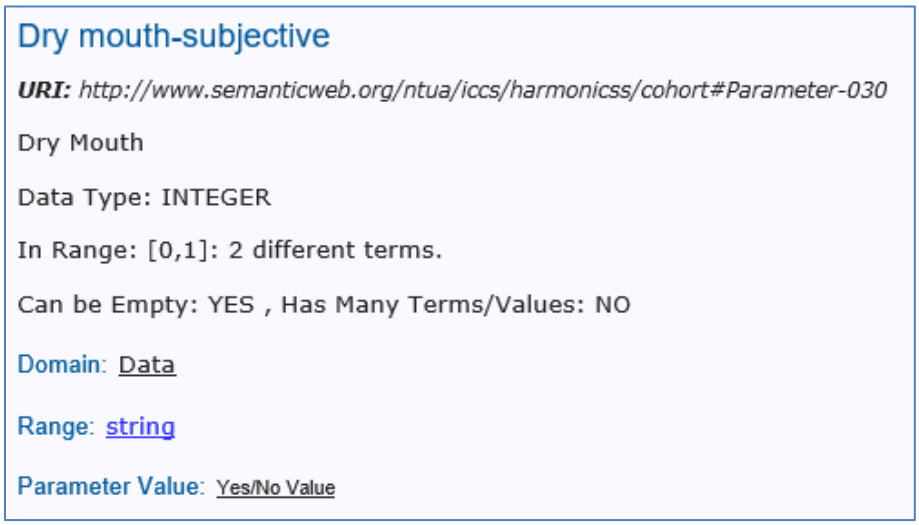
4.2. Συμβολή των Τεχνικών Μηχανικής Μάθησης και Εξόρυξης Γνώσης

4.2.1. Προ-επεξεργασία Δεδομένων

Οι τεχνικές που χρησιμοποιήθηκαν (συμπεριλαμβανομένων των τεχνικών από το ευρύτερο πεδίο της στατιστικής καθώς και των τεχνικών μηχανικής μάθησης και εξόρυξης γνώσης) συνέβαλαν καθοριστικά στην κατανόηση των δεδομένων και κυρίως στον περιορισμό των λαθών που υπήρχαν στα MS Excel αρχεία. Ειδικότερα, ακραίες τιμές που εντοπίστηκαν, σε πάνω από το 10% των πεδίων των δύο κέντρων έρευνας, επισημάνθηκαν στους υπεύθυνους του κάθε ερευνητικού κέντρου, έτσι ώστε να ελέγξουν τα σχετικά δεδομένα και να διορθώσουν ορισμένα από αυτά, που οφείλονταν κυρίως σε ανθρώπινα λάθη. Για παράδειγμα, οι τιμές των λευκών αιμοσφαιρίων στο αίμα (white blood cell - WBC) κυμαίνονταν μεταξύ 6000 και 9000. Ωστόσο, μια τιμή βρισκόταν εκτός του ορίου αυτού, εξαιτίας τυπογραφικού λάθους, την οποία και εντοπίσαμε κατά την οργάνωση των τιμών της παραμέτρου αυτής σε δύο κατηγορίες.

[20] Apache POI, https://poi.apache.org/
[21] Apache Jena, https://jena.apache.org/
[22] jQuery, https://jquery.com/

Η ανάλυση που έλαβε χώρα, μας έδωσε τη δυνατότητα να κατανοήσουμε καλύτερα τις παραμέτρους που είχαν καταγραφεί και κυρίως τις τιμές τους. Ειδικότερα, ο τύπος των δεδομένων και το εύρος των τιμών τους μας βοήθησαν να κατανοήσουμε καλύτερα την πληροφορία που καταγράφηκε στα σχετικά πεδία, όπως σε αυτό που παρουσιάζεται στην Εικόνα 14. Για παράδειγμα, σε ορισμένα πεδία η τιμή ήταν κάποιος ακέραιος αριθμός, αλλά το γεγονός ότι η τιμή τους προερχόταν από ένα πολύ περιορισμένο σύνολο τιμών, μας βοήθησε να κατανοήσουμε το γεγονός ότι οι τιμές αυτές είχαν κάποια ιδιαίτερη σημασία και, σε συνεργασία με τους ειδικούς του κέντρου έρευνας, να κατανοήσουμε τη σημασία των τιμών (για την ακρίβεια κωδικών) αυτών και γενικότερα τη σημασία του πεδίου από το οποίο αυτές προέρχονταν. Επίσης, μας βοήθησε να κατανοήσουμε κατά πόσο το εκάστοτε πεδίο είχε συμπληρωθεί για όλους τους ασθενείς ή όχι και κατ' επέκταση να κατανοήσουμε εάν η παράμετρος ήταν υποχρεωτική ή προαιρετική ή είχε γίνει κάποιο λάθος / παράληψη και δεν είχε ακόμη συμπεριληφθεί η τιμή του ασθενούς στο MS Excel έγγραφο.



Εικόνα 14: Αυτομάτως εξαγόμενα μεταδεδομένα για ένα συγκεκριμένο πεδίο.

Σχετικά με τον αυτόματο εντοπισμό συχνά εμφανιζόμενων μοτίβων, χρησιμοποιήθηκε ο αλγόριθμος FP-growth. Ειδικότερα, μετά την κατανόηση των πεδίων και των τιμών τους, δημιουργήθηκαν ειδικά για κάθε κέντρο έρευνας python scripts που δημιουργούν εγγραφές με δεδομένα για κάθε ασθενή, λαμβάνοντας υπόψη και τη σημασία των τιμών τους (π.χ., η εγγραφή περιείχε τον όρο με το σύμπτωμα όταν η τιμή του πεδίου ήταν 1), έτσι ώστε να εκτελεστεί ακολούθως ο αλγόριθμος. Δύο από τα πιο συχνά εμφανιζόμενα μοτίβα έδειχναν ότι το σύνδρομο προς εξέταση αφορά κυρίως γυναίκες και ότι ένα από τα βασικότερα συμπτώματα τους είναι η ξηρότητα του στόματος και των ματιών.

4.2.2. Δημιουργία Μοντέλου

Το μοντέλο που αναπτύχθηκε για την έκφραση των εναρμονισμένων δεδομένων βασίστηκε τόσο σε διεθνή πρότυπα όσο και στα μετα-δεδομένα που εξήχθησαν αυτόματα από καθένα από τα δύο κέντρα και ειδικότερα τα ονόματα των παραμέτρων και τους διαφορετικούς όρους που χρησιμοποιήθηκαν σε ορισμένα από αυτά. Στον Πίνακας 4 μπορούμε να δούμε το πλήθος των παραμέτρων και το πλήθος των διαφορετικών όρων που εντοπίστηκαν (δηλαδή συμβολοακολουθιών, εξαιρουμένων των αριθμών και των τιμών επιβεβαίωσης, όπως είναι η boolean τιμές) σε καθένα από τα δύο κέντρα έρευνας.

Κέντρο Έρευνας	# Παραμέτρων	# Διαφορετικών Τιμών
Κέντρο Έρευνας 1	186	199
Κέντρο Έρευνας 2	163	228

Πίνακας 4: Παράμετροι και Διαφορετικές Τιμές ανά Κέντρο Έρευνας

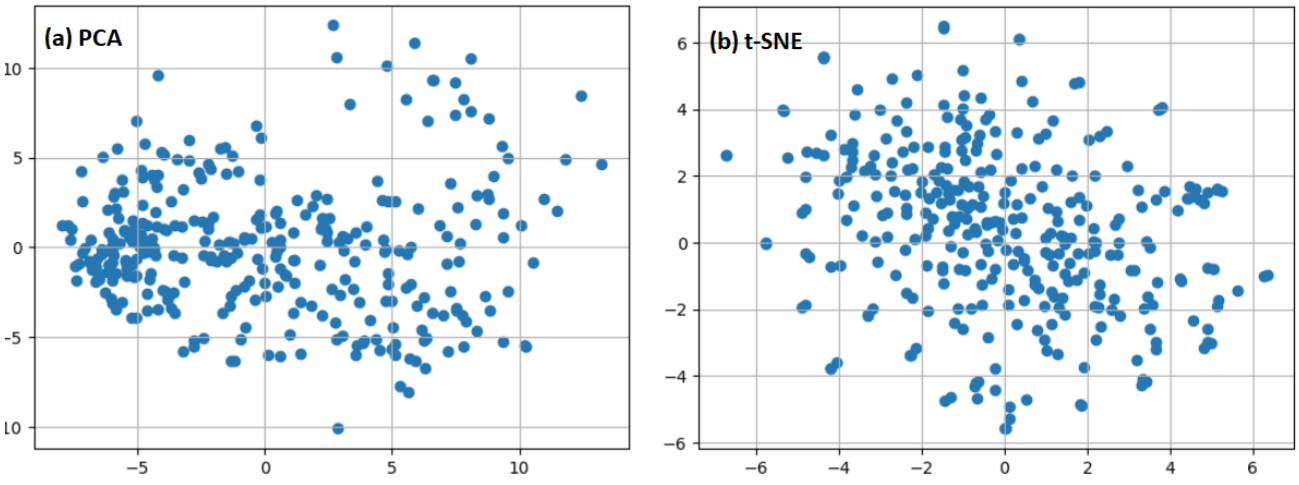
Ακολούθως, εντοπίσαμε τη διανυσματική αναπαράσταση των λέξεων που αναφέρονται τόσο στις παραμέτρους όσο και στις τιμές τους. Ό λόγος που εστιάσαμε στις λέξεις οφείλεται στο γεγονός ότι σημαντικό μέρος της πληροφορίας των ιατρικών όρων καλύπτεται συνήθως μέσω μιας πολλές φορές συμβολοακολουθίας, ειδικά όταν πρόκειται για ένα αρκετά συγκεκριμένο πεδίο γνώσης. Ωστόσο, για την αποφυγή πιθανών λαθών, κατά την επικοινωνία με τους ειδικούς στον χώρο της έρευνας για τη δημιουργία του μοντέλου, εστιάσαμε στους όρους στους οποίους υπήρχαν οι λέξεις αυτές. Στον Πίνακας 5 μπορούμε να δούμε τις διαφορετικές λέξεις που εντοπίσαμε σε κάθε ένα από τα δύο κέντρα έρευνας (εξαιρουμένων των stop words και των αριθμών ή, εν γένει, των αριθμητικών εκφράσεων) καθώς και το πλήθος αυτών για τις οποίες εντοπίσαμε τη διανυσματική τους αναπαράσταση. Για τη διανυσματική αναπαράσταση των λέξεων, βασιστήκαμε σε ένα προ-εκπαιδευμένο Word2Vec[23] μοντέλο, σύμφωνα με το οποίο το διάνυσμα κάθε λέξης αποτελείται από 300 χαρακτηριστικά. Όπως μπορούμε να δούμε και στον πίνακα αυτόν, δεν μπορέσαμε να εντοπίσουμε τη διανυσματική αναπαράσταση για ένα σημαντικό μέρος των λέξεων, το οποίο οφείλεται στο γεγονός ότι χρησιμοποιήσαμε ένα μοντέλο γλώσσας γενικού σκοπού. Μελλοντικά, θα μπορούσαμε να χρησιμοποιήσουμε ένα μοντέλο γλώσσας που ταιριάζει καλύτερα σε αυτό το πεδίο γνώσης, όπως είναι για παράδειγμα το BioBERT.

[23] word2vec, https://code.google.com/archive/p/word2vec/

	# Διαφορετικών Εκφράσεων	# Διαφορετικών Λέξεων	# Με Διανυσμ. Αναπαράσταση	# Χωρίς Διανυσμ. Αναπαράσταση
Κέντρα Έρευνας 1	341	394	255	139
Κέντρα Έρευνας 2	322	367	239	128
Συνολικά	471	521	326	195

Πίνακας 5: Διανυσματική Αναπαράσταση Λέξεων ανά Κέντρο Έρευνας

Στην Εικόνα 15 μπορούμε να δούμε μια γραφική απεικόνιση των λέξεων για τις οποίες εντοπίσαμε τη διανυσματική τους αναπαράσταση, χρησιμοποιώντας δύο διαφορετικές προσεγγίσεις, την PCA ανάλυση και τον t-SNE αλγόριθμο (περισσότερες πληροφορίες για αυτούς τους δύο αλγορίθμους υπάρχουν στην υπο-ενότητα 2.3.1).



Εικόνα 15: Γραφική απεικόνιση των όρων με βάση (α) την PCA ανάλυση και (β) τον t-SNE αλγόριθμο

Όπως μπορούμε να δούμε, οι λέξεις αυτές καλύπτουν ένα ευρύ φάσμα και κατ' επέκταση, για την περαιτέρω εξέταση τόσο των λέξεων αυτών όσο και των όρων από τις οποίες προέρχονται, τις οργανώσαμε σε κατηγορίες, χρησιμοποιώντας τον K-Means αλγόριθμο, με βάση, ωστόσο, τη μεταξύ τους ομοιότητα συνημίτονου (cosine similarity). Επίσης, από κάθε κατηγορία, εντοπίσαμε τις 5 κοντινότερες στο κέντρο λέξεις και ακολούθως τους όρους στους οποίους αυτές υπήρχαν. Για παράδειγμα, στην Εικόνα 16 μπορούμε να δούμε το αποτέλεσμα της εκτέλεσης του αλγορίθμου για K=20.

```
[('lymphadenopathy', 0.8338652), ('erythematous', 0.8290348), ('purpura', 0.82025087), ('telangiectasia', 0.81819147), ('papular', 0.8166022)]
[('positive', 0.51501226), ('still', 0.49046716), ('weakness', 0.45589694), ('damage', 0.45430875), ('failure', 0.45306626)]
[('annular', 0.63804495), ('malar', 0.6213807), ('erythematous', 0.6135747), ('discoid', 0.61217344), ('nodular', 0.605387)]
[('multiple', 0.6378268), ('one', 0.62915117), ('three', 0.62578887), ('minor', 0.5091738), ('plus', 0.48678526)]
[('wb', 0.6480718), ('fs', 0.6463718), ('aza', 0.6359788), ('cp', 0.6350948), ('fc', 0.62544876)]
[('vitiligo', 0.6760287), ('hereditary', 0.6255647), ('tears', 0.53569186), ('visit', 0.49141455), ('alopecia', 0.43219963)]
[('hypertension', 0.7962913), ('dyslipidemia', 0.78190154), ('osteoporosis', 0.77365714), ('hypothyroidism', 0.7714075), ('hyperlipidemia', 0.7540143)]
[('fosamax', 0.8020422), ('lipitor', 0.79818773), ('plavix', 0.7719449), ('parkinson', 0.7249936), ('cns', 0.6653099)]
[('2nd', 0.85263026), ('4th', 0.8516776), ('3rd', 0.84543955), ('1st', 0.8452197), ('first', 0.683182)]
[('baseline', 0.71332633), ('score', 0.68086874), ('low', 0.67853975), ('test', 0.33681047), ('titer', 0.2884486)]
[('lymphadenopathy', 0.77089185), ('fibrosis', 0.7489818), ('adenoma', 0.7486845), ('leucopenia', 0.74791634), ('thrombophlebitis', 0.7452359)]
[('hysterectomy', 0.7527305), ('mastectomy', 0.67803544), ('infertility', 0.65660214), ('thyroid', 0.6347201), ('breast', 0.6300324)]
[('renal', 0.7657382), ('liver', 0.7350439), ('lung', 0.7171182), ('kidney', 0.715429), ('pulmonary', 0.7083419)]
[('azathioprine', 0.8339407), ('leflunomide', 0.83163476), ('anakinra', 0.8312572), ('hydroxychloroquine', 0.8128386), ('corticosteroids', 0.7966805)]
[('skin', 0.71909976), ('ear', 0.6119026), ('facial', 0.60858715), ('telangiectasia', 0.59018767), ('eye', 0.5761472)]
[('monocyte', 0.81238323), ('glycoprotein', 0.8090079), ('antibody', 0.80416334), ('lymphocyte', 0.7944222), ('antibodies', 0.79047674)]
[('myalgias', 0.72473466), ('asthma', 0.6800474), ('chronic', 0.66374797), ('rhinitis', 0.6635312), ('polymyalgia', 0.66138625)]
[('type', 0.59021175), ('date', 0.4932054), ('specific', 0.47766584), ('new', 0.47209778), ('year', 0.46629274)]
[('embolism', 0.66483045), ('ulcer', 0.65832317), ('thrombophlebitis', 0.6556624), ('arrythmias', 0.6461895), ('esophagus', 0.63666475)]
[('lupus', 0.81110525), ('disease', 0.75710297), ('cancer', 0.7569764), ('psoriasis', 0.747944), ('pemphigus', 0.7431188)]
```

Εικόνα 16: Αποτελέσματα εκτέλεσης αλγορίθμου K-means για K=20

Με μία γρήγορη ανάγνωση των αποτελεσμάτων μπορούμε να δούμε ότι το σύνδρομο αυτό σχετίζεται, για παράδειγμα, με θέματα που έχουν να κάνουν με το πρόσωπο (π.χ., μάτια και αυτιά) και το δέρμα. Επίσης, επηρεάζει διάφορα όργανα του σώματος (π.χ., συκώτι και πνεύμονα). Επιπρόσθετα, για το σύνδρομο αυτό αναφέρονται συγκεκριμένες ασθένειες και φάρμακα (και κατηγορίες αυτών), τα οποία και εξετάσαμε σε συνεργασία με ειδικούς στον χώρο αυτό και ακολούθως τα συμπεριλάβαμε στο μοντέλο που αναπτύχθηκε. Τέλος, μπορούμε να δούμε ότι ορισμένες συμβολοακολουθίες όπως «1st», «2nd», κτλ… απομονώθηκαν από τους υπόλοιπους όρους και τοποθετήθηκαν σε ξεχωριστές κατηγορίες, διευκολύνοντας έτσι την ανάλυση.

Στην Εικόνα 17 μπορούμε να δούμε, για παράδειγμα, τους όρους που σχετίζονται με μία από τις 20 κατηγορίες που αναφέρονται στην παραπάνω εικόνα.

```
[('renal', 0.7657382), ('liver', 0.7350439), ('lung', 0.7171182), ('kidney', 0.715429), ('pulmonary', 0.7083419)]
    renal
        interstitial renal disease(0-1)
    liver
        liver disease
        liver involvmt (sclerosing cholangitis)(0-1)
        liver involvmt - (sclerosing cholangitis ) / diagnosisdate(-yr)
        liver involvmt - autoimmune hepatitis (0-1)
        liver involvmt - autoimmune hepatitis / diagnosis date(-yr)
        liver involvmnt- pbc (0 -1)
        liver involvmnt- pbc diagnosis date (0-1)
        liver involvmt (autoimmune cholangitis) (0-1)
        liver involvmt - autoimmune hepatitis /date(-yr)
    lung
        lung involvmt - interstitial disease type(0-1)
        lung involvmt - interstitial disease date(-yr)
        lung involvmt - bronchocentric disease (0-1)
        lung involvmt - bronchocentric disease date(-yr)
        lung involvmt - pleurisy (0-1)
        lung involvmt - pleurisy date(-yr)
    kidney
        chronic kidney disease
        kidney involvmt -gn-biopsy (0-1)
        kidney involvmt -gn-biopsy date(-yr)
        kidney involvmt -rta - nephrocalcinosis (0-1)
        kidney involvmt -rta - nephrocalcinosis date(-yr)
        kidney involvmt -biopsy -interstitial infiltrates (0-1)
        kidney involvmt -biopsy -interstitial infiltrates date(-yr)
    pulmonary
        pulmonary embolism
        pulmonary hypertension
        pulmonary nodule
        chronic obstructive pulmonary disease
```

Εικόνα 17: Εκφράσεις που περιείχαν τις πέντε κοντινότερες στο προς εξέταση κέντρο.

Από την παραπάνω οργάνωση των όρων είναι εμφανές ότι τα όργανα του σώματος και οι σχετικές διαγνώσεις σχετίζονται άμεσα με το προς εξέταση σύνδρομο και κατ' επέκταση φροντίσαμε να συμπεριλάβουμε στο μοντέλο τους σχετικούς όρους (συμπεριλαμβανομένων των σχετικών εργαστηριακών εξετάσεων και ασθενειών), κατόπιν συνεννόησης και με τους ειδικούς στον χώρο αυτό. Τέλος, θα θέλαμε να επισημάνουμε ότι η παραπάνω οργάνωση βασίστηκε στη διανυσματική αναπαράσταση των όρων και όχι των εκφράσεων. Η χρήση διανυσματικών αναπαραστάσεων, που προέρχονται από ένα πιο συγκεκριμένο πεδίο γνώσης, όπως αυτό της βιοϊατρικής, μπορεί να συμβάλει στη δημιουργία καλύτερων αποτελεσμάτων, ενώ μοντέλα γλώσσας, όπως για παράδειγμα είναι το BERT, μπορούν να διευκολύνουν τη δημιουργία διανυσματικών αναπαραστάσεων των εκφράσεων, με βάση τις διανυσματικές αναπαραστάσεις των επιμέρους λέξεων.

4.2.3. Εντοπισμός Συσχετίσεων

Ένα σημαντικό μέρος των συσχετίσεων μεταξύ των όρων των μοντέλων των δύο κέντρων έρευνας και των όρων που συμπεριλάβαμε στο μοντέλο αναφοράς εντοπίστηκε αυτόματα από το σύστημα. Για τον σκοπό αυτό, αρχικά, δημιουργήσαμε τη διανυσματική αναπαράσταση των εκφράσεων που υπάρχουν σε καθένα από τα δύο κέντρα έρευνας (είτε στο όνομα των παραμέτρων, είτε στην τιμή τους) καθώς και τη διανυσματική αναπαράσταση των όρων που ορίσαμε στο μοντέλο που δημιουργήθηκε, έτσι ώστε, ακολούθως, να χρησιμοποιήσουμε την πληροφορία αυτή για τον εντοπισμό των πιθανών συσχετίσεων.

Στο σημείο αυτό, να τονίσουμε ότι η προσέγγιση που ακολουθήσαμε εξαρτάται σε μεγάλο βαθμό από το προ-εκπαιδευμένο μοντέλο που χρησιμοποιήσαμε για τη διανυσματική αναπαράσταση των λέξεων καθώς και τις λέξεις που αυτό περιέχει. Ειδικότερα, παρατηρήσαμε ότι ένας περιορισμένος αριθμός των όρων που υπάρχουν είτε στους όρους που χρησιμοποιούνται στα κέντρα έρευνας είτε σε αυτούς του μοντέλου αναφοράς, υπάρχουν, ως έχουν, στο μοντέλο γλώσσας και κατ' επέκταση μόνο για ένα πολύ μικρό σύνολο όρων μπορέσαμε να εντοπίσουμε άμεσα τη διανυσματική τους αναπαράσταση. Για τη διανυσματική αναπαράσταση των υπολοίπων όρων βασιστήκαμε στις επιμέρους λέξεις και τη διανυσματική αναπαράσταση αυτών, εφόσον υπήρχε. Ειδικότερα, η διανυσματική αναπαράσταση των εκφράσεων αυτών, προέκυψε ως η μέση τιμή των τιμών των χαρακτηριστικών των επιμέρους λέξεων (εξαιρουμένων των stop words και των αριθμών) και κατ' επέκταση υπολογίστηκε μόνο για τους όρους εκείνους, για τους οποίους υπήρχε η διανυσματική αναπαράσταση των λέξεων. Στον Πίνακας 6 μπορούμε να δούμε συνοπτικά το πλήθος των όρων / εκφράσεων που υπήρχαν σε καθένα από τα δύο κέντρα έρευνας και το πλήθος των όρων

που συμπεριλάβαμε στο μοντέλο που δημιουργήθηκε, καθώς επίσης και το πλήθος αυτών για τους οποίους μπορέσαμε να υπολογίσουμε τη διανυσματική τους αναπαράσταση.

Μοντέλο	**# Όρων (Παραμέτρων και Τιμών)**	**# Όροι (αυτούσιοι) του Μοντέλου Γλώσσας**	**# Όροι με Λέξεις του Μοντέλου Γλώσσας**	**# Όρων με Διανυσματική Αναπαράσταση**
Κέντρου Έρευνας 1	341	55	99	154
Κέντρου Έρευνας 2	322	64	88	152
Μοντέλο Αναφοράς	647	96	329	425

Πίνακας 6: Διανυσματική Αναπαράσταση Όρων / Εκφράσεων

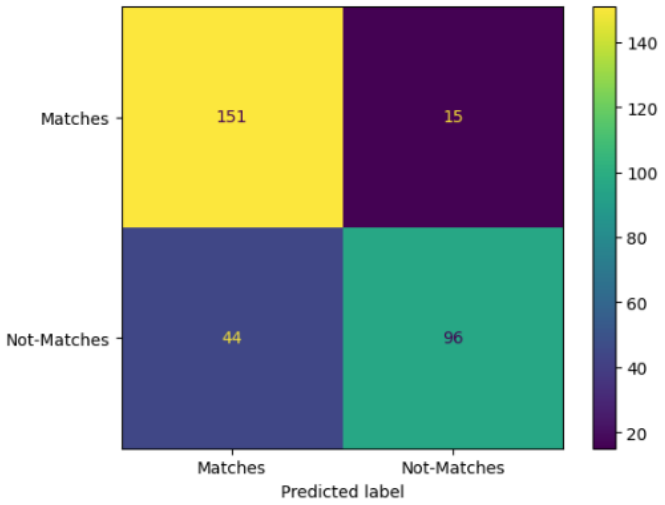
Στη συνέχεια, για τον εντοπισμό των συσχετίσεων μεταξύ των όρων / εκφράσεων που χρησιμοποιούνται στα δύο κέντρα έρευνας και των όρων του μοντέλου αναφοράς, εξετάσαμε όλους τους δυνατούς συνδυασμούς και ακολούθως επιλέξαμε τον όρο εκείνο με τη μεγαλύτερη ομοιότητα συνημίτονου (cosine similarity), εφόσον αυτή ήταν πάνω από ένα προκαθορισμένο όριο. Όπως φαίνεται και στον Πίνακας 7, μέσω της παραπάνω διαδικασίας μπορέσαμε να εντοπίσουμε αυτόματα τον αντίστοιχο όρο μόνο για έναν αρκετά μικρό αριθμό όρων του κάθε κέντρου έρευνας. Ωστόσο, θα πρέπει να λάβουμε υπόψη κατά πόσο υπήρχε ο όρος που ψάχναμε στο μοντέλο αναφοράς

Μοντέλο	**# Όρων (Παραμέτρων και Τιμών)**	**# Όρων με Διανυσματική Αναπαράσταση**	**# Αυτομάτως Προτεινόμενοι Όροι**
Κέντρο Έρευνας 1	341	154	90
Κέντρο Έρευνας 2	322	152	76

Πίνακας 7: Αυτόματα Εντοπισμένες Συσχετίσεις Όρων

Για την αξιολόγηση της προσέγγισης που ακολουθήσαμε, βασιστήκαμε στους κανόνες συσχέτισης που έχουμε καθορίσει με τη μορφή ενός JSON αρχείου, σε συνεργασία με τους υπεύθυνους από το κάθε κέντρο έρευνας, μέσω της χρήσης του γραφικού περιβάλλοντος του Εργαλείου Καθορισμού Συσχέτισης που δημιουργήθηκε και παρουσιάστηκε στην Ενότητα 3.3.2. Ειδικότερα, λαμβάνοντας υπόψη τόσο τους πιθανούς κανόνες συσχέτισης που εντοπίσαμε αυτόματα όσο και αυτούς που ορίστηκαν από τους εκάστοτε χρήστες, μπορέσαμε να διακρίνουμε το πλήθος των προτάσεων που εντοπίσαμε ή δεν εντοπίσαμε σωστά (True Positive – TP και True Negative - TN) καθώς και των λαθών που έγιναν είτε γιατί λανθασμένα

εντοπίσαμε κάποιον όρο (False Positive – FP) είτε γιατί υπήρχε όρος, αλλά δεν τον εντοπίσαμε (False Negative – FN), τα οποία παρουσιάζονται στην Εικόνα 18.



Εικόνα 18: Πίνακας Σύγχυσης του Μοντέλου

Με βάση τις παραπάνω τιμές υπολογίσαμε τις μετρικές accuracy, precision και recall καθώς και τον συνδυασμό των δύο τελευταίων, γνωστό ως F-score ή F-measure. Ο τρόπος υπολογισμού των παραπάνω καθώς και η τιμή που υπολογίσαμε, φαίνονται στον Πίνακας 8.

Μετρικές Απόδοσης	**Εκφραση**	**Τιμή**
Accuracy	(TP + TN) / (TP + FP + FN + TN)	0.8071895424836601
Precision	TP / (TP + FP)	0.9096385542168675
Recall	TP / (TP + FN)	0.7743589743589744
F-measure	2 * (Precision * Recall) / (Precision + Recall)	0.8365650969529085

Πίνακας 8: Αποτελέσματα Αξιολόγησης Μοντέλου

Το μοντέλο που αναπτύχθηκε μπορεί να εντοπίσει, με σχετικά μεγάλη ακρίβεια, τους αντίστοιχους όρους του μοντέλου αναφοράς. Ωστόσο, αρκετοί από τους όρους που αναφέρονται στα δύο MS Excel έγγραφα, τα οποία περιέχουν τα δεδομένα των ασθενών, δεν εντοπίστηκαν αυτόματα. Αυτό οφείλεται και στο γεγονός ότι, σε ορισμένες περιπτώσεις, οι όροι αυτοί ήταν μέρος αρκετά μεγαλύτερων εκφράσεων, το οποίο, με τη σειρά του, επηρέασε σημαντικά τη διαδικασία εντοπισμού τους. Επίσης, αρκετές εκφράσεις δεν υπήρχαν στο λεξιλόγιο που αναπτύχθηκε (έτος γέννησης), λαμβάνοντας υπόψη το γεγονός ότι εστιάσαμε στους

όρους του μοντέλου (π.χ., εντοπισμό σχετικών διαταραχών, φαρμακευτικών ουσιών, κτλ.) και όχι τόσο στις παραμέτρους των κλάσεων που δημιουργήθηκαν για την αποθήκευση των δεδομένων. Τα παραπάνω δείχνουν το γεγονός ότι η παραπάνω διαδικασία, θα μπορούσε να βοηθηθεί από τη χρήση τεχνικών εντοπισμού ονομαστικών όρων (named entity recognition - NER), έτσι ώστε ακολούθως να γίνει η αναζήτηση για σχετικούς όρους στο μοντέλο, λαμβάνοντας υπόψη και τους όρους που ήδη υπάρχουν στο μοντέλο.

4.3. Περαιτέρω Συζήτηση και Μελλοντική Εργασία

Η διαδικασία που θα πρέπει να ακολουθήσουμε για την εναρμόνιση των δεδομένων, εξαρτάται από διάφορους παράγοντες που έχουν να κάνουν με τη μορφή και το είδος των δεδομένων, τη συχνότητα αλλαγής/προσθήκης δεδομένων κτλ. Στην παρούσα εργασία εστιάσαμε σε αρχεία/βάσεις που έχουν έναν περιορισμένο αριθμό εγγραφών, τα οποία δεν μεταβάλλονται. Ωστόσο,όπως θα δούμε και πιο κάτω, η προσέγγιση αυτή θα μπορούσε να χρησιμοποιηθεί ακόμη και όταν δεν πληρούνται ορισμένες από τις παραπάνω προυποθέσεις.

4.3.1. Μορφή Δεδομένων Οργανισμών

Στην εργασία αυτή εστιάσαμε στο γεγονός ότι τα δεδομένα έχουν αναπαρασταθεί με τη μορφή ενός πίνακα, οι στήλες του οποίου αντιστοιχούν στα χαρακτηριστικά των ασθενών. Ωστόσο, πολλές φορές τα δεδομένα των ασθενών αποθηκεύονται σε Βάσεις Δεδομένων που ακολουθούν με μια διαφορετική μορφή/σχήμα. Για παράδειγμα, σε ένα σχεσιακό σύστημα τα δεδομένα ενός και μόνο ασθενή θα ήταν διάσπαρτα σε διάφορους πίνακες. Ωστόσο, ακόμη και στην περίπτωση αυτή, τα δεδομένα των ασθενών (ή μέρος) θα μπορούσαν να εξαχθούν με την παραπάνω μορφή μέσω μιας αυτόματης διαδικασίας και ακολούθως να εφαρμοστεί η προηγούμενη διαδικασία. Για παράδειγμα, οι διαταραχές με τις οποίες έχει διαγνωστεί ένας ασθενής θα μπορούσαν να βρίσκονται σε έναν ξεχωριστό πίνακα μιας σχεσιακής βάσης δεδομένων, ο οποίος συνδέεται με τον πίνακα των ασθενών (όπου έχουν καταγραφεί τα βασικά του στοιχεία όπως για παράδειγμα η ημερομηνία ή έτος γέννησής του) μιας σχεσιακής βάσης μέσω ενός ξένου κλειδιού. Στην περίπτωση αυτή, θα μπορούσαμε να εκφράσουμε όλα τα δεδομένα των ασθενών με τη μορφή ενός πίνακα, στον οποίο είτε θα υπάρχει ένα και μόνο πεδίο στο οποίο θα βάζουμε όλες τις πιθανές διαταραχές (στην περίπτωση που αυτές δεν συνοδεύονται από επιπλέον πληροφορία, όπως το πότε διαγνώστηκε ο ασθενής με αυτή) είτε θα δημιουργούσαμε τόσα πεδία όσες και οι πιθανές διαταραχές και η τιμή του πεδίου θα έδειχνε εάν ο ασθενής διαγνώστηκε με αυτή ή όχι. Στη δεύτερη περίπτωση μπορούμε να καταγράψουμε

επιπλέον πληροφορία (π.χ., ημερομηνία που έγινε η εκάστοτε διάγνωση) είτε μέσα στο ίδιο είτε σε ξεχωριστό κελί – στήλη.

Στο σημείο αυτό, να τονίσουμε ότι, πολλές φορές, τα δεδομένα είναι διαθέσιμα σε μία μορφή που είναι αρκετά κοντά στον τρόπο αποθήκευσής τους σε ένα, για παράδειγμα, σύστημα διαχείρισης σχεσιακών δεδομένων. Κατ' επέκταση, η περιγραφή των πεδίων και των τιμών τους απέχει σημαντικά από ένα μοντέλο αναπαράστασης γνώσης [69], το οποίο δυσχεραίνει την κατανόηση τόσο των παραμέτρων όσο και των τιμών τους. Αυτό γίνεται κατανοητό, εαν λάβουμε υπόψη την εκτεταμένη χρήση των συντομεύσεων στον χώρο της κλινικής έρευνας και περίθαλψης [3]. Συνεπώς, ακόμη και οι κατώτερου επιπέδου τεχνικές επεξεργασίας κειμένου (π.χ., εντοπισμός των λέξεων μιας φράσης) έχουν αυξημένη δυσκολία, εξαιτίας της ευρείας χρήσης των σημείων στίξης στη δημιουργία των ονομάτων και εκφράσεων. Στην παρούσα εργασία, αυτό είχε άμεση επίπτωση στην επεξεργασία των μεταδεδομένων που εξήχθησαν από κάθε ένα από τα δύο κέντρα έρευνας και στον εντοπισμό των διανυσματικών αναπαραστάσεων των επιμέρους λέξεων μιας έκφρασης.

4.3.1.1. Μη Δομημένα Δεδομένα

Στα πλαίσια της εργασίας αυτής ασχοληθήκαμε με την εναρμόνιση δομημένων ή ημι-δομημένων δεδομένων που προέρχονται από διαφορετικούς οργανισμούς και αφορούν ένα συγκεκριμένο πεδίο της γνώσης, λαμβάνοντας υπόψη και πιθανούς περιορισμούς ιδιωτικότητας και ασφάλειας που συχνά συνοδεύουν τα δεδομένα. Ωστόσο, μεγάλος όγκος δεδομένων εξακολουθεί να υπάρχει σε ημι-δομημένη ή μη δομημένη μορφή, περιπλέκοντας τη διαδικασία ενοποίησης των δεδομένων. Τα τελευταία χρόνια υπάρχει αρκετή έρευνα γύρω από την επεξεργασία κειμένου και την εναρμόνιση πληροφορίας που υπάρχει σε αυτό. Ωστόσο, υπάρχει πολύ λιγότερη έρευνα στον τομέα της επεξεργασίας εικόνας και βίντεο και ειδικότερα της εναρμόνισης των δεδομένων αυτών.

4.3.2. Διαδικασία Εναρμόνισης και Πιθανοί Κίνδυνοι

Ακολουθώντας τη μεθοδολογία που παρουσιάσαμε στο έγγραφο αυτό και χρησιμοποιώντας τα εργαλεία που αναπτύχθηκαν, μπορούμε να εκφράσουμε τα αρχικά δεδομένα σε μια νέα – διαφορετική μορφή, με βάση τους κανόνες συσχέτισης που έχουν χρησιμοποιηθεί. Ωστόσο, θα πρέπει να προσέξουμε, έτσι ώστε η αρχική σημασία των δεδομένων να μην μεταβληθεί κατά την μετάβαση από την μία μορφή αναπαράστασης στην άλλη. Αυτό απαιτεί τη βαθύτερη κατανόηση της πληροφορίας που έχει καταγραφεί σε κάθε παράμετρο των αρχικών δεδομένων αλλά και του μοντέλου που έχει αναπτυχθεί, έτσι ώστε να μπορέσουμε στη συνέχεια να καθορίσουμε την ακριβή σχέση που υπάρχει μεταξύ τους. Ωστόσο, όπως είδαμε και στα

αποτελέσματα της εναρμόνισης των δεδομένων, ένα σημαντικό μέρος της πληροφορίας των αρχικών δεδομένων αγνοήθηκε κατά την παραπάνω διαδικασία εναρμόνισης, λόγω αδυναμίας καθορισμού συσχέτισης με τους όρους του μοντέλου που αναπτύχθηκε. Επίσης, ακόμη και οι παράμετροι που συσχετίστηκαν με τους αντίστοιχους όρους του μοντέλου αναφοράς επιτρέπουν τη μετάβαση από την μία αναπαράσταση στην άλλη με τη μικρότερη δυνατή απώλεια γνώσης. Για παράδειγμα, όταν ένα συγκεκριμένο μη-στεροειδές αντιφλεγμονώδες φάρμακο (NSAID) καταγραφόταν στα εναρμονισμένα δεδομένα η κατηγορία αυτή του φαρμάκου και όχι το συγκεκριμένο φάρμακο, καθώς δεν υπήρχε στο μοντέλο. Ωστόσο, σε ορισμένες περιπτώσεις οι έννοιες που υπάρχουν στο αρχικό μοντέλο μπορεί να είναι μερικώς επικαλυπτόμενες, γεγονός που δημιουργεί επιπρόσθετες δυσκολίες στην ακριβή τους μετάφραση, η οποία πρακτικά περιορίζεται στο τμήμα εκείνης της πληροφορίας που μπορεί να συσχετιστεί με τους όρους που υπάρχουν στο μοντέλο που έχει δημιουργηθεί.

4.3.2.1. Πρόωρη και Αργή Εναρμόνιση Δεδομένων

Η χρήση δεδομένων από διαφορετικές πηγές μπορεί να συμβάλει στη δημιουργία καλύτερων μοντέλων. Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν, σύμφωνα με τους συγγραφείς του έργου [70], με τρεις διαφορετικούς τρόπους. Στην πρώτη προσέγγιση, που ονομάζεται πρόωρη ενοποίηση (early integration), θα πρέπει πρώτα να εκφράσουμε τα δεδομένα που προέρχονται από διαφορετικές πηγές δεδομένων με μία κοινή μορφή και έπειτα να χρησιμοποιήσουμε τα δεδομένα αυτά για τη δημιουργία του μοντέλου. Στη δεύτερη προσέγγιση, που ονομάζεται αργή ενοποίηση (late integration), θα πρέπει να αναπτύξουμε διαφορετικά μοντέλα για κάθε μία από τις πηγές δεδομένων και να χρησιμοποιήσουμε τα αποτέλεσμα των εξόδων τους, για να εκπαιδεύσουμε ένα νέο μοντέλο. Στην τρίτη προσέγγιση, που ονομάζεται ενδιάμεση ενοποίηση (intermediate integration), εκπαιδεύουμε ένα αρκετά εκφραστικό μοντέλο (όπως είναι για παράδειγμα ένα βαθύ νευρωνικό δίκτυο), το οποίο να είναι σε θέση να διαχειριστεί αποτελεσματικά την από κοινού αναπαράσταση περισσότερων του ενός συνόλου δεδομένων και κατ' επέκταση να υποστηρίξει τους σκοπούς μας. Η τελευταία αυτή προσέγγιση, σύμφωνα με τους συγγραφείς του παραπάνω έργου, μπορεί να πετύχει καλύτερα αποτελέσματα.

4.3.3. Εφαρμογή σε Διαφορετικό Πεδίο Γνώσης

Στην εργασία αυτή εστιάσαμε στην εναρμόνιση δεδομένων ασθενών που πάσχουν από ένα συγκεκριμένο σύνδρομο. Ωστόσο, η μεθοδολογία που ακολουθήσαμε μπορεί να χρησιμοποιηθεί και σε κάποιο εντελώς διαφορετικό πεδίο γνώσης, για την έκφραση των δεδομένων των οργανισμών με μία κοινή μορφή. Για τον σκοπό αυτό, είναι απαραίτητη η δημιουργία ενός μοντέλου αναφοράς που θα μπορεί να καλύψει τις

ανάγκες του πεδίου αυτού καθώς και τα σχετικά σενάρια καθορισμού συσχετίσεων μεταξύ των όρων των οργανισμών και αυτών του μοντέλου αναφοράς που θα αναπτυχθεί. Για τον σκοπό αυτό, μπορούμε να χρησιμοποιήσουμε το εργαλείο που αναπτύχθηκε για την εξαγωγή μετα-δεδομένων από κάθε έναν οργανισμό, έτσι ώστε να χρησιμοποιήσουμε ακολούθως την πληροφορία αυτή για τη δημιουργία του μοντέλου και τον καθορισμό των σεναρίων συσχέτισης. Ακολούθως, μπορούμε να χρησιμοποιήσουμε τα εργαλεία που αναπτύχθηκαν και παρουσιάστηκαν στην εργασία αυτή, για τον καθορισμό των συσχετίσεων μεταξύ των όρων που χρησιμοποιούνται από κάθε οργανισμό και αυτών που υπάρχουν στο μοντέλο αναφοράς, με τη βοήθεια του εργαλείου καθορισμού συσχέτισης που παρουσιάστηκε στις προηγούμενες παραγράφους. Τέλος, ο κάθε οργανισμός μπορεί να χρησιμοποιήσει το εργαλείο μετατροπής δεδομένων, για να εκφράσει τα δεδομένα που έχει στην επιθυμητή μορφή.

Η σελίδα αυτή είναι σκόπιμα λευκή

5. Σύνοψη

Η εναρμόνιση των δεδομένων που προέρχονται από διάφορους οργανισμούς αποσκοπεί στην έκφρασή τους με ένα κοινό μοντέλο, το οποίο διευκολύνει την περαιτέρω ανάλυση και επεξεργασία των δεδομένων αυτών και την εξαγωγή χρήσιμων συμπερασμάτων, τα οποία θα μπορούσαν να αξιοποιηθούν από τους οργανισμούς για τη βελτίωση των εσωτερικών τους διεργασιών και τη δημιουργία ενός ανταγωνιστικού πλεονεκτήματος. Ειδικά στον χώρο της κλινικής έρευνας, η χρήση δεδομένων από διάφορα κέντρα για τη μελέτη ενός ιατρικού φαινομένου είναι επιτακτική για την αξιοπιστία των αποτελεσμάτων της έρευνας. Για τον σκοπό αυτό αναπτύχθηκε ένα σύστημα αποτελούμενο από τρία διαφορετικά εργαλεία, που επιτρέπουν στους χρήστες των οργανισμών να εκφράσουν τα δεδομένα τους με ασφάλεια, χρησιμοποιώντας τους όρους ενός κοινού μοντέλου που αναπτύχθηκε. Οι τεχνικές μηχανικής μάθησης, και ειδικότερα τα μοντέλα γλώσσας, έχουν εξέχουσα σημασία στην διαδικασία αυτή, καθώς επιτρέπουν την καλύτερη αξιοποίηση των διαθέσιμων δεδομένων και την επιτάχυνση της διαδικασίας της εναρμόνισης. Ειδικότερα, με βάση την χρήση κλασικών τεχνικών μηχανικής μάθησης και προ-εκπαιδευμένων μοντέλων γλώσσας, μπορέσαμε να αυξήσουμε τον όγκο της πληροφορίας που εναρμονίστηκε μέσω του καλύτερου σχεδιασμού του μοντέλου αναφοράς και να επιταχύνουμε τη διαδικασία της εναρμόνισης μέσω του αυτόματου εντοπισμού πιθανών συσχετίσεων μεταξύ των όρων διαφορετικών μοντέλων. Η προσέγγιση που ακολουθήθηκε, και ειδικότερα τα εργαλεία που αναπτύχθηκαν και οι τεχνικές που εφαρμόστηκαν, θα μπορούσαν να χρησιμοποιηθούν για την εναρμόνιση των δεδομένων που προέρχονται από ένα διαφορετικό πεδίο γνώσης, κάνοντας τις απαραίτητες τροποποιήσεις / παραμετροποιήσεις των εργαλείων του συστήματος.

Η σελίδα αυτή είναι σκόπιμα λευκή

6. Παράρτημα

6.1. Δεδομένα

6.1.1. Κατηγορίες Δεδομένων και Μετρικές Απόδοσης

6.1.1.1. Labeled vs Unlabeled Data

Τα δεδομένα που χρησιμοποιούνται στη μηχανική μάθηση μπορούν να ενταχθούν σε δύο μεγάλες κατηγορίες, με βάση την ύπαρξη ή μη ύπαρξη ετικετών (labels). Στην πρώτη κατηγορία ανήκουν τα δεδομένα τα οποία συνοδεύονται από κάποια ετικέτα (annotated/labeled data), η οποία δείχνει την ευρύτερη κατηγορία στην οποία αυτά ανήκουν ή κάποιο αποτέλεσμα που θα μπορούσε να εξαχθεί από αυτά. Στη δεύτερη κατηγορία ανήκουν τα δεδομένα για τα οποία δεν υπάρχει κάποια τέτοια ετικέτα (unlabeled data).

6.1.1.2. Training / Validation / Testing Datasets

Τα δεδομένα ενός επισημασμένου συνόλου δεδομένων (annotated dataset), μπορούν να χωριστούν σε τρεις κατηγορίες, ανάλογα με τον τρόπο χρήσης τους. Τα δεδομένα εκπαίδευσης (training data) χρησιμοποιούνται για τη δημιουργία του μοντέλου και ειδικότερα για τον προσδιορισμό των παραμέτρων του μοντέλου μηχανικής μάθησης, μέσω μιας αυτόματης διαδικασίας. Τα δεδομένα ελέγχου (test data) χρησιμοποιούνται για τον προσδιορισμό της απόδοσης του μοντέλου, μέσω του υπολογισμού ορισμένων μετρικών απόδοσης. Εκτός από τα παραπάνω δύο σύνολα δεδομένων, υπάρχει και ένα τρίτο σύνολο που περιλαμβάνει τα δεδομένα επικύρωσης (validation data) και χρησιμοποιείται συχνά για τον προσδιορισμό των υπερπαραμέτρων (hyper-parameters) του μοντέλου, μέσω δοκιμών.

Τα τρία παραπάνω σύνολα θα πρέπει να είναι ξένα μεταξύ τους. Τα αρχικά δεδομένα θα πρέπει να ενταχθούν σε ένα από τα τρία παραπάνω σύνολα, πριν χρησιμοποιηθούν για τις ανάγκες της μηχανικής μάθησης. Κατ' επέκταση, ένα μόνο μέρος των αρχικών δεδομένων χρησιμοποιείται για την εκπαίδευση του μοντέλου (π.χ., το 70% των αρχικών δεδομένων), ενώ ένα άλλο μέρος (π.χ. το 30%) μόνο στο τέλος, κατά τη φάση μέτρησης της απόδοσης του μοντέλου (hold-out). Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε την τεχνική K-Fold Cross Validation[24], σύμφωνα με την οποία χωρίζουμε τα δεδομένα μας σε K ομάδες/σύνολα ίδιου μεγέθους και χρησιμοποιούμε κάθε φορά τα K-1 σύνολα για έναν σκοπό (π.χ.,

[24] A Gentle Introduction to k-fold Cross-Validation, https://machinelearningmastery.com/k-fold-cross-validation/

training) και το εναπομείναν σύνολο για έναν άλλον σκοπό (π.χ., testing). Η μέτρηση της απόδοσης ενός μοντέλου προκύπτει από τη μέση τιμή των μετρικών, που έχουν υπολογιστεί καθεμία από τις K φορές που έχουμε επαναλάβει την παραπάνω διαδικασία.

6.1.1.3. Μετρικές Απόδοσης

Για τη μέτρηση της απόδοσης ενός μοντέλου μηχανικής μάθησης και ειδικότερα ενός μοντέλου δυαδικής κατηγοριοποίησης δεδομένων (binary classification problem), υπάρχουν διάφορες μετρικές[25], όπως είναι τα accuracy, precision, recall και F-score/F-measure (συνδυασμός των δύο προηγούμενων), τα οποία βασίζονται στο πλήθος των δειγμάτων της κάθε κατηγορίας που προβλέπονται σωστά (true positive και true negative) ή λάθος (false positive και false negative), τα οποία αποτελούν τα βασικά συστατικά ενός πίνακα σύγχυσης (confusion matrix). Οι παραπάνω μετρικές θα μπορούσαν να χρησιμοποιηθούν και για μοντέλα πολλαπλής κατηγοριοποίησης (multi-class classification). Ωστόσο, στην περίπτωση αυτή χρειάζεται να συνοψίσουμε (micro/macro averaging)[26] τους επιμέρους δείκτες που υπολογίζονται για τα precision και recall κάθε κλάσης. Οι καμπύλες του Precision-Recall (PR) και Receiver operating characteristic (ROC) μας βοηθούν να καταλάβουμε καλύτερα πώς συμπεριφέρεται το μοντέλο για διάφορες τιμές του κατωφλιού (threshold). Στην περίπτωση της παλινδρόμησης (regression) μπορούμε να χρησιμοποιήσουμε τη ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error - RMSE) καθώς και τους δείκτες R^2 και Adjusted R^2.

6.1.2. Εξισορρόπηση Δεδομένων

Ένα σύνολο δεδομένων είναι μη-ισορροπημένο (imbalanced), όταν οι κατηγορίες στις οποίες ανήκουν τα δεδομένα δεν αντιπροσωπεύονται με τον ίδιο τρόπο σε αυτό. Πρακτικά, αυτό συμβαίνει, όταν δεν υπάρχει ο ίδιος ή σχεδόν ίδιος αριθμός δειγμάτων από κάθε κατηγορία. Το φαινόμενο αυτό παρατηρείται συχνά στα δεδομένα που συλλέγονται από διαδικασίες ή συστήματα που υπάρχουν στον πραγματικό κόσμο, όπως για παράδειγμα τα δεδομένα που εκφράζουν τη φυσιολογική ή μη συμπεριφορά ενός συστήματος. Το γεγονός αυτό (ότι δηλαδή το σύνολο δεδομένων είναι μη ισορροπημένο) έχει συχνά άμεση επίδραση στα μοντέλα μηχανικής μάθησης που αναπτύσσονται με βάση τα δεδομένα αυτά.

[25] 12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2023), https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/
[26] Accuracy, precision, and recall in multi-class classification, https://www.evidentlyai.com/classification-metrics/multi-class-metrics

Ένας τρόπος για να περιοριστεί το πρόβλημα αυτό (της ύπαρξης μη ισορροπημένων δεδομένων), είναι να επιλέξουμε τυχαία ορισμένα από δεδομένα που ανήκουν στην κατηγορία που υπερτερεί, ώστε, εν τέλει, να πετύχουμε την εξισορρόπηση των δεδομένων (majority class under sampling). Εναλλακτικά, θα μπορούσαμε να προσθέσουμε (είτε τα ίδια είτε νέα, μέσω της προσθήκης κάποιου θορύβου) επιπλέον στοιχεία που να ανήκουν στην κατηγορία με τα λιγότερα δείγματα (minority class oversampling), έτσι ώστε να προσεγγίσουμε τον αριθμό των στοιχείων της άλλης κατηγορίας (στην περίπτωση που έχουμε μόνο δύο κατηγορίες). Στο έγγραφο [71][72] περιγράφεται μία μέθοδος, γνωστή ως Synthetic Minority Oversampling Technique (SMOTE), σύμφωνα με την οποία μπορούμε να φτιάξουμε ένα εξισορροπημένο σύνολο δεδομένων συνδυάζοντας τις δύο παραπάνω προσεγγίσεις. Ειδικότερα, δημιουργούνται νέα δεδομένα (synthetic data) από την κατηγορία που μειοψηφεί (synthetic minority oversampling) και ακολούθως επιλέγονται τυχαία ορισμένα από τα στοιχεία της κατηγορίας που υπερτερεί (majority under sampling), έτσι ώστε το σύνολο που προκύπτει να έχει τα ίδια ή σχεδόν τα ίδια στοιχεία από κάθε κατηγορία. Το σύνολο δεδομένων που προκύπτει από την προηγούμενη διαδικασία μπορεί να συμβάλει στη δημιουργία ενός καλύτερου μοντέλου, εν συγκρίσει με αυτό που θα πετυχαίναμε, εάν ακολουθούσαμε αποκλειστικά και μόνο μία από τις δύο παραπάνω προσεγγίσεις.

Ένα μοντέλο μηχανικής μάθησης, για την κατηγοριοποίηση των δεδομένων (classification model), συχνά ακολουθεί μια διακριτική προσέγγιση (discriminative approach), επιδιώκοντας πρακτικά να βρει / μάθει τη γραμμή / υπερ-πλάνο / σύνορο μεταξύ των κλάσεων. Προς αυτήν την κατεύθυνση, μια διαφορετική έκδοση του παραπάνω αλγορίθμου εστιάζει στα δεδομένα που είναι κοντά στο σύνορο κατά τη διαδικασία επιλογής/δημιουργίας νέων δειγμάτων από την κατηγορία που μειοψηφεί γνωστή ως Borderline-SMOTE [73] και με βάση τα πειράματα που έγιναν μπορεί να συμβάλει στη δημιουργία μοντέλων που πετυχαίνουν ακόμη καλύτερα αποτελέσματα. Οι συγγραφείς της δημοσίευσης [74] δίνουν περισσότερη έμφαση στα δεδομένα που ανήκουν στην κατηγορία που μειοψηφεί και τα οποία είναι πιο δύσκολο να μάθουν κατά τη δημιουργία των συνθετικών δεδομένων. Η μέθοδος αυτή, γνωστή ως Adaptive Synthetic Sampling Approach (ADASYN), προσπαθεί επιπρόσθετα να περιορίσει την προκατάληψη (bias) που υπάρχει μεταξύ των δεδομένων της ιδίας κλάσης και με βάση τα πειράματα που έγιναν (και παρουσιάζονται στο προαναφερθέν έγγραφο) μπορεί να πετύχει ικανοποιητικά αποτελέσματα.

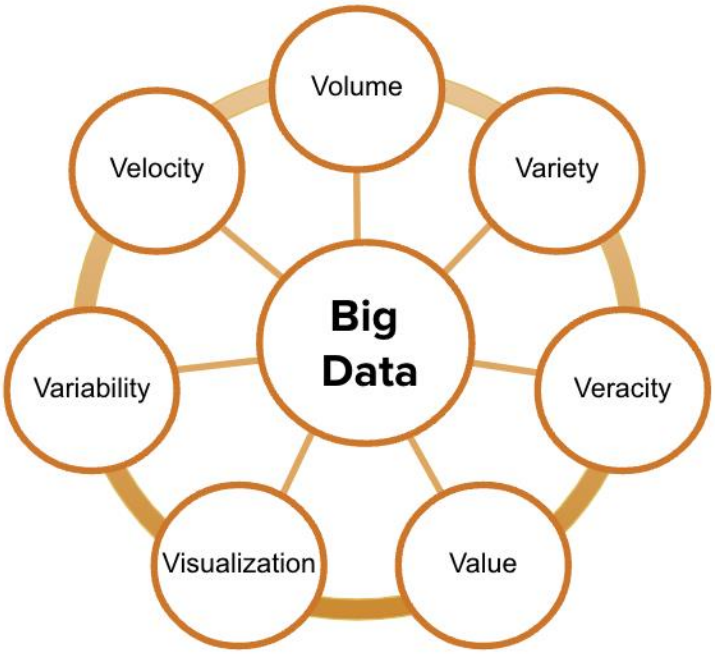
6.1.3. Ανεξάρτητα Ομοιόμορφα Κατανεμημένα Δεδομένα

Μια άλλη σημαντική παράμετρος, που σχετίζεται άμεσα με τα δεδομένα που χρησιμοποιούνται για τις ανάγκες της μηχανικής μάθησης (είτε για λόγους εκπαίδευσης είτε επικύρωσης είτε ελέγχου), είναι η

υπόθεση της ανεξαρτησίας και ομοιόμορφης κατανομής των δεδομένων (independent identically distributed – i.i.d.). Σύμφωνα με την υπόθεση αυτή, τα δεδομένα θα πρέπει να είναι ανεξάρτητα μεταξύ τους και όλα αυτά να προέρχονται από την ίδια (αρχικά άγνωστη) κατανομή, την οποία αρκετές φορές επιδιώκει να μάθει ένα μοντέλο μηχανικής μάθησης. Ωστόσο, η υπόθεση αυτή πολλές φορές παραβιάζεται είτε γιατί τα δεδομένα σχετίζονται μεταξύ τους (όπως είναι για παράδειγμα τα δεδομένα μιας χρονοσειράς) είτε γιατί υπάρχει κάποια αλλαγή στην κατανομή των δεδομένων (όπως για παράδειγμα τα δεδομένα που δημιουργούνται από διαφορετικές πηγές δεδομένων ή τα δεδομένα ροής – stream data) και είναι γνωστή και ως μετατόπιση δεδομένων (data shift).

6.2. Μεγάλα Δεδομένα

Ο όρος «μεγάλα δεδομένα» (big data) [75] χρησιμοποιείται ευρέως, για να αναφερθούμε σε έναν υπερβολικά μεγάλο όγκο δεδομένων (που συνήθως είναι μερικώς ή μη δομημένα). Ωστόσο, ο όρος αυτός δηλώνει πολλά περισσότερα από αυτό που αναφέρει το όνομά του. Για την καλύτερη κατανόηση του όρου αυτού θα αναφερθούμε. εν συντομία. στα χαρακτηριστικά που έχουν τα μεγάλα δεδομένα, γνωστά και ως 3/4/5/7-Vs, λόγω του αρχικού χαρακτήρα του ονόματός τους) – Εικόνα 19.



Εικόνα 19: Τα χαρακτηριστικά[27] των μεγάλων δεδομένων

[27] The 7 V's of Big Data, https://impact.com/marketing-intelligence/7-vs-big-data/

6.2.1. Τα Χαρακτηριστικά των Μεγάλων Δεδομένων

Τρία από τα βασικότερα χαρακτηριστικά των μεγάλων δεδομένων είναι ο όγκος (volume), η ταχύτητα (velocity) και η ποικιλία (variety), γνωστά και ως 3Vs. Ο όγκος αναφέρεται στο μέγεθος των δεδομένων που παράγονται και συλλέγονται καθημερινά (σύμφωνα με μια παλαιότερη έρευνα του 2011, τα δεδομένα που δημιουργούνται μέσα σε 2 ημέρες ξεπερνούν τα 1.8ZB). Η ταχύτητα εστιάζει στον χρόνο που παράγονται τα δεδομένα και την ανάγκη για γρήγορη και έγκαιρη συλλογή και επεξεργασία τους. Η ποικιλία εστιάζει στον τρόπο αναπαράστασης των δεδομένων, τα οποία μπορεί να είναι δομημένα, ημι-δομημένα ή μη δομημένα (π.χ., εικόνες και βίντεο). Στα παραπάνω χαρακτηριστικά των δεδομένων σύντομα προστέθηκε ένα ακόμη χαρακτηριστικό, σχετικά με την αξία (value) των δεδομένων, δημιουργώντας μία λίστα από 4 πλέον χαρακτηριστικά, γνωστά ως 4Vs. Αργότερα προστέθηκε ένα επιπλέον χαρακτηριστικό, το οποίο έχει να κάνει με την εγκυρότητα ή ακρίβεια των δεδομένων και είναι γνωστό ως ειλικρίνεια (veracity), αυξάνοντας τα χαρακτηριστικά των δεδομένων στα 5Vs. Στη συνέχεια, προστέθηκαν αρκετά ακόμη χαρακτηριστικά για την καλύτερη περιγραφή/κατανόηση των μεγάλων δεδομένων, όπως για παράδειγμα η ποικιλομορφία των δεδομένων (variability) και η απεικόνιση (visualization) σχηματίζοντας τα 7Vs (Εικόνα 19).

6.2.1.1. Δομημένα – Ημιδομημένα και Μη-δομημένα Δεδομένα

Σε κάθε μορφή μπορούμε να πούμε ότι τα δεδομένα έχουν εκφραστεί με κάποιο τρόπο που διευκολύνει την πρόσβαση προς αυτά (μέσω κάποιας συγκεκριμένης γλώσσας/πρωτόκολλο) και ακολουθούν κάποιο μοντέλο, το οποίο καθορίζει τη σημασία των επιμέρους στοιχείων και τη σχέση αυτών.

Στην περίπτωση των δομημένων δεδομένων (Structured Data), όπως για παράδειγμα είναι τα δεδομένα που αποθηκεύονται σε μία σχεσιακή βάση, τα δεδομένα ακολουθούν ένα συγκεκριμένο σχήμα που εστιάζει στην αποθήκευση (και γενικότερα διαχείριση) των δεδομένων και σε συνεργασία με τη γλώσσα / σύνταξη που υποστηρίζεται (στην προκειμένη περίπτωση SQL) μπορεί κάποιος να έχει πρόσβαση σε αυτά. Στο παραπάνω παράδειγμα, η γλώσσα αυτή επιτρέπει στον χρήστη να εκφράσει με ακρίβεια τα δεδομένα που αναζητεί και ακολούθως να λάβει τα αντίστοιχα αποτελέσματα με βάση και τη σημασιολογία της γλώσσας αυτής. Ένα άλλο μοντέλο (πιο κοντά στον τρόπο με τον οποίο αντιλαμβάνεται ο άνθρωπος τις οντότητες που υπάρχουν στον πραγματικό κόσμο) είναι το μοντέλο που βασίζεται σε γράφους, όπως είναι για παράδειγμα το RDF[28]. Και σε αυτήν την περίπτωση τα δεδομένα χρειάζεται να ακολουθούν ένα συγκεκριμένο μοντέλο (RDF σχήμα / OWL οντολογία) και η πρόσβαση επιτρέπεται σε αυτά μέσω κάποιας συγκεκριμένης γλώσσας (π.χ., SPARQL).

[28] Resource Description Framework (RDF), https://www.w3.org/RDF/

Στην περίπτωση των μη-δομημένων δεδομένων (Unstructured Data), όπως για παράδειγμα είναι τα κείμενα, οι εικόνες και τα βίντεο, μπορούμε να πούμε ότι το μοντέλο είναι κατά κάποιον τρόπο εκφυλισμένο, ενώ η πρόσβαση σε αυτά είναι στοιχειώδης, υπό την έννοια ότι μπορούμε να έχουμε πρόσβαση στο σύνολο της πληροφορίας (π.χ., μια συγκεκριμένη εικόνα ή έστω μια συγκεκριμένη πρόταση ενός κειμένου), χωρίς, ωστόσο, να υπάρχει άμεση δυνατότητα να εντοπίσουμε με ευκολία και ακρίβεια τα επιμέρους στοιχεία και τη σημασία τους (π.χ., τις έννοιες/οντότητες στις οποίες αναφέρονται, και τις μεταξύ τους σχέσεις).

Στα ημι-δομημένα δεδομένα (Semi-structured Data) ανήκουν παραδοσιακά τα δεδομένα που είναι εκφρασμένα χρησιμοποιώντας τις γλώσσες XML, JSON (ή έστω και YAML) και έχουν ως στόχο να επιτρέψουν στον χρήστη να έχει πρόσβαση στα επιμέρους στοιχεία ενός τέτοιου εγγράφου, ενώ συχνά τα δεδομένα αυτά ακολουθούν ένα προκαθορισμένο σχήμα που διέπει τον τρόπο δόμησης αυτών. Στην κατηγορία αυτή ενδεχομένως να μπορούσαμε να εντάξουμε ακόμη και τα δεδομένα που προέρχονται από έναν συνδυασμό των παραπάνω. Για παράδειγμα, θα μπορούσαμε να έχουμε μια σχεσιακή βάση δεδομένων, στην οποία να αποθηκεύουμε τις προτάσεις ενός κειμένου ως ξεχωριστές εγγραφές.

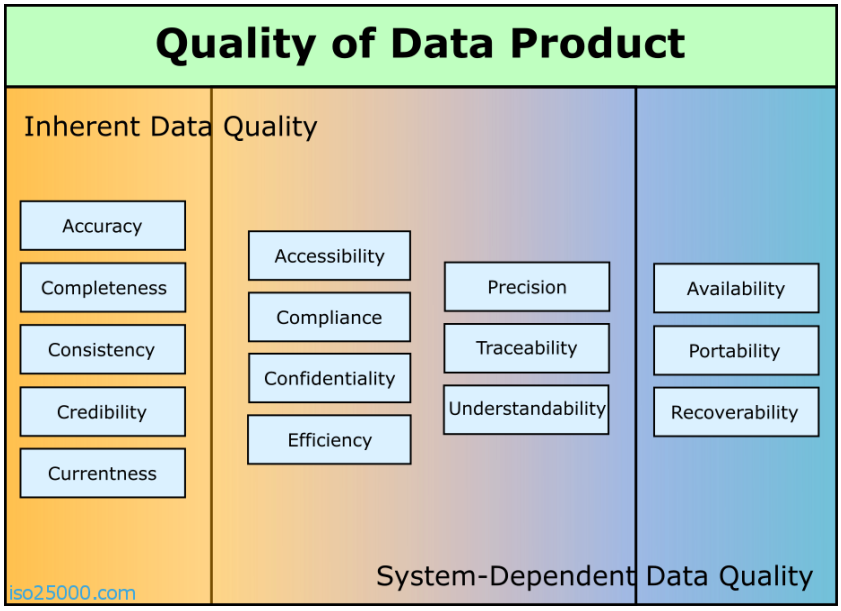
6.2.2. Ποιότητα Δεδομένων

6.2.2.1. Χαρακτηριστικά / Διαστάσεις Ποιότητας Δεδομένων

Στη βιβλιογραφία, η ποιότητα των δεδομένων (Data Quality) ορίζεται συχνά ως η καταλληλότητά τους για κάποια χρήση (fitness for use) [76][77]. Σε άλλα site στο διαδίκτυο (όπως για παράδειγμα η σελίδα της IBM[29]) η ποιότητα των δεδομένων έχει να κάνει με το κατά πόσο πληρούν τα δεδομένα κάποια κριτήρια που έχουν να κάνουν με ακρίβεια, πληρότητα, εγκυρότητα, επικαιρότητα και καταλληλότητα για έναν σκοπό.

Σύμφωνα με το ISO/IEC 25012[30], τα χαρακτηριστικά των δεδομένων που σχετίζονται με την ποιότητά τους (γνωστά επίσης και ως διαστάσεις ποιότητας δεδομένων – Data Quality Dimensions) μπορούμε να τα εντάξουμε σε δύο κατηγορίες (Εικόνα 20). Η πρώτη κατηγορία περιλαμβάνει τα χαρακτηριστικά που έχουν τα δεδομένα από τη φύση τους (inherent data quality), όπως είναι για παράδειγμα η ακρίβεια (accuracy), η πληρότητα (completeness) και η συνοχή (consistency). Η δεύτερη κατηγορία περιλαμβάνει τα χαρακτηριστικά εκείνα που σχετίζονται με τη χρήση των δεδομένων από τα συστήματα (system-dependent data quality), όπως είναι για παράδειγμα η διαθεσιμότητα των δεδομένων (availability) και η συμμόρφωσή τους με κάποια πρότυπα (compliance).

[29] What is data quality?, https://www.ibm.com/topics/data-quality
[30] ISO/IEC 25012, https://iso25000.com/index.php/en/iso-25000-standards/iso-25012



Εικόνα 20: Χαρακτηριστικά Ποιότητας Δεδομένων του ISO/IEC 25012

Τα παραπάνω χαρακτηριστικά μπορούν να μας παρέχουν μια καλύτερη εικόνα για τα δεδομένα. Συνεπώς, είναι ιδιαίτερα σημαντικό να καθορίσουμε τον τρόπο ή τα βήματα που πρέπει να ακολουθήσουμε ή τις μετρικές που χρειάζεται να υπολογίσουμε (data quality metrics) για την ποιοτική ή ποσοτική αξιολόγηση των παραπάνω χαρακτηριστικών. Οι συγγραφείς του έργου [78] εστιάζουν στα χαρακτηριστικά που έχουν τα δεδομένα από μόνα τους (intrinsic characteristics), τα οποία σχετίζονται άμεσα με τη διαδικασία που ακολουθήθηκε για τη δημιουργία τους. Στο έγγραφο αυτό οι συγγραφείς παρουσιάζουν 34 δείκτες (indicators) που μπορούν να χρησιμοποιηθούν για τον σκοπό αυτό, οι οποίοι είναι οργανωμένοι σε 4 κατηγορίες: ακεραιότητα (integrity), πληρότητα (completeness), συνοχή (consistency) και ακρίβεια (accuracy). Ωστόσο, η ακριβής σημασία των δεικτών αυτών έχει να κάνει και με τον τρόπο υλοποίησής τους.

6.2.2.2. Προβλήματα Ποιότητας Δεδομένων

Τα προβλήματα, που μπορούν να προκύψουν σχετικά με την ποιότητα των δεδομένων, μπορούμε να τα χωρίσουμε σε δύο ευρύτερες κατηγορίες. Τα προβλήματα ποιότητας που προκύπτουν, όταν εξετάζουμε μία μόνο πηγή δεδομένων και αυτά που προκύπτουν, όταν εξετάζουμε πολλές πηγές δεδομένων μαζί [77]. Στην πρώτη περίπτωση θα πρέπει να εξετάσουμε τόσο το μοντέλο των δεδομένων και τους κανόνες/περιορισμούς που υπάρχουν όσο και τα δεδομένα αυτά καθ' αυτά και πιθανά λάθη που μπορεί να υπάρχουν, αντιφάσεις μεταξύ των δεδομένων και διπλο-εγγραφές. Στη δεύτερη περίπτωση θα πρέπει να λάβουμε υπόψη την ετερογένεια των δεδομένων τόσο σε επίπεδο σχήματος όσο και οντοτήτων και τη δυνατότητα ευθυγράμμισης αυτών.

Στο έγγραφο [76] οι συγγραφείς εστιάζουν στον κύκλο ζωής των μεγάλων δεδομένων και τα διάφορα στάδια από τα οποία αυτά περνούν, επισημαίνοντας ότι είναι σημαντικό να εξετάζουμε την ποιότητα των δεδομένων σε κάθε στάδιο, λαμβάνοντας υπόψη το γεγονός ότι η ποιότητα των δεδομένων σε ένα επόμενο στάδιο εξαρτάται από την ποιότητα των δεδομένων που είχαν σε προηγούμενο στάδιο. Λαμβάνοντας υπόψη το γεγονός ότι η εναρμόνιση των δεδομένων είναι και αυτή μια διαδικασία που δέχεται ως είσοδο μια πηγή δεδομένων και ακολούθως εκφράζει τα δεδομένα αυτά στην επιθυμητή μορφή, θα πρέπει να λάβουμε υπόψη τόσο την ποιότητα της αρχικής μορφής των δεδομένων, όσο και της εναρμονισμένης τους μορφής.

6.3. Ανάλυση Δεδομένων και Προγραμματιστικά Μοντέλα

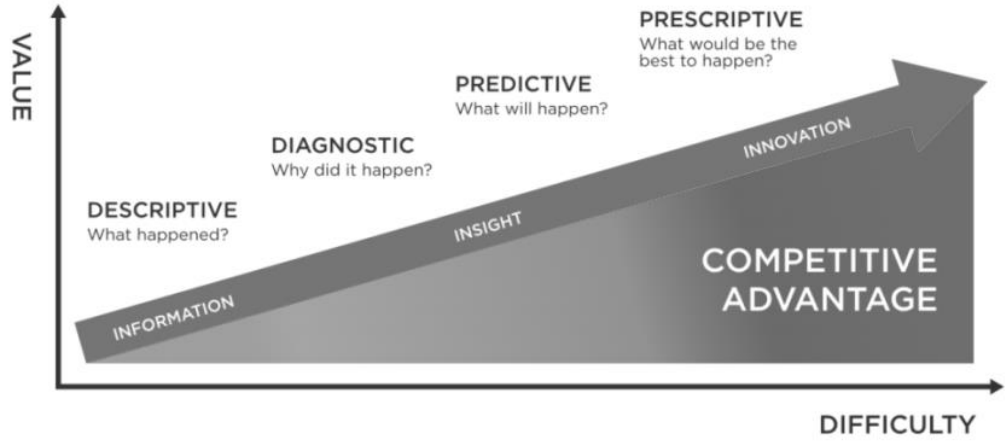
6.3.1. Ανάλυση Δεδομένων

Σύμφωνα με τους συγγραφείς του άρθρου [79], οι μέθοδοι που υπάρχουν για την ανάλυση των δεδομένων (data analytics) μπορούν να οργανωθούν σε τρείς κατηγορίες.

Η πρώτη κατηγορία ονομάζεται περιγραφική ανάλυση (descriptive analytics) και περιλαμβάνει μεθόδους που έχουν ως στόχο τη συνοπτική περιγραφή των δεδομένων. Στην κατηγορία αυτή ανήκουν μέθοδοι από το ευρύτερο πεδίο της στατιστικής, που έχουν ως στόχο τον προσδιορισμό της μέσης τιμής των δεδομένων (ή γενικότερα το μέσο) και τη διασπορά τους (ή γενικότερα απόκλιση/παραμόρφωση) και είναι ιδιαίτερα χρήσιμες, όταν πρόκειται για ομογενοποιημένα δεδομένα. Στην κατηγορία αυτή μπούμε επίσης να εντάξουμε μεθόδους από το πεδίο της μηχανικής μάθησης και της εξόρυξης γνώσης. Για παράδειγμα τεχνικές ομαδοποίησης (clustering) μπορούν να βοηθήσουν τους χρήστες στην οργάνωση των δεδομένων, ενώ τεχνικές εντοπισμού συχνά εμφανιζόμενων μοτίβων και κανόνων (frequent pattern and association rules mining) στην καλύτερη μελέτη του τρόπου σύνδεσης μεταξύ των δεδομένων. Ωστόσο, οι μέθοδοι αυτές ορισμένες φορές περιγράφονται σε μια ξεχωριστή κατηγορία, που ονομάζεται διαγνωστική ανάλυση (diagnostic analytics) και αποσκοπεί να βοηθήσει τον χρήστη να καταλάβει γιατί κάτι συμβαίνει.

Η δεύτερη κατηγορία ονομάζεται προβλεπτική ανάλυση (predictive analysis) και περιλαμβάνει μεθόδους που αποσκοπούν να προβλέψουν τι θα συμβεί στο μέλλον, με βάση τα δεδομένα που έχουν συλλεχθεί στο παρελθόν και με την προϋπόθεση ότι η κατάσταση θα είναι παρόμοια και στο μέλλον. Για τον σκοπό αυτό, μπορούν να χρησιμοποιηθούν επιβλεπόμενοι μέθοδοι μηχανικής μάθησης, όπως είναι για παράδειγμα τα μοντέλα παλινδρόμησης και κατηγοριοποίησης.

Η Τρίτη κατηγορία ονομάζεται συνταγογραφική ανάλυση (prescriptive analytics) και αποσκοπεί στην εύρεση του βέλτιστου πλάνου που θα μπορούσε να ικανοποιήσει τις μελλοντικές μας ανάγκες. Εδώ ανήκουν μέθοδοι που ανήκουν στο πεδίο της κυρτής βελτιστοποίησης (convex optimization) και της ευρετικής αναζήτησης (heuristic search), όπως για παράδειγμα η προσομοιωμένη ανόρθωση (simulated annealing).



Εικόνα 21: Κατηγορίες Ανάλυσης/Επεξεργασίας[31] Δεδομένων

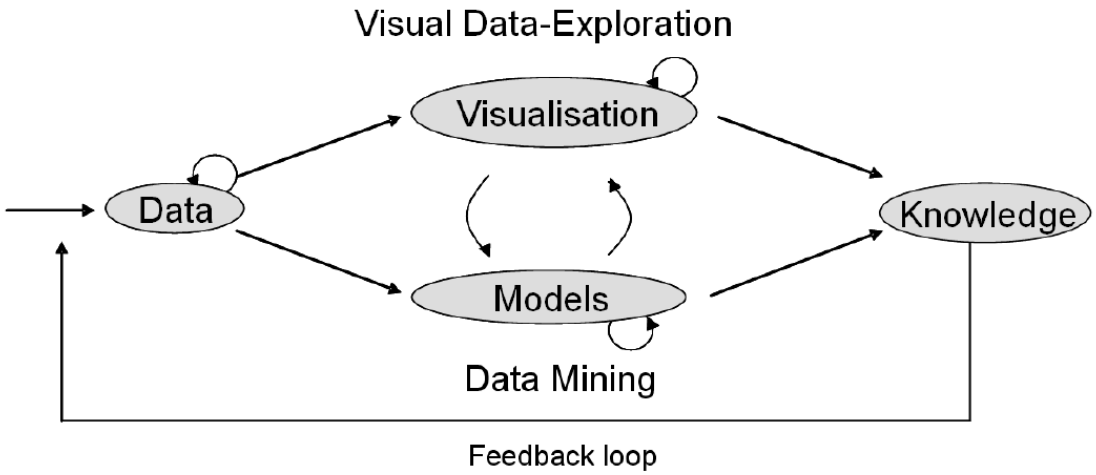
Οι μέθοδοι ανάλυσης δεδομένων φαίνονται και στο παραπάνω σχήμα (Εικόνα 21), όπου μπορούμε να δούμε τη συμβολή που έχουν οι τεχνικές αυτές στην δημιουργία αξίας καθώς και τη δυσκολία που υπάρχει στην εφαρμογή τους. Ειδικότερα οι τεχνικές που εστιάζουν στην κατανόηση των δεδομένων (descriptive & diagnostic analysis) είναι πιο εύκολο να εφαρμοστούν, εν συγκρίσει με τις τεχνικές που αποσκοπούν στην πρόβλεψη του μέλλοντος και τη λήψη των απαραίτητων αποφάσεων (predictive & prescriptive analysis). Ωστόσο, οι τεχνικές αυτές έχουν μεγαλύτερη επίδραση στη δημιουργία ενός ανταγωνιστικού πλεονεκτήματος ενός οργανισμού και κατ' επέκταση είναι αυτές οι οποίες μπορούν να του προσδώσουν αξία, παίρνοντας τις σωστές αποφάσεις, με βάση τη σωστή ανάλυση και ερμηνεία των δεδομένων που συλλέχθηκαν.

6.3.2. Οπτική Ανάλυση Δεδομένων

Οι παραπάνω μέθοδοι ανάλυσης δεδομένων λειτουργούν αυτόματα (χωρίς ανθρώπινη παρέμβαση) και παράγουν χρήσιμα αποτελέσματα, τα οποία μπορούν να μας βοηθήσουν να κατανοήσουμε καλύτερα τα δεδομένα και να εξάγουμε χρήσιμα συμπεράσματα. Παρά το γεγονός ότι αποτελούν ένα σημαντικό και καθοριστικό βήμα για τη διαδικασία της απόκτησης γνώσης, το πρόβλημα της ανάλυσης/κατανόησης των

[31] What is Predictive Analytics?, https://www.tibco.com/reference-center/what-is-predictive-analytics

δεδομένων αυτών και των αποτελεσμάτων των παραπάνω μεθόδων παραμένει. Ο σκοπός της οπτικής ανάλυσης των δεδομένων (Visual Data Analytics) [80] είναι να καλύψει το παραπάνω κενό και να συμβάλει στην απόκτηση γνώσης μέσω της ανάλυσης των δεδομένων και της αλληλεπίδρασης με τον χρήστη (Εικόνα 22). Η οπτική αναπαράσταση των δεδομένων αποτελεί το μέσο για να πετύχουμε τον σκοπό αυτό.



Εικόνα 22: Αλληλεπίδραση μεταξύ των βασικών οντοτήτων της οπτικής ανάλυσης των δεδομένων

Στο παραπάνω μοντέλο (Εικόνα 22), ο χρήστης έχει κομβικό ρόλο στο να κατευθύνει την ανάλυση των δεδομένων για την επίτευξη ενός συγκεκριμένου σκοπού. Ειδικότερα, ο χρήστης εισέρχεται στον κύκλο ανάλυσης των δεδομένων με τα οποία μπορεί να αλληλεπιδρά, προκειμένου να αποκτήσει καλύτερη «εικόνα» των δεδομένων και των αναπαραστάσεων και κατ' επέκταση να κατευθύνει σωστά τη διαδικασία της ανάλυσης και απόκτησης γνώσης.

6.3.3. Προγραμματιστικά Μοντέλα

Δύο σημαντικοί παράγοντες κατά την επεξεργασία και ανάλυση των δεδομένων είναι το μέγεθός τους και η ταχύτητα με την οποία αυτά καταφθάνουν. Όταν το μέγεθος των δεδομένων είναι υπερβολικά μεγάλο και αυξάνεται συνεχώς, η επεξεργασία τους από έναν και μόνο κόμβο είναι πρακτικά αδύνατη. Για αυτόν τον σκοπό, έχουν αναπτυχθεί μοντέλα που επιτρέπουν την επεξεργασία των δεδομένων από διάφορους κόμβους και διευκολύνουν τη μεταξύ τους επικοινωνία, αναλαμβάνοντας να διαχειριστούν προβλήματα που συχνά προκύπτουν κατά τη διαδικασία αυτή (π.χ., ένας κόμβος δεν αποκρίνεται).

Το Map-Reduce μοντέλο [81] επιτρέπει την παράλληλη επεξεργασία μεγάλου όγκου δεδομένων από τους κόμβους του δικτύου, χωρίς να χρειάζεται ο χρήστης να ασχολείται με τον διαμοιρασμό των δεδομένων ή τα

προβλήματα που μπορούν να προκύψουν κατά την εκτέλεση των σχετικών διεργασιών. Για τον σκοπό αυτό, ο χρήστης καλείται να εκφράσει και ακολούθως υλοποιήσει το προς επίλυση πρόβλημα με τη μορφή δύο διαφορετικών συναρτήσεων, τη «map» και τη «reduce». Η συνάρτηση «map» λαμβάνει ως είσοδο ένα ζευγάρι από δεδομένα (key-1, value-1) και παράγει ως έξοδο ένα σύνολο από ενδιάμεσα αποτελέσματα (list/set of key-2, value-2), ενώ η συνάρτηση «reduce» λαμβάνει ως είσοδο ένα κλειδί (key-2) και ένα σύνολο από δεδομένα που σχετίζονται με το κλειδί αυτό, τα οποία ακολούθως επεξεργάζεται, για να βγάλει ένα άλλο αποτέλεσμα. Το ευρύτερο πλαίσιο (framework) το οποίο υποστηρίζει το παραπάνω μοντέλο (π.χ., Apache Hadoop[32]) αναλαμβάνει τόσο τον διαμοιρασμό και την οργάνωση των δεδομένων όσο και την εκτέλεση των παραπάνω διεργασιών.

Το Spark [82] βασίζεται στην ύπαρξη ενός ανθεκτικού κατανεμημένου συνόλου δεδομένων (Resilient Distributed Dataset - RDD) [83] για την παράλληλη επεξεργασία των δεδομένων μας. Το μοντέλο αυτό μας επιτρέπει την παράλληλη εκτέλεση διάφορων μετασχηματισμών (π.χ., map, join, filter), από τους οποίους προκύπτουν νέα τέτοια σύνολα, ενώ το ευρύτερο πλαίσιο (framework) αναλαμβάνει την εκτέλεση των σχετικών διεργασιών και τη διαχείριση των δεδομένων. Στο μοντέλο αυτό, η επεξεργασία των δεδομένων λαμβάνει χώρα μόνο όταν αυτό είναι εντελώς απαραίτητο (lazy evaluation), το οποίο με τη σειρά του επιτρέπει στο σύστημα την εύρεση ενός βέλτιστου πλάνου εκτέλεσης των μετασχηματισμών καθώς επίσης και την ανασύνθεση του τελικού αποτελέσματος στην περίπτωση ύπαρξης κάποιου προβλήματος (fault-tolerance).

6.4. Cloud, Edge and Fog Computing

6.4.1. Υπολογιστικό Νέφος

Σύμφωνα με τον ορισμό του NIST [84], το υπολογιστικό νέφος (Cloud Computing) είναι ένα μοντέλο που μας επιτρέπει να έχουμε πρόσβαση, μέσω του διαδικτύου, σε μια λίμνη από παραμετροποίησιμους υπολογιστικούς πόρους (π.χ., αποθηκευτικό χώρο), όταν τους χρειαζόμαστε, τους οποίους μπορούμε να τους δεσμεύσουμε και αποδεσμεύσουμε εύκολα. Ειδικότερα, το μοντέλο αυτό επιτρέπει στους χρήστες να έχουν πρόσβαση σε διάφορες υπηρεσίες (Software as a Service - SaaS), πλατφόρμες (Platform as a Service - PaaS) και υπολογιστικούς πόρους (Infrastructure as a Service - IaaS) ανάλογα με τις ανάγκες τους.

[32] Apache Hadoop, https://hadoop.apache.org/

Για να γίνει καλύτερα κατανοητή η έννοια του υπολογιστικού νέφους, θα πρέπει να λάβουμε υπόψη ορισμένα από τα βασικά χαρακτηριστικά του. Καταρχάς, το υπολογιστικό νέφος επιτρέπει στους χρήστες να έχουν πρόσβαση στους πόρους/υπηρεσίες που χρειάζονται, χωρίς την ανάγκη ανθρώπινης παρέμβασης (on-demand self-services), μέσω του διαδικτύου και των σχετικών μηχανισμών / πρωτοκόλλων επικοινωνίας (broad network access). Για να το πετύχουν αυτό, οι πάροχοι επιτρέπουν στο υπολογιστικό νέφος να μπορεί να υποστηρίξει ταυτόχρονα πολλούς χρήστες (multi-tenant model), δεσμεύοντας και αποδεσμεύοντας τους απαραίτητους πόρους εύκολα και γρήγορα, ανάλογα με τις ανάγκες των χρηστών (rapid elasticity), δίνοντας τους έτσι την εικόνα ύπαρξης αφθονίας υπολογιστικών πόρων. Τέλος, ένα τέτοιο σύστημα παρακολουθεί την κατάσταση των υπολογιστικών πόρων και υπηρεσιών (measured service), με σκοπό την καλύτερη διαχείρισή τους.

Το υπολογιστικό νέφος γνωρίζει ιδιαίτερη άνθηση στον χώρο των επιχειρήσεων, καθώς δίνει τη δυνατότητα στους οργανισμούς αυτούς να έχουν πρόσβαση σε έναν απεριόριστο αριθμό υπολογιστικών πόρων, όταν και όποτε τους χρειάζονται με πολύ μικρότερο κόστος έναντι αυτού που θα απαιτούνταν για την αγορά και διατήρηση του σχετικού εξοπλισμού (συμπεριλαμβανομένου του υλικού και λογισμικού) στον χώρο τους (on premise hosting). Σχετικά με την επεξεργασία των δεδομένων, το παραπάνω μοντέλο προϋποθέτει την μεταφορά των δεδομένων στο υπολογιστικό νέφος και ακολούθως την επεξεργασία τους. Συνεπώς, θα πρέπει να λάβουμε υπόψη τον χρόνο που απαιτείται για τη μεταφορά των δεδομένων από τον χώρο δημιουργίας τους στο σύννεφο καθώς και τους σχετικούς κινδύνους που υπάρχουν σχετικά με την ασφάλεια των δεδομένων, ειδικά όταν πρόκειται για ευαίσθητα ή προσωπικά δεδομένα χρηστών.

6.4.2. Υπολογισμός στα Άκρα

Το υπολογιστικό αυτό μοντέλο (Edge Computing) είναι σχετικά νέο και κατ' επέκταση δεν υπάρχει ένας κοινώς αποδεκτός ορισμός [85]. Το μοντέλο αυτό δημιουργήθηκε, για να καλύψει τις αδυναμίες που υπάρχουν στο υπολογιστικό νέφος και εστιάζει στην επεξεργασία των δεδομένων κοντά στα σημεία / συσκευές όπου αυτά παράγονται, τα οποία βρίσκονται, εν γένει, στα άκρα του δικτύου, με απώτερο στόχο να καλύψουν τις ανάγκες που έχουν οι εφαρμογές που απαιτούν επεξεργασία των δεδομένων σε πραγματικό χρόνο (real-time applications), λαμβάνοντας υπόψη θέματα ιδιωτικότητας και ασφάλειας (security and privacy), που σχετίζονται με τα δεδομένα αυτά και φυσικά την ενέργεια που απαιτείται για την ανάλυση των δεδομένων.

Σύμφωνα με το μοντέλο αυτό υπάρχουν τρία επίπεδα. Στο πρώτο επίπεδο (terminal layer) ανήκουν όλων των ειδών οι συσκευές που είναι συνδεδεμένες στο δίκτυο και μπορούν να λειτουργούν τόσο ως πάροχοι

όσο και ως καταναλωτές δεδομένων. Στο δεύτερο επίπεδο (boundary layer) ανήκουν οι κόμβοι των άκρων (edge nodes), οι οποίοι υποδέχονται τα δεδομένα από το προηγούμενο επίπεδο για την περαιτέρω επεξεργασία τους, ενώ τα αποτελέσματά τους μπορούν να σταλούν στο επόμενο και τελευταίο επίπεδο. Στο τρίτο επίπεδο βρίσκεται το υπολογιστικό νέφος (cloud layer), το οποίο είναι και το πιο υπολογιστικά ισχυρό επίπεδο και έχει τη δυνατότητα αποθήκευσης και ανάλυσης των δεδομένων που καταφθάνουν σε αυτό.

Το μοντέλο αυτό εστιάζει στα άκρα και στην ικανοποίηση συγκεκριμένων αναγκών των εφαρμογών, παρά στην παροχή υπολογιστικών πόρων, όπως συμβαίνει στην περίπτωση του υπολογιστικού νέφους.

6.4.3. Υπολογισμός Ομίχλης

Το υπολογιστικό αυτό μοντέλο (Fog Computing) δίνει τη δυνατότητα στους χρήστες να έχουν πρόσβαση στους απαραίτητους υπολογιστικούς πόρους (π.χ., δίκτυο, αποθηκευτικό χώρο, μονάδες επεξεργασίας) που χρειάζονται, για να καλύψουν τις ανάγκες των σύγχρονων εφαρμογών (π.χ., εφαρμογές που απαιτούν επεξεργασία δεδομένων σε πραγματικό χρόνο), λαμβάνοντας υπόψη τους περιορισμούς και τις αδυναμίες που υπάρχουν στο υπολογιστικό νέφος [86]. Στο μοντέλο αυτό οι κόμβοι ομίχλης (fog nodes) βρίσκονται μεταξύ του υπολογιστικού νέφους (cloud computing) και των συσκευών των άκρων (edge device) και αναλαμβάνουν να παρέχουν τους απαραίτητους πόρους (IaaS) / πλατφόρμα (PaaS) / υπηρεσίες (SaaS) στις συσκευές αυτές.

Τα κύρια χαρακτηριστικά του υπολογιστικού αυτού μοντέλου είναι η αποκεντροποιημένη αρχιτεκτονική, η μικρή καθυστέρηση στην επικοινωνία, η δυνατότητα επεξεργασίας δεδομένων σε πραγματικό χρόνο, η γνώση της θέσης/τοποθεσίας, η προσαρμοστικότητα και η συμβατότητα των κόμβων ομίχλης, η αποτελεσματική διαχείριση της ετερογένειας των πόρων και η παροχή των σχετικών υπηρεσιών στα άκρα. Για την παροχή των παραπάνω δυνατοτήτων, το ευρύτερο πλαίσιο (framework), που υπάρχει για τον σκοπό αυτό, αποτελείται από διάφορα επίπεδα που αναλαμβάνουν την εικονική παροχή (virtualization) των σχετικών πόρων, την περιγραφή των επιμέρους οντοτήτων του μοντέλου, την παρακολούθηση/διαχείριση αυτών καθώς και την πριν/μετά-επεξεργασία των σχετικών δεδομένων και την ασφάλεια των εφαρμογών.

Η σελίδα αυτή είναι σκόπιμα λευκή

7. Βιβλιογραφικές Αναφορές

[1] Fortier, Isabel, et al. "Maelstrom Research guidelines for rigorous retrospective data harmonization." International journal of epidemiology 46.1 (2017): 103-105.

[2] Leal, Gabriel da Silva Serapião, Wided Guédria, and Hervé Panetto. "Interoperability assessment: A systematic literature review." Computers in Industry 106 (2019): 111-132.

[3] Chondrogiannis, Efthymios, et al. "Semantically-enabled context-aware abbreviations expansion in the clinical domain." Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology. 2017.

[4] Chondrogiannis, Efthymios, et al. "Using blockchain and semantic web technologies for the implementation of smart contracts between individuals and health insurance organizations." Blockchain: Research and Applications 3.2 (2022): 100049.

[5] Voigt, Paul, and Axel Von dem Bussche. "The eu general data protection regulation (gdpr)." A Practical Guide, 1st Ed., Cham: Springer International Publishing 10.3152676 (2017): 10-5555.

[6] Fortier, Isabel, et al. "Maelstrom Research guidelines for rigorous retrospective data harmonization." International journal of epidemiology 46.1 (2017): 103-105.

[7] Spjuth, Ola, et al. "Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research." European Journal of Human Genetics 24.4 (2016): 521-528.

[8] Firnkorn, D., et al. "A generic data harmonization process for cross-linked research and network interaction." Methods of information in medicine 54.05 (2015): 455-460.

[9] Dong, Xin Luna, and Theodoros Rekatsinas. "Data integration and machine learning: A natural synergy." Proceedings of the 2018 international conference on management of data. 2018.

[10] Kumar, Ganesh, et al. "Data harmonization for heterogeneous datasets: A systematic literature review." Applied Sciences 11.17 (2021): 8275.

[11] Keele, Staffs. "Guidelines for performing systematic literature reviews in software engineering." (2007).

[12] Borthakur, Dhruba. "The hadoop distributed file system: Architecture and design." Hadoop Project Website 11.2007 (2007): 21.

[13] White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.

[14] Zaharia, Matei, et al. "Apache spark: a unified engine for big data processing." Communications of the ACM 59.11 (2016): 56-65.

[15] Ben-Gal, I.: Bayesian networks. Encyclopedia of statistics in quality and reliability 1 (2008).

[16] Ali, J., et al.: Random forests and decision trees. International Journal of Computer Science Issues 9(5),

[17] Noble, W.S.: What is a support vector machine?. Nature biotechnology 24(12), 1565-1567 (2006).

[18] Frazier, Peter I. "A tutorial on Bayesian optimization." arXiv preprint arXiv:1807.02811 (2018).

[19] Thornton, Chris, et al. "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013.

[20] Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern recognition letters 31(8), 651-666 (2010).

[21] Schubert, E., et al.: DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems 42(3), 1-21 (2017).

[22] Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. ACM sig-mod record 29(2), 1-12 (2000).

[23] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Proceedings of the 20th VLDB conf., pp. 487-499 (1994).

[24] Kurita, T. (2020). Principal Component Analysis (PCA). In: Computer Vision. Springer, Cham.

[25] Martinez, Aleix M., and Avinash C. Kak. "Pca versus lda." IEEE transactions on pattern analysis and machine intelligence 23.2 (2001): 228-233.

[26] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

[27] Van Engelen, Jesper E., and Holger H. Hoos. "A survey on semi-supervised learning." Machine learning 109.2 (2020): 373-440.

[28] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[29] Watkins, Christopher JCH, and Peter Dayan. "Q-learning." Machine learning 8 (1992): 279-292.

[30] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.

[31] Tan, Chuanqi, et al. "A survey on deep transfer learning." Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer International Publishing, 2018.

[32] Gu, J., et. al.: Recent advances in convolutional neural networks. Pattern Recognition 77, 354-377 (2018).

[33] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[34] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[35] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747 (2016).

[36] Qian, Ning. "On the momentum term in gradient descent learning algorithms." Neural networks 12.1 (1999): 145-151.

[37] Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." Journal of machine learning research 12.7 (2011).

[38] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[39] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. pmlr, 2015.

[40] Bengio, Yoshua, et al. "Curriculum learning." Proceedings of the 26th annual international conference on machine learning. 2009.

[41] Cho, Kyunghyun, et al. "On the properties of neural machine translation: Encoder-decoder approaches." arXiv preprint arXiv:1409.1259 (2014).

[42] Baldi, P.: Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning, pp. 37-49 (2012).

[43] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[44] Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization." arXiv preprint arXiv:1607.06450 (2016).

[45] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[46] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." International conference on machine learning. PMLR, 2017.

[47] Donahue, Jeff, Philipp Krähenbühl, and Trevor Darrell. "Adversarial feature learning." arXiv preprint arXiv:1605.09782 (2016).

[48] Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5):544–551, 2011.

[49] Guarino, N., Oberle, D., & Staab, S. What Is an Ontology? S. Staab & R. Studer (Eds.), Handbook on Ontologies, 1–17, 2009.

[50] Princeton University "About WordNet." WordNet. Princeton University. 2010.

[51] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. The Journal of Machine Learning Research, 3:1137–1155, 2003.

[52] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013: 3111-3119, 2013.

[53] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781, 2013.

[54] Xin Rong. word2vec parameter learning explained. arXiv preprint arXiv:1411.2738, 2014.

[55] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics, 87:12–20, 2018.

[56] Google, Word2Vec (2022), Available online at: https://code.google.com/archive/p/word2vec/

[57] Piotr Bojanowski et al. Enriching Word Vectors with Subword Information. CoRR, abs/1607.04606, 2016.

[58] Yijia Zhang et al. BioWordVec, Improving biomedical word embeddings with subword information and MeSH. Scientific data 6.1, 1–9, 2019.

[59] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543, 2014, Doha, Qatar.

[60] Sarzynska-Wawer, Justyna, et al. "Detecting formal thought disorder by deep contextualized word representations." Psychiatry Research 304 (2021): 114135.

[61] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[62] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

[63] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

[64] Chondrogiannis, Efthymios, et al. "A novel semantic representation for eligibility criteria in clinical trials." Journal of biomedical informatics 69 (2017): 10-23.

[65] Chondrogiannis, E., Andronikou, V., Karanastasis, E. and Varvarigou, T., 2019, February. A novel approach for clinical data harmonization. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-8). IEEE.

[66] Chondrogiannis, E., Andronikou, V., Karanastasis, E. and Varvarigou, T.A., 2014, December. An Intelligent Ontology Alignment Tool Dealing with Complicated Mismatches. In SWAT4LS.

[67] Chondrogiannis, Efthymios, et al. "Bridging the Gap among Cohort Data Using Mapping Scenarios." Journal of Advances in Information Technology Vol 12.3 (2021).

[68] Grau, Bernardo Cuenca, et al. "OWL 2: The next step for OWL." Journal of Web Semantics 6.4 (2008): 309-322.

[69] Spyns, Peter, Robert Meersman, and Mustafa Jarrar. "Data modelling versus ontology engineering." ACM SIGMod Record 31.4 (2002): 12-17.

[70] Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

[71] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[72] Fernández, Alberto, et al. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary." Journal of artificial intelligence research 61 (2018): 863-905.

[73] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." International conference on intelligent computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[74] He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). Ieee, 2008.

[75] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: A survey." Mobile networks and applications 19 (2014): 171-209.

[76] Taleb, Ikbal, Mohamed Adel Serhani, and Rachida Dssouli. "Big data quality: A survey." 2018 IEEE International Congress on Big Data (BigData Congress). IEEE, 2018.

[77] Sidi, Fatimah, et al. "Data quality: A survey of data quality dimensions." 2012 International Conference on Information Retrieval & Knowledge Management. IEEE, 2012.

[78] Schmidt, Carsten Oliver, et al. "Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R." BMC medical research methodology 21.1 (2021): 1-15.

[79] Duan, Lian, and Li Da Xu. "Data analytics in industry 4.0: A survey." Information Systems Frontiers (2021): 1-17.

[80] Keim, Daniel, et al. Visual analytics: Definition, process, and challenges. Springer Berlin Heidelberg, 2008.

[81] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.

[82] Zaharia, Matei, et al. "Spark: Cluster computing with working sets." 2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10). 2010.

[83] Zaharia, Matei, et al. "Resilient distributed datasets: A {Fault-Tolerant} abstraction for {In-Memory} cluster computing." 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12). 2012.

[84] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011).

[85] Cao, Keyan, et al. "An overview on edge computing research." IEEE access 8 (2020): 85714-85728.

[86] Sabireen, H., and V. J. I. E. Neelanarayanan. "A review on fog computing: Architecture, fog with IoT, algorithms and research challenges." Ict Express 7.2 (2021): 162-176.

Η σελίδα αυτή είναι σκόπιμα λευκή

8. Συντομεύσεις

Στον Πίνακας 9 παραθέτουμε τη πλήρη έκφραση (long forms) των συντομογραφιών (short forms aka acronyms or abbreviations) που χρησιμοποιήθηκαν στη εργασία αυτή:

Συντόμευση	**Πλήρης Έκφραση**
ADAM	Adaptive Moment Estimation
ADASYN	Adaptive Synthetic Sampling Approach
ANN	Artificial Neural Network
ATC	Anatomical Therapeutic Chemical
BERT	Bidirectional Encoder Representation from Transformer
BiGAN	Bidirectional Generative Adversarial Network
CASH	Combined Algorithm Selection and Hyperparameters Optimization
CBOW	Continuous Bag of Words
CDISC	Clinical Data Interchange Standards Consortium
CNN	Convolutional Neural Network
CSV	Comma-separated values
DBN	Deep Belief Network
DOM	Document Object Model
EIF	European Interoperability Framework
ELMo	Embeddings from Language Models
ETL	Extract-Transform-Load
FCN	Fully Convolutional Network
FFNN	Feed Forward Neural Network
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GDPR	General Data Privacy Regulation

Συντόμευση	**Πλήρης Έκφραση**
GRU	Gated Recurrent Units
HDFS	Hadoop Distributed File System
HL7	Health Level Seven
HTML	HyperText Markup Language
i.i.d.	independent identically distributed
IaaS	Infrastructure as a Service
ICD	International Classification of Diseases
INAS	Interoperability Assessment
JSON	JavaScript Object Notation
kNN	K nearest neighbor
LDA	Linear Discrimination Analysis
LSTM	Long-Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
NER	Named Entity Recognition
NN	Neural Network
NSAID	Non-steroidal anti-inflammatory drug
OWL	Web Ontology Language
PaaS	Platform as a Service
PCA	Principal Component Analysis
PR	Precision-Recall
pSS	primary Sjögren Syndrome
RDD	Resilient Distributed Dataset
RDF	Resource Description Framework
ReLU	Rectified Linear Unit
RMM	Recurrent Neural Network
RMSE	Root Mean Squared Error

Συντόμευση	**Πλήρης Έκφραση**
ROC	Receiver operating characteristic
SaaS	Software as a Service
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Oversampling Technique
SPARQL	SPARQL Query Language for RDF
SQL	Structured query language
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
t-SNE	t-distributed Stochastic Neighbor Embedding
WBC	White Blood Cell
WGAN	Wasserstein Generative Adversarial Network
XML	Extensible Markup Language
YAML	Yet Another Markup Language

Πίνακας 9: Πίνακας Συντομεύσεων