ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Ομαδοποίηση Ιατρικών Προφίλ από Δημιουργημένη Βάση Δεδομένων με Τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ντουντούμι Α. Κλέιντα

Επιβλέπων : Δημήτριος – Διονύσιος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2020

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΥΛΙΚΩΝ

Ομαδοποίηση Ιατρικών Προφίλ από Δημιουργημένη Βάση Δεδομένων με Τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ντουντούμι Α. Κλέιντα

Επιβλέπων : Δημήτριος – Διονύσιος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 03/11/2020

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
Δ. Κουτσούρης	Π. Τσανάκας	Γ. Ματσόπουλος
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2020

(Υπογραφή)

.....

ΝΤΟΥΝΤΟΥΜΙ ΚΛΕΪΝΤΑ

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

Περίληψη

Η χρήση των τεχνικών μηχανικής μάθησης και εξόρυξης γνώσης διαδραματίζουν ολοένα και πιο καταλυτικό ρόλο στη βελτιστοποίηση των δυνατοτήτων της εξατομικευμένης ιατρικής φροντίδας, ιδιαίτερα σε μια εποχή που η διαθεσιμότητα δεδομένων υγείας από self-monitoring εφαρμογές συνεχώς αυξάνεται. Αντικείμενο της παρούσας διπλωματικής εργασίας ήταν η διερεύνηση της δυνατότητας ομαδοποίησης δημιουργημένων ιατρικών προφίλ με παρεμφερή χαρακτηριστικά με τεχνικές μηχανικές μάθησης.

Τα δεδομένα των ιατρικών προφίλ δημιουργήθηκαν με τυχαίο τρόπο και αποθηκεύτηκαν σε μια σχεσιακή βάση δεδομένων. Τα μεγέθη αυτά περιλαμβάνουν γενικές πληροφορίες για κάθε χρήστη (ηλικία, φύλο, μορφωτικό επίπεδο) καθώς και ιατρικά μεγέθη κωδικοποιημένα κατά τα πρότυπα ICD-10 και SNOMED CT. Συγκεκριμένα, συναντώνται πληροφορίες ιατρικού ιστορικού (ιστορικό διαβήτη, κατάθλιψης, απώλειας ακοής, εγκεφαλοαγγειακής νόσου, γνωστικής δυσλειτουργίας κ.α.), αλλά και πληθώρα μετρήσεων όπως το βάρος, το ύψος, η μυϊκή μάζα, η χρήση αλκοόλ, ο εθισμός στη νικοτίνη, το επίπεδο φυσικής άσκησης και ο κορεσμός του οξυγόνου.

Στα δεδομένα αυτά πραγματοποιήθηκε προ επεξεργασία και επιλογή χαρακτηριστικών ενώ εφαρμόστηκε ο αλγόριθμος των κ-Μέσων με χρήση της μεθόδου του αγκώνα για επιλογή του αριθμού των συστάδων, ο αλγόριθμος Ιεραρχικής συγκεντρωτικής ομαδοποίησης, ο αλγόριθμος t-SNE και ο αλγόριθμος DBSCAN. Τα αποτελέσματα αξιολογήθηκαν με χρήση των συντελεστών Davies-Bouldin, Calinski-Harabasz και Silhouette. Επιπλέον, χρησιμοποιήθηκε 10-fold cross validation για αξιολόγηση της δυνατότητας ταξινόμησης μεγαλύτερου αριθμού ιατρικών προφίλ με βάση την υπάρχουσα ομαδοποίηση. Από την παραπάνω διαδικασία, η οποία πραγματοποιήθηκε σε Python με χρήση της βιβλιοθήκης scikit-learn, προέκυψαν ενδείξεις επιτυχούς ομαδοποίησης για δύο από τους τέσσερις αλγορίθμους που χρησιμοποιήθηκαν παρά τον υψηλό αριθμό διαστάσεων στο σύνολο δεδομένων εισόδου.

Λέξεις Κλειδιά: Μηχανική Μάθηση, Προσωποποιημένη ιατρική, Ομαδοποίηση, Ιατρικά προφίλ, Python, Scikit-learn, Αλγόριθμος των κ-Μέσων, Μέθοδος του αγκώνα, Αλγόριθμος Ιεραρχικής συγκεντρωτικής ομαδοποίησης, Αλγόριθμος t-SNE , Αλγόριθμος DBSCAN, Συντελεστής Davies-Bouldin, Συντελεστής Calinski-Harabasz, Συντελεστής Silhouette,

10-fold cross validation, ICD-10, SNOMED CT.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

The use of machine learning and knowledge mining techniques, in conjunction with the increased availability of health data generated by self-monitoring apps, are considered important catalysts in the advancement of personalized medicine capabilities.

The scope of this thesis was the clustering of generated medical profiles with similar characteristics by utilizing various clustering algorithms.

Medical profiles were generated randomly and stored in a relational database. The data consisted of demographic information for each user (age, gender, educational level) as well as medical measurements encoded using the ICD-10 and SNOMED CT protocols. Information on medical history (history of diabetes, depression, hearing loss, cerebrovascular disease, cognitive impairment, etc.) but also a variety of measurements such as weight, height, muscle mass, alcohol use, nicotine addiction, level of exercise and oxygen saturation were included in our study.

Initially, preprocessing and feature selection were performed and then k-Means was applied using the elbow method to select the appropriate number of clusters. Hierarchical Agglomerative Clustering, t-SNE and DBSCAN were also used. Results were evaluated using the Davies-Bouldin Index, the Calinski-Harabasz Index and the Silhouette Score. In addition, 10-fold cross validation was performed to evaluate the performance of a classifier based on the existing clustering. The aforementioned task was performed in Python using the scikit-learn library and the results showed evidence of successful clustering for two of the four algorithms used, despite the high dimensionality of our dataset.

Keywords: Machine learning, Clustering, Personalized medicine, , Medical profiles, Generated database, Python, Scikit-learn, k-Means, Elbow method, Hierachical Agglomerative Clustering, t-SNE, DBSCAN, Calinski-Harabasz Index, Davies-Bouldin Index, Silhouette Score, 10-fold cross validation, ICD-10, SNOMED CT.

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα καταρχήν να ευχαριστήσω θερμά τον καθηγητή κ. Δημήτριο Κουτσούρη για τη συνεργασία και την εμπιστοσύνη του στην ανάθεση του θέματος της παρούσας διπλωματικής εργασίας. Επιπλέον, ευχαριστώ ιδιαίτερα τον Δρ. Ιωάννη Κουρή για την εξαιρετική καθοδήγηση, την πολύτιμη βοήθεια και υπομονή κατά τη διάρκεια της εκπόνησης της.

Σε προσωπικό επίπεδο, χρωστάω απίστευτη αγάπη και ευγνωμοσύνη στην οικογένεια μου για τη υποστήριξη και τις θυσίες τους καθ' όλη τη διάρκεια της ζωής μου και ιδιαίτερα κατά την περίοδο των σπουδών μου. Θέλω επιπλέον να ευχαριστήσω θερμά τους φίλους μου και τα αγαπημένα μου πρόσωπα για την αμέριστη στήριξη και το προνόμιο να βρίσκονται στη ζωή μου.

Το αποτέλεσμα της προσπάθειας αυτής, αφιερώνεται στη μνήμη της μητέρας μου Μαριάνθης Κύρκο με απεριόριστη αγάπη, θαυμασμό και ευγνωμοσύνη για όσα μου προσέφερε.

Πίνακας περιεχομένων

1 Εισαγωγή

1.1 Αντικείμενο και συνεισφορά

Η επέλαση των νέων τεχνολογιών στο χώρο της υγείας και η αναγνώριση του μοντέλου της εξατομικευμένης ή «προσωποποιημένης» ιατρικής τα τελευταία έτη, επιτείνουν το ενδιαφέρον για την αξιοποίηση των διαθέσιμων μέσων και δεδομένων και τη δημιουργία ιατρικών συστημάτων προτάσεων(recommendation systems) που θα επιτρέπουν:

- πιο στοχευμένες τακτικές ιατροφαρμακευτικής περίθαλψης,
- ενίσχυση της προληπτικής προσέγγισης σε άτομα με προδιάθεση ορισμένων νόσων,
- έγκαιρες παρεμβάσεις σε συνήθειες ή στοιχεία του τρόπου ζωής που λειτουργούν ως παράγοντες κινδύνου παθήσεων και
- πιο αποτελεσματικές μεθόδους θεραπείας με καλύτερη ανταπόκριση στη μεθοδολογία και μείωση των παρενεργειών της φαρμακευτικής αγωγής.

Τα οφέλη είναι ποικίλα και πολυδιάστατα με καταλυτική συνεισφορά στο βιοτικό επίπεδο των ασθενών, αλλά και στο οικονομικό κόστος για τα συστήματα υγείας.

Η παρούσα διπλωματική εργασία πραγματεύεται το πρόβλημα της ομαδοποίησης ιατρικών προφίλ, με τη βοήθεια της μηχανικής μάθησης, και συγκεκριμένα με τη χρήση αλγορίθμων συσταδοποίησης (clustering algorithms), διαδικασία η οποία αποτελεί βάση για την υλοποίηση ενός αξιόπιστου συστήματος προτάσεων σε ιατρικό πλαίσιο. Στα πλαίσια της διερεύνησης της δυνατότητας αξιόπιστης ομαδοποίησης, εμπεριέχονται και τα εξής στοιχεία:

- Η δομή του ιατρικού προφίλ και η κατάλληλη προτυποποίηση του ώστε να επιτευχθεί μια χρηστική σε πρώτο επίπεδο ομαδοποίηση.
- Η σύγκριση ορισμένων διαδεδομένων αλγορίθμων ομαδοποίησης και των αποτελεσμάτων που παρέχουν ως προς αυτήν την κατεύθυνση.

- Η προσέγγιση της εύρεσης διαθέσιμων ιατρικών προφίλ για την εκτέλεση της μελέτης και των πειραμάτων μας, μέσω της δημιουργίας κατάλληλης βάσης δεδομένων με ικανό αριθμό χρηστών και λοιπών καταγεγραμμένων μεγεθών.
- Η σωστή χρήση των τεχνικών και εργαλείων μηχανικής μάθησης, της κατάλληλης μεθόδου επιλογής χαρακτηριστικών και των αξιόπιστων μετρικών αξιολόγησης του αποτελέσματος.
- Η πρόταση μελλοντικών επεκτάσεων ανάλογα με το επίπεδο εξειδίκευσης των ομάδων και τυχόν υπό-ομάδων ιατρικών προφίλ που στοχεύει το σύστημα.

1.2 Σχετικές μελέτες

Ορισμένες γενικές αρχές που προκύπτουν από διαδικασίες παρόμοιων μελετών για ομαδοποίηση και εξαγωγή κάποιων εσωτερικών δομών σε ιατρικά δεδομένα και ακολουθούνται και στην παρούσα διπλωματική εργασία είναι οι εξής:

Αλγόριθμος ομαδοποίησης και μέτρα ομοιότητας: Οι μέθοδοι ομαδοποίησης αποτελούνται σε μεγάλο βαθμό από ιεραρχικές διαδικασίες (Ward, 1963) και επαναλαμβανόμενες τεχνικές διαμερίσεων (MacQueen, 1965). Ωστόσο, δεν υφίστανται περιορισμοί στο ποιοι αλγόριθμοι μπορούν να χρησιμοποιηθούν. Μάλιστα, προκειμένου να αντιμετωπιστούν οι περιορισμοί των αλγορίθμων ιεραρχικής και μη ιεραρχικής ομαδοποίησης, υπάρχει η πρόταση να χρησιμοποιούνται ταυτόχρονα και να συγκρίνονται τα αποτελέσματα καθώς και να χρησιμοποιούνται διαφορετικά μέτρα για την εκτίμηση της ομοιότητας μεταξύ των οντοτήτων που συγκεντρώνονται. Καθώς διαφορετικά μέτρα απόστασης μπορεί να παράγουν διαφορετικές λύσεις συστάδων, συνιστάται η χρήση διαφόρων μετρήσεων απόστασης και η σύγκριση των υπολογισμένων ομάδων με θεωρητικά ή γνωστά μοτίβα (Hair et al., 2006).

Αριθμός συστάδων: Διαφορετικά μέτρα, όπως ο συντελεστής συσσωμάτωσης ή το κριτήριο κυβικής ομαδοποίησης, για τον προσδιορισμό του αριθμού των συστάδων. Συνιστάται να εφαρμόζετε πρακτική κρίση, κοινή λογική ή θεωρητικά θεμέλια κατά τον ορισμό της τελικής λύσης συμπλέγματος.

Επικύρωση των αποτελεσμάτων: Η αξιοπιστία της λύσης ομαδοποίησης θα πρέπει να επαληθεύεται αξιολογώντας τη σταθερότητα των συστάδων χρησιμοποιώντας πολλαπλούς αλγόριθμους (Ketchen and Shook, 1996) ή χωρίζοντας το δείγμα (Punj and Stewart, 1983). Η εξωτερική εγκυρότητα ελέγχεται με ομαδοποίηση σε ένα δείγμα αναμονής χρησιμοποιώντας τις ίδιες μεταβλητές και αξιολογώντας την ομοιότητα των δύο λύσεων. Ωστόσο, αυτό ισχύει

μόνο σε γενικευμένες ή μη ειδικές ρυθμίσεις. Η ερμηνεία και αξιολόγηση των συστάδων που προκύπτουν απαιτεί συνδυασμό θεωρητικής και κριτικής προσέγγισης με βάση τις συγκεκριμένες παραμέτρους που εξετάζονται και το προσδοκώμενο αποτέλεσμα.[2]

Στη συνέχεια, παρουσιάζονται ορισμένες μελέτες που έχουν πραγματοποιηθεί σε συναφή αντικείμενα με το τρέχον. Καθώς το αντικείμενο είναι σχετικά ευρύ και πολύπλοκο, συναντάται ικανός αριθμός από μελέτες όπως ορισμένες οι οποίες πραγματοποιούνται για ομαδοποίηση εξειδικευμένη σε κάποια συγκεκριμένη πάθηση και τα χαρακτηριστικά της [3][4] ή χρήση τεχνικών που επικεντρώνονται στο θέμα από την σκοπιά της στατιστικής ανάλυσης [5]. Εστιάζουμε, επομένως, από εδώ και στο εξής, σε ορισμένες μελέτες που ακολουθούν παρόμοια μεθοδολογία με την παρούσα διπλωματική εργασία και υπάγονται στο πεδίο της μηχανικής μάθησης.

Εστιάζοντας σε παρόμοιες μελέτες, συναντάται η μελέτη των Newcomer et al [6], η οποία πραγματοποιεί ανάλυση συστάδων με χρήση του αλγορίθμου ελάχιστης διακύμανσης του Ward. Λαμβάνονται υπόψη περίπλοκα προφίλ ασθενών που παρουσιάζουν 2 ή περισσότερες από τις 17 συνολικά παθήσεις που θεωρούνται οι πιο κοστοβόρες για τα συστήματα υγείας. Αναγνωρίζονται 10 υπό-ομάδες στα δεδομένα, ενώ η αξιολόγηση των αποτελεσμάτων γίνεται με τη μέθοδο flexible beta.

Ενδιαφέρον παρουσιάζει επιπλέον, η μελέτη που πραγματοποιήθηκε από τους Vuik et al. [7] στην οποία χρησιμοποιείται τυχαίο δείγμα 300.000 ασθενών και λαμβάνονται χαρακτηριστικά κοινωνικά , δημογραφικά, απαιτήσεων φροντίδας και θνησιμότητας. Γίνεται χρήση του αλγορίθμου k-Means και αναγνωρίζονται 8 υπό-ομάδες ασθενών, ενώ η αξιολόγηση πραγματοποιείται με χρήση ιεραρχικών αλγορίθμων σε υπό-ομάδες των 3000 ατόμων και σύγκριση των αποτελεσμάτων.

Ιδιαίτερο ενδιαφέρον παρουσιάζει και η μελέτη των Busch et al. [8] καθώς μελετάει, μέσω ερωτηματολογίου, συμπεριφορές που σχετίζονται με την υγεία (πχ. Κάπνισμα, διατροφικές συνήθειες, επίπεδο φυσικής άσκησης, κατανάλωση αλκοόλ, ψυχολογικά προβλήματα, χρήση υπολογιστή και video games κ.α.) σε 2690 μαθητές λυκείου ως παράμετρο ομαδοποίησης. Εφαρμόζεται ο αλγόριθμος TCA (Two Step Cluster Analysis) ενώ η αξιολόγηση πραγματοποιείται με το δείκτη KMO (Kaiser-Meyer-Olken Measure of Adequacy) και τη δοκιμή σφαιρικότητας του Bartlett.

1.3 Διάρθρωση της διπλωματικής εργασίας

Περιγράφεται στη συνέχεια η διάρθρωση των ενοτήτων της τρέχουσας διπλωματικής εργασίας ώστε να δοθεί μια συμπτυγμένη εικόνα της δομής της.

Ξεκινάμε τη προσέγγιση του θέματος που μελετάται με τη θεωρητική περιγραφή του αντικειμένου της Μηχανικής Μάθησης, στο δεύτερο κεφάλαιο, που έπεται της εισαγωγής. Εξερευνώνται οι κύριες κατηγοριοποιήσεις που συναντώνται στο ευρύ αυτό πεδίο, οι οποίες μας φανερώνουν και τα ποικίλα και πολύπλοκα είδη προβλημάτων τα οποία μπορεί να επιλύσει. Τέλος, περιγράφονται οι σημαντικότεροι αλγόριθμοι ταξινόμησης και κυρίως ομαδοποίησης, με παρουσίαση τους σε μορφή ψευδοκώδικα, επεξήγηση του τρόπου λειτουργίας και των ιδιαιτεροτήτων τους, καθώς και των βασικών πλεονεκτημάτων και μειονεκτημάτων κατά περίπτωση.

Συνεχίζουμε με περιγραφή της βάσης δεδομένων (κεφάλαιο 3^{ο}) που δημιουργήθηκε στα πλαίσια της διπλωματικής εργασίας, η οποία εμπεριέχει στοιχεία δημογραφικών μεγεθών, μετρήσεων και ιατρικού ιστορικού που συνθέτουν ένα βασικό ιατρικό προφίλ. Ταυτόχρονα, αναλύονται περαιτέρω τα βασικά πρωτόκολλα που μας βοηθούν στην κωδικοποίηση των ιατρικών ποσοτήτων που μας ενδιαφέρουν.

Στο επόμενο κεφάλαιο (κεφάλαιο 4^{ο}), γίνεται αναφορά στον τρόπο υλοποίησης της διπλωματικής, στα βασικά εργαλεία και βιβλιοθήκες της γλώσσας Python που χρησιμοποιήθηκαν, στη δομή δεδομένων που αξιοποιήσαμε καθώς και στη διαδικασία προεπεξεργασίας ορισμένων μεγεθών και εκλογής χαρακτηριστικών από τα δεδομένα που έχουμε στη διάθεσή μας.

Τέλος, παρουσιάζονται οι δείκτες που χρησιμοποιούνται για σύγκριση των αποτελεσμάτων σε θεωρητικό επίπεδο, καθώς και τα αποτελέσματα τους στα πειραματικά δεδομένα και τους αλγορίθμου ομαδοποίησης που χρησιμοποιήσαμε. Η διπλωματική εργασία κλείνει με παρουσίαση των μελλοντικών επεκτάσεων και βελτιώσεων που μπορούν να υλοποιηθούν για την επίτευξη ενός πληρέστερου συστήματος ομαδοποίησης ιατρικών προφίλ και την επέκταση που αυτό συνεπάγεται.

2 Μηχανική Μάθηση

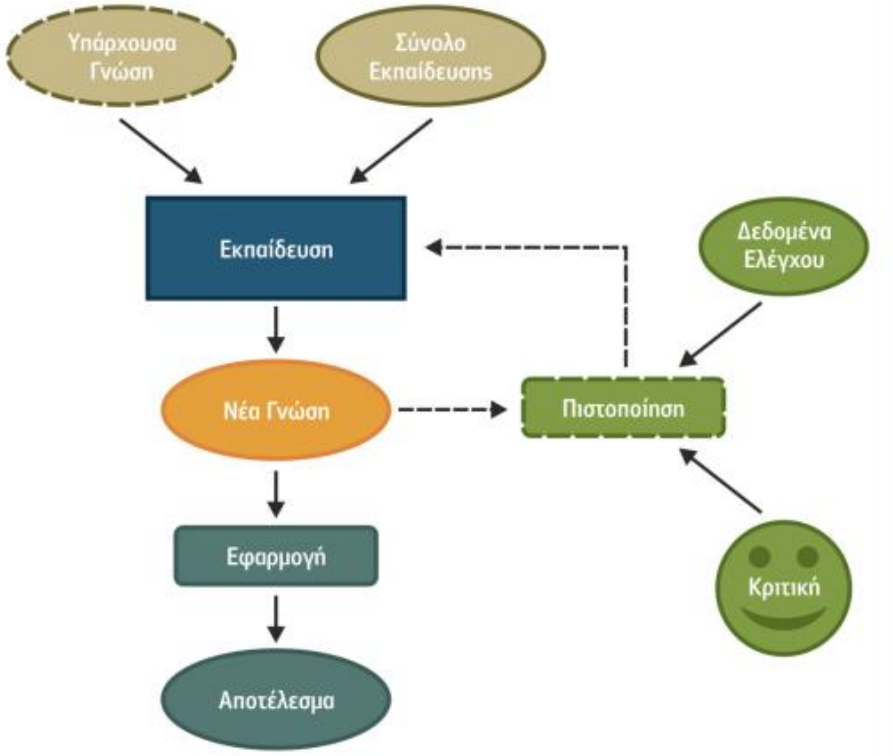
2.1 Εισαγωγή

Η μηχανική μάθηση ανήκει στον επιστημονικό κλάδο της Τεχνητής Νοημοσύνης και μελετά την ικανότητα δημιουργίας μοντέλων ή προτύπων από ένα υπολογιστικό σύστημα μέσω της χρήσης ενός συνόλου δεδομένων. Τα τελευταία χρόνια παρουσιάζεται εφαρμογή τεχνικών μηχανικής μάθησης σε πληθώρα επιστημονικών τομέων με ουσιαστική συνεισφορά στην επίλυση διαφόρων προβλημάτων. Ο τομέας της υγείας επωφελείται ιδιαίτερα από αυτές τις εξελίξεις μέσω της τάση προσανατολισμού προς εφαρμογές που δίνουν τη δυνατότητα αυτοπαρακολούθησης (self-monitoring) διαφόρων παραμέτρων και μεγεθών (HRV, τιμές ζαχάρου στο αίμα, κλπ) για πρόληψη και αποτελεσματικότερη παρακολούθηση προβλημάτων τόσο σωματικής όσο και ψυχικής υγείας καθώς και με την ανάπτυξη αποτελεσματικότερων εργαλείων διάγνωσης μέσω εξόρυξης γνώσης. Τέτοια παραδείγματα αποτελούν η χρήση τεχνικών μηχανικής μάθησης για πρόβλεψη και διαχείριση κατάθλιψης, για αναγνώριση ομιλίας για άτομα με αναπηρία που επηρεάζει την ομιλία, για προσωποποιημένη απώλεια βάρους, η ανίχνευση αλλαγών στη ψυχοκοινωνική κατάσταση ασθενών με πολλαπλή νοσηρότητα, για διαχείριση του στρες, για θεραπεία του πόνου των φανταστικών άκρων και για διακοπή του καπνίσματος και εξατομικευμένη διατροφή με βάση τη γλυκαιμική απόκριση.[9]

Στο παρόν κεφάλαιο γίνεται αναφορά στον τρόπο λειτουργίας και εφαρμογής των αλγορίθμων μηχανικής μάθησης και στις κατηγορίες στις οποίες διακρίνουμε τη μηχανική μάθηση, ενώ παρουσιάζονται αναλυτικότερα οι κυριότεροι αλγόριθμοι που μπορούν να αξιοποιηθούν στα πλαίσια της παρούσας διπλωματικής εργασίας.

2.2 *Γενικός τρόπος λειτουργίας των αλγορίθμων μηχανικής μάθησης*

Στο παραπάνω σχήμα απεικονίζεται ο τρόπος λειτουργίας των αλγορίθμων μηχανικής μάθησης.



Εικόνα 2.1: Διάγραμμα ροής γενικού τρόπου λειτουργίας αλγορίθμων ML [1]

Ο αλγόριθμος δέχεται ως είσοδο ένα σύνολο από δεδομένα τα οποία αποτελούν τα στιγμιότυπα εκπαίδευσης(training set), τα οποία αξιοποιούνται προαιρετικά και σε συνδυασμό με κάποια υπάρχουσα γνώση.

Ακολουθεί η διαδικασία της εκπαίδευσης που οδηγεί στην παραγωγή νέας γνώσης, η οποία με τη σειρά της πρέπει να αξιολογηθεί από ένα πλαίσιο ελέγχου και πιστοποίησης, προτού γίνει αποδεκτή η εφαρμογή της στον τομέα που χρησιμοποιείται. Στο πλαίσιο αυτό της πιστοποίησης χρησιμοποιείται συνήθως ως είσοδος στον αλγόριθμο ένα νέο σύνολο δεδομένων που δεν έχει συμπεριληφθεί στη διαδικασία εκπαίδευσης (test dataset) ώστε μέσω

διαδικασιών ανάκλησης (recall) του αλγορίθμου να γίνει δυνατή η μέτρηση της πληρότητας και της αποτελεσματικότητας του. Επιπλέον, επιστρατεύεται σε πολλές περιπτώσεις και η κριτική ικανότητα του αναγνώστη για την αξιολόγηση των αποτελεσμάτων που παρέχει ο αλγόριθμος.[10]

2.3 Είδη Μηχανικής Μάθησης

Η μηχανική μάθηση μπορεί να διαχωριστεί σε τομείς με βάση την διαδικασία που ακολουθείται για να επιτευχθεί η δημιουργία κάποιου μοντέλου ή προτύπου.

2.3.1 Κατηγοριοποίηση με βάση το βαθμό χρήσης υπάρχουσας γνώσης

Λαμβάνοντας υπόψη το βαθμό χρήσης υπάρχουσας γνώσης παρουσιάζονται σε φθίνουσα σειρά οι εξής περιοχές[10][11]:

1. **Μάθηση Βασισμένη σε Επεξηγήσεις (Explanation-based Learning)**

   Σε αυτή την κατηγορία χρησιμοποιείται υπάρχουσα γνώση για να πραγματοποιηθεί ανάλυση ή επεξήγηση του πως τα δεδομένα και οι συσχετίσεις μεταξύ τους επιβεβαιώνουν τη γνώση. Στη συνέχεια, η επεξήγηση αυτή συμβάλλει στο διαχωρισμό των παραμέτρων των στιγμιότυπων εκπαίδευσης σε σχετικά και μη σχετικά, έτσι ώστε τα παραδείγματα να μπορούν να γενικευτούν βάσει λογικής αντί για στατιστικής. Η Βασισμένη σε Επεξηγήσεις Μάθηση χρησιμοποιείται ευρέως σε προβλήματα σχεδιασμού και δρομολόγησης, ενώ χαρακτηριστικό παράδειγμα εφαρμογής της, αποτελεί η εκμάθηση σκακιού σε έναν υπολογιστή.

2. **Μάθηση Βασισμένη σε Περιπτώσεις (Instance-based Learning)**

   Οι μαθησιακές μέθοδοι που βασίζονται σε περιπτώσεις αποτελούν εννοιολογικά απλούς τρόπους για προσέγγιση συναρτήσεων συνεχών ή διακριτών τιμών. Η διαδικασία της μάθησης σε αυτούς τους αλγορίθμους συνίσταται στην απλή αποθήκευση των δεδομένων εκπαίδευσης και όταν παρουσιάζεται μια νέα περίπτωση, ένα σύνολο παρόμοιων σχετικών περιπτώσεων ανακτάται από τη μνήμη και χρησιμοποιείται για την ταξινόμηση του. Ένα σημαντικό χαρακτηριστικό αυτών

των μεθόδων είναι ότι μπορούν να κατασκευάσουν μια διαφορετική προσέγγιση της συνάρτησης που θέλουμε να εκτιμήσουμε για κάθε ξεχωριστή παρουσία ερωτήματος που γίνεται στο σύστημα. Αυτό έχει ως αποτέλεσμα να διευκολύνεται σημαντικά η προσέγγιση όταν η συνάρτηση που κτίζεται είναι περίπλοκη, καθώς μπορεί να περιγραφεί από ένα σύνολο λιγότερο σύνθετων τοπικών προσεγγίσεων. Η μάθηση Βασισμένη σε Περιπτώσεις έχει εφαρμοστεί σε εργασίες όπου η αποθήκευση και η επαναχρησιμοποίηση προηγούμενης εμπειρίας από προηγουμένως επιλυμένα προβλήματα είναι ιδιαίτερα χρήσιμη όπως σε συλλογιστικές διαδικασίες για νομικές υποθέσεις ή σε εργασίες κέντρων εξυπηρέτησης ή γραφείων πληροφοριών.

3. Μάθηση από Επαγωγές μέσω στιγμιότυπων και παραδειγμάτων (Inductive Learning)

Η μάθηση από επαγωγές μέσω στιγμιότυπων και παραδειγμάτων έγκειται στη χρήση αλγορίθμων πάνω σε ένα σύνολο πληροφοριών ή δεδομένων που έχουν συγκεντρωθεί από παρατηρήσεις και στη λήψη κάποιων αποφάσεων για τις επικρατούσες σχέσεις που παρατηρούνται ανάμεσα σε αυτά. Όπως και στη διαδικασία της επαγωγής, όπου ο άνθρωπος κατανοεί το περιβάλλον με βάση τις παρατηρήσεις που κάνει, έτσι και στη μάθηση από επαγωγές συμβαίνει το ίδιο με αποτέλεσμα τη δημιουργία απλοποιημένων αφαιρετικών εκδοχών του περιβάλλοντος, γνωστά ως νοητικά μοντέλα. Είναι η μέθοδος μάθησης που συναντάται συνήθως σε προβλήματα ταξινόμησης ή παρεμβολής, όπως θα τα δούμε στη συνέχεια.

4. Νευρωνικά Δίκτυα (Neural Networks)

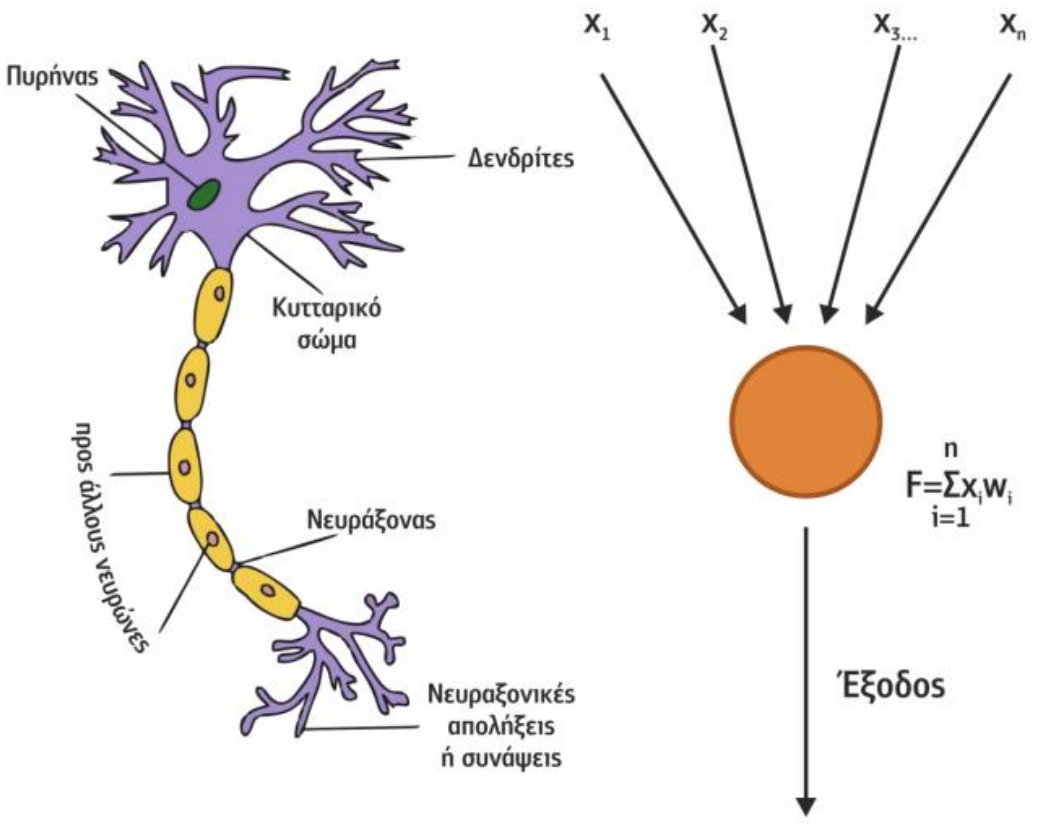
Αποτελούνται από ένα σύνολο από διαφορετικά μαθηματικά μοντέλα, εμπνευσμένα από αντίστοιχα βιολογικά μοντέλα, δηλαδή μοντέλα που προσπαθούν να μιμηθούν τη συμπεριφορά των νευρώνων του ανθρώπινου εγκεφάλου.

Τα φυσικά νευρωνικά δίκτυα του εγκεφάλου αποτελούνται από χιλιάδες διασυνδεδεμένους νευρώνες το καθένα, οι οποίοι επικοινωνούν μεταξύ τους μέσω της μεταφοράς ηλεκτρικών σημάτων. Για την παραγωγή σήματος, ο νευρώνας δέχεται σήματα εισόδου που επιδρούν στο δυναμικό του αυξομειώνοντάς το. Όταν αθροιστικά το δυναμικό ξεπεράσει κάποιο όριο (ποικίλλει από κατηγορία σε κατηγορία κυττάρου μεταξύ – 40 mV και – 75 mV), τότε ο νευρώνας διεγείρεται και παράγει το ηλεκτρικό σήμα, το οποίο μεταφέρεται πάντοτε προς μια προβλέψιμη και σταθερή κατεύθυνση.

Τα τεχνητά νευρωνικά δίκτυα κατά αντιστοιχία με τα βιολογικά, μοντελοποιούνται μαθηματικά από έναν αριθμό απλών και με υψηλό βαθμό εσωτερικής διασύνδεσης επεξεργαστικών μονάδων, οργανωμένων σε στρώματα. Τα Τεχνητά Νευρωνικά Δίκτυα(ΤΝΔ) πραγματοποιούν επεξεργασία πληροφοριών ανταποκρινόμενα δυναμικά σε εξωτερικά ερεθίσματα (εισόδους). Κάθε τεχνητός νευρώνας αποτελείται από πολλές εισόδους xi και μία μόνο έξοδο y. Κάθε είσοδος xi αντιστοιχίζεται με ένα βάρος wi και τα αποτελέσματα αθροίζονται μέσω της συνάρτησης αθροίσματος (summation function) F:

$$F = \sum_{i}^{n} x_i w_i$$

Ο τεχνητός νευρώνας δίνει έξοδο μέσω της συνάρτησης μετάβασης (transfer function), μόνο όταν το ζυγισμένο άθροισμα των εισόδων είναι μεγαλύτερο μιας ορισμένης τιμής κατωφλίου (threshold value) θ. Ένας τεχνητός νευρώνας αποτελεί απλοποιημένο μοντέλο του φυσικού νευρώνα κατά το ότι τα βάρη διασύνδεσης σχηματίζουν τα ηλεκτρικά χαρακτηριστικά της επαφής της σύναψης και η τιμή κατωφλίου προσομοιώνει τη συμπεριφορά κορεσμού του φυσικού νευρώνα, όπως φαίνεται και στην παρακάτω εικόνα.



Εικόνα 2.2: Παρουσίαση δομής φυσικού και τεχνητού νευρώνα[1]

Η έξοδος που δίνει το νευρωνικό δίκτυο είναι μια περίπλοκη συνάρτηση - πρόβλεψη με βάση τις εισόδους που έχει δεχθεί ως εισροή, με αποτέλεσμα τα νευρωνικά δίκτυα να δίνουν λύση σε πολύπλοκα προβλήματα όπου η εύρεση απλούστερων συσχετίσεων μέσω εναλλακτικών μεθόδων μηχανικής μάθησης δεν επαρκεί.

2.3.2 Κατηγοριοποίηση με βάση τον τρόπο λειτουργίας

Διακρίνουμε τις εξής κατηγορίες ανάλογα με τον τρόπο λειτουργίας :

1. **Αλγόριθμοι Επιτηρούμενης Μάθησης**

   Οι αλγόριθμοι επιτηρούμενης μηχανικής μάθησης συνάγουν τις υποκείμενες σχέσεις ανάμεσα στα δεδομένα εισόδου και τα αντίστοιχα επιθυμητά αποτελέσματα (στόχους πρόβλεψης). Η διαδικασία με την οποία επιτυγχάνουν το στόχο αυτό, χρησιμοποιεί κάποια είδη κατηγοριοποιημένα ή επισημασμένα δεδομένα εκπαίδευσης για τη σύνθεση της συνάρτησης που θα γενικεύει τη σχέση ανάμεσα στα δεδομένα εισόδου και εξόδου. Τα δεδομένα εκπαίδευσης αποτελούνται από ένα διάνυσμα χαρακτηριστικών και κάποια επιθυμητή τιμή εξόδου (class label). Η διαδικασία της εκπαίδευσης έχει ως στόχο την ελαχιστοποίηση του σφάλματος πρόβλεψης και εξαρτάται σε πολύ μεγάλο βαθμό από την ποιότητα των προσημασμένων δεδομένων εισόδου.

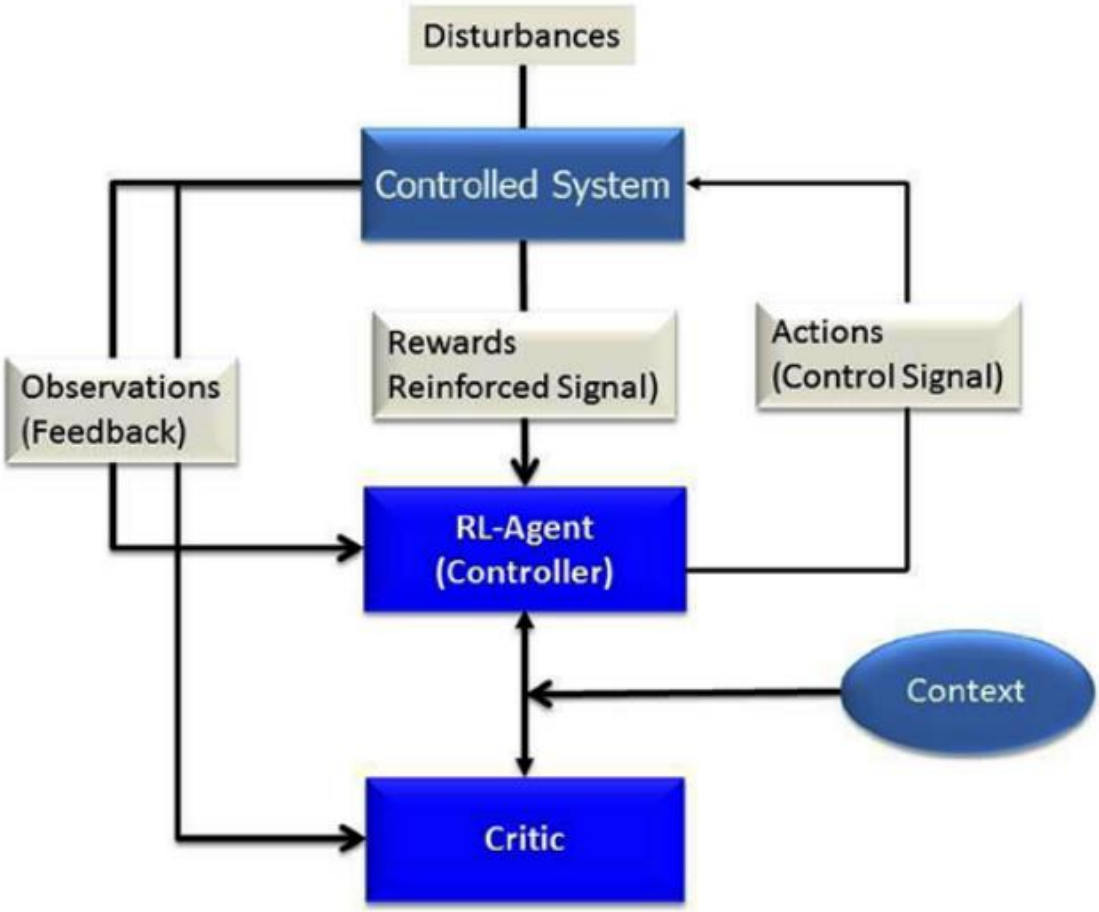
2. **Αλγόριθμοι μη Επιτηρούμενης Μάθησης**

   Οι μη εποπτευόμενοι αλγόριθμοι μάθησης έχουν σχεδιαστεί για να ανακαλύπτουν κρυφές δομές σε μη επισημασμένα σύνολα δεδομένων, στα οποία η επιθυμητή έξοδος είναι άγνωστη. Η γενική προσέγγιση στη μη επιτηρούμενη μάθηση περιλαμβάνει την εκπαίδευση μέσω πιθανοτικών μοντέλων δεδομένων. Σε γενικές γραμμές, σε έναν αλγόριθμο μη επιτηρούμενης μάθησης δίνονται κάποια δεδομένα εισόδου, αλλά δεν υπάρχουν ούτε στόχοι πρόβλεψης (όπως στην εποπτευόμενη μάθηση) ούτε

ανταμοιβές από το περιβάλλον. Ο στόχος της μάθησης σε αυτή την περίπτωση είναι να ανακαλύψουμε τις αναπαραστάσεις των δεδομένων εισαγωγής για αποτελεσματική λήψη αποφάσεων, πρόβλεψη, φιλτράρισμα και ομαδοποίηση πληροφοριών. Οι αλγόριθμοι που στηρίζονται στη θεωρία πιθανοτήτων κάνουν συνήθως χρήση των μεθόδων Bayes και της Εκτιμήτριας Μέγιστης Πιθανοφάνειας, ενώ χρησιμοποιούνται κατά κόρον μεθοδολογίες και από το πεδίο της στατιστικής.

3. Αλγόριθμοι Εξελικτικής Μάθησης

Η μεθοδολογία των αλγορίθμων εξελικτικής μάθησης περιλαμβάνει την εξερεύνηση μιας προσαρμοστικής ακολουθία ενεργειών ή συμπεριφορών από έναν έξυπνο πράκτορα (RL-agent) σε ένα δεδομένο περιβάλλον με κίνητρο τη μεγιστοποίηση μιας αθροιστικής συνάρτησης ανταμοιβής. Οι ενέργειες του ευφυούς πράκτορα προκαλούν παρατηρήσιμες αλλαγές στην κατάσταση του περιβάλλοντος. Η μαθησιακή τεχνική αυτή, συνθέτει ένα μοντέλο προσαρμογής μέσω της εκπαίδευσης σε ένα δεδομένο σύνολο πειραματικών ενεργειών και των αντίστοιχων παρατηρούμενων ανταποκρίσεων στην κατάσταση του περιβάλλοντος. Η εξελικτική μάθηση μπορεί να θεωρηθεί ένα θεωρητικό μοντέλο δοκιμών και σφαλμάτων με ανταμοιβές και τιμωρίες που σχετίζονται με μια ακολουθία ενεργειών. Ο έξυπνος πράκτορας αλλάζει την πολιτική του με βάση τη συλλογική εμπειρία και συνακόλουθες ανταμοιβές και έτσι η εκάστοτε μεθοδολογία μηχανικής μάθησης αναζητά προηγούμενες ενέργειες που εξερεύνησε και κατέληξαν σε ανταμοιβές. Για να δημιουργηθεί μια εξαντλητική βάση δεδομένων ή μοντέλο όλων των πιθανών προβολών δράσης, πρέπει να δοκιμαστούν πολλές μη αποδεδειγμένες ενέργειες ενώ οι ενέργειες αυτές μπορεί να πρέπει να δοκιμαστούν πολλές φορές πριν εξακριβωθεί η αποτελεσματικότητά τους. Επομένως, πρέπει να εξισορροπήσει την εξερεύνηση νέων πιθανών ενεργειών και η πιθανότητα αποτυχίας που προκύπτει από αυτές τις ενέργειες. Στο παρακάτω διάγραμμα ροής διακρίνουμε τα κυριότερα στοιχεία και τον τρόπο λειτουργίας των αλγορίθμων εξελικτικής μάθησης.



Εικόνα 2.3: Διάγραμμα ροής υψηλού επιπέδου εξελικτικής μηχανικής μάθησης [12]

Με βάση αυτή την απεικόνιση, διακρίνουμε τα εξής κρίσιμα στοιχεία της εξελικτικής μηχανικής μάθησης:

- Η πολιτική(Policy) αποτελεί βασικό συστατικό ενός αλγορίθμου εξελικτικής μάθησης καθώς αντιστοιχεί στις ενέργειες ελέγχου την αντιληπτή κατάσταση του περιβάλλοντος.
- Ο κριτικός(Critic) αντιπροσωπεύει μια συνάρτηση αξιολόγησης που ελέγχει τις ενέργειες που γίνονται σύμφωνα με την υπάρχουσα πολιτική. Εναλλακτικά, ο κριτικός αξιολογεί την απόδοση της τρέχουσας κατάστασης ως απόκριση σε μια ενέργεια που πραγματοποιήθηκε σύμφωνα με την τρέχουσα πολιτική. Ο κριτικός-πράκτορας διαμορφώνει την πολιτική κάνοντας συνεχείς προσαρμογές και διορθώσεις.
- Η συνάρτηση ανταμοιβής εκτιμά το κατά πόσο είναι επιθυμητή η κατάσταση του περιβάλλοντος για μια διεργασία ελέγχου.

- Τα μοντέλα σχεδιάζουν εργαλεία που βοηθούν στην πρόβλεψη της πορείας δράσης μελετώντας πιθανές μελλοντικές καταστάσεις.

2.3.3 Κατηγοριοποίηση με βάση τη φύση του επιλυόμενου προβλήματος

Ανάλογα με τη φύση του προβλήματος που επιχειρούν να επιλύσουν έχουμε την εξής κατηγοριοποίηση:

1. Αλγόριθμοι Ταξινόμησης (Classification)
2. Αλγόριθμοι Παρεμβολής (Regression)
3. Αλγόριθμοι Ομαδοποίησης (Clustering)
4. Αλγόριθμοι Συνεργατικού Φιλτραρίσματος (Collaborative Filtering)
5. Αλγόριθμοι Μείωσης Διαστάσεων (Dimensionality Reduction)

Στη συνέχεια, παρουσιάζονται αναλυτικότερα οι παραπάνω κατηγορίες.

2.3.3.1 Αλγόριθμοι Ταξινόμησης (Classification)

Οι αλγόριθμοι ταξινόμησης ανήκουν στους αλγορίθμους επιτηρούμενης μάθησης και χρησιμοποιούνται σε προβλήματα με καλά καθορισμένα όρια, όπου οι είσοδοι καθορίζονται από ένα συγκεκριμένο σύνολο χαρακτηριστικών και οι έξοδοι είναι διακριτές κλάσεις ή κατηγορίες.

Η διαδικασία ταξινόμησης αναπτύσσει ένα αρχείο εμπειριών που οδηγεί σε αξιολόγηση νέων εισροών, ταιριάζοντάς τα με προηγούμενα παρατηρούμενα πρότυπα. Εάν ένα μοτίβο μπορεί να ταξινομηθεί, η είσοδος σχετίζεται με το προκαθορισμένο μοτίβο. Εάν ένα μοτίβο δεν μπορεί να ταξινομηθεί, φυλάσσεται για περαιτέρω αξιολόγηση για να προσδιοριστεί εάν είναι ένα έγκυρο μοτίβο που δεν έχει ανακαλυφθεί ακόμη ή ένα σπάνιο μοτίβο.

Τα τρία κύρια βήματα που εμπλέκονται στην ταξινόμηση είναι η σύνθεση ενός μοντέλου, η χρήση ενός αλγορίθμου εκμάθησης και η χρήση του μοντέλου για την κατηγοριοποίηση νέων δεδομένων.

Οι συνηθέστεροι αλγόριθμοι που συναντάμε κατά την επίλυση προβλημάτων ταξινόμησης στη Μηχανική Μάθηση είναι α)τα δέντρα αποφάσεων(decision trees), β) K-

Πλησιέστεροι-Γείτονες (KNN), γ) Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine – SVM) κ.α.

2.3.3.2 Αλγόριθμοι Παρεμβολής ή Παλινδρόμησης (Regression)

Οι αλγόριθμοι παρεμβολής χρησιμοποιούνται για την πρόβλεψη κάποιας αριθμητικής τιμής, όπως για παράδειγμα ηλικίας, τιμής, κόστους κλπ. Ανήκουν μαζί με τους αλγορίθμους ταξινόμησης στην επιτηρούμενη μηχανική μάθηση, ενώ η διαφορά τους από τους πρώτους είναι ότι προσεγγίζουν συνεχείς αριθμητικές τιμές σε αντιδιαστολή με τις διακριτές τιμές των κλάσεων της ταξινόμησης. Οι μέθοδοι μηχανικής μάθησης με παρεμβολή επιχειρούν να προσεγγίσουν τη βέλτιστη σχέση που συνδέει μία ή περισσότερες μεταβλητές εισόδου με τη ζητούμενη έξοδο, εφαρμόζοντας τη θεωρία της παλινδρόμησης, επιχειρώντας δηλαδή την πρόβλεψη της τιμής εξόδου έτσι ώστε η διαφορά του σφάλματος μεταξύ της προβλεπόμενης τιμής και της πραγματικής τιμής να είναι ελάχιστη. Υπάρχουν διάφοροι αλγόριθμοι παρεμβολής ανάλογα με α) τον αριθμό των μεταβλητών εισόδου β)τον αριθμό των μεταβλητών εξόδου και γ) το σχήμα της καμπύλης παρεμβολής. Ενδεικτικά, αναφερόμαστε στη γραμμική παλινδρόμηση (η καμπύλη παλινδρόμησης είναι ευθεία), στην πολυωνυμική παλινδρόμηση(η καμπύλη παλινδρόμησης εκφράζεται από πολυωνυμική συνάρτηση δευτέρου βαθμού και πάνω) και γ) η λογιστική παλινδρόμηση η οποία χρησιμοποιείται συχνά ως αλγόριθμος ταξινόμηση, καθώς χρησιμοποιείται όταν η εξαρτημένη μεταβλητή εκφράζεται με δυαδικό κατηγορικό τρόπο πχ. σε ερωτήματα κατάφασης/άρνησης, σε προβλήματα με δύο πιθανές κλάσεις ως εξόδους κλπ.
[13]

2.3.3.3 Αλγόριθμοι Ομαδοποίησης (Clustering)

Η ομαδοποίηση είναι μια διαδικασία ανακάλυψης γνώσης που ομαδοποιεί στοιχεία από μια δεδομένη συλλογή, με βάση παρόμοια χαρακτηριστικά ή γνωρίσματα Τα μέλη του ίδιου συμπλέγματος έχουν παρόμοια χαρακτηριστικά, σε σχέση με αυτά ανήκουν σε διαφορετικές συστάδες. Η ομαδοποίηση ανήκει στο πεδίο της μη επιτηρούμενης μηχανικής μάθησης.

Γενικά, η ομαδοποίηση περιλαμβάνει έναν επαναληπτικό αλγόριθμο δοκιμής και σφάλματος που λειτουργεί με την υπόθεση ομοιότητας (ή ανομοιότητας) και σταματά όταν ικανοποιηθεί ένα κριτήριο τερματισμού. Η πρόκληση είναι να βρεθεί μια συνάρτηση που μετρά το βαθμό ομοιότητας μεταξύ δύο αντικειμένων (ή σημεία δεδομένων) ως αριθμητική

τιμή. Οι παράμετροι ομαδοποίησης, όπως ο αλγόριθμος ομαδοποίησης, η συνάρτηση απόστασης, το όριο πυκνότητας και ο αριθμός των συστάδων, εξαρτώνται από τις εφαρμογές και το μεμονωμένο σύνολο δεδομένων. Γνωστοί αλγόριθμοι της συγκεκριμένης κατηγορίας είναι ο Αλγόριθμος των K-μέσων(K-means), ο αλγόριθμος Mean-Shift, ο αλγόριθμος DBSCAN κ.α.

2.3.4 Αλγόριθμοι Συνεργατικού Φιλτραρίσματος (Collaborative Filtering)

Το συνεργατικό φιλτράρισμα (CF) είναι μια διαδικασία φιλτραρίσματος για πληροφορίες ή μοτίβα, χρησιμοποιώντας συνεργατικές μεθόδους μεταξύ πολλαπλών πηγών δεδομένων. Η CF εξερευνά μια περιοχή ενδιαφέροντος συγκεντρώνοντας προτιμήσεις από πολλούς χρήστες με παρόμοια ενδιαφέροντα και προτάσεις που βασίζονται σε αυτές τις προτιμήσεις. Οι αλγόριθμοι αυτοί κάνουν ικανοποιητικές συστάσεις σε σύντομο χρονικό διάστημα, παρά τα πολύ αραιά δεδομένα, τον αυξανόμενο αριθμό χρηστών και αντικειμένων, συνωνύμων, το θόρυβο δεδομένων και τα ζητήματα απορρήτου.

Η μηχανική μάθηση πραγματοποιεί προγνωστική ανάλυση, με βάση τις καθιερωμένες ιδιότητες που αντλήθηκαν από το δεδομένα και βοηθά στην εξερεύνηση χρήσιμων γνώσεων ή προηγουμένως άγνωστης γνώσης, συνδυάζοντας νέες πληροφορίες με ιστορικές πληροφορίες που υπάρχουν με τη μορφή προτύπων. Αυτά τα μοτίβα χρησιμοποιούνται για να φιλτράρουν νέες πληροφορίες ή μοτίβα. Μόλις επικυρωθούν αυτές οι νέες πληροφορίες σε ένα σύνολο συνδεδεμένων προτύπων συμπεριφοράς, ενσωματώνονται στην υπάρχουσα βάση δεδομένων γνώσεων. Οι νέες πληροφορίες μπορούν επίσης να διορθώσουν τα υπάρχοντα μοντέλα ενεργώντας ως πρόσθετα δεδομένα εκπαίδευσης

2.3.5 Αλγόριθμοι Μείωσης Διαστάσεων (Dimensionality Reduction)

Η μείωση διαστάσεων αποτελεί τη διαδικασία μείωσης τυχαίων μεταβλητών μέσω επιλογής παραμέτρων και εξαγωγής χαρακτηριστικών. Η μείωση διαστάσεων επιτρέπει μικρότερους χρόνους εκπαίδευσης και βελτιωμένη γενίκευση και μειώνει την υπερμοντελοποίηση (overfitting).

Η επιλογή παραμέτρων είναι η διαδικασία σύνθεσης ενός υποσυνόλου των αρχικών μεταβλητών για κατασκευή μοντέλου εξαλείφοντας περιττά ή άσχετα χαρακτηριστικά. Η

εξαγωγή χαρακτηριστικών αντίθετα, είναι η διαδικασία μετατροπής ενός πολυδιάστατου χώρου σε χώρο με λιγότερες διαστάσεις συνδυάζοντας χαρακτηριστικά.

[12]

2.4 Αλγόριθμοι Ταξινόμησης

2.4.1 Δέντρα αποφάσεων(Decision Trees)

Οι μέθοδοι μηχανικής μάθησης που στηρίζονται σε δέντρα αποφάσεων στοχεύουν στη δημιουργία ενός μοντέλου / στην εύρεση μιας συνάρτησης για την επίλυση ενός κατηγορικού προβλήματος. Για να επιτύχουν το στόχο τους χρησιμοποιούν μία άπληστη (greedy) αναδρομική προσέγγιση κατά την οποία κατασκευάζεται ένα δέντρο από πάνω προς τα κάτω που χωρίζει τα δεδομένα εκπαίδευση σε υποσύνολα καθώς αυξάνεται το βάθος του δέντρου. Το δέντρο που δομείται με αυτή τη διαδικασία ισοδυναμεί με ένα σύνολο κανόνων (κανόνες ταξινόμησης) ενώ στο τελευταίο επίπεδο του δέντρου συναντάμε τις κατηγορίες ταξινόμησης ως φύλλα του.

Η διαδικασία «διαίρει και βασίλευε» που ακολουθείται με συνεχείς υποδιαιρέσεις των δεδομένων εκπαίδευσης σταματάει όταν ενεργοποιηθεί ένα κριτήριο διακοπής. Κοινά κριτήρια διακοπής είναι τα παρακάτω:

1. Όλα τα ενδεχόμενα στο σύνολο των εκπαιδευτικών δεδομένων ανήκουν σε μία μόνο τιμή του y
2. Το μέγιστο βάθος του δέντρου έχει επιτευχθεί.
3. Ο αριθμός των περιπτώσεων στο τερματικό κόμβο είναι μικρότερος από τον ελάχιστο αριθμό των υποθέσεων των μητρικών κόμβων.
4. Εάν ο κόμβος χωρίστηκε, ο αριθμός των περιπτώσεων, σε ένα ή περισσότερους κόμβους «παιδιά του αρχικού κόμβου», είναι μικρότερος από τον ελάχιστο αριθμό των περιπτώσεων για τους κόμβους «γονείς».
5. Το καλύτερο κριτήριο διαχωρισμού δεν είναι μεγαλύτερο από ένα ορισμένο όριο.

Ακολουθεί ενδεικτικά ένας αλγόριθμος που κατασκευάζει δέντρο απόφασης από δεδομένα εκπαίδευσης σε ψευδογλώσσα όπου είσοδος είναι ένα σύνολο εκπαίδευσης D, το οποίο είναι

ένα σύνολο παρατηρήσεων και οι σχετικές τιμές της αντίστοιχης κλάσης, η λίστα χαρακτηριστικών Α και ένα επιλεγμένο κριτήριο διακοπής, ενώ ως έξοδος προκύπτει το δέντρο απόφασης:

```
(1)Δημιουργία κόμβου N

(2) Αν όλες οι περιπτώσεις του συνόλου εκπαίδευσης έχουν
    την ίδια τιμή της τάξης C, τότε επέστρεψε το N σαν
    φύλλο με την ετικέτα C


(3)Αν η λίστα των χαρακτηριστικών είναι άδεια, τότε
    επέστρεψε N σαν φύλλο με την ετικέτα με τη μεγαλύτερη
    τιμή της τάξης στην έξοδο στο σύνολο της εκπαίδευσης

(4) Εφάρμοσε το επιλεγμένο κριτήριο διαχωρισμού στο σύνολο
    της εκπαίδευσης για να βρεθεί το καλύτερο
    χαρακτηριστικό για διαχωρισμό


(5)Ετικέτα κόμβου N με το χαρακτηριστικό του κριτηρίου
    διαχωρισμο

(6)Αφαίρεση του χαρακτηριστικού του κριτηρίου διαχωρισμού
    από τη λίστα των χαρακτηριστικών


(7)Για κάθε τιμή j στο χαρακτηριστικό του κριτηρίου
    διαχωρισμού
    • Ας είναι Dj οι περιπτώσεις στο σύνολο εκπαίδευσης που
    ικανοποιούν το χαρακτηριστικό με τιμή j
    • Αν Dj είναι άδειο (καμιά περίπτωση), τότε πάρε σαν
    φύλλο με την ετικέτα με τη μεγαλύτερη τιμή της τάξης
    στην έξοδο στον κόμβο N
    • διαφορετικά πάρε τον κόμβο που έδωσε το Generate
    Decision Tree (Dj,λίστα χαρακτηριστικών, επιλεγμένο
    κριτήριο διαχωρισμού) στον κόμβο N


(8)Τέλος για ( for)

(9)Δώσε τον κόμβο N
```

[14]

2.4.2 k-Κοντινότεροι-Γείτονες (k-Nearest-Neighbor, kNN)

Ο αλγοριθμος των k πλησιέστερων γειτόνων στηρίζεται στην παραδοχή ότι η τιμή της συνάρτησης-στόχου για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των k «κοντινών» του στιγμιότυπων εκπαίδευσης, τα οποία αποτελούν τους «γείτονές» του. Με άλλα λόγια, κάθε απόφαση για ταξινόμηση ενός δείγματος x παίρνεται αποκλειστικά με βάση τις ετικέτες των k πλησιέστερων γειτόνων του. Όπως προείπαμε, η αναζήτηση σύμφωνα με τον κανόνα του k πλησιέστερου γείτονα ξεκινάει από το σημείο ελέγχου x και μεγαλώνει μία σφαιρική περιοχή ώσπου να περιλαμβάνει k δείγματα εκπαίδευσης. Το σημείο ελέγχου παίρνει την ετικέτα που έχει η πλειοψηφία αυτών των δειγμάτων. Αν k=1 τότε απλά το αντικείμενο αντιστοιχείται στην κλάση του κοντινότερου γείτονα του.

Τρία ζητήματα πρέπει να αποφασιστούν προκειμένου να καθοριστεί πλήρως ο αλγόριθμος:

• Ο ορισμός της απόστασης μεταξύ δύο στιγμιότυπων, δηλαδή μιας μετρικής πάνω στο χώρο των στιγμιότυπων (instance space), που θα εκφράζει την εγγύτητα, ή αλλιώς την «ομοιότητα» μεταξύ των στιγμιότυπων.

• Ο τρόπος συνδυασμού των τιμών των k κοντινότερων γειτόνων.

• Η τιμή του k.

Ένα μεγάλο πλεονέκτημα των ταξινομητών kNN σε σχέση με τα δέντρα αποφάσεων και η σταθερότητα τους (Breiman L. 1996). Μια μέθοδος μάθησης χαρακτηρίζεται «ασταθής» εάν μικρές αλλαγές στα δεδομένα εκπαίδευσης έχουν σαν αποτέλεσμα μεγάλες αλλαγές στα αποτελέσματα του ταξινομητή.[15]

```
Είσοδος: T // Σύνολο δεδομένων εκπαίδευσης
         K // Αριθμός κοντινότερων γειτόνων
         t // πλειάδα προς κατηγοριοποίηση
         Έξοδος: c // Κλάση όπου θα κατηγοριοποιηθεί η t

(1) N = Ø

(2) Για κάθε d ∈ T επανέλαβε

(3) Αν |N| ≤ K τότε
        N = N ∪ {d};
        Αν ∃ u ∈N τέτοιο ώστε dist(t,u) ≤ dist (t, d), τότε
           N = N − {u};
           N = N ∪ {d};
     Τέλος _ αν
    Τέλος _ επανάληψης

(4) c = κλάση όπου τα περισσότερα u ∈ N κατηγοριοποιούνται

(5) Τέλος αλγορίθμου
```

[16]

Ευρέως διαδεδομένη είναι και μια παραλλαγή του κλασσικού αλγορίθμου kNN, όπου κάθε γείτονας έχει διαφορετική συνεισφορά στην κατηγοριοποίηση ανάλογα με ένα βάρος που του έχει εκχωρηθεί. Σε αυτή την περίπτωση οι πιο μακρινοί γείτονες έχουν μικρότερο βάρος, ενώ οι πιο κοντινοί μεγαλύτερο. Στην παραπάνω μέθοδο, όπου συμμετέχουν όλες οι πλειάδες των δεδομένων εκπαίδευσης, αναφερόμαστε πλέον στον αλγόριθμο κοντινότερου γείτονα *σταθμισμένης* απόστασης.

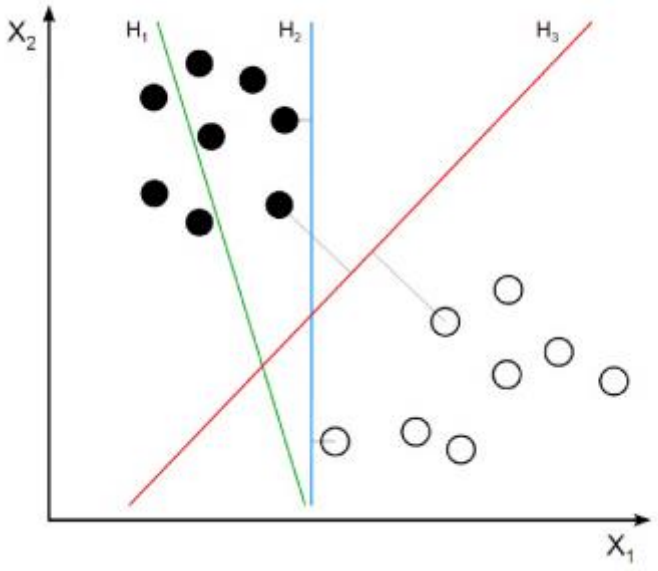
2.4.3 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης πραγματοποιούν το στόχο της κατηγοριοποίησης αναπαριστώντας τα δεδομένα εκπαίδευσης ως σημεία στο χώρο και μελετώντας την τοπολογία τους ώστε να ανακαλύψουν το βέλτιστο χωρικό διαχωρισμό ανάμεσα σε δεδομένα

που ανήκουν σε διαφορετικές κατηγορίες. Η γεωμετρική αυτή κατηγοριοποίηση στηρίζεται σε δύο κεντρικές ιδέες:

α) τα διανύσματα που βρίσκονται σε χώρο υψηλής διάστασης χαρτογραφούνται με μη γραμμικό τρόπο ενώ η κατηγοριοποίηση στο νέο χώρο πραγματοποιείται με γραμμικό τρόπο.

β) ο αλγόριθμος προσπαθεί να βρει τα υπέρεπίπεδα (hyperplanes) που διασπούν τα δεδομένα όσο το δυνατόν περισσότερο. Για τον υπολογισμό των βέλτιστων υπερεπιπέδων, ιδιαίτερα χρήσιμη είναι η έννοια του περιθωρίου (margin), το οποίο ορίζεται ως η μικρότερη απόσταση ενός σημείου από το επίπεδο διαχωρισμού και επιλέγεται πάντα το υπερεπίδεδο που μεγιστοποιεί το περιθώριο. [18]



Εικόνα 2.4: SVM διαχωρισμός [17]

2.4.4 Συγκριτική παρουσίαση πλεονεκτημάτων και μειονεκτημάτων προαναφερθέντων αλγορίθμων ταξινόμησης

ΑΛΓΟΡΙΘΜΟΣ	ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
Δέντρα Αποφάσεων	• Φυσικός και κατανοητός τρόπος αναπαράστασης της γνώσης για αποσαφήνιση • Παραγόμενος αυτόματα εκτελέσιμος κώδικας, εφόσον κάθε δέντρο μεταφράζεται σε μια ακολουθία if-then εντολών, • Αποτελεί παραδοσιακό μοντέλο μηχανικής μάθησης.	• Υπάρχει δυσκολία στη μετατροπή αριθμητικών τιμών χαρακτηριστικών σε αντίστοιχα κατηγορήματα • Προκύπτει ανάγκη κλαδέματος κλαδιών που αναπτύσσονται συνεχώς. • Η υπερβολική ανάπτυξη μειώνει μεν το σφάλμα στο σύνολο εκπαίδευσης, αλλά με κόστος την αύξηση του πλήθους των απαιτούμενων ελέγχων σφάλματος. • Υπάρχει δυσχέρεια στη διαχείριση χαρακτηριστικών με πολλές τιμές • Δυσκολεύεται ο αλγόριθμος να λειτουργήσει σωστά, όταν στο σύνολο εκπαίδευσης υπάρχουν δείγματα χωρίς τιμές σε ορισμένα από τα χαρακτηριστικά τους, σύνηθες φαινόμενο κατά τη συγκέντρωση και καταγραφή δειγμάτων
KNN	• Είναι αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών. • Διαθέτουν απλό αλγόριθμο. • Σε πολλές περιπτώσεις πετυχαίνουν υψηλές επιδόσεις κατηγοριοποίησης.	• Το γεγονός ότι γίνονται πολλές συγκρίσεις μεταξύ παρατηρήσεων απαιτεί πολύ αποτελεσματικές τεχνικές καταλογοποίησης (indexing). • Η κατηγοριοποίηση νέων παρατηρήσεων διαρκεί πολύ περισσότερο χρόνο, ειδικά στις περιπτώσεις όπου ο αριθμός των εν δυνάμει «γειτόνων» είναι μεγάλος.

		• Τα αποτελέσματα τους μπορούν να επηρεαστούν σε σημαντικό βαθμό από το πλήθος των γειτόνων k. • Είναι ευαίσθητοι σε τοπικά χαρακτηριστικά των δεδομένων. • Είναι ευαίσθητοι στην ύπαρξη μη σημαντικών μεταβλητών εισόδου
SVM	• Η χρήση της συνάρτησης πυρήνα τις καθιστά πολύ αποτελεσματικές σε περιπτώσεις όπου υπάρχουν μη γραμμικές σχέσεις στα δεδομένα. • Επιτυγχάνουν υψηλές επιδόσεις κατηγοριοποίησης, κυρίως στην περίπτωση δυαδικών κλάσεων. • Διαθέτουν στιβαρή θεωρητική θεμελίωση. • Είναι ανθεκτικές στην υπερπροσαρμογή και διαθέτουν πολύ καλή δυνατότητα γενίκευσης με κατάλληλη ρύθμιση κάποιας παραμέτρου. • Δεν παγιδεύονται σε τοπικά ελάχιστα. • Είναι αποτελεσματικές σε περιπτώσεις συνόλων δεδομένων με πολλές στήλες και σχετικά λίγες γραμμές.	• Δεν υπάρχει κάποια μεθοδολογία για την επιλογή της συνάρτησης πυρήνα καθώς και των παραμέτρων του πυρήνα. • Δεν παρέχουν ερμηνεύσιμα μοντέλα. Η συμβολή της εκάστοτε μεταβλητής εισόδου στο τελικό αποτέλεσμα κατηγοριοποίησης είναι αδιαφανής. • Έχουν σχετικά μεγάλους χρόνους εκπαίδευσης, αν και σημαντικά χαμηλότερους από αυτούς των Νευρωνικών Δικτύων. • Έχουν μεγάλες απαιτήσεις σε μνήμη υπολογιστή. • Σε περίπτωση κλάσεων με πολλαπλές τιμές το πρόβλημα διατυπώνεται σαν συνδυασμός προβλημάτων δυαδικών κλάσεων.

Πίνακας 2.1: Πλεονεκτήματα και μειονεκτήματα αλγορίθμων kNN, SVM & Δέντρων Αποφάσεων [19]

2.4.5 Ταξινόμηση πολλαπλών κλάσεων (Multiclass Classification)

Η ταξινόμηση πολλαπλών κλάσεων χρησιμοποιείται σε εργασίες ταξινόμησης όπου οι διαθέσιμες κατηγορίες/κλάσεις είναι παραπάνω από δύο και αποτελεί μια γενίκευση του προβλήματος της ταξινόμησης. Στην ταξινόμηση πολλαπλών κλάσεων γίνεται η παραδοχή ότι κάθε είσοδος μπορεί να αντιστοιχίζεται σε μία μόνο προβλεπόμενη κλάση/έξοδο.

Πολλοί αλγόριθμοι ταξινόμησης μπορούν να προσαρμοστούν ώστε να επιλύουν προβλήματα πολλαπλών κλάσεων. Αυτή η προσαρμογή μπορεί να επιτευχθεί με τους εξής τρόπους:

- **One-Versus-All (OVA)**

  Δεδομένου ενός προβλήματος ταξινόμησης με m διακριτές κλάσεις, αυτό μπορεί να μετασχηματιστεί σε ισάριθμα προβλήματα δύο κλάσεων. Κάθε υποπρόβλημα σε αυτή τη διαδικασία έχει μια θετική κλάση, η οποία αντιπροσωπεύει την αρχική κατηγορία και μια αρνητική κλάση, όπου συμπεριλαμβάνονται οι υπόλοιπες (m-1) αρχικές κατηγορίες. Για να ταξινομηθεί μια νέα είσοδος απαιτείται η συμμετοχή όλων των ταξινομητών. Για παράδειγμα, αν ο πρώτος ταξινομητής αποφανθεί θετικά τότε λαμβάνει μία ψήφο. Στη συνέχεια, η είσοδος περνάει και από τους υπόλοιπους ταξινομητές όπου και πάλι αν υπάρξει πρόβλεψη στη θετική κλάση κάποιου λαμβάνει την αντίστοιχη ψήφο, ενώ στο τέλος επιλέγεται ο ταξινομητής με τις περισσότερες ψήφους ως η τελική πρόβλεψη της ταξινόμησης πολλαπλών κλάσεων.

- **All-Versus-All (AVA)**

  Η μέθοδος αυτή αποτελεί μία εναλλακτική της One-Versus-All μεθόδου. Κατασκευάζονται $\frac{m(m-1)}{2}$ προβλήματα ταξινόμησης και στο καθένα οι δύο κλάσεις αποτελούνται από συνδυασμό των αρχικών κλάσεων ανά δύο. Και εδώ για να ταξινομηθεί μια νέα είσοδος απαιτείται η συμμετοχή όλων των ταξινομητών και επιλέγεται τελικά η πρόβλεψη με τις περισσότερες ψήφους. Η All-Versus-All μεθοδολογία θεωρείται ανώτερη της One-Versus-All.

- **Κωδικοί διόρθωσης σφαλμάτων (Error-correcting codes)**

Καθώς οι ταξινομητές δύο κλάσεων είναι ευαίσθητοι σε σφάλματα είναι πιθανό να επηρεαστούν οι ψήφοι στις μεθοδολογίες που περιγράψαμε προηγουμένως. Έτσι πολλές φορές χρησιμοποιούνται κωδικοί διόρθωσης σφαλμάτων ως μια εναλλακτική μέθοδος ταξινόμησης πολλαπλών κλάσεων ή για τη βελτίωση της ακρίβειας ταξινόμησης στις προηγούμενες περιπτώσεις. Στη μεθοδο αυτή χρησιμοποιείται ένα διάνυσμα από bits που δημιουργούν μια λέξη και εκπαιδεύουμε έναν ταξινομητή σε κάθε θέση bit. Στη συνέχεια, εάν έχουμε ταξινόμηση σε παραπάνω από μία κλάσεις, γεγονός που συνεπάγεται λάθος, γίνεται μια μέτρηση απόστασης μεταξύ των θετικών κλάσεων για να υπολογιστεί η πλησιέστερη. Η απόσταση που υπολογίζεται ανάμεσα

στα bit λέγεται απόσταση Hamming και αποτελεί τον αριθμό των διαφορετικών bits μεταξύ δύο κωδικών διόρθωσης.[20]

2.5 *Αλγόριθμοι Ομαδοποίησης*

Στην παρούσα ενότητα, αναφερόμαστε αναλυτικότερα στα χαρακτηριστικά των αλγορίθμων ομαδοποίησης(clustering), περιγράφοντας τους βασικούς άξονες στους οποίους κινούνται οι αλγόριθμοι αυτοί για το σχηματισμό των ομάδων ενώ περιγράφουμε και τους αλγορίθμους ομαδοποίησης που θα χρησιμοποιήσουμε στα πλαίσια της παρούσης διπλωματικής εργασίας.

2.5.1 *Βασικοί τρόποι ομαδοποίησης/συσταδοποίησης*

- Μέσω διαμερίσεων (partitioning)

  Οι περισσότερες μέθοδοι διαμέρισης βασίζονται σε απόστασεις των δεδομένων. Με δεδομένο τον αριθμό των διαμερίσεων {k}, η μέθοδος δημιουργεί την πρώτη διαμέριση,. Στη συνέχεια χρησιμοποιεί μια επαναληπτική τεχνική μετεγκατάστασης που προσπαθεί να βελτιώσει τη διαμέριση μετακινώντας αντικείμενα από τη μια ομάδα στην άλλη με βάση τις αποστάσεις τους από τις κοντινές διαμερίσεις.

- Με μεθόδους ιεραρχίας

  Μια ιεραρχική μέθοδος μπορεί να χαρακτηριστεί ως είτε προσθετική ή διαιρετική. Η πρώτη προσέγγιση, που ονομάζεται επίσης προσέγγιση από κάτω προς τα πάνω, ξεκινά με κάθε ένα αντικείμενο που σχηματίζει μια ξεχωριστή ομάδα. Συγχωνεύει διαδοχικά τα αντικείμενα ή τις ομάδες που είναι κοντά το ένα στο άλλο, έως ότου όλες οι ομάδες συγχωνευτούν σε μία (το ανώτατο επίπεδο της ιεραρχίας), ή ισχύει μια συνθήκη τερματισμού. Η διαιρετική προσέγγιση, που ονομάζεται επίσης το προσέγγιση από πάνω προς τα κάτω, ξεκινά με όλα τα αντικείμενα στην ίδια ομάδα. Σε κάθε διαδοχική επανάληψη, ένα σύμπλεγμα χωρίζεται σε μικρότερα σμήνη, έως ότου τελικά κάθε αντικείμενο είναι ένα σύμπλεγμα ή με βάση μια συνθήκη τερματισμού. Βασική αδυναμία αυτών των μεθόδων είναι η αδυναμία αναίρεσης μιας διάσπασης ή συγχώνευσης αφού αυτή πραγματοποιηθεί.

- Βάσει πυκνότητας

  Βασική ιδέα των μεθόδων που λειτουργούν βάσει πυκνότητας είναι να συνεχίσει να αναπτύσσεται ένα δεδομένο σύμπλεγμα όσο η πυκνότητα

(αριθμός αντικειμένων ή σημεία δεδομένων) στη «γειτονιά» υπερβαίνει κάποιο όριο. Για παράδειγμα, για καθένα σημείο δεδομένων εντός ενός δεδομένου συμπλέγματος, η γειτονιά μιας δεδομένης ακτίνας πρέπει να περιέχει τουλάχιστον έναν ελάχιστο αριθμό σημείων.

- Με μεθόδους πλέγματος(grid-based methods)

  Οι μέθοδοι που βασίζονται σε πλέγμα κβαντοποιούν το χώρο του αντικειμένου σε ένα πεπερασμένο αριθμό κελιών που σχηματίζουν δομή πλέγματος. Όλες οι εργασίες ομαδοποίησης εκτελούνται στη δομή του πλέγματος (δηλαδή στον κβαντοποιημένο χώρο). Το κύριο πλεονέκτημα αυτής τη; προσέγγισης είναι ο γρήγορος χρόνος επεξεργασίας, ο οποίος είναι συνήθως ανεξάρτητος του αριθμού των δεδομένων και εξαρτάται μόνο από τον αριθμό των κελιών σε κάθε διάσταση στον κβαντοποιημένο χώρο. Η χρήση πλεγμάτων είναι συχνά μια αποτελεσματική προσέγγιση σε πολλά προβλήματα εξόρυξης χωρικών δεδομένων ενώ οι μέθοδοι που βασίζονται στο πλέγμα μπορούν να ενσωματωθούν με άλλες μεθόδους ομαδοποίησης. [19]

Ακολουθεί η περιγραφή της λειτουργίας του αλγορίθμου Κ-μέσων.

2.5.2 Αλγόριθμος των Κ-μέσων (K-means)

Για την εφαρμογή του αλγορίθμου K-means, θεωρείται προκαθορισμένος από πριν ο αριθμός των ομάδων Κ. Αρχικά, Κ τυχαία σημεία επιλέγονται ως τα κέντρα των ομάδων από τα διαθέσιμα δεδομένα. Στη συνέχεια, κάθε σημείο ανατίθεται στην ομάδα στης οποίας το κέντρο είναι πιο κοντά, δηλαδή στην ομάδα από την οποία το σημείο έχει τη μικρότερη απόσταση . □ Έπειτα, υπολογίζεται το μέσο διάνυσμα των σημείων κάθε ομάδας (ο μέσος όρος των σημείων της) και ορίζεται εκ νέου αυτό ως κέντρο της κλάσης. Τα δύο τελευταία βήματα επαναλαμβάνονται για ένα προκαθορισμένο αριθμό βημάτων ή μέχρι να μην υπάρχει αλλαγή στο διαχωρισμό των σημείων σε ομάδες. Ακολουθεί ο αλγόριθμος των Κ-μέσων σε μορφή ψευδοκώδικα, όπου είσοδος είναι τα δεδομένα (D = {x1, …, xn}) και ο αριθμός των ομάδων (k) ενώ έξοδος οι ομάδες (Ci):

```
(1) )//ανάθεση τυχαίων κέντρων
    Για i= 1,…,k κάνε:
      θεώρησε mi ως ένα τυχαίο στοιχείο από το D;

(2) //δημιουργία ομάδων
    Για i = 1, …, k κάνε
       Ci = {x ∈ D| d(mi,x) ≤ d(mj,x) για όλα τα j = 1,
…,k, j ≠ k};

(3) // υπολογισμός νέων κέντρων
    Για i = 1,…,k κάνε
      mi = το μέσο διάνυσμα των σημείων που ανήκουν στην
    ομάδα Ci;

(4) Επανάληψη βημάτων (2) και (3) μέχρι την επίτευξη
    ομαδοποίησης
```

[20]

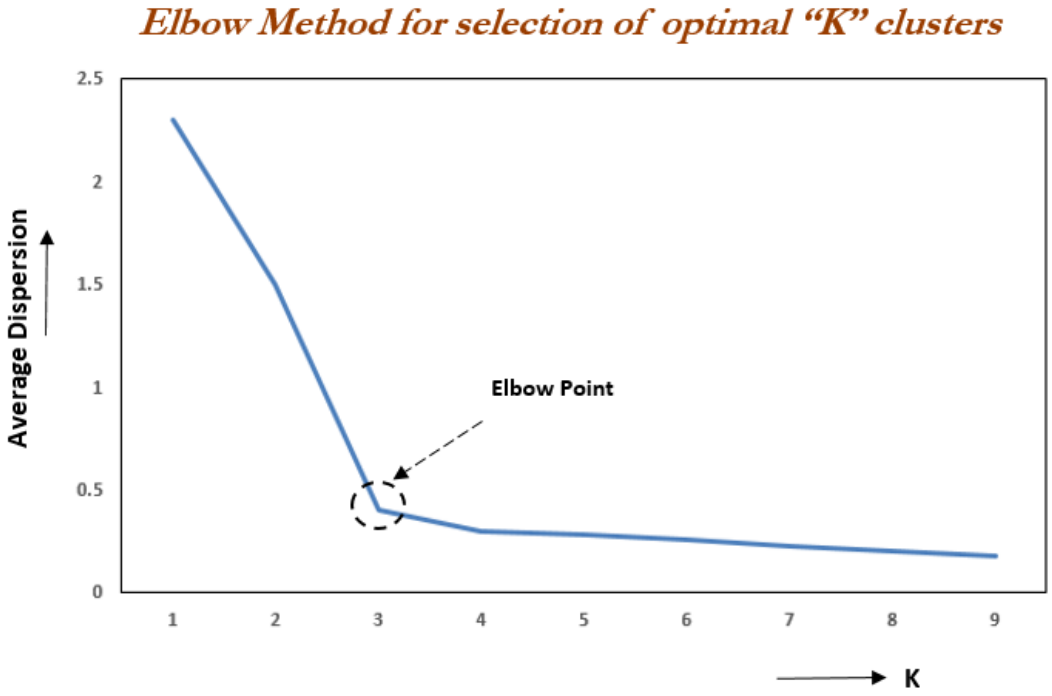
2.5.3 Προσδιορισμός του βέλτιστου αριθμού των κλάσεων

Ο προσδιορισμός του βέλτιστου αριθμού των κλάσεων σε μια διεργασία ομαδοποίησης συνιστά ένα απαραίτητο βήμα σε περιπτώσεις που χρησιμοποιείται ο αλγόριθμος k-Means ο οποίος έχει προαπαιτούμενη είσοδο τον αριθμό των κλάσεων. Ταυτόχρονα όμως, διαδραματίζει ιδιαίτερα σημαντικό ρόλο σε κάθε πρόβλημα ομαδοποίησης, καθώς καθορίζει το βαθμό λεπτομέρειας και ακρίβειας που θέτουμε ως στόχο. Έτσι, θέτει τα θεμέλια για να βρεθεί η σωστή ισορροπία για μια όσο το δυνατόν πιο ορθή συσταδοποίηση ανάλογα με τη φύση του προβλήματος ανάμεσα στις δύο ακραίες περιπτώσεις: το να θεωρηθεί κάθε παρατήρηση του dataset μια ξεχωριστή κλάση και στο να αποτελέσουν όλες οι παρατηρήσεις μία και μόνο κλάση.

Μία από τις συνηθέστερες μεθόδους που χρησιμοποιείται κατά τον προσδιορισμό του αριθμού των κλάσεων είναι η **μέθοδος του αγκώνα** (Elbow Method). Η μέθοδος του αγκώνα είναι μια γραφική μέθοδος η οποία απεικονίζει τη διακύμανση των παρατηρήσεων συναρτήσει του αριθμού των ομάδων και επιτρέπει την οπτική αναγνώριση του σημείου όπου το οριακό κέρδος από περαιτέρω ομαδοποίηση μειώνεται. Η μέθοδος αυτή στηρίζεται στην

παρατήρηση ότι στα αρχικά στάδια ομαδοποίησης, οι παρατηρήσεις τείνουν να σχηματίζουν όλο και πιο ομοιογενείς κατηγορίες με αποτέλεσμα να μειώνεται το άθροισμα των διακυμάνσεων των παρατηρήσεων εντός των ομάδων (γνωστό και ως WCSS –Within Cluster Sum of Squares). Όσο αυξάνεται ο αριθμός των ομάδων όμως, η επίδραση τους στη μείωση του WCSS μειώνεται, καθώς η κατάτμηση ενός ήδη καλά ομαδοποιημένου συνόλου δεδομένων αποφέρει μικρό κέρδος στις διακυμάνσεις τους. Έτσι, επιλέγεται ως βέλτιστος αριθμός ομάδων k εκείνος μετά τον οποίον δεν εντοπίζεται απότομη αλλαγή κλίσης της καμπύλης, όπως φαίνεται στην εικόνα που ακολουθεί.

[19]



Εικόνα 2.5: Παράδειγμα εφαρμογής της μεθόδου του αγκώνα, όπου ο βέλτιστος αριθμός κλάσεων είναι k=3 [https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml]

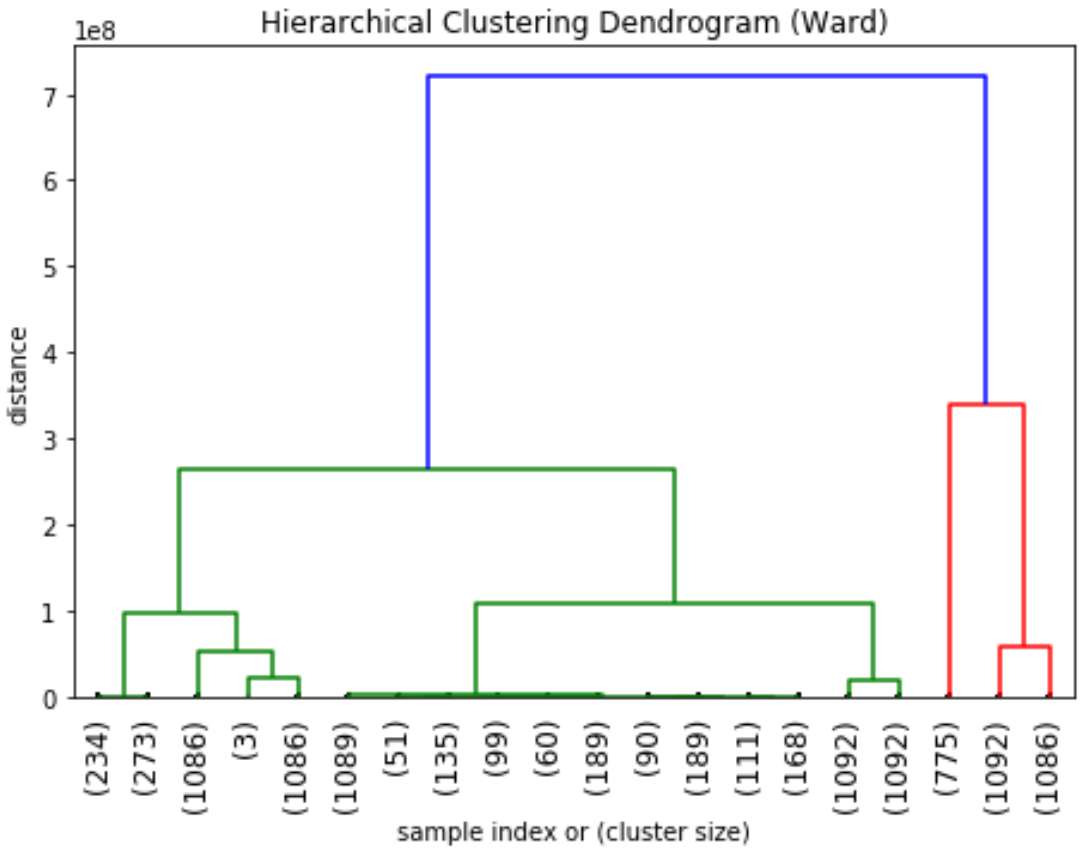
Ως εύρος για την εύρεση του k χρησιμοποιείται ο αριθμός $\sqrt{n/2}$, όπου n ο αριθμός των παρατηρήσεων εισόδου.

2.5.4 Αλγόριθμος Ιεραρχικής συγκεντρωτικής ομαδοποίησης – Hierarchical Agglomerative Clustering(HAC)

Ο αλγόριθμος Agglomerative Clustering αποτελεί μια τεχνική ιεραρχικής ομαδοποίησης με τα χαρακτηριστικά των μεθόδων αυτών όπως παρουσιάστηκαν νωρίτερα στο κεφάλαιο.

ο αλγόριθμος διατηρεί ένα «ενεργό σύνολο» κλάσεων και σε κάθε στάδιο αποφασίζει ποιες δύο κλάσεις θα πρέπει να συγχωνεύσει. Όταν συγχωνεύονται δύο κλάσεις, η κάθε μία αφαιρείται από το ενεργό σύνολο ενώ η ένωσή τους προστίθεται σε αυτό. Αυτό επαναλαμβάνεται έως ότου υπάρχει μόνο μία κλάση στο ενεργό σύνολο ενώ ως έξοδος προκύπτει το δέντρο που σχηματίστηκε παρακολουθώντας τη διαδικασία των διαδοχικών αυτών συγχωνεύσεων.

Αξιοσημείωτος τρόπος οπτικοποίησης του δέντρου αυτού αποτελεί το δενδρόγραμμα (dendrogram). Ένα δενδρόγραμμα εμφανίζει στοιχεία δεδομένων κατά μήκος ενός άξονα και αποστάσεις κατά μήκος του άλλου άξονα. Ένα δενδρόγραμμα δείχνει μια συλλογή μονοπατιών σε σχήμα δέντρου, όπου στις διακλαδώσεις εμφανίζονται οι ομάδες που έχουν ενωθεί. Αυτές οι ομάδες μπορεί να είναι η βάση ενός άλλου δέντρου ή μπορεί να είναι ομάδες που αντιπροσωπεύονται ως δεδομένα κατά μήκος του άξονα. Μια βασική ιδιότητα του δενδρογράμματος είναι ότι η βάση του δέντρου βρίσκεται κατά μήκος του άξονα των αποστάσεων αναλογικά με την απόσταση μεταξύ των ομάδων που συγχωνεύονται. Για να οδηγήσει σε μια λογική ομαδοποίηση - και ένα έγκυρο δενδρόγραμμα - αυτές οι αποστάσεις πρέπει να αυξάνονται μονοτονικά. Δηλαδή, η απόσταση μεταξύ δύο συγχωνευμένων ομάδων πρέπει πάντα να είναι μεγαλύτερη ή ίση με την απόσταση μεταξύ οποιονδήποτε υπο-ομάδων που συγχωνεύτηκαν σε πρότερη φάση.



Εικόνα 2.6: Παράδειγμα δενδρογράμματος που δημιουργήθηκε μέσω του scikit-learn.

Ακολουθεί ο αλγόριθμος HAC σε μορφή ψευδοκώδικα, όπου είσοδος είναι τα δεδομένα (D = {x1, …, xn}) ενώ έξοδος το δενδρόγραμμα T.

```
(1) //αρχικοποίηση του ενεργού συνόλου
    A ← ∅

(2) //δημιουργία ομάδων με κάθε παρατήρηση ως ξεχωριστή
    κλάση αρχικά
    Για i = 1, …, N κάνε
      A ← A ∪ {{xn}}

(3) T ← A
```

```
(4) Όσο |A|>1 κάνε        // επανέλαβε μέχρι το ενεργό σύνολο
    να αποτελείται από ένα μόνο στοιχείο

       G1′ , G2′ ← arg min dist G1,G2∈A ; // Επίλεξε ζεύγος
    στο A με την καλύτερη απόσταση.


       A ← (A\{G1′ })\{G2′ } // Αφαίρεσε τα από το ενεργό
    σύνολο

       A ← A ∪ {G1′ ∪ G2′ } // Πρόσθεσε την ένωσή τους στο
    ενεργό σύνολο

       T ← T ∪ {G1′ ∪ G2′ } // και στο δένδρόγραμμα



(5) Επιστροφή T
```

[21]

2.5.5 Αλγόριθμος t-SNE

Ο αλγόριθμος t-Distributed Stochastic Neighbor Embedding (t-SNE) χρησιμοποιείται για απεικόνιση πολυδιάστατων δεδομένων σε χώρο χαμηλών διαστάσεων (συνήθως δύο ή τριών). Ανήκει στις τεχνικές μείωσης διαστάσεων της μηχανικής μάθησης αξιοποιώντας μαθηματικές μεθόδους για να ανακαλύψει το χώρο χαμηλών διαστάσεων που υπάρχει ενσωματωμένος στις πολυδιάστατες εισόδους του αλγορίθμου με πολύπλοκους μη γραμμικούς τρόπους, ενώ αναπαριστά (είτε σε δύο ή σε τρεις διαστάσεις) τα σημεία υψηλής ομοιότητας σε κοντινή μεταξύ τους απόσταση και τα σημεία χαμηλής ομοιότητας σε πιο μακρινή απόσταση.

Ο αλγόριθμος t-SNE εκτελείται σε δύο στάδια και κατασκευάζει μια κατανομή πιθανότητας στον υψηλό χώρο διαστάσεων αρχικά και στο χαμηλό χώρο διαστάσεων στη συνέχεια. Στο πρώτο στάδιο, κατασκευάζεται μια κατανομή πιθανότητας σε ζεύγη αντικειμένων υψηλών διαστάσεων με τέτοιο τρόπο ώστε τα παρόμοια αντικείμενα να έχουν

υψηλότερη πιθανότητα, ενώ τα διαφορετικά σημεία να έχουν χαμηλότερη πιθανότητα. Έπειτα, ορίζεται μια παρόμοια κατανομή πιθανότητας πάνω από τα σημεία του χάρτη χαμηλών διαστάσεων, η οποία ελαχιστοποιεί την απόκλιση Kullback-Leibler (απόκλιση KL) μεταξύ των δύο κατανομών σε σχέση με τις θέσεις των σημείων στο χάρτη. Η επιλογή των τιμών των παραμέτρων αποτελεί βασικό πρόβλημα κατά την εκτέλεση του t-SNE, καθώς κατά η ομαδοποίηση σημείων στα γραφήματα του αλγορίθμου μπορεί να είναι παραπλανητική δηλαδή οι οπτικές συστάδες που δημιουργούνται να αντιστοιχούν σε ομαδοποίηση που δεν υπάρχει στην πραγματικότητα. Επομένως είναι απαραίτητη η προσεκτική επιλογή παραμέτρων και η επικύρωση των αποτελεσμάτων. Ιδιαίτερα σημαντική είναι η επιλογή της παραμέτρου της πολυπλοκότητας(perplexity) με προτεινόμενο εύρος τιμών από 5 έως 50 για πιο αξιόπιστα αποτελέσματα.

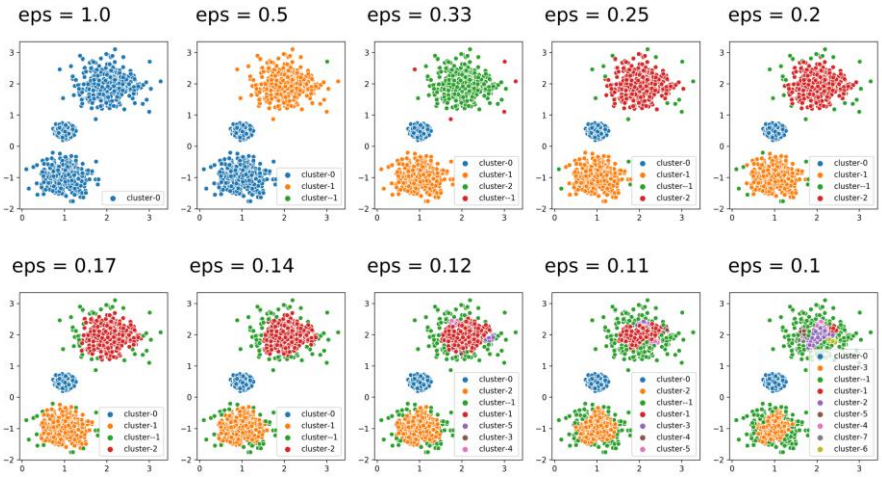
Ακόμη ένας παράγοντας που πρέπει να ληφθεί υπόψη κατά την εκτέλεση του t-SNE είναι το μέγεθος του διανύσματος εισόδου, καθώς η χρονική και υπολογιστική πολυπλοκότητα του αλγορίθμου είναι O(η^2), με αποτέλεσμα να καθίσταται δυσχερής ή και αδύνατη η χρήση του σε σύνολα δεδομένων με περισσότερα από 10000 σημεία. Στην παρούσα διπλωματική εργασία χρησιμοποιείται η παραλλαγή Barnes-Hut του αλγορίθμου t-SNE, η οποία παρουσιάζει αισθητή βελτίωση στο χρόνο εκτέλεσης [πολυπλοκότητα O($n \bullet \log n$)] , επιτρέποντας εισόδους με εκατομμύρια σημεία.

[22] [23][24]

2.5.6 Αλγόριθμος DBSCAN (Density Based Spatial Clustering of Applications with Noise)

Ο αλγόριθμος DBSCAN αποτελεί αντιπροσωπευτικό παράδειγμα μεθόδου ομαδοποίησης βάσει πυκνότητας, καθώς δημιουργεί συστάδες γειτονικών σημείων που βρίσκονται σε κοντινή απόσταση μεταξύ τους ενώ αντιλαμβάνεται τα υπόλοιπα σημεία ως θόρυβο ή απόκλιση. Για να δημιουργηθούν οι ομάδες χρησιμοποιείται η έννοια της πυκνότητας ως εξής: ορίζονται κάποια κεντρικά σημεία p τα οποία είναι απαραίτητο να περιβάλλονται από έναν ελάχιστο αριθμό σημείων(γνωστός και ως η παράμετρος minPts του αλγορίθμου) που απέχουν από το κάθε p απόσταση μικρότερη από την παράμετρο eps του αλγορίθμου. Οι ομάδες σχηματίζονται από όλα τα σημεία που είναι άμεσα προσβάσιμα από κεντρικά σημεία ή έμμεσα προσβάσιμα από ένα μονοπάτι που διέρχεται από αυτά, ενώ δεν ανήκουν στην ομάδα σημεία που είναι προσβάσιμα μέσω μη κεντρικών μονοπατιών(μονοπατιών δηλαδή που σχηματίζονται από μη κεντρικά σημεία).

Ο τρόπος λειτουργίας του αλγορίθμου οδηγεί στην αναγνώριση ομάδων που μπορεί να περιβάλλονται από άλλες ομάδες χωρίς να συνδέονται μεταξύ τους, καθώς και στην καλή απόδοση του ακόμη και σε εισόδους με πολύ θόρυβο. Παράλληλα όμως είναι απαραίτητη και σε αυτή την περίπτωση μια προσεκτική επιλογή των δύο παραμέτρων που προυποθέτει καλή κατανόηση των δεδομένων εισόδου. [25]



Εικόνα 2.7: Διαφοροποίηση στα αποτελέσματα της ομαδοποίησης με χρήση του αλγορίθμου DBSCAN με μεταβολή της παραμέτρου eps.[https://towardsdatascience.com/how-to-use-dbscan-effectively-ed212c02e62]

3 Βάση Δεδομένων

3.1 Εισαγωγή

Στο παρόν κεφάλαιο περιγράφουμε τη βάση δεδομένων που δημιουργήθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας με κάποια τεχνητά δεδομένα για τη μοντελοποίηση των ιατρικών προφίλ χρηστών .Τα δεδομένα αυτά πληρούν συγκεκριμένες προϋποθέσεις στηρίζονται σε διαδεδομένα ιατρικά πρωτόκολλα και μεθόδους και φέρουν χαρακτηριστικά που αναλύονται στη συνέχεια του κεφαλαίου.

Η αρχική μοντελοποίηση της βάσης δεδομένων έγινε με γνώμονα την ένταξη ενός όσο το δυνατόν πιο πλήρους συνόλου δεδομένων με βασικές ποιοτικές και ποσοτικές παραμέτρους ώστε να μπορούν να επιλεχθούν οι κατάλληλες και να εκτελεστούν με ρεαλιστικές χρονικές πολυπλοκότητες οι αλγόριθμοι μηχανικής μάθησης για ομαδοποίηση των χρηστών. Ταυτόχρονα τίθενται τα θεμέλια για μελλοντική επέκταση της βάσης δεδομένων σε ένα όσο το δυνατόν πιο πλήρες σύνολο από ιατρικά προφίλ.

Η αρχική μας βάση περιλαμβάνει 5.000 δημιουργημένα προφίλ χρηστών με περίπου 35.000 συνολικές καταχωρήσεις ιατρικού ιστορικού και 25 εκατομμύρια μετρήσεις ιατρικών μεγεθών .Τα μεγέθη που υπάρχουν στη βάση δεδομένων αφορούν πιο εξειδικευμένα ιατρικά μεγέθη και διαγνώσεις (πχ. πίεση, κορεσμός οξυγόνου, ιστορικό διαβήτη, κατάθλιψης κ.α.) αλλά και πληθώρα πληροφοριών όπως το βάρος, το ύψος, η μυϊκή μάζα, η χρήση αλκόολ , ο εθισμός στη νικοτίνη, το επίπεδο φυσικής άσκησης κ.α., οι οποίες μπορούν να τροφοδοτούνται στο ιατρικό προφίλ ενός χρήστη μέσω εφαρμογών self-monitoring.

3.2 Το σχήμα της βάσης δεδομένων (database schema)

Η αποθήκευση των δεδομένων που παρήχθησαν έγινε σε μία σχεσιακή βάση δεδομένων της mysql που περιλαμβάνει τις εξής κύριες οντότητες:

3.2.1 Άτομο (Person)

Στον πίνακα της οντότητας αυτής καταγράφονται οι βασικές προσωπικές πληροφορίες του κάθε ιατρικού προφίλ. τα πεδία που συμπεριλαμβάνονται είναι:

Person ID: αποτελεί το πρωτεύον κλειδί του πίνακα ως ένας μοναδικός αριθμός για κάθε προφίλ χρήστη. Χρησιμοποιείται με σχέση "1-προς-πολλά" για τη σύνδεση με τους υπόλοιπους πίνακες της βάσης δεδομένων

Identifier: ένα ακόμη αναγνωριστικό πεδίο με μοναδική τιμή για κάθε προφίλ χρήστη

Ημερομηνία γέννησης (Date of Birth): Χρησιμοποιείται σε μορφή ακριβείας με δυνατότητα καταχώρησης του έτους, μήνα, ημέρας, ώρας, λεπτού και δευτερολέπτου γέννησης. Χρησιμεύει ιδιαίτερα καθώς εξάγεται η ηλικία του κάθε ατόμου που αντιπροσωπεύεται στο προφίλ και χρησιμοποιείται ως μια από τις παραμέτρους που χρησιμοποιούνται κατά την εκτέλεση των αλγορίθμων ομαδοποίησης.

Gender: το φύλο του χρήστη του ιατρικού προφίλ.

Educational Level (Μορφωτικό Επίπεδο): Λαμβάνει μία εκ των τιμών: βασική εκπαίδευση, δευτεροβάθμια εκπαίδευση και ανώτερη εκπαίδευση.

Creation Datetime: ο χρόνος δημιουργίας της καταχώρησης.

Modification Datetime: ο χρόνος τελευταίας μεταβολής της καταχώρησης.

IsActive: εάν πρόκειται για ενεργό προφίλ ή ανενεργό.

Revision: αριθμός που καταμετρά τις συνολικές φορές επεξεργασίας και αλλαγής στην καταχώρηση.

Για τη διαδικασία της ομαδοποίησης εισάγουμε ως παραμέτρους (feautures) ομαδοποίησης τα πεδία DateOfBirth, Gender, Educational Level, καθώς και το πεδίο PersonId ως αναγνωριστικό του προφίλ.

3.2.2 Medical History (Ιατρικό Ιστορικό)

Η συγκεκριμένη οντότητα περιλαμβάνει καταχωρήσεις ιατρικού ιστορικού που αντιστοιχούν στους χρήστες της οντότητας Person. Η σύνδεση με την οντότητα Person γίνεται με το foreign key PersonId όπως προαναφέραμε, με μία σχέση "ένα προς πολλά"

καθώς έχουμε πολυάριθμες καταχωρήσεις ιστορικού για κάθε χρήστη. Τα υπόλοιπα πεδία του πίνακα είναι τα εξής:

MedicalHistoryId: Primary Key του πίνακα και το Unique Id της κάθε καταχώρησης ιστορικού.

IcdCode: Ο κωδικός της νόσου ή διάγνωσης που καταγράφεται με βάση την κωδικοποίηση ICD-10 (η οποία περιγράφεται στη συνέχεια της υπο-ενότητας)

Details: Προαιρετικό πεδίο που περιλαμβάνει περαιτέρω λεπτομέρειες/διευκρινήσεις που μπορεί να απαιτηθούν για πληρότητα της καταχώρησης.

Creation Datetime: ο χρόνος δημιουργίας της καταχώρησης.

Modification Datetime: ο χρόνος τελευταίας μεταβολής της καταχώρησης.

IsActive: εάν πρόκειται για ενεργή ή ανενεργή καταχώρηση.

Revision: αριθμός που καταμετρά τις συνολικές φορές επεξεργασίας και αλλαγής στην καταχώρηση.

3.2.2.1 Διεθνής Στατιστική Ταξινόμηση Νόσων και Συναφών Προβλημάτων Υγείας - 10η έκδοση (ICD-10)

Η Διεθνής Στατιστική Ταξινόμηση Νόσων και Συναφών Προβλημάτων Υγείας (ICD-10 - International Classification of Diseases and Related Health Problems 10th Revision) αποτελεί ένα σύστημα κατηγοριών στις οποίες είναι καταχωρημένες οι παθολογικές οντότητες σύμφωνα με προκαθορισμένα κριτήρια, το οποίο έχει θεμελιωθεί από τον Παγκόσμιο Οργανισμό Υγείας. Έχει γίνει το διεθνές πρότυπο ταξινόμησης διαγνώσεων για όλους τους γενικούς επιδημιολογικούς σκοπούς και για πολλούς σκοπούς που αφορούν στη διαχείριση της υγείας. Αυτοί περιλαμβάνουν την ανάλυση της γενικής κατάστασης της υγείας των πληθυσμιακών ομάδων και την παρακολούθηση της επίπτωσης και του επιπλασμού των νόσων και άλλων προβλημάτων υγείας σε σχέση με άλλες παραμέτρους, όπως τα χαρακτηριστικά και οι συνθήκες των προσβληθέντων ατόμων. Το πεδίο της έχει διευρύνθει για να περιλάβει διαγνώσεις νοσηρότητας και μπορεί να χρησιμοποιηθεί για την ταξινόμηση δεδομένων που έχουν καταχωρηθεί κάτω από τίτλους όπως "διάγνωση", "αιτία εισαγωγής", "καταστάσεις που αντιμετωπίστηκαν" και "αιτία επίσκεψης", οι οποίοι εμφανίζονται σε ένα ευρύ φάσμα αρχείων υγείας. Η κωδικοποίηση ICD χρησιμοποιείται για να μετατρέψει τις

διαγνώσεις νόσων και άλλων προβλημάτων υγείας από λέξεις σε έναν αλφαριθμητικό κωδικό που επιτρέπει την εύκολη αποθήκευση, ανάκτηση και ανάλυση των δεδομένων.

Οι ταξινομήσεις αναφοράς του Π.Ο.Υ., συμπεριλαμβανομένης της Διεθνής Στατιστική Ταξινόμηση Νόσων και Συναφών Προβλημάτων Υγείας , είναι αποτέλεσμα διεθνών συμφωνιών, ενώ χαίρουν ευρείας αποδοχής και επίσημης συμφωνίας για χρήση. Επίσης, έχουν εγκριθεί και συστήνονται ως κατευθυντήριες οδηγίες για διεθνείς αναφορές και εκθέσεις υγείας. Μπορούν να χρησιμοποιηθούν ως πρότυπα για την ανάπτυξη ή την αναθεώρηση άλλων ταξινομήσεων όσον αφορά στη δομή αλλά και το χαρακτήρα και τον καθορισμό των κατηγοριών.

Η ICD-10 χωρίζεται σε 21 κατηγορίες, καθεμία από τις οποίες αποτελεί ένα κεφάλαιο στην ταξινόμηση και την κωδικοποίηση και δεν υπάρχει ισοκατανομή του αριθμού των νοσημάτων σε κάθε κεφάλαιο. Σε δεύτερο επίπεδο κάθε κεφάλαιο αποτελείται από ομάδες νοσημάτων, οι οποίες με τη σειρά τους καταλαμβάνουν κάποιους από τους κωδικούς του κεφαλαίου. Οι κωδικοί αυτοί ξεκινούν από το πρώτο κεφάλαιο με τους λατινικούς χαρακτήρες Α ή Β, ενώ ακολουθούν 2 ακέραιοι αριθμοί που αφορούν σε επίπεδο αιτιολογίας μεγαλύτερης λεπτομέρειας, και καθορίζουν την ομάδα και συγκεκριμένα το νόσημα. Στο πλαίσιο της παρούσας διπλωματικής χρησιμοποιούνται κωδικοί κεφαλαίων του πρώτου και δευτέρου επιπέδου.

Αναλυτικότερα, το πρότυπο ICD-10 ακολουθεί την εξής κωδικοποίηση: το όνομα της κάθε ασθένειας, όπως αυτή ορίζεται από τον Παγκόσμιο Οργανισμό Υγείας στη Διεθνή Ονοματολογία Νοσημάτων (International Nomenclature of Diseases - IND), αντιστοιχεί και σε ένα τριψήφιο αλφαριθμητικό, το οποίο αποτελεί τον κωδικό της ασθένειας. Η IND ονοματολογία παρέχει ένα μοναδικό, απλό και όσο το δυνατό περιγραφικό όνομα για κάθε νόσο. Ενώ ο κωδικός είναι τριψήφιος και αποτελείται από ένα γράμμα του λατινικού αλφαβήτου (εκτός από το γράμμα U) και ένα διψήφιο αριθμό. Δηλαδή, η έκταση των αριθμών είναι από το A00 έως το Z99 και αν υπολογίσουμε την εξαίρεση αποτελούν 2500 διαφορετικούς κωδικούς. Οι κωδικοί που ξεκινούν από U δεν χρησιμοποιούνται από την επίσημη ταξινόμηση, αλλά οι U00-U49 χρησιμοποιούνται στην προσωρινή κωδικοποίηση νέων ασθενειών και οι U50-U99 για την έρευνα. Στη συνέχεια, κωδικοποιούνται οι υποδιαιρέσεις των ασθενειών και προστίθεται και τέταρτος αριθμός μετά από μία τελεία. Για παράδειγμα, το E00 είναι Συγγενή σύνδρομα έλλειψης ιωδίου και το E00.0 Συγγενή σύνδρομο έλλειψης ιωδίου, νευρολογικός τύπος. Οι κωδικοί M00-M99, S00-T88 και V01-Y98 που αποτελούν τα κεφάλαια 13, 19 και 20 αντίστοιχα έχουν πολλές φορές πέμπτο, έκτο και έβδομο χαρακτήρα, προσθέτοντας περισσότερες πληροφορίες σε μία νόσο. Με αυτό τον τρόπο έχουμε τα εργαλεία εκείνα που απαιτούνται για να καλυφθεί όλο το φάσμα των ασθενειών που γνωρίζουμε και να κωδικοποιηθεί.

Σε άλλες περιπτώσεις χρησιμοποιείται μετά των κωδικών των τριών ή των τεσσάρων γραμμάτων ένας σταυρός (+) ή ένας αστερίσκος (*). Ο σταυρός δηλώνει ότι ο κωδικός αναφέρεται στον βασικό ορισμό μιας ασθένειας και ο αστερίσκος δηλώνει ότι αναφέρεται στην περίπτωση που αυτή η ασθένεια προσβάλει ένα συγκεκριμένο όργανο. Αυτού του είδους η διάκριση είναι απαραίτητη για τη σωστή συγκέντρωση στατιστικών στοιχείων και δεν είναι ιδιαίτερα διαδεδομένη στο πρότυπο, παρά μόνο σε πολύ συγκεκριμένους κωδικούς. Τέτοιο παράδειγμα είναι ο κωδικός Α18.1+ φυματίωση του ουρογεννητικού συστήματος και ο κωδικός Ν33.0* φυματίωση της ουροδόχου κύστης. Μερικοί κωδικοί ή ασθένειες, επίσης, εμφανίζονται μέσα σε παρενθέσεις. Αυτό μπορεί να συμβαίνει γιατί:

- η περιγραφή είναι προαιρετική, όπως ο κωδικός Ι10 που αναφέρεται ως υπέρταση (αρτηριακή) (καλοήθης) (ιδιοπαθής) (κακοήθης) (πρωτοπαθής) (συστηματική), οπότε Ι10 είναι μόνο η υπέρταση ή η υπέρταση σε συνδυασμό με οποιαδήποτε από τις μορφές των παρενθέσεων,
- για να εξαιρέσουμε έναν κωδικό, όπως Η01.0 Βλεφαρίτιδα αλλά εξαιρούνται βλεφαροεπιπεφυκίτιδα (Η10.5),
- στους τίτλους των ενοτήτων αναφερόμαστε στους κωδικούς της ενότητας ή
- όταν αναφέρεται ένας κωδικός με αστερίσκο και σε παρένθεση μπορεί να•εμφανίζεται ο κωδικός με σταυρό που σχετίζεται με αυτόν.

Σε άλλους κωδικούς έχουμε την εμφάνιση αγκυλών που αυτό μπορεί να σημαίνει: συνώνυμο ή επεξήγηση, για παράδειγμα Α30 Λέπρα [Νόσος του Hansen], αναφορά σε σημειώσεις, για παράδειγμα C00.8 Αλληλεπιβαίνουσα αλλοίωση του χείλους ή αναφορά σε κάποια υποδιαίρεση, όπως Κ27 Πεπτικό έλκος, μη καθορισμένη εντόπιση. Τέλος, η αναφορά του «και» σε τίτλους έχει την έννοια του «ή/και», ενώ σε κάποιους κωδικούς μπορεί να εμφανίζεται μία παύλα (-), η οποία μας δείχνει ότι στην κωδικοποίηση τριών χαρακτήρων που κάνουμε μπορεί να υπάρχει και τέταρτος χαρακτήρας σε μια άλλη κατηγορία που καλό είναι να αναζητηθεί, έτσι έχουμε για παράδειγμα τον κωδικό G03 Μηνιγγίτιδα οφειλόμενη σε άλλες και μη καθορισμένες αιτίες, εξαιρούνται: μηνιγγοεγκεφαλίτιδα (G04.-).

Παρόλη την έκταση του ICD-10 δεν μπορεί ακόμα να καλύψει όλες τις ανάγκες και να παρέχει λεπτομέρειες για όλες τις ειδικότητες. Για το γεγονός αυτό ο ΠΟΥ έχει δημιουργήσει μια οικογένεια ταξινομήσεων, δηλαδή μία ομάδα προτύπων ταξινόμησης με κοινά χαρακτηριστικά, που καλύπτουν διαφορετικές πτυχές του χώρου της υγείας και μπορούν να χρησιμοποιούνται είτε μαζί, είτε ως μεμονωμένα πρότυπα. Στον πυρήνα αυτής της οικογένειας είναι το πρότυπο ICD και έχει δημιουργηθεί το δίκτυο Family of

International Classifications, οπότε έχουμε να κάνουμε με την οικογένεια προτύπων WHO-FIC.

Γενικότερα, υπήρξε ένας μακρύς δρόμος εξέλιξης του προτύπου, το οποίο εξελίσσεται και προσαρμόζεται συνεχώς. Απαρχή του προτύπου μπορεί να θεωρηθεί ότι έγινε από το 1770 και τον François Bossier de Lacroix, που με το έργο του Nosologia methodica37 προσπάθησε να ταξινομήση τις βασικότερες ασθένειες. Άλλα παρόμοια έργα ακολούθησαν για να φθάσουμε σήμερα στην 10η αναθεώρησή του διεθνώς αναγνωρισμένου προτύπου ταξινόμησης ICD. Ακολουθώντας την ίδια την ύπαρξη του προτύπου και τις διαφορετικές του εκδόσεις θα λέγαμε ότι αυτές ήταν: ICD, ICD-2, ICD-3, ICD-4 και ICD-5, υπήρξε ταξινόμηση των αιτιών θανάτου, όπου ανά δεκαετία γινόταν αναθεώρηση, οπότε διαφορετικές εκδόσεις του προτύπου εκδόθηκαν τα έτη 1900, 1910, 1921, 1930 και 1939. Το ICD-6 εφαρμόστηκε στις ΗΠΑ , για πρώτη φορά το 1948 και έγινε προσπάθεια ένα πρότυπο ταξινόμησης να αποκτήσει συγκεκριμένη μορφή για τη διεθνή κοινότητα. Υπήρχε ένα κεφάλαιο για τραυματισμούς και άλλο ένα για εξωτερικές αιτίες, ενώ για πρώτη φορά ένα πρότυπο είχε αναφορά και στις ψυχικές ασθένειες. Το ICD-7, εκδόθηκε το 1955 και διόρθωσε όλα τα σφάλματα του προηγούμενου προτύπου.Το ICD-8 εκδόθηκε το 1965 και περιείχε ακόμα περισσότερες διορθώσεις αν και η γενικότερη φιλοσοφία του προτύπου παρέμεινε αναλλοίωτη. Από αυτή την έκδοση και μετά το ICD έγινε δεκτό από τα περισσότερα νοσοκομεία και υπηρεσίες υγείας ως ένα κοινά αποδεκτό πρότυπο ταξινόμησης. Στις ΗΠΑ συγκροτήθηκε ειδική επιτροπή μελέτης του προτύπου, η οποία χρησιμοποίησε την παρούσα έκδοση για τη συλλογή στατιστικών στοιχείων θνησιμότητας και νοσηρότητας, τα ενσωμάτωσε στο πρότυπο και καταλήξαμε το 1968 να δημοσιευθεί η αμερικάνικη έκδοση του ICD-8 η ICD-8a, που υπήρξε η βάση για την επόμενη έκδοση του προτύπου. Το ICD-9, εκδόθηκε το 1975 και ήταν η πρώτη φορά που συμφωνήθηκε το πρότυπο να μην διαφέρει αισθητά από την προηγούμενη έκδοσή του, ώστε να μην χρειάζεται η παγκόσμια κοινότητα να προσαρμόζεται συνεχώς σε νέα δεδομένα, αλλά να αποτελεί την λογική εξέλιξη της προηγούμενης έκδοσης. Σε αυτά τα πλαίσια, διατηρήθηκε η αρχική ταξινόμηση, αλλά προστέθηκαν περισσότερα ψηφία για να προσδώσουν και περισσότερες λεπτομέρειες. Το ICPM, το International Classification of Procedures in Medicine εκδόθηκε το 1978, στηρίχθηκε στο ICD-9 και προσπάθησε να γίνει συμπληρωματικό εργαλείο προς αυτό, περιλαμβάνοντας ταξινόμηση των εργαστηριακών (ακτινολογικών, εργαστηριακών και χειρουργικών) μεθόδων. Το ICD-9-CM, International Classification of Diseases, Clinical Modification, βασίστηκε στο ICD-9 και χρησιμοποιήθηκε από στατιστικά κέντρα των ΗΠΑ στην κωδικοποίηση διαγνώσεων ιατρικών κέντρων και εξωτερικών ιατρών. Το πρότυπο υποστηρίζεται ακόμα και σήμερα από το National Center for Health Statistics38 (NCHS), βασικό κέντρο συλλογής στατιστικών στοιχείων της Αμερικής. Το ICD-10, εκδόθηκε το

1990, αναθεωρήθηκε το 1994 και αποτελεί την κυρίαρχη ταξινόμηση σήμερα. Οι 17.000 κωδικοί του ICD-9 έγιναν πάνω από 155.000 και υιοθετήθηκε από τις περισσότερες υπηρεσίες υγείας παγκοσμίως.Το ICD-10-CM, αποτελεί προσαρμογή του NCHS στην έκδοση ICD-10. Το ICD-10-CA, αποτελεί την Καναδική έκδοση του ICD-10 και περιλαμβάνει παράγοντες και καταστάσεις πέρα από τις ασθένειες που επηρεάζουν την υγεία του ατόμου, όπως το εργαστηριακό στρες και οι κοινωνικές συνθήκες. Το ICD-11, θα αποτελέσει το νέο πρότυπο ταξινόμησης, δεν έχει ακόμα εκδοθεί, ενώ οι συζητήσεις για την έκδοσή του γίνονται βάση μιας διαδικτυακής πλατφόρμα με την ονομασία iCAD39 .
[26][27]

3.2.3 Measurement (Μέτρηση)

Η οντότητα περιλαμβάνει έναν ικανό αριθμό μετρήσεων ορισμένων μεγεθών που αφορούν τους χρήστες όπως βάρος, πίεση, θερμοκρασία, κ.α. που μπορούν να χρησιμοποιηθούν σε ιατρικό διαγνωστικό πλαίσιο. Η σύνδεση με την οντότητα Person γίνεται με το foreign key PersonId όπως προαναφέραμε, με μία σχέση "ένα προς πολλά" καθώς έχουμε πολυάριθμες καταχωρήσεις μετρήσεων για κάθε χρήστη. Τα υπόλοιπα πεδία του πίνακα είναι τα εξής:

MeasurementId: Αύξοντας αριθμός που αποτελεί το primary Key του πίνακα και το Unique Id της κάθε καταχώρησης..

SnomedId: Ο κωδικός του κλινικού όρου που καταγράφεται με βάση το πρότυπο SNOMED CT (το οποίο περιγράφεται στη συνέχεια της υπο-ενότητας)

Value: Η τιμή που λαμβάνει η μέτρηση με το συγκεκριμένο SnomedId.

Unit: Η μονάδα μέτρησης του μεγέθους που εξετάζεται.

Creation Datetime: ο χρόνος δημιουργίας της καταχώρησης.

Modification Datetime: ο χρόνος τελευταίας μεταβολής της καταχώρησης.

IsActive: εάν πρόκειται για ενεργή ή ανενεργή καταχώρηση.

Revision: αριθμός που καταμετρά τις συνολικές φορές επεξεργασίας και αλλαγής στην καταχώρηση.

3.2.3.1 Το Πρότυπο κωδικοποίησης SNOMED (Systematized Nomenclature of Medicine)

Το πρότυπο κωδικοποίησης SNOMED αποτελεί μια οργανωμένη συγκέντρωση ονοματολογίας ιατρικών όρων, που χρησιμοποιείται για τις κλινικές αναφορές και την τεκμηρίωση. Σήμερα έχει επικρατήσει η έκδοση SNOMED CT (SNOMED Clinical Terms), η οποία βασίζεται σε τέσσερα στοιχεία:

- έννοιες: είναι κωδικοί που χαρακτηρίζουν μοναδικά έναν κλινικό όρο,
- περιγραφές: είναι απλό κείμενο που περιγράφει έναν κλινικό όρο,
- σχέσεις: είναι σχέσεις μεταξύ των εννοιών που συσχετίζονται και
- αναφορές: που ομαδοποιούν τις έννοιες σε ομάδες.

Οι έννοιες χωρίζονται σε κάποιους βασικούς άξονες και κωδικοποιούνται με αριθμητικούς χαρακτήρες ώστε μεμονωμένοι ή συνδυαστικοί κωδικοί να παρέχουν πληροφορία για το ιατρικο συμβάν ή μέγεθος στο οποίο αναφέρονται. Οι έννοιες χωρίζονται σε 15 βασικούς άξονες ως εξής:

- Διαδικασία / παρέμβαση: οι διαδικασίες κατά την παροχή φροντίδας υγείας
- Ευρήματα / διαταραχές Μετρήσιμες / παρατηρήσιμες οντότητες: παρατηρήσιμες λειτουργίες όπως η όραση και επίσης μετρήσιμες ποσότητες όπως επίπεδο σακχάρου
- Κοινωνικές / διοικητικές έννοιες: Δομή σώματος ανατομικές έννοιες και μορφολογικές ανωμαλίες
- Οργανισμοί : όλους τους οργανισμούς συμπεριλαμβανομένων των μικροοργανισμών και των φορέων μολύνσεων
- Ουσίες: Φάρμακα
- Φυσικά αντικείμενα: υλικά, αντικείμενα κατασκευασμένα από τον άνθρωπο
- Φυσικές δυνάμεις: οι φυσικές δυνάμεις ως αιτία τραυματισμού
- Γεγονότα: γεγονότα που οδήγησαν σε τραυματισμό
- Περιβάλλον / γεωγραφική τοποθεσία: τύποι περιβάλλοντος και γεωγραφικές τοποθεσίες
- Δείγματα: τμήματα του σώματος τα οποία έχουν ληφθεί για εξέταση
- Βασισμένη στο περιβάλλον κατηγορία: έννοιες οι οποίες αλλάζουν το νόημα των πραγμάτων με τα οποία συσχετίζονται
- Χαρακτηριστικά: τα οποία προσθέτουν περαιτέρω πληροφορία σε μια έννοια

- Προσδιοριστές: περιλαμβάνει εναπομένουσες έννοιες.

Οι σχέσεις μπορεί να είναι δύο ειδών:

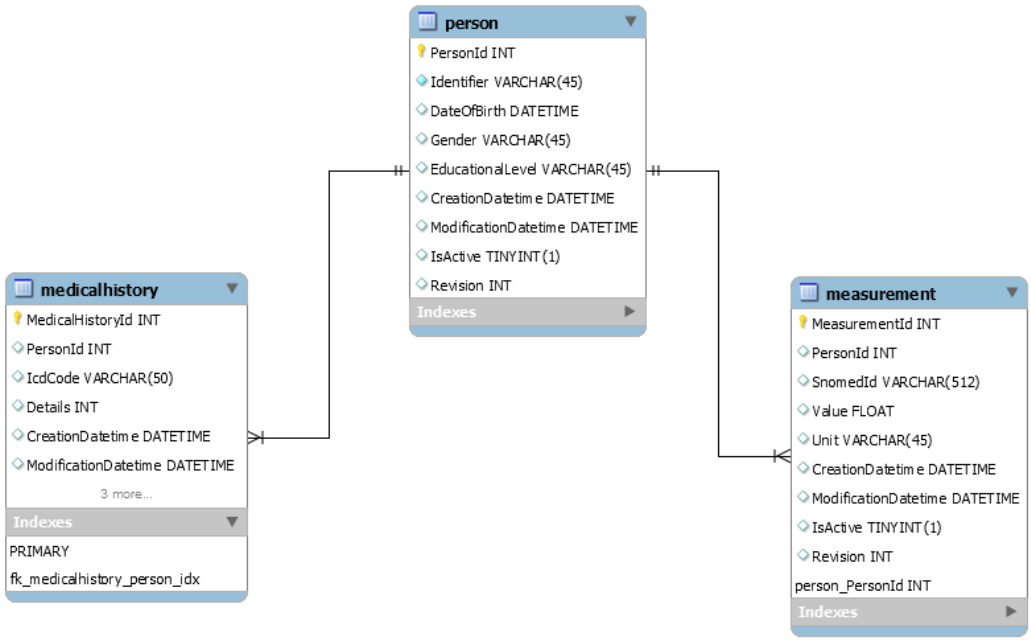
- Ιεραρχικές τύπου IS A, όπου μία έννοια συσχετίζεται με μία γενικότερη, όπως η αρθρίτιδα IS A αρθροπάθεια ή το κοινό κρυολόγημα IS A ιός.
- Μη ιεραρχικές, όπου δύο έννοιες σχετίζονται μέσω κάποιου κοινού χαρακτηριστικού τους, όπως σκωληκοειδίτιδα με σχετιζόμενη μορφολογία τη φλεγμονή.

Τέλος, τα σύνολα αναφορών μπορεί να αποτελούνται από κάποια σύνολα εννοιών ή σύνολα εννοιών με τις σχέσεις τους και τα οποία όλα μαζί περιγράφουν μονοσήμαντα μία κατάσταση.

Το πρότυπο SNOMED CT είναι το πλέον κατάλληλο για ανταλλαγή δεδομένων μεταξύ αυτοματοποιημένων συστημάτων καθώς δίνει περισσότερες λεπτομέρειες, οπότε είναι καταλληλότερο για τη δημιουργία αναφορών, περιέχει σε αρκετά κομμάτια του φυσική γλώσσα, άρα υπό προϋποθέσεις μπορεί να είναι πιο φιλικό και το κυριότερο κάθε περίπτωση που περιγράφει, μπορεί να σπάει σε υποπεριπτώσεις και να καλύπτει περισσότερες περιπτώσεις.

[27][28]

Ακολουθεί το σχεσιακό σχήμα της βάσης δεδομένων που δημιουργήσαμε:



3.2.4 Σύντομη αναφορά στα δεδομένα της βάσης

Στην παρούσα υποενότητα γίνεται αναφορά σε κάποια πιο συγκεκριμένα στοιχεία της βάσης δεδομένων που δημιουργήθηκε καθώς τα μεγέθη αυτά θα αποτελέσουν τις παραμέτρους με βάση της οποίας να προσεγγιστεί η προσπάθεια ομαδοποίησης των ιατρικών προφίλ χρηστών. Τα βασικά στοιχεία είναι:

Συνολικός αριθμός προφίλ χρηστών: 5.000

Συνολικές καταχωρήσεις ιατρικού ιστορικού: 35.187

Μοναδικοί κωδικοί ιατρικού ιστορικού με βάση την κωδικοποίηση ICD-10: 14

Οι κωδικοί αυτοί με τις σύντομες περιγραφές τους παρουσιάζονται στον παρακάτω πίνακα:

Κωδ. ICD-10	**Περιγραφή**
'E08'	Διαβήτης (Diabetes)
'Z91.81'	Ιστορικό πτώσης (History of falling)
'G31.84'	Ήπια γνωστική εξασθένηση , αναφερόμενη (Mild cognitive impairment, so stated)
'H90'	Αγώγιμη και αισθητηριακή απώλεια ακοής (Conductive and sensorineural hearing loss)
'I67.9'	Εγκεφαλοαγγειακή νόσος, μη καθορισμένη (Cerebrovascular disease, unspecified)
'R63.4'	Μη φυσιολογική απώλεια βάρους (Abnormal weight loss)
'F10.9'	Χρήση Αλκοόλ (Alcohol use)
'F17'	Εξάρτηση από νικοτίνη (Nicotine dependence)
'Z82.2'	Οικογενειακό ιστορικό κώφωσης και απώλειας ακοής (Family history of deafness and hearing loss)
'G47.9'	Διαταραχή ύπνου, μη καθορισμένη (Sleep disorder, unspecified)
'F32.9'	Μεμονωμένο επεισόδιο μείζονος κατάθλιψης (Major depressive disorder single episode)
'Z01.11'	Εξέταση αυτιών και ακοής με μη φυσιολογικά ευρήματα (Encounter for examination of ears and hearing with abnormal findings)
'Z77.122'	Επαφή με και (ύποπτη) έκθεση σε θόρυβο (Contact with and (suspected) exposure to noise)
'R26.89'	Άλλες ανωμαλίες βάδισης και κινητικότητας (Other abnormalities of gait and mobility)

Πίνακας 3.1: Κωδικοί ICD-10 που εμφανίζονται στη βάση δεδομένων.

Συνολικός αριθμός μετρήσεων για όλα τα προφίλ χρηστών: 24.906.710

Μοναδικοί κωδικοί του πίνακα μετρήσεων με βάση το πρότυπο SNOMED CT: 11 (Distinct SNOMED Ids)

Οι κωδικοί αυτοί με τις σύντομες περιγραφές τους παρουσιάζονται στον παρακάτω πίνακα:

SNOMED CT	**Περιγραφή**
50373000	Ύψος Body height
386725007	Θερμοκρασία Body temperature
271650006	Διαστολική πίεση Diastolic Blood Pressure
72313002	Συστολική Πίεση - Systolic Arterial Pressure
444981005	Καρδιακοί παλμοί - Resting heart rate
726527001	Βάρος- Weight
103228002	Hemoglobin saturation with oxygen
163636005	Μυική μάζα - On examination - muscle mass
165263003	Απόσταση που περπατήθηκε - Walking distance
250825003	Θερμοκρασία περιβάλλοντος - Ambient temperature
6012004	Βοήθημα Ακοής - Hearing aid, device

Πίνακας 3.2: Κωδικοί SNOMED CT που εμφανίζονται στη βάση δεδομένων.

3.3 Δημιουργία βάσης δεδομένων

Η βάση δεδομένων που χρησιμοποιήθηκε ως είσοδος για την ομαδοποίηση των προφίλ, αποτελείται από τεχνητά δεδομένα και ο τρόπος που δημιουργήθηκε περιγράφεται στη συνέχεια.

Δημιουργία χρήστη: Αρχικοποιούνται οι χρήστες με διαδοχικούς αριθμούς για το μοναδικό PersonId και το αναγνωριστικό του καθενός. Το φύλο και το μορφωτικό επίπεδο δίνονται από μια τυχαία γεννήτρια αριθμών μέσα από μια προκαθορισμένη λίστα επιλογών. Η ηλικία ορίζεται με τον ίδιο τρόπο σε αριθμο ετών που ανήκει στο εύρος 67 έως 80 και σε ημέρες (1 έως 364).

Δημιουργία ιατρικού ιστορικού: Για τη δημιουργία του πίνακα Medical History της βάσης δεδομένων , καταχωρούνται αρχικά οι κωδικοί που έχουμε επιλέξει σε μια λίστα ενώ ορίζεται ακόμη μία γεννήτρια τυχαίων αριθμών και μια γεννήτρια Boolean τιμών. Μέσω αυτών των γεννητριών αποφασίζεται για κάθε PersonId εάν θα υπάρχει ο κάθε κωδικός της λίστας στο ιατρικό ιστορικό του και εκχωρείται μια τυχαία ημερομηνία δημιουργίας της καταχώρησης του ιστορικού τα τελευταία τέσσερα έτη.

Δημιουργία μετρήσεων ιατρικών μεγεθών: Για τη δημιουργία του πίνακα Measurement της βάσης δεδομένων , χρησιμοποιείται ξεχωριστή συνάρτηση για την καταχώρηση μοναδικών ή πολλαπλών μετρήσεων έντεκα μεγεθών σε κάθε προφίλ χρήστη. Αναλυτικότερα για κάθε μία από τις παραπάνω συναρτήσεις έχουμε τα εξής:

- Χρησιμοποιούμε γεννήτρια τυχαίων αριθμών στο εύρος 110-195 για το ύψος και η μέτρηση προστίθεται μία φορά για κάθε προφίλ.
- Αρχικοποιούνται οι μετρήσεις βάρους στο εύρος 50 έως 120 κιλά και στη συνέχεια με τυχαίο τρόπο προσθαφαιρούνται κάθε φορά διαφορετικός αριθμός κιλών στο αρχικό βάρος.
- Ομοίως για τη μυϊκή μάζα, με εύρος τιμών [0.55, 0.95].
- Αρχικοποιούνται οι μετρήσεις καρδιακών παλμών με τυχαίο τρόπο στο εύρος 60 έως 100 για φυσιολογικές τιμές και στο εύρος 95 έως 140 για μη φυσιολογικές τιμές. Οι μετρήσεις εκτός φυσιολογικού ορίου που προστίθενται δεν ξεπερνούν τη 1 στις 7.
- Με παρόμοιο τρόπο αρχικοποιούνται και οι μετρήσεις συστολικής πίεσης με τυχαίο τρόπο στο εύρος 90 έως 145 για φυσιολογικές τιμές και στο εύρος 145 έως 200 για μη φυσιολογικές τιμές, καθώς και διαστολικής πίεσης στα εύρη [70,90] και [90,140] αντίστοιχα. Οι μετρήσεις εκτός φυσιολογικού ορίου που προστίθενται δεν ξεπερνούν τη 1 στις 7 σε κάθε τύπο πίεσης.
- Η απόσταση που διανύεται με περπάτημα λαμβάνει τυχαία τιμή από 1500 έως 12000.
- Η θερμοκρασία ς επίσης δημιουργείται στο εύρος [34,37] για φυσιολογικές τιμές , ενώ για μη φυσιολογικές στο εύρος [37, 40] αντίστοιχα με ακρίβεια δεκάτων.
- Όσον αφορά τη θερμοκρασία περιβάλλοντος αρχικοποιείται με την τιμή των 25 βαθμών και στη συνέχεια προσθαφαιρούνται τυχαία πολλαπλάσια της μισής μονάδας για να προκύψει η τελική τιμή.

- Τα επίπεδα οξυγόνου στο αίμα αρχικοποιούνται με τυχαίο τρόπο στο εύρος 92 έως 100 για φυσιολογικές τιμές και στο εύρος 80 έως 90 για μη φυσιολογικές τιμές
- Τέλος, όσον αφορά το βοήθημα ακοής, αποτελεί έναν τύπο μέτρησης που προστίθεται πολύ σπανιότερα στα προφίλ χρηστών με τις κατάλληλες αποδεκτές τιμές , χρησιμοποιώντας παρόμοια μέθοδο με χρήση δύο συνθηκών τυχαιότητας μέσω γεννητριών τυχαίων αριθμών.

Η συγκεκριμένη διαδικασία επαναλαμβάνεται 600 φορές για κάθε προφίλ χρήστη και κάθε μια από τις 10 μετρήσεις (εξαιρείται το ύψος) που προστίθενται κάθε φορά, ενώ οι μετρήσεις αυτές(το είδος και η συχνότητα εμφάνισης) διαφέρουν σε κάθε επανάληψη με βάση ορισμένες συνθήκες βασισμένες σε δημιουργημένες γεννήτριες τυχαίων αριθμών. Οι τιμές is Active και Revision ορίζονται ίδιες για όλες τις καταχωρήσεις ενώ ως ModificationDateTime ορίζεται πάντα η ώρα δημιουργίας της καταχώρησης.

4 Υλοποίηση

4.1 Εισαγωγή

Στο τρέχον κεφάλαιο περιγράφεται η διαδικασία της υλοποίησης των τεχνικών στοιχείων που απαιτήθηκαν για την εκπόνηση της διπλωματικής εργασίας. Γίνεται αναφορά στις κυριότερες βιβλιοθήκες που χρησιμοποιούνται για τη διαδικασία της μηχανικής μάθησης και στη δομή δεδομένων που αξιοποιήθηκε για την αποθήκευση και επεξεργασία των δεδομένων. Επιπλέον, αναλύεται και η ευρύτερη μέθοδος που ακολουθείται για την προεπεξεργασία των διαθέσιμων πληροφοριών των χρηστών και την επιλογή των χαρακτηριστικών που θα δοθούν ως είσοδος στις επιλεχθείσες τεχνικές ομαδοποίησης.

4.2 Βιβλιοθήκες της Python

Για την υλοποίηση του προγραμματιστικού μέρους της παρούσας διπλωματικής εργασίας επιλέχθηκε η γλώσσα Python, όντας μία εκ των πλέον διαδεδομένων στον τομέα της μηχανικής μάθησης και προσφέροντας ένα δοκιμασμένο οικοσύστημα εργαλείων και χρήσιμων βιβλιοθηκών που αξιοποιήσαμε. Ακολουθεί μια αναφορά στις βασικότερες.

4.2.1 Scikit-learn

Το Scikit-Learn είναι μια βιβλιοθήκη μηχανικής εκμάθησης ανοιχτού κώδικα που υποστηρίζει την εποπτευόμενη και μη εποπτευόμενη μάθηση. Παρέχει επίσης διάφορα εργαλεία για εκπαίδευση μοντέλου, προεπεξεργασία δεδομένων, επιλογή και αξιολόγηση μοντέλου και πολλά άλλα βοηθητικά προγράμματα. Το Scikit-learn παρέχει δεκάδες ενσωματωμένους αλγόριθμους και μοντέλα μηχανικής μάθησης, που ονομάζονται εκτιμητές. Κάθε εκτιμητής μπορεί να προσαρμοστεί σε ορισμένα δεδομένα χρησιμοποιώντας τη μέθοδο fit , η οποία δέχεται γενικά 2 εισόδους:

- τον πίνακα δειγμάτων (ή μήτρα σχεδίασης) X. Το μέγεθος του X είναι συνήθως (n_samples, n_features), πράγμα που σημαίνει ότι τα δείγματα αντιπροσωπεύονται ως σειρές και τα χαρακτηριστικά παρουσιάζονται ως στήλες.
- Οι τιμές στόχου y που είναι πραγματικοί αριθμοί για εργασίες παλινδρόμησης, ή ακέραιοι αριθμοί για την ταξινόμηση (ή οποιοδήποτε άλλο διακριτό σύνολο τιμών). Για μη επιβλεπόμενες μαθησιακές διαδικασίες, δεν χρειάζεται να προσδιοριστεί.

Μόλις εκπαιδευτεί το μοντέλο μηχανικής μάθησης, μπορεί να χρησιμοποιηθεί για την πρόβλεψη των τιμών-στόχων των νέων δεδομένων, χωρίς να χρειάζεται επανεκπαίδευση.

Η βιβλιοθήκη scikit-learn είναι χτισμένη πάνω στις βιβλιοθήκες:

- NumPy, η οποία προσθέτει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες, μαζί με μια μεγάλη συλλογή μαθηματικών συναρτήσεων υψηλού επιπέδου για πράξεις πάνω σε αυτούς τους πίνακες.
- SciPy , η οποία προσφέρει εργαλεία για περίπλοκους υπολογισμούς και μαθηματικές, επιστημονικές και μηχανικές διεργασίες.
- Matplotlib, η οποία διευκολύνει τη γραφική αναπαράσταση των διαφόρων παραμέτρων και αποτελεσμάτων.

Οι υλοποιήσεις των αλγορίθμων ομαδοποίησης με τις όποιες βελτιώσεις που χρησιμοποιούμε (k-Means, HAC, DBSCAN, t-SNE) προέρχονται από τη βιβλιοθήκη scikit-learn.

[29]

4.2.2 Pandas & δομή DataFrame

Η Pandas είναι μια βιβλιοθήκη λογισμικού γραμμένη για τη γλώσσα προγραμματισμού Python για χειρισμό και ανάλυση δεδομένων. Συγκεκριμένα, προσφέρει δομές και λειτουργίες δεδομένων για χειρισμό αριθμητικών πινάκων και χρονοσειρών.

Τα κύρια στοιχεία της βιβλιοθήκης είναι:

- Μια γρήγορη και αποτελεσματική δομή δεδομένων γνωστή ως DataFrame για χειρισμό δεδομένων με ενσωματωμένη ευρετηρίαση (indexing).
- Τα εργαλεία για ανάγνωση και εγγραφή δεδομένων μεταξύ δομών δεδομένων που υπάρχουν στη μνήμη και διαφορετικών μορφών όπως αρχεία

CSV και κειμένου, Microsoft Excel, βάσεις δεδομένων SQL και η γρήγορη μορφή HDF5.

- Το ευφυές data alignment και ο ολοκληρωμένος χειρισμός δεδομένων που λείπουν. Η βιβλιοθήκη προσφέρει δυνατότητες αυτόματου data alignment, δηλαδή αυτόματης τοποθέτησης στη μνήμη δεδομένων διαφορετικών μεγεθών σε byte και της σωστής και γρήγορης ανάκτησής τους . Επιπλέον μπορεί να χειριστεί εύκολα δεδομένα με ακανόνιστη δομή μετατρέποντας τα σε μια ομαλή μορφή.
- Η ευέλικτη αναδιαμόρφωση και δημιουργία δυναμικών συνόλων δεδομένων (pivot tables).
- Η δυνατότητα έξυπνης κατάτμησης βάσει ετικετών, υψηλού επιπέδου indexing και εξαγωγής υποσυνόλων μεγάλων συνόλων δεδομένων.
- Το γεγονός ότι στήλες μπορούν να εισαχθούν και να διαγραφούν από δομές δεδομένων με ευκολία για μεταβλητότητα μεγέθους.
- Η συγκέντρωση (aggregation) και η μετατροπή δεδομένων με ισχυρές δυνατότητες “group by”, που επιτρέπουν λειτουργίες διαχωρισμού των δεδομένων με βάση κάποιο κριτήριο, εφαρμογή κάποιας συνάρτησης στο υποσύνολο που προέκυψε και συγχώνευση των αποτελεσμάτων και των αρχικών δεδομένων σε κοινή δομή δεδομένων.
- Η συγχώνευση και σύνδεση συνόλων δεδομένων με υψηλή απόδοση.
- Η ευρετηρίαση ιεραρχικού άξονα, η οποία παρέχει έναν διαισθητικό τρόπο εργασίας με δεδομένα υψηλής διάστασης σε μια δομή δεδομένων χαμηλότερης διάστασης.
- Η λειτουργικότητα χρονοσειρών η οποία επιτρέπει τη δημιουργία εύρους ημερομηνιών και τη μετατροπή συχνότητας, στατιστικά με τη χρήση της τεχνικής του κινούμενου παραθύρου, αλλαγή ημερομηνίας και lagging.
- Η ιδιαίτερη βελτιστοποίηση για απόδοση, με κρίσιμα κομμάτια κώδικα γραμμένα σε Cython ή C.

Η γλώσσα Python με χρήση της βιβλιοθήκηw pandas βρίσκει εφαρμογή σε μια μεγάλη ποικιλία ακαδημαϊκών και εμπορικών τομέων, συμπεριλαμβανομένων των Χρηματοοικονομικών, Νευροεπιστήμης, Οικονομικών, Στατιστικών, Διαφήμισης, Web Analytics και άλλων.

[30]

Η δομή Dataframe της βιβλιοθήκης Pandas αποτελεί την κύρια δομή δεδομένων που χρησιμοποιήθηκε για την αποθήκευση και επεξεργασία των δεδομένων που αντλήθηκαν από τη σχεσιακή βάση, πριν δοθούν ως είσοδος στους αλγορίθμους ομαδοποίησης.

Η δομή δεδομένων DataFrame είναι μια δισδιάστατη δομή με στήλες δυνητικά διαφορετικών τύπων την οποία μπορούμε να παρομοιάσουμε με υπολογιστικό φύλλο ή έναν πίνακα SQL. Είναι γενικά το πιο συχνά χρησιμοποιούμενο αντικείμενο της βιβλιοθήκης pandas και διαθέτει ισχυρές δυνατότητες ευρετηρίασης (indexing). Η δομή DataFrame μπορεί να σχηματιστεί λαμβάνοντας ως είσοδο μονοδιάστατους ή πολυδιάστατους πίνακες, λίστες, ευρετήρια, σειρές, άλλα DataFrames καθώς και πληθώρα τύπων αρχείων, όπως φαίνεται και στην εικόνα που ακολουθεί.

Type	Data Description	Reader	Writer
text	CSV	read_csv	to_csv
text	Fixed-Width Text File	read_fwf	
text	JSON	read_json	to_json
text	HTML	read_html	to_html
text	Local clipboard	read_clipboard	to_clipboard
	MS Excel	read_excel	to_excel
binary	OpenDocument	read_excel	
binary	HDF5 Format	read_hdf	to_hdf
binary	Feather Format	read_feather	to_feather
binary	Parquet Format	read_parquet	to_parquet
binary	ORC Format	read_orc	
binary	Msgpack	read_msgpack	to_msgpack
binary	Stata	read_stata	to_stata
binary	SAS	read_sas	
binary	SPSS	read_spss	
binary	Python Pickle Format	read_pickle	to_pickle
SQL	SQL	read_sql	to_sql
SQL	Google BigQuery	read_gbq	to_gbq

Εικόνα 4.1: Παρουσίαση τύπων αρχείων που μπορούν να αποτελέσουν είσοδο για τη δημιουργία ενός DataFrame

Χρησιμοποιώντας τη δομή αυτή μας δίνεται η δυνατότητα να κατανοήσουμε καλύτερα το περιεχόμενο των δεδομένων μας με χρήσιμες συναρτήσεις, να λάβουμε περιγραφικές στατιστικές ποσότητες για αυτά, να εφαρμόσουμε συναρτήσεις μαζικά και να αντιμετωπίσουμε περιπτώσεις ελλείψεων σε ορισμένες στήλες. Ταυτόχρονα μπορούμε να εκμεταλλευτούμε τις δυνατότητες βελτιστοποίησης για αποδοτικότερους χειρισμούς μεγάλων DataFrames, να κάνουμε συγχωνεύσεις, διαχωρισμούς , ελέγχους και συγκρίσεις στηλών ή ολόκληρων αντικειμένων DataFrame και να αξιοποιήσουμε τις διαθέσιμες γραφικές μεθόδους για καλύτερη κατανόηση των δεδομένων. [31] Όλα τα παραπάνω καθιστούν τη συγκεκριμένη δομή δεδομένων ιδανική για τις διεργασίες που πραγματοποιούνται στη συγκεκριμένη εργασία, ενώ εγγυούνται την ύπαρξη μιας κατάλληλης δομής δεδομένων για διαχείριση ακόμη μεγαλύτερου αριθμού γραμμών και στηλών για τυχόν επέκταση των δυνατοτήτων της παρούσας πρότασης ομαδοποίησης.

Παρακάτω παρουσιάζονται οι πρώτες γραμμές ενός DataFrame που προέκυψε σε ενδιάμεσο στάδιο προεπεξεργασίας των δεδομένων μας.

	PersonId	Gender	EducationalLevel	Age	IcdCodes	50373000	386725007	271650006	72313002	444981005	726527001	103228002	163636005	165263003	250825003	6012004
0	25670	female	college	78	[E08, Z91.81, G31.84]	127.0	38.5	86.0	96.0	83.0	86.0	99.0	5.66	10128.0	16.0	776.0
1	25671	male	college	74	[H90, I67.9, G31.84, R63.4, F10.9, F17, Z82.2,...	114.0	36.6	134.0	186.0	139.0	65.0	96.0	6.83	6736.0	34.5	NaN
2	25672	male	primary	72	[H90, Z91.81, I67.9, F32.9, Z82.2, Z01.11]	170.0	36.3	96.0	147.0	109.0	66.2	92.0	7.91	6441.0	33.5	NaN
3	25673	male	primary	76	[E08, Z91.81, I67.9, F10.9, F17, Z77.122]	134.0	35.2	75.0	122.0	66.0	72.1	98.0	7.98	7034.0	15.0	485.0
4	25674	female	primary	68	[H90, I67.9, G31.84, F17, Z82.2, Z01.11]	142.0	36.2	80.0	135.0	65.0	61.1	82.0	5.96	7506.0	28.5	NaN

Εικόνα 4.2: Ενδεικτικό DataFrame σε ενδιάμεσο βήμα της προεπεξεργασίας των δεδομένων της βάσης δεδομένων μας.

4.3 Προεπεξεργασία δεδομένων και επιλογή χαρακτηριστικών

4.3.1 Βήματα προεπεξεργασίας

Για να προετοιμάσουμε τα δεδομένα της βάσης δεδομένων για τη διαδικασία της ομαδοποίησης χρηστών, χρειάστηκαν ορισμένα βήματα ώστε να έρθουν σε κατάλληλη μορφή και να αντιπροσωπεύει η κάθε σειρά το προφίλ ενός χρήστη. Ταυτόχρονα, ορισμένα μεγέθη απαιτούσαν μετατροπές σε μορφές που να μπορούν να γίνουν αντιληπτές σε πλαίσια μηχανικές μάθησης, ενώ απαραίτητη ήταν και οι σύμπτυξη κάποιων ποσοτήτων σε

αντιπροσωπευτικό δείγμα τους για να μειωθεί η πολυπλοκότητα του τεράστιου όγκου δεδομένων που είχαμε στη διάθεσή μας και να γίνει χρήση κάποιων τιμών που μπορούν να αποτελέσουν χαρακτηριστικά προς μελέτη κατά τη διαδικασία μηχανικής μάθησης. Τα βήματα αυτά παρουσιάζονται στη συνέχεια.

1. **Εισαγωγή ορισμένων προσωπικών στοιχείων από τον πίνακα Person της βάσης δεδομένων**.

   Οι στήλες που επιλέχθηκαν ήταν το PersonId, ως αναγνωριστικό της κάθε σειράς, η ημερομηνία γέννησης (DateOfBirth), το φύλο (Gender) και το επίπεδο μόρφωσης (Educational Level).

2. **Μετατροπή της στήλης DateOfBirth.**

   Η στήλη DateOfBirth περιείχε σε μορφή κειμένου τις ημερομηνίες γέννησης των χρηστών, η οποία μετατράπηκε σε ακέραια αριθμητική τιμή που αντιστοιχεί στην ηλικία του εκάστοτε ιατρικού προφίλ. Η μετατροπή αυτή της τιμής σε αριθμητική επιτρέπει τη χρήση της ως ένα από τα χαρακτηριστικά(features) που λαμβάνουμε, μαζί με το φύλο και το επίπεδο μόρφωσης, τα οποία επίσης θα υποστούν επεξεργασία στην οποία θα αναφερθούμε στη συνέχεια, στην υποενότητα κωδικοποίησης κατηγορικών μεταβλητών.

3. **Εύρεση λίστας με τους κωδικούς ιατρικού ιστορικού που εμφανίζονται για κάθε χρήστη.**

   Αφού εξαγάγουμε τους κωδικούς ιατρικού ιστορικού με βάση την κωδικοποίηση ICD-10, θα εφαρμόσουμε One Hot Encoding για να λάβουμε έναν πίνακα όπου κάθε κωδικός θα αποτελεί μια ξεχωριστή στήλη και στην ένδειξη της στήλης αυτή για τον κάθε χρήστη θα συμπληρώνεται η τιμή 1 εάν ο χρήστης παρουσιάζει το συγκεκριμένο κωδικό στο ιστορικό του και αντίστοιχα 0 εάν δεν τον παρουσιάζει. Η διαδικασία του Label Encoding είναι ιδιαίτερα συνήθης και άκρως απαραίτητη σε διεργασίες μηχανικής μάθησης όπου παρουσιάζονται κατηγορικές μεταβλητές και παρουσιάζεται αναλυτικά στη συνέχεια του κεφαλαίου (στην υποενότητα κωδικοποίησης κατηγορικών μεταβλητών). Με το τέλος αυτού του βήματος θα έχουμε 14 νέα χαρακτηριστικά που θα αντιστοιχούν στους κωδικούς ιστορικού που μετατρέψαμε σε νέες στήλες. Αξίζει να αναφερθεί ότι η συγκεκριμένη εξαγωγή χαρακτηριστικών θα μπορούσε να αντικατασταθεί από κάποια άλλη που θα θεωρούνταν χρήσιμη με βάση την ερμηνεία των δεδομένων μας, για παράδειγμα θα μπορούσαμε να λάβουμε συγκεκριμένο υποσύνολο των κωδικών ιστορικού, τον αριθμό των ετών που έχουν παρέλθει από την

εμφάνιση κάποιου συγκεκριμένου κωδικού (όπως για παράδειγμα ο κωδικός E08 που αντιστοιχεί σε εμφάνιση διαβήτη) ή τη συχνότητα εμφάνισης ορισμένων άλλων κωδικών ιστορικού(όπως για παράδειγμα ο κωδικός I67.9 που αντιστοιχεί σε εμφάνιση εγκεφαλοαγγειακής νόσου). Παρόλα αυτά, για τους σκοπούς της εργασίας μας και την επίτευξη μιας ομαδοποίησης σε πρώτο επίπεδο, επιλέχθηκε μια απλούστερη κωδικοποίηση όλων των μεγεθών του ιστορικού με βάση την εμφάνιση τους ή μη.

4. **Εξαγωγή χαρακτηριστικών από τον πίνακα Measurement**

   Ο πίνακας Measurement της βάσης δεδομένων περιλαμβάνει μετρήσεις μεγεθών με βάση το πρότυπο SNOMED CT και αποτελείται από περισσότερες από 25 εκατομμύρια καταχωρήσεις. Ως εκ τούτου, κρίθηκε σκόπιμο να επιλέξουμε κάποιο αντιπροσωπευτικό δείγμα εξ αυτών για κάθε χρήστη. Έτσι, δημιουργήθηκαν ξεχωριστές στήλες στο DataFrame για κάθε κωδικό SNOMED CT (είχαμε 11 μεμονωμένους κωδικούς) και στη συνέχεια αντλήθηκαν οι μετρήσεις του κάθε χρήστη ανά κωδικό ώστε να υπολογιστεί ο μέσος όρος τους, με εξαίρεση τον κωδικό της συσκευής ακουστικού βοηθήματος(SNOMED CT Id: 6012004). Καθώς στη βάση δεδομένων μας δεν υπήρχε ιδιαίτερη μεταβλητότητα στα χρονικά διαστήματα κατά τα οποία αντλήθηκαν οι μετρήσεις αυτές, ο αριθμητικός μέσος της κάθε μίας χρησιμοποιήθηκε ως χαρακτηριστικό. Υπάρχει, όπως και στην περίπτωση των κωδικών ICD-10, η δυνατότητα διαφοροποίησης των χαρακτηριστικών αυτών, όπως μια επιλογή του μέσου όρου των πιο πρόσφατων μετρήσεων για έναν συγκεκριμένο αριθμό ετών, τον υπολογισμό της διακύμανσης ή κάποιου άλλου στατιστικού μεγέθους που θεωρείται χρήσιμο, ή και τη συχνότητα εμφάνισης μετρήσεων εκτός φυσιολογικού εύρους σε ένα χρονικό διάστημα. Επιλέξαμε και πάλι μια συνήθη προσέγγιση όπως αυτή του υπολογισμού της μέσης τιμής ως καταλληλότερη στα πλαίσια της τρέχουσας εργασίας.

5. **Μετατροπή όλων των κατηγορικών μεγεθών σε ποσοτικές**

   Οι εναπομένουσες κατηγορικές ποσότητες δηλαδή το φύλο και το μορφωτικό επίπεδο μετατράπηκαν σε αριθμητικούς δείκτες με τη χρήση της μεθόδου Label Encoding, όπως αυτή περιγράφεται στη συνέχεια του κεφαλαίου.

Το τελικό σύνολο χαρακτηριστικών που επιλέχθηκαν ή εξήχθησαν αποτελείται από τις εξής ποσότητες:

- ηλικία,
- φύλο,
- μορφωτικό επίπεδο,
- εμφάνιση ή μη διαβήτη,
- εμφάνιση ή μη ιστορικού πτώσης
- εμφάνιση ή μη ήπιας γνωστικής εξασθένησης,
- εμφάνιση ή μη αγώγιμης και αισθητηριακής απώλειας ακοής,
- εμφάνιση ή μη εγκεφαλοαγγειακής νόσου,
- εμφάνιση ή μη απώλειας βάρους σε μη φυσιολογικό βαθμό ή τρόπο,
- εμφάνιση ή μη χρήσης αλκοόλ,
- εμφάνιση ή μη εξάρτησης από νικοτίνη
- εμφάνιση ή μη οικογενειακού ιστορικού κώφωσης και απώλειας ακοής
- εμφάνιση ή μη διαταραχής ύπνου
- εμφάνιση ή μη μεμονωμένου επεισοδίου μείζονος κατάθλιψης
- εμφάνιση ή μη εξέτασης αυτιών και ακοής με μη φυσιολογικά ευρήματα
- εμφάνιση ή μη επαφής με ύποπτη έκθεση σε θόρυβο
- εμφάνιση ή μη άλλων ανωμαλιών βαδίσεων ή κινητικότητας
- μέση τιμή ύψους
- μέση τιμή βάρους
- μέση τιμή θερμοκρασίας
- μέση τιμή συστολικής πίεσης
- μέση τιμή διαστολικής πίεσης
- μέση τιμή καρδιακού ρυθμού
- μέση τιμή κορεσμού οξυγόνου
- μέση τιμή μυικής μάζας
- μέση τιμή απόστασης που περπατήθηκε
- μέση τιμή θερμοκρασίας περιβάλλοντος

4.4 Κωδικοποίηση κατηγορικών μεταβλητών

Κατηγορικές ονομάζονται οι μεταβλητές που λαμβάνουν τιμές που δεν δίνονται από αριθμητικές ποσότητες αλλά λαμβάνουν τιμές από ένα σύνολο, οι οποίες δεν ιεραρχούνται. Τέτοιες μεταβλητές είναι για παράδειγμα το φύλο, το μορφωτικό επίπεδο, το επάγγελμα, η χώρα καταγωγής, οι διατροφικές προτιμήσεις, η ομάδα αίματος (με τιμές A, B, AB, O), οι συνέπειες του καπνίσματος (με τιμές καρδιακά νοσήματα, καρκίνος κτλ), όπως επίσης και η οικονομική κατάσταση και η υγεία των ανθρώπων (που μπορεί να χαρακτηριστεί ως κακή, μέτρια, καλή ή πολύ καλή κ.α. [32] Οι μεταβλητές αυτές παρουσιάζονται αρκετά συχνά σε προβλήματα μηχανικής μάθησης και απαιτούν έναν ιδιαίτερο χειρισμό καθώς πρέπει να κωδικοποιηθούν με τέτοιο τρόπο ώστε να λαμβάνουν αριθμητικές τιμές ή κάποια αναπαράσταση που να γίνεται αντιληπτή από υπολογιστικά συστήματα. Στην παρούσα υλοποίηση, χειριζόμαστε τις κατηγορικές μεταβλητές με τη χρήση δύο αρκετά διαδεδομένων μεθόδων, της μεθόδου Label Encoding και της μεθόδου One-Hot-Encoding.

4.4.1 Label Encoding

Η κωδικοποίηση ετικέτας (label encoding) εκχωρεί μια ακέραια τιμή σε κάθε πιθανή τιμή μιας κατηγορηματικής μεταβλητής. Για παράδειγμα, εάν ένα σύνολο δεδομένων έχει μια κατηγορηματική μεταβλητή με τιμές που προέρχονται από το σύνολο {"άσπρο", "μαύρο", "μπλε"}, τότε η κωδικοποίηση ετικέτας ενδέχεται να εκχωρήσει τις αντιστοιχισμένες τιμές από το σύνολο {0,1,2} αντίστοιχα. Είναι ο απλούστερος τρόπος μετατροπής κατηγορηματικών τιμών σε αριθμητικές τιμές. Μόλις γνωρίζουμε όλες τις πιθανές τιμές μιας κατηγορηματικής μεταβλητής, τα κωδικοποιημένα ισοδύναμά τους καθορίζονται από τον τρόπο που επιλέγουμε αυθαίρετα να τους αντιστοιχίσουμε ακέραιες τιμές. Παρόλα αυτά, αυτό αποτελεί και ένα μειονέκτημα της κωδικοποίησης καθώς εισάγει μια έννοια ιεραρχίας σε μεταβλητές όπου πιθανόν να μην υφίσταται.[33][34] Χρησιμοποιούμε Label Encoding, όπως προαναφέρθηκε, για τις κατηγορικές τιμές φύλο και μορφωτικό επίπεδο.

4.4.2 One Hot Encoding

Η One Hot Encoding αποτελεί την πιο ευρέως χρησιμοποιούμενη μέθοδο κωδικοποίησης. Μετατρέπει τις κατηγορικές μεταβλητές σε ένα σταθερό επίπεδο αναφοράς. Μια μεμονωμένη μεταβλητή με ν παρατηρήσεις και μ διακριτές τιμές μετατρέπεται σε έως μ ξεχωριστές δυαδικές μεταβήτές με ν παρατηρήσεις η κάθε μία. Κάθε μία από αυτές τις παρατηρήσεις δείχνει την παρουσία (με την τιμή 1) ή την απουσία(με την τιμή 0) της διακριτής παρατήρησης που μετατρέπεται σε δυαδική μορφή.[33]

Όπως αναφέρθηκε νωρίτερα στο κεφάλαιο, χρησιμοποιείται One Hot Encoding για την κατηγορική μεταβλητή IcdCodes. Παρακάτω, απεικονίζεται γραφικά η μετατροπή των IcdCodes με One Hot Encoding, όπως πραγματοποιήθηκε για τις ανάγκες της εργασίας μας.

IcdCodes	E08	F10.9	F17	F32.9	G31.84	G47.9	H90	I67.9	R26.89	R63.4	Z01.11	Z77.122	Z82.2	Z91.81
[E08, Z91.81, G31.84]	1	0	0	0	1	0	0	0	0	0	0	0	0	1
[H90, I67.9, G31.84, R63.4, F10.9, F17, Z82.2,...	0	1	1	0	1	1	1	1	0	1	0	0	1	0
[H90, Z91.81, I67.9, F32.9, Z82.2, Z01.11]	0	0	0	1	0	0	1	1	0	0	1	0	1	1
[E08, Z91.81, I67.9, F10.9, F17, Z77.122]	1	1	1	0	0	0	0	1	0	0	0	1	0	1
[H90, I67.9, G31.84, F17, Z82.2, Z01.11]	0	0	1	0	1	0	1	1	0	0	1	0	1	0

Εικόνα 4.3: Παράδειγμα εφαρμογής One Hot Encoding στη στήλη IcdCodes

5 Αποτελέσματα

5.1 Θεωρητικό Υπόβαθρο

Για την αξιολόγηση των αποτελεσμάτων και τη σύγκριση των αλγορίθμων ομαδοποίησης που χρησιμοποιήσαμε για την επίτευξη του σκοπού της παρούσας εργασίας, θα αξιοποιήσουμε ορισμένους δείκτες που προτείνονται για εκτίμηση της επίδοσης των μοντέλων μηχανικής μάθησης σε περιπτώσεις που οι κλάσεις και οι ετικέτες αλήθειας δεν είναι γνωστές. Σε προβλήματα ομαδοποίησης, δεν είναι εύκολο να προσδιοριστεί η αποδοτικότητα ενός αλγορίθμου γεγονός που έχει οδηγήσει στη δημιουργία πολλαπλών τεχνικών αξιολόγησης. Πολύ συχνά, η διαδικασία αξιολόγησης που εκτελείται εξαρτάται από τον αλγόριθμο που χρησιμοποιείται για τη λήψη των αποτελεσμάτων ομαδοποίησης.

Οι μέθοδοι που συναντώνται για την αξιολόγηση των αποτελεσμάτων μπορούν να χωριστούν σε δύο κατηγορίες, τις εσωτερικές και τις εξωτερικές Οι εσωτερικές μέθοδοι επικύρωσης (internal validation methods) καθιστούν δυνατή την διαπίστωση της ποιότητας ομαδοποίησης χωρίς πρόσβαση σε εξωτερικές πληροφορίες (δηλαδή βασίζονται αποκλειστικά στις πληροφορίες που παρέχονται από τα δεδομένα που χρησιμοποιούνται ως είσοδος στον αλγόριθμο ομαδοποίησης). Διακρίνονται περαιτέρω δύο τύποι μετρήσεων εσωτερικής επικύρωσης, τα μέτρα συνοχής και τα μέτρα διαχωρισμού. Τα μέτρα συνοχής αξιολογούν πόσο κοντά είναι τα στοιχεία του ίδιου συμπλέγματος το ένα το άλλο, ενώ τα μέτρα διαχωρισμού ποσοτικοποιούν το επίπεδο διαχωρισμού μεταξύ των συστάδων που έχουν προκύψει. Αυτά τα μέτρα είναι επίσης γνωστοί ως εσωτερικοί δείκτες επειδή υπολογίζονται από τα δεδομένα εισαγωγής χωρίς εξωτερικές πληροφορίες. Στην παρούσα διπλωματική εργασία , χρησιμοποιούνται κυρίως εσωτερικοί δείκτες για την αξιολόγηση των αποτελεσμάτων καθώς δεν έχουμε εξωτερικές πληροφορίες για τα δεδομένα μας και συγκεκριμένα ο Συντελεστής Davies-Bouldin, ο Συντελεστής Calinski-Harabasz και ο Συντελεστής Silhouette. Κρίνεται σκόπιμο να πραγματοποιηθεί μια παρουσίαση των δεικτών αυτών και των κύριων χαρακτηριστικών τους στην αρχή του κεφαλαίου, πριν προχωρήσουμε

στην περιγραφή των αποτελεσμάτων με βάση τις τιμές που προέκυψαν κατά την εκτέλεση των ομαδοποιήσεων μετέπειτα.

5.1.1 Συντελεστής Davies-Bouldin

Αυτός ο δείκτης υποδηλώνει τη μέση «ομοιότητα» μεταξύ συστάδων, όπου η ομοιότητα είναι ένα μέτρο που συγκρίνει την απόσταση μεταξύ συστάδων με το μέγεθος των ίδιων των συστάδων. Το μηδέν είναι η χαμηλότερη τιμή ενώ όσο πλησιέστερα στο μηδέν είναι ο αριθμός που θα προκύψει, τόσο καλύτερος ο διαχωρισμός σε ομάδες και η απόδοση του αλγορίθμου ομαδοποίησης. Ανήκει στα μέτρα που αξιολογούν το διαχωρισμό μεταξύ των ομάδων ενώ βασικό πλεονέκτημα είναι η ταχύτητα υπολογισμού του. Ορίζεται από τη σχέση:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max[\, R(i, j)], $$

όπου k ο αριθμός των συστάδων, ενώ η ποσότητα $R(i, j)$ είναι ο λόγος της μέσης απόστασης μεταξύ κάθε σημείου της συστάδας και του κέντρου της, ως προς την απόσταση μεταξύ των κέντρων των συστάδων i,j.

5.1.2 Συντελεστής Calinski-Harabasz

Ο Συντελεστής Calinski-Harabasz (CH), , επίσης γνωστός ως κριτήριο αναλογίας διακύμανσης, είναι ένα μέτρο που βασίζεται στην εσωτερική διασπορά των συστάδων(SSE) καθώς και στη διασπορά ή απόσταση μεταξύ συστάδων(SSB). Δίνεται από τον παρακάτω τύπο, όπου k είναι ο αριθμός των συστάδων:

$$CH = \frac{\frac{SSB}{k-1}}{\frac{SSE}{k}}$$

Όσο υψηλότερη τιμή έχει ο συγκεκριμένος δείκτης, τόσο καλύτερα διαχωρισμένες είναι οι ομάδες που έχουν προκύψει και επομένως πιο επιτυχημένη η προσπάθεια ομαδοποίησης. Βασικό του πλεονέκτημα είναι η σχετικά μικρή πολυπλοκότητα που επιτρέπει το γρήγορο υπολογισμό της ποσότητας ενώ στα αρνητικά του συγκαταλέγεται η χαμηλή αξιοπιστία του σε τεχνικές ομαδοποίησης που στηρίζονται στην πυκνότητα όπως ο αλγόριθμος DBSCAN.

5.1.3 Συντελεστής Silhouette

Ο συντελεστής Silhouette(SS) αποτελεί ακόμη ένα συνδυαστικό μέτρο συνοχής και διαχωρισμού. Το εύρος τιμών που λαμβάνει είναι [-1,1]. Οι θετικές τιμές δείχνουν υψηλό διαχωρισμό μεταξύ συστάδων. Οι αρνητικές τιμές αποτελούν ένδειξη ότι η ομαδοποίηση δεν είναι εφικτή.. Όταν ο συντελεστής είναι κοντά στο μηδέν, είναι μια ένδειξη ότι τα δεδομένα είναι ομοιόμορφα κατανεμημένα σε όλο τον Ευκλείδειο χώρο και οι συστάδες αναμιγνύονται μεταξύ τους (δηλ ένδειξη αλληλεπικαλυπτόμενων συστάδων). Ένα από τα κύρια μειονεκτήματα του συντελεστή είναι η υψηλή υπολογιστική του πολυπλοκότητα, ενώ αποτελεί τοn πιο χρονοβόρο σε υπολογισμούς συντελεστή από όσους χρησιμοποιούμε, γεγονός που καθιστά μη πρακτική την εφαρμογή του σε μεγάλες εισόδους. Αντιθέτως, ένα από τα πλεονεκτήματα του είναι η καλή απόδοση σε τεχνικές ομαδοποίησης που στηρίζονται στην πυκνότητα όπως ο αλγόριθμος DBSCAN. Δίνεται από τον παρακάτω τύπο:

$$SS = \frac{B - A}{\max(A, B)}$$

όπου Α είναι η μέση απόσταση μεταξύ ενός σημείου και όλων των άλλων σημείων στην ίδια ομάδα, ενώ Β η μέση απόσταση μεταξύ ενός σημείου και όλων των άλλων σημείων στο επόμενο κοντινότερο σύμπλεγμα.

[35][36]

5.1.4 Επικύρωση μέσω k-fold cross validation

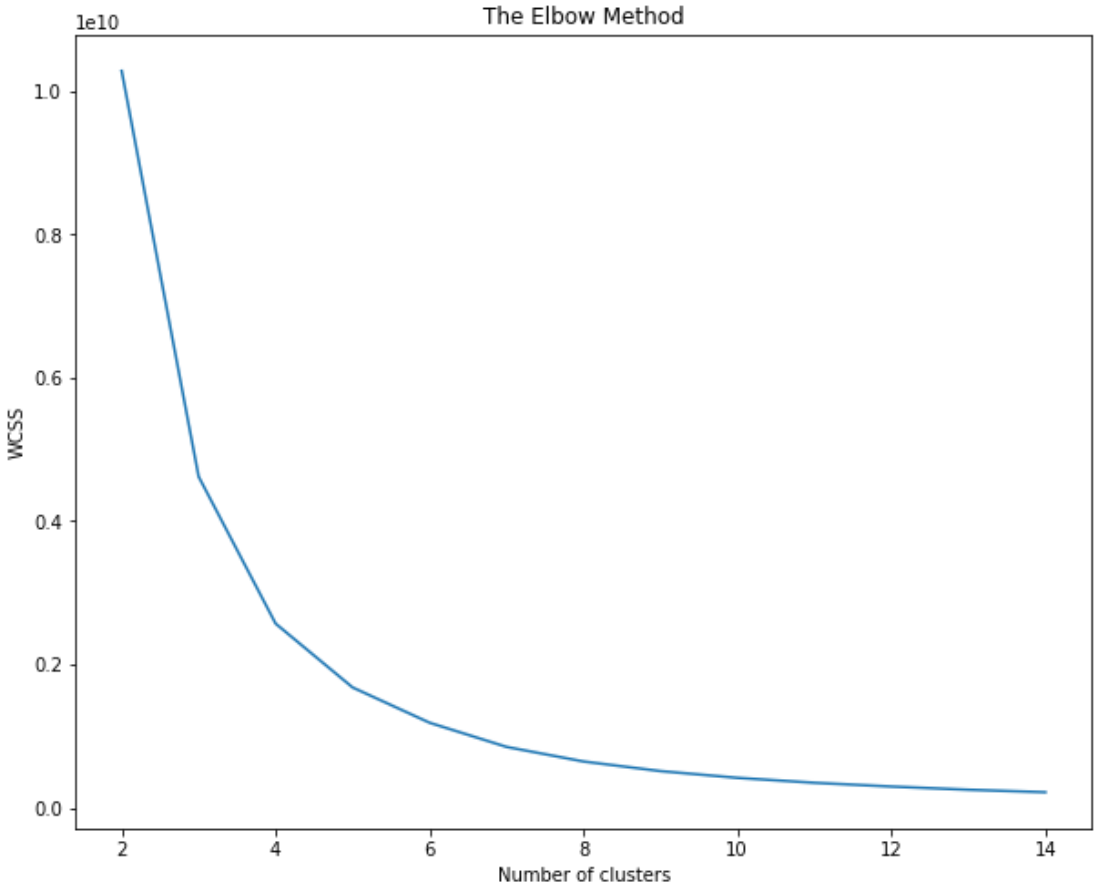
Η μέθοδος πολλαπλής επικύρωσης k-fold (k-fold cross validation) χρησιμοποιείται για αξιολόγηση σε περιπτώσεις που έχουμε δημιουργήσει το μοντέλο ομαδοποίησης σε ορισμένα προηγούμενα διαθέσιμα δεδομένα και στη συνέχεια θέλουμε να αντιστοιχίσουμε νέα σημεία δεδομένων στα συμπλέγματα που δημιουργήθηκαν προηγουμένως. Με βάση τη μέθοδο αυτή το σύνολο των δεδομένων Χ χωρίζεται σε k υποσύνολα ίσου μεγέθους. Κάθε φορά , το k - οστό υποσύνολο σχηματίζει ένα σύνολο ελέγχου , ενώ τα υπόλοιπα αποτελούν το σύνολο εκπαίδευσης που χρησιμοποιείται για την ανάπτυξη ενός μοντέλου ταξινόμησης. Αξιολογείται έτσι η ακρίβεια της πρόβλεψης με βάση το μοντέλο που χρησιμοποιείται. Η έννοια της ακρίβειας της πρόβλεψης αποτελεί ένα μέτρο σταθερότητας των συστάδων και μπορεί να χρησιμοποιηθεί για την αξιολόγηση της απόδοσης μιας μεθόδου ομαδοποίησης. Εάν οι προβλέψεις στα δεδομένα εκπαίδευσης και δοκιμών είναι παρόμοιες, η ομαδοποίηση μπορεί να θεωρηθεί καλή. Αυτή η προσέγγιση εφαρμόζεται συχνά σε ιατρικά δεδομένα όταν αξιολογείται η πιθανότητα εμφάνισης διαφόρων προβλημάτων υγείας σε διαφορετικές ομάδες ασθενών, όπως συμβαίνει και στην παρούσα διπλωματική εργασία. [37]

5.2 Αποτελέσματα

Για κάθε έναν από τους αλγορίθμους που δοκιμάζονται παρουσιάζουμε τα αποτελέσματα που μας δίνουν τις καλύτερες μετρικές μετά από κατάλληλη αλλαγή και προσαρμογή των παραμέτρων τους ανά περίπτωση.

5.2.1 k-Means

Αρχικά εφαρμόζουμε τον αλγόριθμο k-Meansσε μη κανονικοποιημένα δεδομένα. Χρησιμοποιούμε το διάγραμμα του αγκώνα για εκτίμηση του βέλτιστου αριθμού συστάδων και λαμβάνουμε το παρακάτω αποτέλεσμα.



Εικόνα 5.1: Μέθοδος του αγκώνα για μη κανονικοποιημένα δεδομένα

Επομένως, επιλέγουμε να τρέξουμε τον αλγόριθμο για k=4, δηλαδή για τέσσερις ομάδες προφίλ. Τα αντίστοιχα μεγέθη που λαμβάνουμε είναι τα εξής:

```
Συντελεστής DB = 0.5
Συντελεστής CH = 22839.44
Συντελεστής Silhouette = 0.57
```

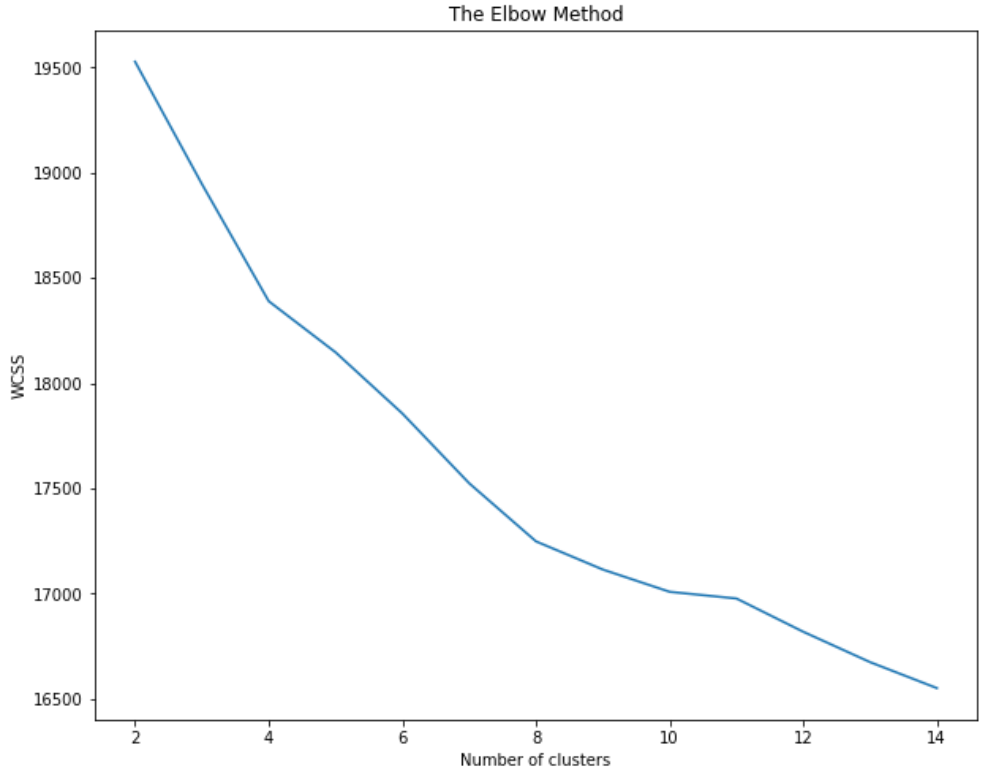
Παρατηρούμε ότι έχουμε μια καλή απόδοση στη διαδικασία ομαδοποίησης με βάση τους δείκτες Davies-Bouldin και Chalinski-Harabaszh. Ο δείκτης Silhouette έχει τιμή περίπου στη μέση του εύρους καλής ομαδοποίησης, γεγονός που φανερώνει ότι υπάρχει κάποια αλληλοεπικάλυψη των συστάδων, η οποία όμως δεν είναι απαγορευτική στην περίπτωσή μας καθώς είναι λογικό να υπάρχουν προφίλ ασθενών που το υποσύνολο τυχών ασθενειών ή χαρακτηριστικών να εντάσσεται σε περισσότερες από μία περιπτώσεις ενώ θεωρείται εξίσου χρήσιμη μια ομαδοποίηση που θα κατατάσσει κάποιον χρήστη σε πολλαπλές ομάδες.

Τέλος, εφαρμόζοντας 10-fold cross validation με χρήση του αλγορίθμου ταξινόμησης SVM προκύπτει αναλυτικά για κάθε κύκλο εκπαίδευσης η παρακάτω ακρίβεια πρόβλεψης

[0.99115044 0.99557522 0.9800885 0.99115044
0.99115044 0.99336283 1. 1.
0.99334812 0.99556541]

Η μέση ακρίβεια επομένως είναι 0.9934 (+/- 0.010)

Στη συνέχεια, πραγματοποιούμε κανονικοποίηση των δεδομένων με προσαρμογή όλων των παραμέτρων στο εύρος τιμών [0,1] με χρήση του MinMaxScaler της βιβλιοθήκης scikit-learn. Τα αποτελέσματα που λαμβάνονται αντίστοιχα είναι τα εξής:



Εικόνα 5.2: Μέθοδος του αγκώνα για κανονικοποιημένα δεδομένα

Καθώς η καμπύλη δεν παρουσιάζει ξεκάθαρη εικόνα για την επιλογή του αριθμού των ομάδων, συγκρίνουμε και τις τιμές του συντελεστή Silhouette και επιλέγουμε τη βέλτιστη λύση η οποία προκύπτει επίσης για k=4.

```
Συντελεστής DB = 2.84
Συντελεστής CH = 652.03
Συντελεστής Silhouette = 0.11
```

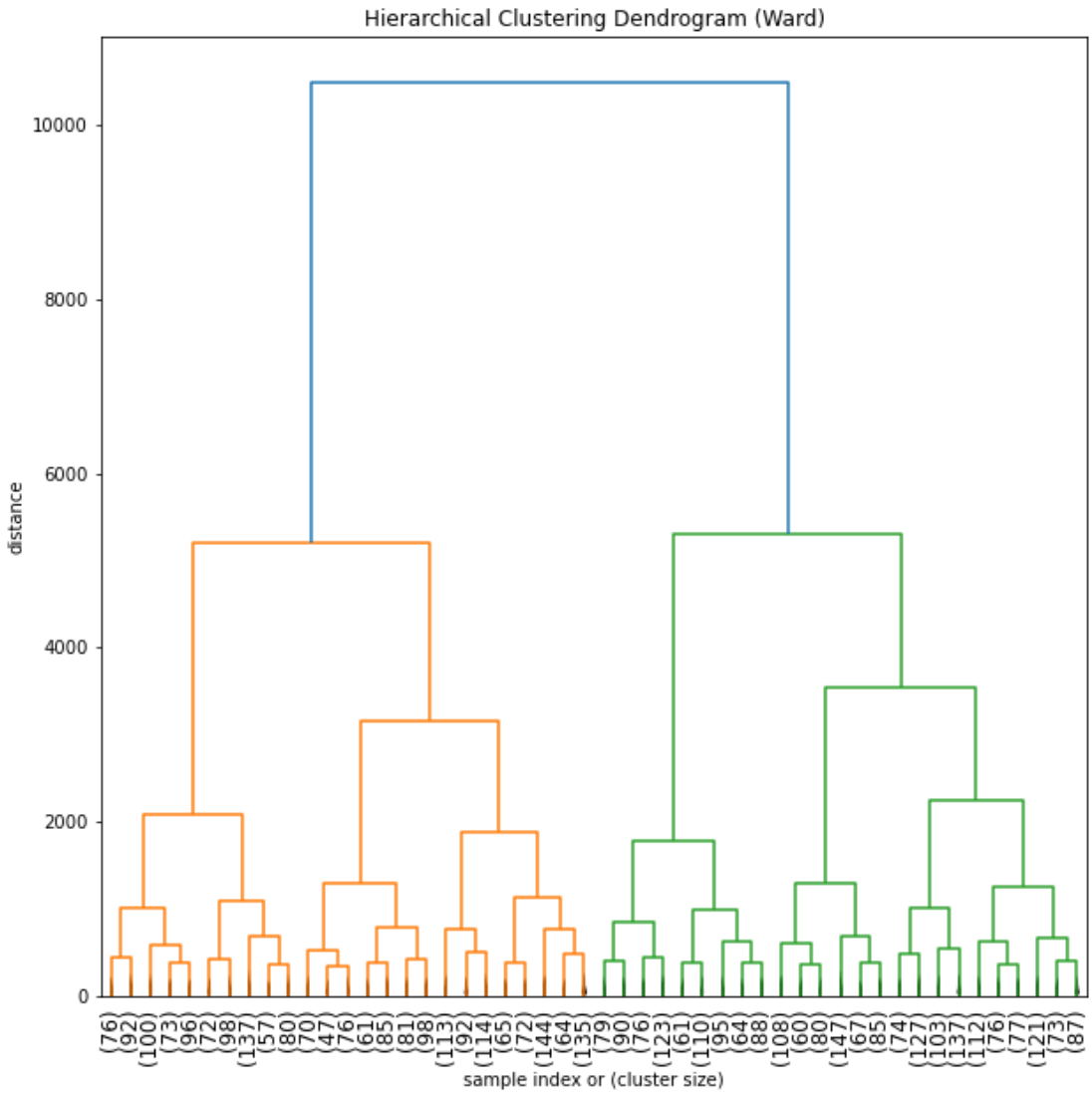
Όσον αφορά την ακρίβεια πρόβλεψης με εφαρμογή 10-fold cross validation και αλγόριθμο ταξινόμησης τον SVM έχουμε:

```
[0.99778761          1                    1                0.99115044
1                    0.99557522        0.99778761       0.99778761
0.99556541           1]

Μέση ακρίβεια 0.9976 (+/- 0.0054)
```

Η μεταβολή των παραμέτρων του αλγορίθμου και δοκιμές με διαφορετικό αριθμό ομάδων αποφέρουν ελαφρά χειρότερα αποτελέσματα συγκριτικά με τα παραπάνω, ταυτόχρονα όμως συμπεραίνουμε και πάλι ότι είναι εφικτή η ομαδοποίηση με χρήση του αλγορίθμου k-Means και προκύπτουν τέσσερις ομάδες χρηστών.

5.2.2 *Hierachical Agglomerative Clustering*



Εικόνα 5.3: Δενδρογράφημα

Στο δενδρογράφημα που εμφανίζεται παραπάνω, κάθε φύλλο αντιστοιχεί σε ένα ιατρικό προφίλ. Καθώς ανεβαίνουμε το δέντρο, αντικείμενα που είναι παρόμοια μεταξύ τους συνδυάζονται σε κλαδιά, τα οποία συγχωνεύονται σε υψηλότερο ύψος. Το ύψος της σύντηξης, που παρέχεται στον κατακόρυφο άξονα, δείχνει την ανομοιογένεια ή απόσταση μεταξύ δύο αντικειμένων / συστάδων. Όσο υψηλότερο είναι το ύψος αυτό, τόσο λιγότερο παρόμοια είναι τα προφίλ που ανήκουν σε διαφορετικές κλάσεις. Παρατηρούμε ότι η μεγαλύτερη ανομοιογένεια εντοπίζεται στο διαχωρισμό των προφίλ σε δύο κλάσεις, γεγονός που επιβεβαιώνεται και από τους συντελεστές αξιολόγησης που ακολουθούν:

Αριθμός συστάδων (k)	DB (Davies Bouldin Index)	CH (Calinski-Harabasz Index)	SS (Silhouette Score)
2	0.5	13210.76	0.56
3	0.5	14850.47	0.55
4	0.51	20530.79	0.5
5	0.51	20761.53	0.52
6	0.51	33107.1	0.51

Πίνακας 5.1: Μετρικές αξιολόγησης που παρατηρούνται κατά την προσπάθεια ομαδοποίησης των δεδομένων μας με χρήση του αλγορίθμου Hierarchical Agglomerative Clustering

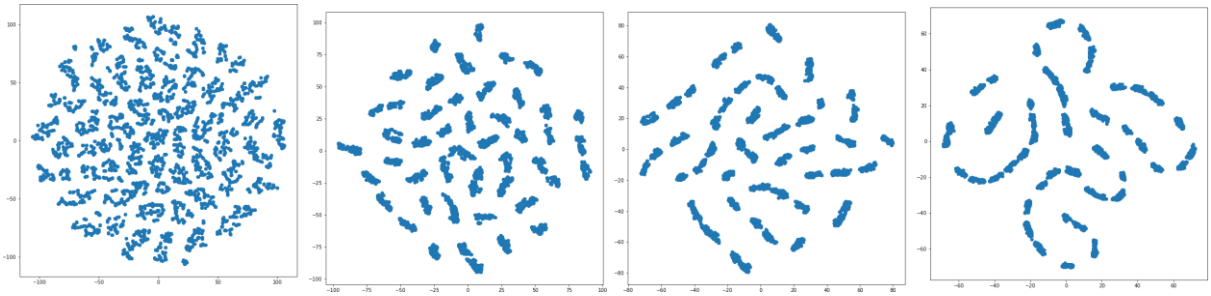
Παράλληλα, η ακρίβεια ταξινόμησης με 10-fold cross validation είναι ίση με:

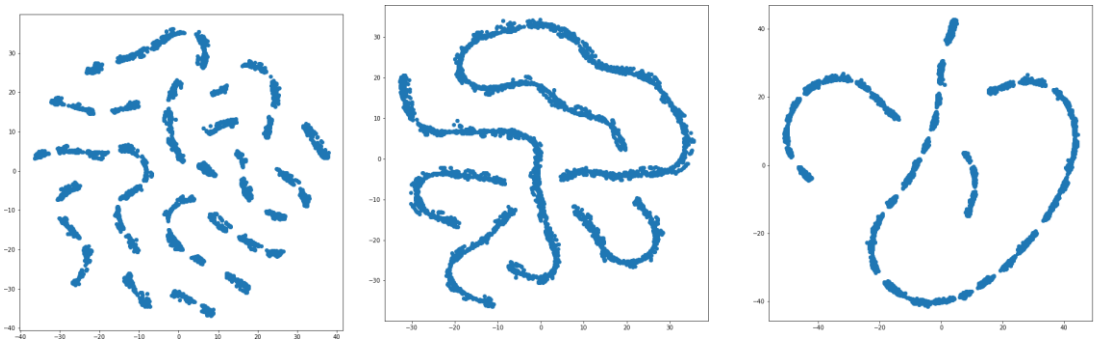
0.9954 (+/- 0.0037)

Τα αποτελέσματα που λαμβάνουμε με κανονικοποίηση των δεδομένων μας δίνουν τον ίδιο αριθμό κλάσεων και παρόμοιες τιμές στις μετρικές αξιολόγησης, με μια μικρή χειροτέρευση των δεικτών όπως και στην περίπτωση του αλγορίθμου k-Means. Στη συγκεκριμένη περίπτωση οι βέλτιστη ομαδοποίηση πραγματοποιείται με αριθμό κλάσεων k=2.

5.2.3 t-SNE

Ο αλγόριθμος t-SNE μας παρέχει μια οπτική αναγνώριση των συστάδων που δημιουργούνται κατά την εκτέλεσή του. Δεν είναι εφικτή η χρήση των μεθόδων αξιολόγησης που περιγράφηκαν νωρίτερα στο κεφάλαιο για το συγκεκριμένο αλγόριθμο καθώς δεν διατηρεί πραγματικές αποστάσεις των σημείων, αλλά τα αποτελέσματα που προκύπτουν είναι ιδιαίτερα χρήσιμα για τον προσδιορισμό του αριθμού των συστάδων που εντοπίζονται σε ένα σύνολο δεδομένων και την επιβεβαίωση της δυνατότητας επιτυχούς ομαδοποίησής του. Όπως προαναφέρθηκε, είναι ιδιαίτερα σημαντική η διερεύνηση της συμπεριφοράς του αλγορίθμου σε μεταβολές της παραμέτρου perplexity. Ακολουθεί μια απεικόνιση ενός υποσυνόλου των αποτελεσμάτων για τιμές της παραμέτρου perplexity στο εύρος 5 έως 150.





Εικόνα 5.4: Οπτικοποιήσεις t-SNE – σταδιακή αύξηση της τιμής του perplexity από 5 έως 150

Τα αποτελέσματα που λαμβάνουμε από την απεικόνιση του αποτελέσματος του t-SNE δεν μας παρέχουν πληροφορίες που να μπορούμε να ερμηνεύσουμε για το συγκεκριμένο σύνολο δεδομένων. Φαίνεται πως η προσπάθεια μείωσης των διαστάσεων δεν δίνει συγκεκριμένο αποτέλεσμα καθώς σε μικρότερες τιμές του perplexity παρατηρείται αρκετός θόρυβος ενώ οι σχηματισμοί που προκύπτουν για μεγαλύτερες τιμές δεν διατηρούν κάποια δομή που να υποδηλώνει ομαδοποίηση. Επομένως δεν μπορούμε να αξιοποιήσουμε το συγκεκριμένο αλγόριθμο στην προσπάθεια ομαδοποίησης ιατρικών προφίλ με τα συγκεκριμένα χαρακτηριστικά ούτε ως ανεξάρτητο εργαλείο αλλά ούτε και σε συνδυασμό με κάποιον άλλο αλγόριθμο ομαδοποίησης.

5.2.4 DBSCAN

Πραγματοποιήσαμε βελτιστοποίηση της παραμέτρου eps του αλγορίθμου DBSCAN δοκιμάζοντας διαφορετικές τιμές για την εκτέλεση του στο σύνολο δεδομένων μας. Ο αλγόριθμος δεν επέστρεψε κανένα μοντέλο ομαδοποίησης σε καμία από αυτές τις δοκιμές. Καθώς ο αριθμός των διαστάσεων που υπάρχουν στο σύνολο δεδομένων μας είναι αρκετά υψηλός, είναι αναμενόμενη η δυσκολία επίτευξης ομαδοποίησης από ορισμένους αλγορίθμους.

5.2.5 Συμπεράσματα

Από το σύνολο των αλγορίθμων που δοκιμάστηκαν στο σύνολο δεδομένων μας, ο αλγόριθμος k-Means και ο αλγόριθμος Hierarchical Agglomerative Clustering έδειξαν δυνατότητα ομαδοποίησης και αποδεκτές τιμές στις μετρικές αξιολόγησης. Ελαφρώς καλύτερη ομαδοποίηση φαίνεται να επιτυγχάνεται με τον αλγόριθμο Hierarchical Agglomerative Clustering ενώ σε αυτή την περίπτωση ο αριθμός συστάδων που προέκυψε

ήταν ίσος με 2. Η κανονικοποίηση φαίνεται να επηρεάζει σε μικρό ποσοστό το αποτέλεσμα της ομαδοποίησης, ενώ η χρήση της προτείνεται καθώς υπάρχουν πολλά μεγέθη σε διαφορετικές κλίμακες στο σύνολο δεδομένων μας τα οποία μπορούν να συσχετιστούν αποτελεσματικότερα όταν αναπροσαρμόζονται στο ίδιο εύρος τιμών.

Καθώς στο πρόβλημα ομαδοποίησης που αντιμετωπίσαμε ήταν πολλά τα χαρακτηριστικά που χρησιμοποιήσαμε, γεγονός που αυξάνει την πολυπλοκότητα του προβλήματος και δυσκολεύει την ομαδοποίηση σε παρόμοιες περιπτώσεις, παρουσιάστηκε αδυναμία ομαδοποίησης και δυσκολία ερμηνείας των αποτελεσμάτων στους αλγορίθμους DBSCAN και t-SNE αντίστοιχα.

5.3 Μελλοντικές Επεκτάσεις

Η παρούσα διπλωματική εργασία θα μπορούσε να επεκταθεί σε αρχικό επίπεδο μέσω διαφορετικής επιλογής των χαρακτηριστικών εισόδου (features) ομαδοποίησης. Έτσι, δύναται να χρησιμοποιηθεί κάποιο υποσύνολο δεδομένων όπως μόνο κωδικοί ιατρικού ιστορικού , ή να εξαχθούν διαφορετικά και πιο σύνθετα μεγέθη από τις υπάρχουσες μετρήσεις με χρήση παραγόμενων ιατρικών ποσοτήτων όπως ο Δείκτης Μάζας Σώματος ή διαφορετικών στατιστικών μεγεθών όπως η διακύμανση. Υπάρχει επιπλέον η δυνατότητα δημιουργίας ομάδων ανά ηλικία ή κάποια συγκεκριμένη πάθηση και η εφαρμογή των τεχνικών που χρησιμοποιήθηκαν σε αυτές τις υπό-ομάδες. Επίσης, δυνατή είναι και η διατήρηση πολύ μεγάλου αριθμού μετρήσεων ανά χρήστη, με επιλογή πιθανόν των πιο πρόσφατων ή όσων παρατηρούνται σε συγκεκριμένη χρονική περίοδο για πιο στοχευμένη διαγνωστική ομαδοποίηση αλλά και εφικτή χρονική πολυπλοκότητα εκτέλεσης των τεχνικών ομαδοποίησης.

Ένας επιπλέον τρόπος μελλοντικής επέκτασης αποτελεί η ενσωμάτωση πιο εμπλουτισμένου συνόλου δεδομένων εισόδου, ώστε να δημιουργηθεί ακόμη πληρέστερο προφίλ χρήστη. Τέτοιου είδους δεδομένα θα μπορούσαν να προέρχονται από wearable συσκευές για συνεχή τροφοδότηση του προφίλ χρήστη με χρήσιμες πληροφορίες όπως βηματομετρητή, θερμίδες που καίγονται ή/και καταναλώνονται, ώρες και ποιότητα ύπνου κ.α. Επιπλέον, θα μπορούσαν να ενσωματωθούν πληροφορίες φαρμακευτικής αγωγής ή άλλου είδους θεραπείας καθώς και κάποιου είδους ερωτηματολόγιο για αξιολόγηση ψυχικής κατάστασης και υγείας. Τέλος, θα ήταν εφικτό να διευρυνθεί και η λίστα των κωδικών ICD-10 και SNOMED CT που χρησιμοποιούνται τόσο σε ποικιλία αλλά και σε εξειδίκευση σε ακόμη πιο λεπτομερή επίπεδα.

Τέλος, προτείνεται περαιτέρω πειραματισμός και με εναλλακτικές μεθόδους ομαδοποίησης καθώς και διαφορετικές μετρικές αξιολόγησης, για να κριθεί η καταλληλότητα μεγαλύτερου αριθμού αλγορίθμων σε πλαίσιο παρόμοιων εφαρμογών.

ΠΑΡΑΡΤΗΜΑ: Κώδικας σε Python

```
1    # import των πακέτων που χρησιμοποιούνται
2    import datetime
3    import numpy as np
4    import pandas as pd
5    import seaborn as sns
6    from sqlalchemy import create_engine
7    import matplotlib.pyplot as plt
8    pd.set_option('display.max_columns', None)
9    pd.set_option('display.max_rows', None)
10   from sklearn.cluster import KMeans
11   import sklearn.metrics as sm
12   # from IPython.core.display import display, HTML
13   from sklearn.metrics.cluster import contingency_matrix
14   from scipy.cluster.hierarchy import linkage, dendrogram
15   from sklearn.cluster import AgglomerativeClustering
16   from sklearn.metrics import pairwise_distances
17   # display(HTML("<style>.container { width:100% !important;
                                  }</style>"))
18   from sklearn.manifold import TSNE
19   from sklearn.metrics import davies_bouldin_score
20   from sklearn.cluster import DBSCAN
21   from sklearn.decomposition import PCA
22   from sklearn.preprocessing import LabelEncoder
23   import os
24   import mysql.connector
25
26
27   print("PLEASE ENTER THE DATABASE PASSWORD")
28   db_password = int(input())
29   #Σύνδεση στη βάση δεδομένων
30   connection = mysql.connector.connect(host = '127.0.0.1',user
                                  ='root', passwd = db_password,
                                  db='patientdata')
31   print(connection)
32
33   #ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ
34
35   #Εκτέλεση SQL query που επιστρέφει PersonId, DateOfBirth,
Gender, EducationalLevel για όλα τα προφίλ χρηστών και αποθήκευση σε
dataframe
```

```
36  df_patient = pd.read_sql_query('''select PersonId, DateOfBirth,
                                    Gender, EducationalLevel from
                                    patientdata.person
37  ''',con=connection)
38
39  #Μετατροπή string DateOfBirth στη μορφή DateTime για περαιτέρω
                                    επεξεργασία
40  df_patient['DateOfBirth'] =
                                    pd.to_datetime(df_patient.DateOfBir
                                    th)
41
42  #Ορισμός συνάρτησης που μετατρέπει την ημερομηνία γέννησης σε
                                    ηλικία (ακέραιο αριθμό)
43  def calculate_age(born):
44      born = datetime.datetime.strptime(str(born), "%Y-%m-%d
                                    %H:%M:%S").date()
45      today = datetime.date.today()
46      return today.year - born.year - ((today.month, today.day) <
                                    (born.month, born.day))
47
48  #Εκτέλεση συνάρτησης για όλα τα rows και αποθήκευση σε νέα στήλη
49  df_patient['Age'] = df_patient['DateOfBirth'].apply(lambda x:
                                    calculate_age(x))
50  #Διαγραφή στήλης DateOfBirth
51  df_patient = df_patient.drop('DateOfBirth', 1)
52  print(df_patient.head())
53  #Αρχικοποίηση στήλης IcdCode με άδεια λίστα
54  df_patient['IcdCodes'] =  np.empty((len(df_patient),
                                    0)).tolist()
55
56  #Συνάρτηση που επιστρέφει τα IcdCodes ανά χρήστη
57  def icd_by_person(pid):
58      tmp_list = pd.read_sql_query('''select IcdCode from
                                    patientdata.medicalhistory where
                                    patientdata.medicalhistory.PersonId
                                    = %s ''',  con=connection, params =
                                    [pid])
59      return list(tmp_list.values.flatten())
60  #Εκτέλεση συνάρτησης για όλα τα rows και αποθήκευση του
                                    αποτελέσματος
61  df_patient['IcdCodes'] = df_patient['PersonId'].apply(lambda x:
                                    icd_by_person(x))
62
63  #OneHotEncoding για τη στήλη IcdCodes
64  df_icd_code_patient= pd.read_sql_query('''select distinct
                                    IcdCode, PersonId from
                                    patientdata.medicalhistory''',
                                    con=connection )
65  df_icd_multicoded =
                                    pd.crosstab(df_icd_code_patient['Pe
                                    rsonId'],df_icd_code_patient['IcdCo
                                    de']).rename_axis(None,axis=1)
66  #Ενσωμάτωση One Hot Encoded στήλης στο αρχικό DataFrame
67  df_patient_icd=pd.merge(df_patient,df_icd_multicoded, on
                                    ="PersonId")
68
```

```
69  # #Εκτέλεση SQL query που επιστρέφει τα distinct snomed ids που
                                    υπάρχουν στη βάση
70  df_snomed_ids = pd.read_sql_query('''select distinct SnomedId
                                    from patientdata.measurement
71  ''',con=connection)
72
73  #Δημιουργία στήλης για κάθε distinct snomed id και αρχικοποίηση
                                    με την τιμή 0
74  for col in df_snomed_ids.values.flatten():
75      df_patient[col] = 0
76  #Ορισμός συνάρτησης που επιστρέφει τη μέση τιμή ανά Snomed Id
                                    και ανά χρήστη
77  def measurement_median_by_person(df_patient):
78      i=0
79      for pid in df_patient['PersonId'].values.flatten():
80          pid = str(pid)
81          for mid in df_snomed_ids.values.flatten():
82              tmp_meas_list = pd.read_sql_query('''select Value
                                    from patientdata.measurement where
                                    patientdata.measurement.PersonId =
                                    %s AND
                                    patientdata.measurement.SnomedId =
                                    %s LIMIT 1''',  con=connection,
                                    params = [pid, mid])
83              df_patient.loc[i,'{}'.format(mid)] =
                                    tmp_meas_list.mean(axis =
                                    0).get('Value')
84          i = i + 1
85          if i%50==0:
86                  df_patient.to_pickle("./dataset_full.pkl")
87                  #Αποθήκευση αποτελεσμάτων ανά 50 χρήστες
88      return df_patient
89  measurement_median_by_person(df_patient)
90  # Αφαίρεση της στήλης 6012004(hearing aid), του Person Id και
                                    της λίστας των IcdCodes
91  df_patient = df_patient.drop(['6012004', 'PersonId','IcdCodes'],
                                    axis=1)
92
93  #Label Encoding στο φύλο και στο επίπεδο μόρφωσης
94  from sklearn import preprocessing
95  le = preprocessing.LabelEncoder()
96  le.fit(['male', 'female'])
97  df_patient['Gender'] = le.transform(df_patient['Gender'])
98  le.fit(['primary', 'secondary', 'college'])
99  df_patient['EducationalLevel'] =
                                    le.transform(df_patient['Educationa
                                    lLevel'])
100 df_patient.to_pickle("./dataset_full.pkl")
101
102 # ΜΕΘΟΔΟΣ ΤΟΥ ΑΓΚΩΝΑ
103
104 # Προαιρετική εφαρμογή κανονικοποίησης με MinMaxScaler
105 # from sklearn.preprocessing import MinMaxScaler
106 # scaler = MinMaxScaler(feature_range=(0, 1))
107 # df_kmeans = pd.DataFrame(scaler.fit_transform(df_patient))
```

```
108
109 #Εισαγωγή προεπεξεργασμένου dataset σε DataFrame εκ νέου
110 df_kmeans = pd.read_pickle("./dataset_full.pkl")
111 plt.figure(figsize=(10, 8))
112 wcss = []
113 for i in range(2, 15):
114     kmeans = KMeans(n_clusters = i, init = 'k-means++',
                                random_state = 42)
115     kmeans.fit(df_kmeans)
116     wcss.append(kmeans.inertia_)
117 plt.plot(range(2, 15), wcss)
118 plt.title('The Elbow Method')
119 plt.xlabel('Number of clusters')
120 plt.ylabel('WCSS')
121 plt.show()
122
123 #K-MEANS
124 #Ορισμός παραμέτρων εκτέλεσης αλγορίθμου
125 kmeans = KMeans(n_clusters = 4, init = 'k-means++', random_state
                                = 42)
126 #Εφαρμογή του μοντέλου
127 y_kmeans = kmeans.fit_predict(df_kmeans)
128 #Αρίθμηση των ομάδων από το 1 αντί για το 0
129 y_kmeans1=y_kmeans
130 y_kmeans1=y_kmeans+1
131 # Ορισμός νέου dataframe με το cluster που προέκυψε
132 cluster = pd.DataFrame(y_kmeans1)
133 # Προσθήκη του αποτελέσματος ομαδοποίησης κάθε προφίλ σε νέα
                                στήλη
134 df_kmeans['cluster'] = cluster
135 #Mean of clusters
136 kmeans_mean_cluster =
                                pd.DataFrame(round(df_kmeans.groupb
                                y('cluster').mean(),1))
137 kmeans_mean_cluster
138 #score
139 #Υπολογισμός δεικτών αξιολόγησης
140 labels = kmeans.labels_
141 dbs=davies_bouldin_score(df_kmeans, labels)
142 dbs=round(dbs,2)
143 ch = sm.calinski_harabasz_score(df_kmeans, labels)
144 ch=round(ch,2)
145 ss=sm.silhouette_score(df_kmeans, labels, metric='euclidean')
146 ss=round(ss,2)
147 print(" db score-", dbs, "|", " ch score - ", ch, "|", " ss
                                score - ", ss)
148
149 #10-FOLD CROSS VALIDATION
150 import numpy as np
151 from sklearn.model_selection import train_test_split
152 from sklearn import svm
153 #Εισοδος τα features
154 X = df_kmeans.iloc[:,
                                [0,1,2,3,4,5,6,7,8,9,10,11,12,13,14
```

```
                                    ,15,16,17,18,19,20,21,22,23,24,25,2
                                    6]]
#Εξοδος/πρόβλεψη το label της ομαδοποίησης
y = df_kmeans.iloc[:, 27]
X.shape, y.shape
from sklearn.model_selection import cross_val_score
clf = svm.SVC(kernel='linear', C=1)
scores = cross_val_score(clf, X, y, cv=10)
print(scores)
print("Accuracy: %0.4f (+/- %0.4f)" % (scores.mean(),
                                    scores.std() * 2))

#HIERARCHICAL AGGLOMERATIVE CLUSTERING
#Εισαγωγή προεπεξεργασμένου dataset σε DataFrame εκ νέου
df_hac = pd.read_pickle("./dataset_full.pkl")
mergings = linkage(df_hac, method='complete')
figure = plt.figure(figsize=(10, 10))
#Ορισμός παραμέτρων εκτέλεσης αλγορίθμου
dendrogram(
    mergings,
    truncate_mode='lastp',  # show only the last p merged
                                    clusters
    p=50,  # show only the last p merged clusters
    leaf_rotation=90.,
    leaf_font_size=12.,
    show_contracted=True,  # to get a distribution impression in
                                    truncated branches
)
plt.title('Hierarchical Clustering Dendrogram (Ward)')
plt.xlabel('sample index or (cluster size)')
plt.ylabel('distance')
plt.show()
for k in range(2,50):
    cluster = AgglomerativeClustering(n_clusters=k,
                                    affinity='euclidean',
                                    linkage='ward').fit(df_hac)
    labels=cluster.labels_
    #Υπολογισμός δεικτών αξιολόγησης
    dbs=davies_bouldin_score(df_hac, labels)
    dbs=round(dbs,2)
    ch = sm.calinski_harabasz_score(df_hac, labels)
    ch=round(ch,2)
    ss=sm.silhouette_score(df_hac, labels, metric='euclidean')
    ss=round(ss,2)
    print("Cluster count-", k, "|", " db score-", dbs, "|", " ch
                                    score - ", ch, "|", " ss score - ",
                                    ss)
    k=str(k)
    # Προσθήκη του αποτελέσματος ομαδοποίησης κάθε προφίλ για κ
                                    ομάδες  σε νέα στήλη
    df_hac['cluster'+k]=labels

#T-SNE
df_tsne = pd.read_pickle("./dataset_full.pkl")
```

```
#Ορισμός παραμέτρων εκτέλεσης αλγορίθμου
model = TSNE(learning_rate=10, perplexity=50)
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0, 1))
df_tsne = pd.DataFrame(scaler.fit_transform(df_patient))
transformed = model.fit_transform(df_tsne)
print(transformed)

x_axis = transformed[:, 0]
y_axis = transformed[:, 1]
plt.figure(figsize=(10,10))
plt.scatter(x_axis, y_axis)

plt.show()

#DBSCAN
#Εισαγωγή προεπεξεργασμένου dataset σε DataFrame εκ νέου
df_dbscan = pd.read_pickle("./dataset_full.pkl")
#Ορισμός παραμέτρων εκτέλεσης αλγορίθμου
dbscan = DBSCAN(eps=111.5, min_samples=5, metric='euclidean',
                                metric_params=None,
                                algorithm='auto', leaf_size=30,
                                p=None, n_jobs=None)
db = dbscan.fit(df_dbscan)
labels = db.labels_
np.unique(labels)
#score
labels = db.labels_
#Υπολογισμός δεικτών αξιολόγησης
dbs=davies_bouldin_score(df_dbscan, labels)
dbs=round(dbs,2)
ch = sm.calinski_harabasz_score(df_dbscan, labels)
ch=round(ch,2)
print(" db score-", dbs, "|", " ch score - ", ch, "|")
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Tim Wilsdon, Anthony Barron, Guy Edwards, Ryan Lawlor. The benefits of personalised medicine to patients, society and healthcare systems. European Biopharmaceutical Enterprises (EBE), and the European Federation of Pharmaceutical Industries and Associations (EFPIA). July 2018
2. Mettler, Tobias. Explorative Clustering of Clinical User Profiles: A First Step towards User-centered Health Information Systems. ECIS 2013 - Proceedings of the 21st European Conference on Information Systems. 2013
3. Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Front Comput Neurosci. 2019;13:31. Published 2019 May 24
4. Christian Lopez , Scott Tucker, Tarik Salameh, Conrad Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. Journal of Biomedical Informatics. Volume 85, Pages 30-39. September 2018
5. Claire Mawditt, Kiriko Sasayama, Kota Katanoda, Stuart Gilmour. The clustering of health-related behaviours in the adult Japanese population. Journal of epidemiology. July 2020
6. Sophia R. Newcomer, MPH , John F. Steiner, MD, MPH , Elizabeth A. Bayliss. Identifying Subgroups of Complex Patients With Cluster Analysis. MSPH Volume 17, Issue 8. August 9, 2011
7. Sabine I. Vuik, Erik Mayer & Ara Darzi. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. Population Health Metrics. December 2016.
8. Busch, V., Van Stel, H.F., Schrijvers, A.J. et al. Clustering of health-related behaviors, health outcomes and demographics in Dutch adolescents: a cross-sectional study. BMC Public Health 13, 1118 (2013).

9. Triantafyllidis AK, Tsanas A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature J Med Internet Res 2019;21(4):e12286 URL: https://www.jmir.org/2019/4/e12286 DOI: 10.2196/12286 PMID: 30950797 PMCID: 6473205
10. Κατερίνα Γεωργούλη. Τεχνητή Νοημοσύνη. Μια Εισαγωγική Προσέγγιση, Κεφάλαιο 4. 2015.
11. Tom M. Mitchell. Machine Learning. Σελίδες 230-233 & 307-325. 1997.
12. Mariette Awad, Rahul Khanna. Efficient Learning Machines – Theories, Concepts, and Applications for Engineers and System Designers , Κεφάλαιο 1. Απρίλιος 2015
13. https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/ Πρόσβαση 22 Αυγούστου 2020
14. Μηνάς Καραολής. Εξόρυξη Γνώσης Με Εξαγωγή Κανόνων Σε Καρδιαγγειακές Βάσεις Δεδομένων, Σελ. 94. Πανεπιστήμιο Κύπρου. 2010.
15. Δονάτος Παπανικολάου. Εφαρμογή Τεχνικών Εξόρυξης Γνώσης στην Εκπαίδευση. Σελ. 62-63. Πανεπιστήμιο Πατρών. 2015.
16. Γεροθανάσης Εμμανουήλ, Μπέκος Ευάγγελος. Κατασκευή ταξινομητών weighted kNN με metric ball trees για εφαρμογές ανακάλυψης γνώσης από βάσεις δεδομένων Oracle, Σελ. 18. ΤΕΙ Κεντρικής Μακεδονίας. 2012.
17. Κωσιωρή Αφροδίτη. Μελέτη τεχνικών, μεθοδολογιών και εφαρμογών στον τομέα της εξόρυξης γνώσης από δεδομένα, σελ. 91-93. ΤΕΙ Καβάλας. 2013.
18. Κύρκος Ευστάθιος. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. Κεφάλαιο 10. 2015.
19. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining (Third Edition) The Morgan Kaufmann Series in Data Management Systems. Σελ. 393-442, 486-487. 2012.
20. Παύλος Μαθιουδάκης, Γιοβάννα Τόσιτς. Ευφυές σύστημα πρόβλεψης τιμών μετοχών. Σελ. 29-30. ΤΕΙ Κρήτης. Σεπτέμβριος 2016.
21. Ryan P. Adams. Hierarchical Agglomerative Clustering. Proc. Ninth SIAM Data Mining Conf. (SDM ’09), Σελ. 510-516. 2009.
22. Laurens van der Maaten & Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research, 9 (Nov) : 2579–2605, 2008
23. Laurens van der Maaten. "Barnes-Hut-SNE" Pattern Recognition and Bioinformatics Group, Delft University of Technology Mekelweg 4, 2628 CD Delft, The Netherlands
24. Wattenberg, Martin; Viégas, Fernanda; Johnson, Ian How to Use t-SNE Effectively. 13-10-2016.

25. M.Ester, H.–P.Kriegel, J.Sander, X.XU. A Density–Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD vol.2, pp.226-231. 1996.
26. ICD-10: Διεθνής Στατιστική Ταξινόμηση Νόσων και Συναφών Προβλημάτων Υγείας. Τόμος 2: Εγχειρίδιο Οδηγιών Παγκόσμιος Οργανισμός Υγείας. Γενεύη. Δέκατη Αναθεώρηση Έκδοση. 2008
27. Τσικρίκα Αθηνά. ICD10 vs SNOMED CT: Συστήματα προτυποποίησης πληροφοριών και ανταλλαγής πληροφοριών στον τομέα της υγείας. ΤΕΙ Πελοποννήσου. 2018
28. Ιατρικά Δεδομένα και Πρότυπα. https://repository.kallipos.gr/bitstream/11419/2977/1/02_chapter_02.pdf Πρόσβαση 10 Οκτωβρίου 2020
29. https://scikit-learn.org/stable/ Πρόσβαση 10 Οκτωβρίου 2020
30. https://pandas.pydata.org/about/index.html Πρόσβαση 10 Οκτωβρίου 2020
31. https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html Πρόσβαση 10 Οκτωβρίου 2020
32. Αδαμόπουλος Λεωνίδας, Δαμιανού Χαράλαμπος, Σβέρκος Ανδρέας. Μαθηματικά και Στοιχεία Στατιστικής (Γ Λυκείου Γενικής Παιδείας) – Ενότητα 2.1. 2011
33. Potdar, Kedar & Pardawala, Taher & Pai, Chinmay. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications. 2017
34. Hancock, J.T., Khoshgoftaar, T.M. Survey on categorical data for neural networks. J Big Data 7, 28 . 2020
35. Julio-Omar Palacio-Nino. Evaluation Metrics for Unsupervised Learning Algorithms. Dept. Computer Science and Artificial Intelligence. Universidad de Granada. 14 May 2019
36. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation Πρόσβαση 10 Οκτωβρίου 2020
37. Tarekegn, A.N., Michalak, K. & Giacobini, M. Cross-Validation Approach to Evaluate Clustering Algorithms: An Experimental Study Using Multi-Label Datasets. *SN COMPUT. SCI.* **1,** 263 . 2020