

Ατομική Διπλωματική Εργασία

**ΑΝΑΛΥΣΗ ΙΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΗΓΗΣΙΜΗ
ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ**

Φλωρεντία Παττίδου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάλυση ιατρικών δεδομένων και εξηγήσιμη τεχνητή νοημοσύνη

Φλωρεντία Παττίδου

Επιβλέπων Καθηγητής

Δρ. Κωνσταντίνος Παττίχης

Δηλώνω ότι η παρούσα διατριβή και οποιοδήποτε συνοδευτικό λογισμικό αποτελούν δικό μου πρωτότυπο έργο, εκπονημένο σε πλήρη συμμόρφωση με τον Κώδικα Δεοντολογίας για την Έρευνα του Πανεπιστημίου Κύπρου (<https://www.ucy.ac.cy/legislation/kodikas-deontologias-kai-politikes/>), και με πλήρη ευθύνη εκ μέρους μου για την ακρίβεια, την ακεραιότητα και την εγκυρότητα όλου του περιεχομένου.

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου.

Μάιος 2026

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τον επιβλέποντα καθηγητή μου, Δρ. Κωνσταντίνο Παττίχη, για την καθοδήγηση, την υποστήριξη και τις πολύτιμες συμβουλές του καθ' όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Επιπλέον, θα ήθελα να ευχαριστήσω τις Δρα Ειρήνη Σχίζα και Δρα Νικολέττα Πρέντζα για την πολύτιμη καθοδήγηση τους και την βοήθεια τους.

Ιδιαίτερες ευχαριστίες απευθύνω σε όλους όσοι συνέβαλαν με οποιονδήποτε τρόπο στην ολοκλήρωση της εργασίας αυτής, είτε μέσω επιστημονικής καθοδήγησης είτε μέσω παροχής δεδομένων και τεχνικής υποστήριξης.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την κατανόηση, την υπομονή και τη συνεχή στήριξή τους καθ' όλη τη διάρκεια των σπουδών μου.

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στην ανάλυση ιατρικών δεδομένων ασθενών με στένωση καρωτίδας, με στόχο την αναγνώριση προτύπων που σχετίζονται με τον κίνδυνο εμφάνισης εγκεφαλικού επεισοδίου. Η μελέτη βασίζεται σε δεδομένα που προέρχονται από υπερηχογραφικά χαρακτηριστικά πλακών και κλινικές μεταβλητές, όπως ο βαθμός στένωσης (STENOSIS), το Gray Scale Median (GSM), η επιφάνεια πλάκας (PLAREA), η παρουσία discrete white areas (DWA), ο δείκτης Juxtaluminal Black Area (JBA), καθώς και το ιστορικό παροδικών ισχαιμικών επεισοδίων ή/και εγκεφαλικού στο ετερόπλευρο ημισφαίριο (CTIASTR), το οποίο αποτελεί σημαντικό κλινικό παράγοντα κινδύνου.

Για την ανάλυση των δεδομένων εφαρμόστηκαν οι αλγόριθμοι ομαδοποίησης K-Modes, ο οποίος είναι κατάλληλος για κατηγορικά δεδομένα και ο αλγόριθμος K-Prototypes που συνδυάζει τόσο κατηγορικά, όσο και αριθμητικά δεδομένα. Η επιλογή του βέλτιστου αριθμού συστάδων για τους 2 αλγόριθμους πραγματοποιήθηκε με τη χρήση της καμπύλης elbow και δεικτών αξιολόγησης, όπως οι δείκτες Silhouette, Davies-Bouldin και Calinski-Harabasz. Η ανάλυση πραγματοποιήθηκε τόσο στο σύνολο των ασθενών όσο και σε επιμέρους υποσύνολα, όπως οι συμπτωματικοί και οι ασυμπτωματικοί ασθενείς.

Επιπλέον, για την ενίσχυση της ερμηνευσιμότητας των αποτελεσμάτων εφαρμόστηκε η τεχνική εξηγήσιμης τεχνητής νοημοσύνης (Explainable Artificial Intelligence – XAI) TE2Rules, με στόχο την εξαγωγή κανόνων που περιγράφουν με σαφήνεια τα χαρακτηριστικά κάθε συστάδας.

Τα αποτελέσματα της ανάλυσης ανέδειξαν διακριτές ομάδες ασθενών με διαφορετικά μορφολογικά προφίλ και επίπεδα κινδύνου, ενώ η χρήση ερμηνεύσιμων μοντέλων συνέβαλε ουσιαστικά στην κατανόηση των βασικών παραγόντων που επηρεάζουν την κατηγοριοποίηση των ασθενών. Τα ευρήματα της εργασίας μπορούν να συμβάλουν στη βελτίωση της αξιολόγησης κινδύνου και στην υποστήριξη της λήψης κλινικών αποφάσεων.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ.....	VI
ΠΕΡΙΕΧΟΜΕΝΑ	VII
ΣΧΗΜΑΤΑ	XI
ΠΙΝΑΚΕΣ	XV
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ.....	1
1.1 ΚΛΙΝΙΚΟ ΥΠΟΒΑΘΡΟ	2
1.2 ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ ΣΤΗΝ ΙΑΤΡΙΚΗ	4
1.3 ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ	6
1.4 ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ	8
ΚΕΦΑΛΑΙΟ 2: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΙΑΤΡΙΚΗ.....	11
2.1 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ	12
2.2 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΤΗΝ ΙΑΤΡΙΚΗ	15
2.3 ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ (CLUSTER ANALYSIS)	17
2.4 ΑΛΓΟΡΙΘΜΟΣ K-MODES.....	20
2.5 ΑΛΓΟΡΙΘΜΟΣ K-PROTOTYPES.....	23
2.6 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ	26
2.6.1 Δείκτης Silhouette	26
2.6.2 Δείκτης Davies–Bouldin	27
2.6.3 Δείκτης Calinski–Harabasz	28
2.6.4 Δείκτης Adjusted Rand Index (ARI)	28
2.6.5 Μεθοδολογική προσέγγιση αξιολόγησης στην παρούσα εργασία	30
2.7 ΕΞΗΓΗΣΙΜΗ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ (EXPLAINABLE ARTIFICIAL INTELLIGENCE – XAI)	31
2.8 TE2RULES	34
ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΛΟΓΙΑ.....	37

3.1	ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ	38
3.2	ΔΙΑΓΡΑΜΜΑ ΡΟΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ	41
3.3	ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ	44
3.4	ΔΙΑΧΩΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ.....	46
3.5	ΜΕΘΟΔΟΣ CLUSTERING (K-MODES & K-PROTOTYPES)	47
3.6	ΕΠΙΛΟΓΗ ΑΡΙΘΜΟΥ ΣΥΣΤΑΔΩΝ.....	50
3.7	ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ	53
3.8	ΜΕΘΟΔΟΣ TE2RULES.....	55
ΚΕΦΑΛΑΙΟ 4: ΑΠΟΤΕΛΕΣΜΑΤΑ		58
4.1	ΣΥΝΟΛΙΚΟ ΣΥΝΟΛΟ ΑΣΘΕΝΩΝ.....	59
4.1.1	<i>Αποτελέσματα ομαδοποίησης στο σύνολο των ασθενών με τον αλγόριθμο K-Modes ...</i>	<i>59</i>
4.1.2	<i>Αποτελέσματα κανόνων TE2Rules στο σύνολο των ασθενών με τον αλγόριθμο K-Modes</i>	<i>66</i>
4.1.3	<i>Αποτελέσματα ομαδοποίησης στο σύνολο ασθενών με τον αλγόριθμο K-Prototypes....</i>	<i>67</i>
4.1.4	<i>Αποτελέσματα κανόνων TE2Rules στο σύνολο ασθενών με τον αλγόριθμο K-Prototypes</i>	<i>75</i>
4.2	ΣΥΝΟΛΟ ΣΥΜΠΤΩΜΑΤΙΚΩΝ ΑΣΘΕΝΩΝ	76
4.2.1	<i>Αποτελέσματα ομαδοποίησης στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K -Modes</i>	<i>76</i>
4.2.2	<i>Αποτελέσματα κανόνων TE2Rules στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K-Modes</i>	<i>84</i>
4.2.3	<i>Αποτελέσματα ομαδοποίησης στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K – Prototypes</i>	<i>86</i>
4.2.4	<i>Αποτελέσματα κανόνων TE2Rules στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K-Prototypes</i>	<i>95</i>
4.3	ΣΥΝΟΛΟ ΑΣΥΜΠΤΩΜΑΤΙΚΩΝ ΑΣΘΕΝΩΝ	97
4.3.1	<i>Αποτελέσματα ομαδοποίησης στο σύνολο των ασυμπτωματικών ασθενών με τον αλγόριθμο K-Modes</i>	<i>97</i>
4.3.2	<i>Αποτελέσματα κανόνων με TE2Rules στο σύνολο των ασυμπτωματικών ασθενών με K-Modes.....</i>	<i>105</i>

4.3.3 Αποτελέσματα ομαδοποίησης στο σύνολο των ασυμπτωματικών ασθενών με τον αλγόριθμο K-Prototypes	106
4.3.4 Αποτελέσματα κανόνων TE2Rules στο σύνολο των ασυμπτωματικών ασθενών με τον αλγόριθμο K-Prototypes	115

ΚΕΦΑΛΑΙΟ 5: ΑΝΑΛΥΣΗ ΚΑΙ ΣΥΖΗΤΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

116

5.1 ΣΥΝΟΛΙΚΗ ΕΡΜΗΝΕΙΑ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΣΤΑΔΩΝ	117
5.2 ΑΝΑΛΥΣΗ ΜΟΡΦΟΛΟΓΙΚΩΝ ΠΡΟΤΥΠΩΝ ΣΤΟ ΣΥΝΟΛΟ ΤΩΝ ΑΣΘΕΝΩΝ	119
5.2.1 Αποτελέσματα K-Modes στο σύνολο των ασθενών	119
5.2.2 Αποτελέσματα K-Prototypes στο σύνολο των ασθενών	121
5.2.3 Σύγκριση K-Modes & K-Prototypes στο σύνολο ασθενών.....	124
5.3 ΑΝΑΛΥΣΗ ΜΟΡΦΟΛΟΓΙΚΩΝ ΠΡΟΤΥΠΩΝ ΣΤΟ ΣΥΝΟΛΟ ΤΩΝ ΣΥΜΠΤΩΜΑΤΙΚΩΝ ΑΣΘΕΝΩΝ	126
5.3.1 Αποτελέσματα K-Modes στους συμπτωματικούς ασθενείς.....	126
5.3.2 Αποτελέσματα K-Prototypes στους συμπτωματικούς ασθενείς.....	129
5.3.3 Σύγκριση K-Modes & K-Prototypes στους συμπτωματικούς ασθενείς.....	132
5.4 ΑΝΑΛΥΣΗ ΜΟΡΦΟΛΟΓΙΚΩΝ ΠΡΟΤΥΠΩΝ ΣΤΟ ΣΥΝΟΛΟ ΤΩΝ ΑΣΥΜΠΤΩΜΑΤΙΚΩΝ ΑΣΘΕΝΩΝ	134
5.4.1 Αποτελέσματα K-Modes στους ασυμπτωματικούς ασθενείς.....	134
5.4.2 Αποτελέσματα K-Prototypes στους ασυμπτωματικούς ασθενείς.....	137
5.4.3 Σύγκριση K-Modes & K-Prototypes στους ασυμπτωματικούς ασθενείς.....	140
5.5 ΣΥΓΚΡΙΣΗ ΜΕΤΑΞΥ ΣΥΜΠΤΩΜΑΤΙΚΩΝ ΚΑΙ ΑΣΥΜΠΤΩΜΑΤΙΚΩΝ ΑΣΘΕΝΩΝ	142
5.6 ΕΡΜΗΝΕΙΑ ΤΗΣ ΕΠΙΚΑΛΥΨΗΣ ΤΩΝ ΣΥΣΤΑΔΩΝ	145
5.7 ΣΥΝΔΕΣΗ ΜΕ ΤΗ ΒΙΒΛΙΟΓΡΑΦΙΑ	147

ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

150

6.1 ΣΥΝΟΛΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ ΤΗΣ ΕΡΓΑΣΙΑΣ.....	151
6.2 ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ.....	154
6.3 ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ	157

ΒΙΒΛΙΟΓΡΑΦΙΑ.....I

ΠΑΡΑΡΤΗΜΑ Α.....	1
ΠΑΡΑΡΤΗΜΑ Β.....	1
Α.1	1
Β.1	2
ΠΑΡΑΡΤΗΜΑ Γ	1

ΣΧΗΜΑΤΑ

Σχήμα 2.1 Σχηματική αναπαράσταση βασικών κατηγοριών μηχανικής μάθησης.	14
Σχήμα 2.2.1 Σχηματική αναπαράσταση της μη επιβλεπόμενης μάθησης.	16
Σχήμα 2.3.1 Γραφική απεικόνιση της διαδικασίας ανάλυσης συστάδων, όπου τα δεδομένα ομαδοποιούνται σε διακριτές περιοχές του χώρου χαρακτηριστικών.	19
Σχήμα 2.7.1: Ρόλος της εξηγήσιμης τεχνητής νοημοσύνης (XAI) στον τομέα της υγείας και η αλληλεπίδρασή της με τους βασικούς εμπλεκόμενους φορείς.	33
Σχήμα 3.2.1 Διάγραμμα ροής των διαδοχικών σταδίων της προτεινόμενης μεθοδολογίας (Data Preparation, Clustering Process, Evaluation, Explainability, Results Analysis).	41
Σχήμα 3.6.1 Καμπύλη elbow για την επιλογή του αριθμού k	50
Σχήμα 3.6.2 Μετρικές ποιότητας clustering ως συνάρτηση του k.....	51
Σχήμα 4.1.1 Καμπύλη elbow για τον αλγόριθμο K-Modes στο σύνολο ασθενών	59
Σχήμα 4.1.2 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο ασθενών K-Modes.....	60
Σχήμα 4.1.3 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο ασθενών K-Modes.....	60
Σχήμα 4.1.4 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο ασθενών K-Modes	61
Σχήμα 4.1.5 Κατανομή των ασθενών ανά συστάδα στο σύνολο ασθενών K-Modes.....	62
Σχήμα 4.1.6 Κατανομή CTIASTR ανά συστάδα στο σύνολο ασθενών K-Modes	62
Σχήμα 4.1.7 Κατανομή DWA ανά συστάδα στο σύνολο ασθενών K-Modes	63
Σχήμα 4.1.8 Κατανομή GSM ανά συστάδα στο σύνολο ασθενών K-Modes.....	63
Σχήμα 4.1.9 Κατανομή JBA ανά συστάδα στο σύνολο ασθενών K-Modes	64
Σχήμα 4.1.10 Κατανομή PLAREA ανά συστάδα στο σύνολο ασθενών K-Modes	64
Σχήμα 4.1.11 Κατανομή STENOSIS ανά συστάδα στο σύνολο ασθενών K-Modes.....	65
Σχήμα 4.1.12 Καμπύλη elbow για τον αλγόριθμο K-Prototypes στο σύνολο ασθενών	67
Σχήμα 4.1.13 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο ασθενών K-Prototypes.....	68

Σχήμα 4.1.14 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο ασθενών K-Prototypes.	68
Σχήμα 4.1.15 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο ασθενών K-Prototypes	69
Σχήμα 4.1.16 Κατανομή των ασθενών ανά συστάδα στο σύνολο ασθενών K-Prototypes.....	70
Σχήμα 4.1.17 Κατανομή CTIASTR ανά συστάδα στο σύνολο των ασθενών K-Prototypes.....	70
Σχήμα 4.1.18 Κατανομή DWA ανά συστάδα στο σύνολο των ασθενών K-Prototypes.....	71
Σχήμα 4.1.19 Κατανομή GSM ανά συστάδα στο σύνολο των ασθενών K-Prototypes.	71
Σχήμα 4.1.20 Κατανομή JBA ανά συστάδα στο σύνολο των ασθενών K-Prototypes.	72
Σχήμα 4.1.21 Κατανομή PLAREA ανά συστάδα στο σύνολο των ασθενών K-Prototypes.	72
Σχήμα 4.1.22 Κατανομή STENOSIS ανά συστάδα στο σύνολο των ασθενών K-Prototypes.	73
Σχήμα 4.2.1 Καμπύλη elbow για τον αλγόριθμο K-Modes στο σύνολο των συμπτωματικών ασθενών	76
Σχήμα 4.2.2 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των συμπτωματικών ασθενών K-Modes.	77
Σχήμα 4.2.3 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των συμπτωματικών ασθενών K-Modes.	77
Σχήμα 4.2.4 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των συμπτωματικών ασθενών K-Modes.....	78
Σχήμα 4.2.5 Κατανομή των ασθενών ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.....	79
Σχήμα 4.2.6 Κατανομή CTIASTR ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.	79
Σχήμα 4.2.7 Κατανομή DWA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes. ...	80
Σχήμα 4.2.8 Κατανομή GSM ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.....	80
Σχήμα 4.2.9 Κατανομή JBA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.....	81
Σχήμα 4.2.10 Κατανομή PLAREA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.	81
Σχήμα 4.2.11 Κατανομή STENOSIS ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.	82
Σχήμα 4.2.12 Καμπύλη elbow για τον αλγόριθμο K-Prototypes στο σύνολο των συμπτωματικών ασθενών	86

Σχήμα 4.2.13 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των συμπτωματικών ασθενών K-Prototypes.	87
Σχήμα 4.2.14 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των συμπτωματικών ασθενών K-Prototypes.	88
Σχήμα 4.2.15 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των συμπτωματικών ασθενών K-Prototypes.....	89
Σχήμα 4.2.16 Κατανομή των ασθενών ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.....	90
Σχήμα 4.2.17 Κατανομή CTIASTR ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.....	90
Σχήμα 4.2.18 Κατανομή DWA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.	91
Σχήμα 4.2.19 Κατανομή GSM ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.	91
Σχήμα 4.2.20 Κατανομή JBA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.	92
Σχήμα 4.2.21 Κατανομή PLAREA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.....	92
Σχήμα 4.2.22 Κατανομή STENOSIS ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.....	93
Σχήμα 4.3.1 Καμπύλη elbow για τον αλγόριθμο K-Modes στο σύνολο των ασυμπτωματικών ασθενών	97
Σχήμα 4.3.2 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των ασυμπτωματικών ασθενών K-Modes	98
Σχήμα 4.3.3 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των ασυμπτωματικών ασθενών K-Modes	98
Σχήμα 4.3.4 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των ασυμπτωματικών ασθενών K-Modes.....	99
Σχήμα 4.3.5 Κατανομή των ασθενών ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes.....	100
Σχήμα 4.3.6 Κατανομή CTIASTR ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes	100
Σχήμα 4.3.7 Κατανομή DWA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes.	101
Σχήμα 4.3.8 Κατανομή GSM ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes	101

Σχήμα 4.3.9 Κατανομή JBA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes...	102
Σχήμα 4.3.10 Κατανομή PLAREA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes	102
Σχήμα 4.3.11 Κατανομή STENOSIS ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes.....	103
Σχήμα 4.3.12 Καμπύλη elbow για τον αλγόριθμο K-Prototypes στο σύνολο των ασυμπτωματικών ασθενών.	106
Σχήμα 4.3.13 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.	107
Σχήμα 4.3.14 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.	108
Σχήμα 4.3.15 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.....	109
Σχήμα 4.3.16 Κατανομή των ασθενών ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.....	110
Σχήμα 4.3.17 Κατανομή CTIASTR ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.....	110
Σχήμα 4.3.18 Κατανομή DWA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.	111
Σχήμα 4.3.19 Κατανομή GSM ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.	111
Σχήμα 4.3.20 Κατανομή JBA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.	112
Σχήμα 4.3.21 Κατανομή PLAREA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.....	112
Σχήμα 4.3.22 Κατανομή STENOSIS ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.....	113

ΠΙΝΑΚΕΣ

Πίνακας 3.1.1 Αριθμητικές μεταβλητές του συνόλου δεδομένων	39
Πίνακας 3.1.2 Κατηγορικές μεταβλητές που χρησιμοποιήθηκαν στην ανάλυση συστάδων	40
Πίνακας 3.1.3 Μεταβλητές που χρησιμοποιήθηκαν για την αξιολόγηση και ερμηνεία των συστάδων.	40
Πίνακας 4.1.1 Εσωτερικές μετρικές ομαδοποίησης K -modes στο σύνολο ασθενών	65
Πίνακας 4.1.2 Σταθερότητα της ομαδοποίησης K-modes στο σύνολο ασθενών	65
Πίνακας 4.1.3 Αντιπροσωπευτικοί κανόνες TE2Rules στο σύνολο ασθενών K-Modes	66
Πίνακας 4.1.4 Εσωτερικές μετρικές ομαδοποίησης K -Prototypes στο σύνολο ασθενών	74
Πίνακας 4.1.5 Σταθερότητα της ομαδοποίησης K-Prototypes στο σύνολο ασθενών	74
Πίνακας 4.1.6 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο ασθενών K-Prototypes	75
Πίνακας 4.2.1 Εσωτερικές μετρικές ομαδοποίησης K -Modes στο σύνολο των συμπτωματικών ασθενών	83
Πίνακας 4.2.2 Σταθερότητα της ομαδοποίησης K-Modes στο σύνολο των συμπτωματικών ασθενών	83
Πίνακας 4.2.3 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των συμπτωματικών ασθενών K- Modes.....	84
Πίνακας 4.2.4 Εσωτερικές μετρικές ομαδοποίησης K -Prototypes στο σύνολο των συμπτωματικών ασθενών	94
Πίνακας 4.2.5 Σταθερότητα της ομαδοποίησης K-Prototypes στο σύνολο των συμπτωματικών ασθενών	94
Πίνακας 4.2.6 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των συμπτωματικών ασθενών με K-Prototypes.....	95
Πίνακας 4.3.1 Εσωτερικές μετρικές ομαδοποίησης K-Modes στο σύνολο των ασυμπτωματικών ασθενών	104
Πίνακας 4.3.2 Σταθερότητα της ομαδοποίησης K-Modes στο σύνολο των ασυμπτωματικών ασθενών	104
Πίνακας 4.3.3 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των ασυμπτωματικών ασθενών με K-Modes.	105
Πίνακας 4.3.4 Εσωτερικές μετρικές ομαδοποίησης K -Prototypes στο σύνολο των ασυμπτωματικών ασθενών	114

Πίνακας 4.3.5 Σταθερότητα της ομαδοποίησης K-Prototypes στο σύνολο των ασυμπτωματικών ασθενών	114
Πίνακας 4.3.6 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των ασυμπτωματικών ασθενών με K-Prototypes.	115

Κεφάλαιο 1: Εισαγωγή

1.1	Κλινικό υπόβαθρο	2
1.2	Μη Επιβλεπόμενη Μάθηση και Ανάλυση Συστάδων στην Ιατρική	4
1.3	Σκοπός της εργασίας	6
1.4	Δομή της εργασίας	8

1.1 Κλινικό Υπόβαθρο

Η αθηρωματική νόσος των καρωτίδων αποτελεί μία από τις κυριότερες αιτίες ισχαιμικού αγγειακού εγκεφαλικού επεισοδίου και συνιστά σημαντικό παράγοντα νοσηρότητας και θνησιμότητας παγκοσμίως. Η ανάπτυξη αθηρωματικών πλακών στην περιοχή του διχασμού της κοινής καρωτίδας οδηγεί σε προοδευτική στένωση του αυλού της αρτηρίας, αλλά και σε μεταβολές της σύστασης και της μορφολογίας της πλάκας, οι οποίες σχετίζονται άμεσα με την ευπάθειά της και τον κίνδυνο εμφάνισης αγγειακών συμβάντων.

Παρόλο που ο βαθμός στένωσης αποτελεί διαχρονικά το βασικό κριτήριο αξιολόγησης της καρωτιδικής νόσου, έχει αποδειχθεί ότι δεν επαρκεί από μόνος του για την ακριβή εκτίμηση του κινδύνου εμφάνισης εγκεφαλικού επεισοδίου. Ειδικότερα, οι ασθενείς με ασυμπτωματική στένωση της έσω καρωτίδας παρουσιάζουν σημαντική ετερογένεια ως προς την πιθανότητα εμφάνισης μελλοντικού αγγειακού επεισοδίου, γεγονός που υποδηλώνει ότι επιπρόσθετοι μορφολογικοί και κλινικοί παράγοντες διαδραματίζουν καθοριστικό ρόλο.

Σύγχρονες μελέτες έχουν αναδείξει τη σημασία συγκεκριμένων υπερηχογραφικών και μορφολογικών χαρακτηριστικών της αθηρωματικής πλάκας, όπως η ηχογένεια (Gray Scale Median – GSM), η παρουσία discrete white areas (DWA), η επιφάνεια της πλάκας (PLAREA), καθώς και η παρουσία Juxtaluminal Black Area (JBA). Τα χαρακτηριστικά αυτά σχετίζονται με τη σύσταση της πλάκας, την παρουσία λιπώδους πυρήνα και τον βαθμό σταθερότητας του ινώδους καλύμματος. Παράλληλα, κλινικοί παράγοντες, όπως το ιστορικό παροδικών ισχαιμικών επεισοδίων ή εγκεφαλικού επεισοδίου στο ετερόπλευρο ημισφαίριο (CTIASTR), έχουν επίσης αναγνωριστεί ως σημαντικοί δείκτες αυξημένου κινδύνου.

Η σημασία των παραπάνω παραμέτρων επιβεβαιώνεται και από τη μελέτη Asymptomatic Carotid STENOSIS and Risk of Stroke (ACSRS) [1], η οποία έδειξε ότι συγκεκριμένα απεικονιστικά και κλινικά χαρακτηριστικά, όπως η παρουσία JBA και DWA, αποτελούν ανεξάρτητους προγνωστικούς παράγοντες για την εμφάνιση αγγειακού εγκεφαλικού επεισοδίου. Τα ευρήματα αυτά αναδεικνύουν την ανάγκη για μία περισσότερο πολυπαραγοντική προσέγγιση στην εκτίμηση του κινδύνου, πέρα από προσεγγίσεις που βασίζονται αποκλειστικά στον βαθμό στένωσης (STENOSIS).

Ωστόσο, η ταυτόχρονη αλληλεπίδραση πολλαπλών μορφολογικών και κλινικών παραγόντων καθιστά δύσκολη την πλήρη κατανόηση της νόσου μέσω συμβατικών στατιστικών προσεγγίσεων. Οι παραδοσιακές στατιστικές μέθοδοι βασίζονται συνήθως σε προκαθορισμένες υποθέσεις και γραμμικές σχέσεις μεταξύ των μεταβλητών, γεγονός που περιορίζει την ικανότητά τους να ανιχνεύουν πολύπλοκες, μη γραμμικές αλληλεπιδράσεις και λανθάνουσες δομές στα δεδομένα.

Στο πλαίσιο αυτό, οι τεχνικές μηχανικής μάθησης (Machine Learning) προσφέρουν τη δυνατότητα ανάλυσης πολυδιάστατων δεδομένων και αναγνώρισης προτύπων που δεν είναι άμεσα εμφανή μέσω συμβατικών μεθόδων. Ιδιαίτερα, η μη επιβλεπόμενη μάθηση (unsupervised learning) επιτρέπει την αναγνώριση υποκρυπτόμενων ομάδων ασθενών με παρόμοια χαρακτηριστικά, συμβάλλοντας στην καλύτερη κατανόηση της μορφολογικής ετερογένειας της νόσου και στη βελτίωση της εκτίμησης του κινδύνου. Παράλληλα, οι τεχνικές αυτές διευκολύνουν την ανάδειξη διακριτών μορφολογικών προτύπων της αθηρωματικής πλάκας, τα οποία δεν είναι εύκολα ανιχνεύσιμα με συμβατικές μεθόδους ανάλυσης.

1.2 Μη επιβλεπόμενη μάθηση και ανάλυση συστάδων στην ιατρική

Η μη επιβλεπόμενη μάθηση (unsupervised learning) αποτελεί έναν από τους βασικούς κλάδους της μηχανικής μάθησης, καθώς επιτρέπει την ανάλυση και οργάνωση δεδομένων χωρίς τη χρήση προκαθορισμένων ετικετών ή γνωστών κατηγοριών κατά την εκπαίδευση των μοντέλων [3, 4, 14]. Στο πλαίσιο αυτό, η ανάλυση συστάδων (cluster analysis) χρησιμοποιείται για την ομαδοποίηση δεδομένων σε ομάδες (clusters), με βάση την ομοιότητα των χαρακτηριστικών τους, επιτρέποντας την ανάδειξη υποκρυπτόμενων προτύπων και σχέσεων μέσα στα δεδομένα [7, 8].

Στον ιατρικό τομέα, οι τεχνικές μη επιβλεπόμενης μάθησης παρουσιάζουν ιδιαίτερο ενδιαφέρον, καθώς μπορούν να συμβάλουν στην αναγνώριση υποομάδων ασθενών με κοινά χαρακτηριστικά, στην ανάδειξη διαφορετικών φαινότυπων νόσου και στη βελτίωση της διαστρωμάτωσης κινδύνου [5, 6]. Η ανάγκη αυτή είναι ιδιαίτερα σημαντική στην καρωτιδική νόσο, όπου οι ασθενείς εμφανίζουν μεγάλη μορφολογική ετερογένεια και συχνά δεν μπορούν να περιγραφούν επαρκώς μόνο με βάση τον βαθμό στένωσης ή την παρουσία συμπτωμάτων.

Σύγχρονες μελέτες έχουν δείξει ότι η αξιοποίηση τεχνικών τεχνητής νοημοσύνης και μηχανικής μάθησης μπορεί να βελτιώσει σημαντικά την ανάλυση της μορφολογίας της αθηρωματικής πλάκας και την αναγνώριση ασθενών αυξημένου κινδύνου [16]. Ειδικότερα, η εφαρμογή τεχνικών ομαδοποίησης (clustering) στην καρωτιδική νόσο μπορεί να συμβάλει στην αναγνώριση διαφορετικών μορφολογικών προτύπων πλάκας και στην καλύτερη κατανόηση της ετερογένειας των ασθενών, πέρα από τις συμβατικές κατηγοριοποιήσεις που βασίζονται αποκλειστικά στην παρουσία ή απουσία αγγειακών συμβάντων.

Δεδομένου ότι, σημαντικό μέρος των εξεταζόμενων δεδομένων είναι κατηγορικής μορφής (π.χ. low, medium, high), ο αλγόριθμος K-Modes αποτελεί κατάλληλη επιλογή, καθώς βασίζεται σε μετρική ασυμφωνίας (matching dissimilarity), αντί της ευκλείδειας απόστασης [9]. Παράλληλα, για την αξιοποίηση τόσο κατηγορικών, όσο και συνεχών αριθμητικών μεταβλητών εφαρμόζεται και ο αλγόριθμος K-Prototypes, ο οποίος συνδυάζει χαρακτηριστικά των K-Means και K-Modes και επιτρέπει την ομαδοποίηση δεδομένων μικτού τύπου (mixed-type data) [15].

Αξιοσημείωτο είναι το γεγονός ότι, τα τελευταία χρόνια έχει αναδειχθεί ιδιαίτερα η σημασία της ερμηνευσιμότητας των μοντέλων μηχανικής μάθησης στην ιατρική. Η δυνατότητα κατανόησης των αποφάσεων και των προτύπων που προκύπτουν από ένα μοντέλο θεωρείται κρίσιμη για την ασφαλή και αξιόπιστη αξιοποίησή του στην κλινική πράξη. Σύμφωνα με τη μελέτη των Prentzas et al. [2], η αξιοποίηση μεθοδολογιών *XAI* επιτρέπει τη βαθύτερη κατανόηση των μοντέλων πρόβλεψης αναφορικά με τον κίνδυνο εμφάνισης ισχαιμικού επεισοδίου. Το γεγονός αυτό καθιστά τις αποφάσεις των μοντέλων κλινικά διαφανείς, διατηρώντας παράλληλα υψηλά τα επίπεδα της προγνωστικής τους ισχύος.

Στο πλαίσιο αυτό, η χρήση τεχνικών Explainable AI, όπως η μέθοδος TE2Rules [10], αποκτά ιδιαίτερη σημασία, καθώς επιτρέπει την εξαγωγή σαφών και ερμηνεύσιμων κανόνων από τα αποτελέσματα της ομαδοποίησης. Με τον τρόπο αυτό, οι υπολογιστικές συστάδες μπορούν να συνδεθούν με συγκεκριμένα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας, διευκολύνοντας την κλινική ερμηνεία και ενισχύοντας την εμπιστοσύνη των ιατρών στα συστήματα τεχνητής νοημοσύνης.

1.3 Σκοπός της εργασίας

Σκοπός της παρούσας εργασίας είναι η διερεύνηση της μορφολογικής ετερογένειας της καρωτιδικής αθηρωματικής πλάκας μέσω τεχνικών μη επιβλεπόμενης μάθησης, με στόχο την αναγνώριση διακριτών μορφολογικών προτύπων που σχετίζονται με διαφορετικό βαθμό μορφολογικού κινδύνου και τη βαθύτερη κατανόηση της καρωτιδικής νόσου.

Για την επίτευξη του σκοπού αυτού, εφαρμόζονται μέθοδοι ανάλυσης συστάδων (clustering) σε υπερηχογραφικά, μορφολογικά και κλινικά χαρακτηριστικά ασθενών με καρωτιδική νόσο. Ιδιαίτερη έμφαση δίνεται στους αλγορίθμους K-Modes και K-Prototypes. Ο K-Modes χρησιμοποιείται για την ομαδοποίηση κατηγορικών χαρακτηριστικών [9], ενώ ο K-Prototypes επιτρέπει την ταυτόχρονη αξιοποίηση κατηγορικών και συνεχών αριθμητικών μεταβλητών, προσφέροντας πιο ολοκληρωμένη αναπαράσταση της μορφολογίας της αθηρωματικής πλάκας [15].

Στο πλαίσιο της μελέτης επιλέχθηκε αριθμός συστάδων $k = 2$, με στόχο την ανάδειξη βασικών μορφολογικών προτύπων ασθενών και τον διαχωρισμό ομάδων με διαφορετικά χαρακτηριστικά μορφολογικής επιβάρυνσης. Η ποιότητα της ομαδοποίησης αξιολογείται μέσω κατάλληλων δεικτών εγκυρότητας συστάδων, όπως οι δείκτες Silhouette [11], Davies–Bouldin [12] και Calinski–Harabasz [13], καθώς και μέσω ανάλυσης σταθερότητας των αποτελεσμάτων με χρήση του δείκτη Adjusted Rand Index (ARI).

Παράλληλα, η ανάλυση εφαρμόζεται τόσο στο σύνολο των ασθενών όσο και σε επιμέρους υποομάδες, οι οποίες προκύπτουν από τον διαχωρισμό των ασθενών σε συμπτωματικούς και ασυμπτωματικούς, καθώς και από την κατηγοριοποίηση με βάση τον χρόνο παρακολούθησης (follow-up time). Η προσέγγιση αυτή επιτρέπει τη διερεύνηση πιθανών διαφοροποιήσεων στα μορφολογικά πρότυπα και στη δομή των συστάδων μεταξύ διαφορετικών κλινικών ομάδων.

Επιπλέον, σημαντικό μέρος της εργασίας αποτελεί η αξιοποίηση τεχνικών Explainable Artificial Intelligence (XAI) και ειδικότερα της μεθόδου TE2Rules [10], με στόχο την καλύτερη ερμηνεία των αποτελεσμάτων της ομαδοποίησης. Μέσω της εξαγωγής ερμηνεύσιμων κανόνων επιδιώκεται η σύνδεση των clusters με συγκεκριμένους συνδυασμούς μορφολογικών χαρακτηριστικών της αθηρωματικής πλάκας, ενισχύοντας τη διαφάνεια και τη δυνατότητα κλινικής αξιοποίησης των αποτελεσμάτων.

Συνολικά, η παρούσα εργασία επιδιώκει να συμβάλει στην καλύτερη κατανόηση της μορφολογικής πολυπλοκότητας της καρωτιδικής νόσου και στην ανάδειξη μορφολογικών προτύπων που θα μπορούσαν μελλοντικά να υποστηρίξουν πιο αξιόπιστες και εξατομικευμένες προσεγγίσεις διαστρωμάτωσης κινδύνου και λήψης κλινικών αποφάσεων.

1.4 Δομή της εργασίας

Η παρούσα εργασία οργανώνεται σε έξι κεφάλαια, τα οποία ακολουθούν μία σταδιακή και λογική πορεία, ξεκινώντας από την εισαγωγή στο αντικείμενο της μελέτης και το θεωρητικό υπόβαθρο και καταλήγοντας στην ανάλυση, ερμηνεία και αξιολόγηση των αποτελεσμάτων. Η δομή αυτή επιτρέπει τη συστηματική παρουσίαση της μεθοδολογίας και διευκολύνει την κατανόηση των επιμέρους σταδίων της έρευνας.

Συγκεκριμένα, το Κεφάλαιο 1 αποτελεί την εισαγωγή της εργασίας και παρουσιάζει το γενικό πλαίσιο της καρωτιδικής νόσου, τη σημασία του προβλήματος, τον σκοπό της μελέτης και τη συνολική δομή της εργασίας.

Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο της εργασίας. Αρχικά, αναλύονται βασικές έννοιες της μηχανικής μάθησης και της ανάλυσης δεδομένων στον ιατρικό τομέα, με ιδιαίτερη έμφαση στις τεχνικές μη επιβλεπόμενης μάθησης και στην ανάλυση συστάδων. Στη συνέχεια παρουσιάζονται οι αλγόριθμοι K-Modes και K-Prototypes, καθώς και οι βασικές μετρικές αξιολόγησης clustering. Επιπρόσθετα, γίνεται αναφορά στις τεχνικές Explainable Artificial Intelligence (XAI) και στη σημασία της ερμηνευσιμότητας των μοντέλων στην κλινική πράξη.

Στο Κεφάλαιο 3 περιγράφεται η μεθοδολογία που ακολουθήθηκε στην παρούσα μελέτη. Παρουσιάζονται τα δεδομένα που χρησιμοποιήθηκαν, τα μορφολογικά και κλινικά χαρακτηριστικά των ασθενών, όπως και η διαδικασία προεπεξεργασίας και μετασχηματισμού των δεδομένων. Επιπλέον, αναλύεται η εφαρμογή των αλγορίθμων K-Modes και K-Prototypes, η διαδικασία επιλογής του αριθμού συστάδων και οι μέθοδοι αξιολόγησης της ποιότητας και της σταθερότητας της ομαδοποίησης.

Στο Κεφάλαιο 4 παρουσιάζονται λεπτομερώς τα αποτελέσματα της ανάλυσης συστάδων. Η εφαρμογή των αλγορίθμων πραγματοποιείται τόσο στο σύνολο των ασθενών όσο και στις δύο βασικές κλινικές υποομάδες, δηλαδή στους συμπτωματικούς και στους ασυμπτωματικούς ασθενείς. Για κάθε ανάλυση παρουσιάζονται οι συστάδες που προέκυψαν, οι κατανομές των μορφολογικών χαρακτηριστικών, οι μετρικές αξιολόγησης της ομαδοποίησης και οι κανόνες που εξήχθησαν μέσω της μεθόδου TE2Rules. Παράλληλα, μέσω των γραφικών απεικονίσεων εξετάζεται και η κατανομή του χρόνου παρακολούθησης (last follow-up) μεταξύ των clusters, με στόχο τη διερεύνηση πιθανών σχέσεων μεταξύ των μορφολογικών προτύπων και της χρονικής εξέλιξης της νόσου.

Στο Κεφάλαιο 5 πραγματοποιείται η συζήτηση και ερμηνεία των αποτελεσμάτων. Συγκεκριμένα αναλύονται τα μορφολογικά πρότυπα που αναδείχθηκαν από τις διαφορετικές αναλύσεις, εξετάζονται οι διαφοροποιήσεις μεταξύ συμπτωματικών και ασυμπτωματικών ασθενών και αξιολογείται η απόδοση των αλγορίθμων K-Modes και K-Prototypes. Συγχρόνως, γίνεται συσχέτιση των ευρημάτων με τη διεθνή βιβλιογραφία και συζητείται η σημασία τους για την κατανόηση της μορφολογικής ετερογένειας της καρωτιδικής νόσου.

Τέλος, στο Κεφάλαιο 6 παρουσιάζονται τα συνολικά συμπεράσματα της εργασίας, οι βασικοί περιορισμοί της μελέτης και πιθανές κατευθύνσεις για μελλοντική έρευνα. Επιπροσθέτως, συζητούνται οι δυνατότητες περαιτέρω αξιοποίησης των τεχνικών clustering και Explainable AI στην ανάλυση της καρωτιδικής νόσου και στην υποστήριξη της κλινικής λήψης αποφάσεων.

Κεφάλαιο 2: Μηχανική μάθηση και ανάλυση δεδομένων στην ιατρική

2.1	Μηχανική μάθηση	12
2.2	Μηχανική μάθηση στην ιατρική	15
2.3	Ανάλυση συστάδων	17
2.4	Αλγόριθμος K-Modes	20
2.5	Αλγόριθμος K-Prototypes	23
2.6	Μετρικές αξιολόγησης	26
2.6.1	Δείκτης Silhouette	26
2.6.2	Δείκτης Davies-Bouldin	27
2.6.3	Δείκτης Calinski - Harabasz	28
2.6.4	Δείκτης Adjusted Rand Index (ARI)	28
2.6.5	Μεθοδολογική προσέγγιση αξιολόγησης στην παρούσα εργασία	30
2.7	Εξηγήσιμη τεχνητή νοημοσύνη (Explainable AI)	31
2.8	TE2Rules	34

2.1 Μηχανική μάθηση

Η μηχανική μάθηση (Machine Learning) αποτελεί έναν από τους σημαντικότερους κλάδους της τεχνητής νοημοσύνης (Artificial Intelligence – AI) [3, 4]. Αντικείμενό της είναι η ανάπτυξη αλγορίθμων και υπολογιστικών μοντέλων που μπορούν να μαθαίνουν από δεδομένα και να βελτιώνουν την απόδοσή τους, χωρίς να απαιτείται ρητός προγραμματισμός για κάθε πιθανό σενάριο. Η τεχνητή νοημοσύνη, ως ευρύτερο επιστημονικό πεδίο, στοχεύει στη δημιουργία συστημάτων ικανών να προσομοιώνουν ανθρώπινες γνωστικές λειτουργίες, όπως η μάθηση, η αναγνώριση προτύπων και η λήψη αποφάσεων.

Οι προσεγγίσεις της τεχνητής νοημοσύνης διακρίνονται σε διαφορετικές κατηγορίες. Η συμβολική τεχνητή νοημοσύνη (Symbolic AI) βασίζεται κυρίως σε λογικούς κανόνες και προκαθορισμένες δομές γνώσης, ενώ η στατιστική ή υπολογιστική προσέγγιση - στην οποία εντάσσεται η μηχανική μάθηση - αξιοποιεί δεδομένα για την εξαγωγή γνώσης και την αναγνώριση συσχετίσεων και προτύπων.

Η συνεχής αύξηση της διαθεσιμότητας μεγάλου όγκου δεδομένων, σε συνδυασμό με τη σημαντική πρόοδο της υπολογιστικής ισχύος, έχει συμβάλει καθοριστικά στη ραγδαία ανάπτυξη της μηχανικής μάθησης και στην εφαρμογή της σε πολλούς επιστημονικούς τομείς, συμπεριλαμβανομένου του τομέα της υγείας. Οι τεχνικές αυτές επιτρέπουν την ανάλυση πολύπλοκων και πολυδιάστατων δεδομένων, συμβάλλοντας στην αναγνώριση χρήσιμων προτύπων και στην υποστήριξη της διαδικασίας λήψης αποφάσεων.

Οι αλγόριθμοι μηχανικής μάθησης διακρίνονται συνήθως σε τρεις βασικές κατηγορίες: την επιβλεπόμενη μάθηση (Supervised Learning), τη μη επιβλεπόμενη μάθηση (Unsupervised Learning) και την ενισχυτική μάθηση (Reinforcement Learning) [3].

Στην επιβλεπόμενη μάθηση, τα μοντέλα εκπαιδεύονται με δεδομένα που συνοδεύονται από γνωστές ετικέτες (labels). Στόχος είναι η εκμάθηση μιας συνάρτησης που συσχετίζει τα χαρακτηριστικά εισόδου με την επιθυμητή έξοδο, ώστε το μοντέλο να μπορεί να πραγματοποιεί προβλέψεις σε νέα δεδομένα. Οι τεχνικές αυτές εφαρμόζονται κυρίως σε προβλήματα ταξινόμησης και παλινδρόμησης.

Αντίθετα, στη μη επιβλεπόμενη μάθηση δεν υπάρχουν προκαθορισμένες ετικέτες και ο στόχος είναι η ανακάλυψη υποκρυπτόμενων δομών και προτύπων μέσα στα δεδομένα. Μία από τις πιο διαδεδομένες τεχνικές της κατηγορίας αυτής είναι η ανάλυση συστάδων (cluster analysis), η οποία στοχεύει στην ομαδοποίηση δεδομένων σε συστάδες με βάση την ομοιότητά τους.

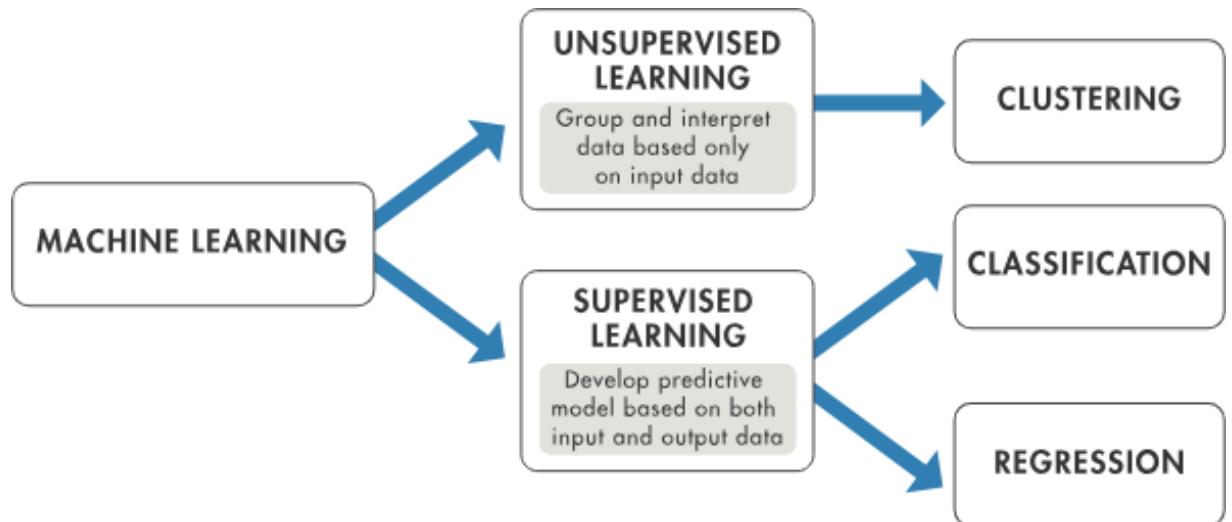
Η ενισχυτική μάθηση (Reinforcement Learning), βασίζεται στην αλληλεπίδραση ενός παράγοντα (agent) με το περιβάλλον του, με στόχο τη μεγιστοποίηση μιας συνάρτησης ανταμοιβής μέσα από διαδοχικές αποφάσεις και ενέργειες.

Η μη επιβλεπόμενη μάθηση παρουσιάζει ιδιαίτερο ενδιαφέρον σε περιπτώσεις όπου επιδιώκεται η διερεύνηση της εσωτερικής δομής των δεδομένων και η αναγνώριση λανθανόντων προτύπων χωρίς τη χρήση προκαθορισμένων κατηγοριών ή ετικετών κινδύνου. Στο πλαίσιο της παρούσας εργασίας, η προσέγγιση αυτή επιλέχθηκε με στόχο τη διερεύνηση της μορφολογικής ετερογένειας των ασθενών με καρωτιδική νόσο και την αναγνώριση διαφορετικών μορφολογικών προτύπων αθηρωματικής πλάκας.

Λόγω της φύσης του συνόλου δεδομένων, το οποίο περιλαμβάνει τόσο κατηγορικά, όσο και συνεχόμενα αριθμητικά χαρακτηριστικά, εφαρμόζονται οι αλγόριθμοι K-Modes και K-Prototypes. Ο K-Modes χρησιμοποιείται για την ομαδοποίηση κατηγορικών μεταβλητών [9], ενώ ο K-Prototypes επιτρέπει την ταυτόχρονη αξιοποίηση κατηγορικών και αριθμητικών χαρακτηριστικών στον ίδιο χώρο δεδομένων [15]. Η μεταβλητή της κλινικής κατάστασης των ασθενών (patstat) χρησιμοποιείται αποκλειστικά για τον διαχωρισμό του δείγματος σε συμπτωματικούς και ασυμπτωματικούς ασθενείς και δεν αποτελεί χαρακτηριστικό εισόδου κατά την εκπαίδευση των αλγορίθμων clustering, διατηρώντας έτσι τον μη επιβλεπόμενο χαρακτήρα της διαδικασίας.

Η εφαρμογή των τεχνικών μηχανικής μάθησης στον τομέα της υγείας παρουσιάζεται αναλυτικότερα στην επόμενη ενότητα.

Η βασική διάκριση των μεθόδων μηχανικής μάθησης παρουσιάζεται στο Σχήμα 2.1.



Σχήμα 2.1 Σχηματική αναπαράσταση βασικών κατηγοριών μηχανικής μάθησης.

2.2 Μηχανική μάθηση στην ιατρική

Η εφαρμογή της μηχανικής μάθησης στον τομέα της ιατρικής έχει αναπτυχθεί σημαντικά τα τελευταία χρόνια, συμβάλλοντας ουσιαστικά στην ανάλυση πολύπλοκων δεδομένων και στην υποστήριξη της κλινικής λήψης αποφάσεων. Η δυνατότητα επεξεργασίας μεγάλου όγκου ετερογενών δεδομένων, όπως ιατρικές εικόνες, βιοϊατρικά σήματα και κλινικά χαρακτηριστικά ασθενών, καθιστά τις τεχνικές αυτές ιδιαίτερα χρήσιμες στη σύγχρονη ιατρική πρακτική [5, 6].

Τα ιατρικά δεδομένα χαρακτηρίζονται συχνά από υψηλή πολυπλοκότητα, μεγάλη διάσταση και παρουσία θορύβου, γεγονός που περιορίζει τις δυνατότητες των παραδοσιακών στατιστικών μεθόδων. Στο πλαίσιο αυτό, η μηχανική μάθηση προσφέρει ισχυρά εργαλεία για την αναγνώριση σύνθετων προτύπων και μη γραμμικών σχέσεων, συμβάλλοντας στην καλύτερη κατανόηση της εξέλιξης των ασθενειών και στη βελτίωση της προγνωστικής ικανότητας [6].

Οι εφαρμογές της μηχανικής μάθησης στην ιατρική είναι ιδιαίτερα εκτεταμένες και περιλαμβάνουν τη διάγνωση ασθενειών, την πρόβλεψη κινδύνου, την ανάλυση ιατρικών εικόνων, την επεξεργασία βιοϊατρικών σημάτων και την υποστήριξη της εξατομικευμένης ιατρικής. Ειδικότερα, στον τομέα των καρδιαγγειακών παθήσεων, και κυρίως στην καρωτιδική νόσο, οι τεχνικές μηχανικής μάθησης επιτρέπουν την πολυπαραγοντική αξιολόγηση του κινδύνου εμφάνισης ισχαιμικού εγκεφαλικού επεισοδίου, λαμβάνοντας υπόψη συνδυαστικά μορφολογικά και κλινικά χαρακτηριστικά των ασθενών.

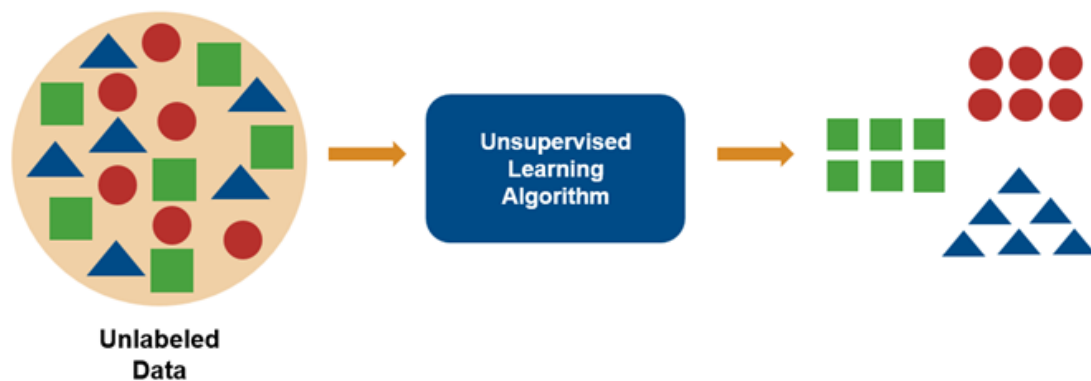
Σύγχρονες μελέτες έχουν δείξει ότι η αξιοποίηση τεχνικών τεχνητής νοημοσύνης μπορεί να βελτιώσει σημαντικά την ανάλυση της μορφολογίας της αθηρωματικής πλάκας και την αναγνώριση ασθενών αυξημένου κινδύνου [1]. Παράλληλα, η χρήση υπερηχογραφικών χαρακτηριστικών σε συνδυασμό με αλγορίθμους μηχανικής μάθησης έχει συμβάλει στην καλύτερη κατανόηση της μορφολογικής ετερογένειας της καρωτιδικής νόσου και στην ανάπτυξη πιο αξιόπιστων προσεγγίσεων διαστρωμάτωσης κινδύνου.

Παρά τα σημαντικά πλεονεκτήματα, η εφαρμογή της μηχανικής μάθησης στην ιατρική συνοδεύεται και από σημαντικές προκλήσεις, με κυριότερη την ερμηνευσιμότητα των μοντέλων. Πολλά σύγχρονα μοντέλα χαρακτηρίζονται ως «black-box», καθώς δεν παρέχουν σαφή και κατανοητή εξήγηση για τον τρόπο με τον οποίο λαμβάνονται οι αποφάσεις. Το γεγονός αυτό, περιορίζει την εμπιστοσύνη των κλινικών ιατρών και δυσχεραίνει την ενσωμάτωση τέτοιων συστημάτων στην καθημερινή κλινική πρακτική [6].

Για τον λόγο αυτό, τα τελευταία χρόνια δίνεται ιδιαίτερη έμφαση στην ανάπτυξη προσεγγίσεων που συνδυάζουν υψηλή απόδοση με δυνατότητα ερμηνείας, οδηγώντας στην ανάπτυξη του πεδίου της Εξηγήσιμης Τεχνητής Νοημοσύνης (Explainable Artificial Intelligence – XAI). Η αξιοποίηση Explainable AI τεχνικών επιτρέπει την καλύτερη κατανόηση των αποτελεσμάτων που παράγουν τα μοντέλα μηχανικής μάθησης και ενισχύει τη διαφάνεια και την αξιοπιστία τους στην κλινική πράξη.

Στο πλαίσιο αυτό, ιδιαίτερο ενδιαφέρον παρουσιάζουν οι τεχνικές μη επιβλεπόμενης μάθησης, όπως η ανάλυση συστάδων (cluster analysis), οι οποίες επιτρέπουν την αναγνώριση υποκρυπτόμενων δομών στα δεδομένα, χωρίς την ύπαρξη προκαθορισμένων ετικετών. Οι τεχνικές αυτές αποτελούν βασικό άξονα της παρούσας εργασίας, αφού επιτρέπουν τη διερεύνηση μορφολογικών προτύπων αθηρωματικής πλάκας και την ομαδοποίηση ασθενών με βάση κοινά χαρακτηριστικά.

Η βασική αρχή λειτουργίας της μη επιβλεπόμενης μάθησης απεικονίζεται στο Σχήμα 2.2.1.



Σχήμα 2.2.1 Σχηματική αναπαράσταση της μη επιβλεπόμενης μάθησης.

2.3 Ανάλυση συστάδων (Cluster Analysis)

Η ανάλυση συστάδων (cluster analysis) αποτελεί μία από τις σημαντικότερες τεχνικές της μη επιβλεπόμενης μάθησης (unsupervised learning) και χρησιμοποιείται για την ομαδοποίηση δεδομένων σε διακριτές ομάδες (clusters) με βάση την ομοιότητα των χαρακτηριστικών τους [3, 7]. Βασικός στόχος της είναι η δημιουργία συστάδων με υψηλή εσωτερική ομοιογένεια (intra-cluster similarity) και ταυτόχρονη σαφή διαφοροποίηση μεταξύ τους (inter-cluster dissimilarity), ώστε τα δεδομένα κάθε ομάδας να παρουσιάζουν κοινά χαρακτηριστικά και να διαχωρίζονται όσο το δυνατόν καλύτερα από τις υπόλοιπες ομάδες.

Η τεχνική αυτή είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου δεν υπάρχουν προκαθορισμένες κατηγορίες ή ετικέτες, καθώς επιτρέπει την αναγνώριση υποκρυπτόμενων δομών και προτύπων μέσα στα δεδομένα. Για τον λόγο αυτό, η ανάλυση συστάδων χρησιμοποιείται ευρέως σε επιστημονικά πεδία όπως η ιατρική, η βιοπληροφορική και η επιστήμη δεδομένων [7].

Στον ιατρικό τομέα, η ανάλυση συστάδων συμβάλλει στην αναγνώριση υποομάδων ασθενών με κοινά χαρακτηριστικά, στην ανάδειξη διαφορετικών φαινοτύπων μιας νόσου και στη βελτίωση της διαστρωμάτωσης κινδύνου. Μέσω της ομαδοποίησης καθίσταται δυνατή η καλύτερη κατανόηση της ετερογένειας που χαρακτηρίζει τους ασθενείς, η οποία συχνά δεν μπορεί να αποτυπωθεί επαρκώς με συμβατικές ή μονοδιάστατες κατηγοριοποιήσεις.

Οι μέθοδοι ανάλυσης συστάδων διαφοροποιούνται ανάλογα με τον τρόπο υπολογισμού της ομοιότητας μεταξύ των δεδομένων. Οι κυριότερες κατηγορίες περιλαμβάνουν τις μεθόδους διαχωρισμού (partitioning methods), τις ιεραρχικές μεθόδους (hierarchical clustering) και τις μεθόδους βασισμένες σε πυκνότητα (density-based clustering).

Οι μέθοδοι διαχωρισμού, όπως ο αλγόριθμος K-Means, στοχεύουν στη διάσπαση των δεδομένων σε έναν προκαθορισμένο αριθμό συστάδων και βασίζονται συνήθως σε μετρικές απόστασης, όπως η ευκλείδεια απόσταση [8]. Ωστόσο, οι προσεγγίσεις αυτές είναι περισσότερο κατάλληλες για αριθμητικά δεδομένα, καθώς η ευκλείδεια απόσταση δεν αποτυπώνει ικανοποιητικά τις σχέσεις μεταξύ κατηγορικών μεταβλητών.

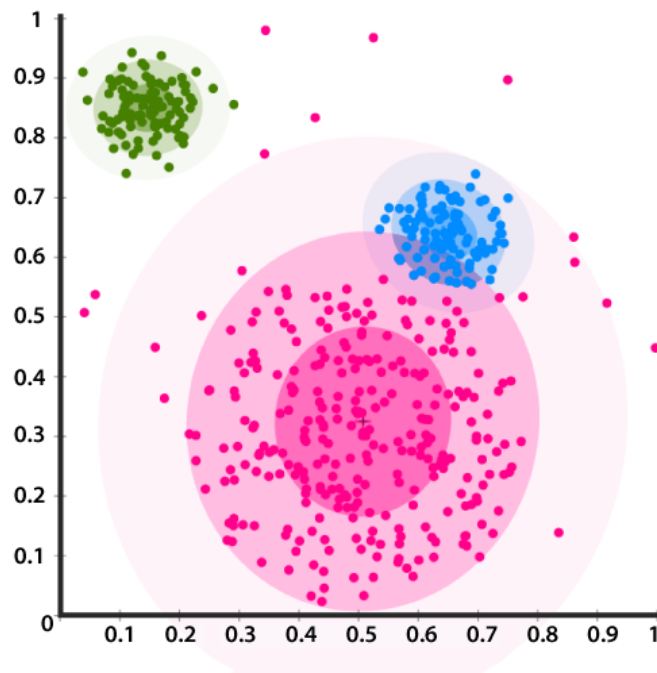
Για τον λόγο αυτό, σε περιπτώσεις όπου τα δεδομένα περιλαμβάνουν κατηγορικά χαρακτηριστικά απαιτούνται εξειδικευμένες τεχνικές clustering. Ο αλγόριθμος K-Modes αποτελεί επέκταση του K-Means για κατηγορικά δεδομένα και βασίζεται σε μετρικές ασυμφωνίας (matching dissimilarity), επιτρέποντας την ομαδοποίηση χωρίς την ανάγκη μετασχηματισμού των μεταβλητών σε συνεχή αριθμητική μορφή [9].

Παράλληλα, ο αλγόριθμος K-Prototypes αποτελεί επέκταση των K-Means και K-Modes και χρησιμοποιείται σε δεδομένα μικτού τύπου (mixed-type data), όπου συνυπάρχουν αριθμητικές και κατηγορικές μεταβλητές [15]. Ο αλγόριθμος αυτός συνδυάζει την ευκλείδεια απόσταση για τα αριθμητικά χαρακτηριστικά με μετρικές ασυμφωνίας για τα κατηγορικά δεδομένα, επιτρέποντας πιο ολοκληρωμένη αναπαράσταση της δομής των δεδομένων.

Παρότι υπάρχουν και άλλες τεχνικές clustering, όπως οι ιεραρχικές μέθοδοι ή τα στατιστικά μοντέλα μειγμάτων (Mixture Models), οι προσεγγίσεις αυτές συχνά συνοδεύονται από αυξημένο υπολογιστικό κόστος ή από την ανάγκη ισχυρών στατιστικών υποθέσεων σχετικά με την κατανομή των δεδομένων. Αντίθετα, οι αλγόριθμοι K-Modes και K-Prototypes παρουσιάζουν σχετικά απλή υλοποίηση, καλή υπολογιστική αποδοτικότητα και δυνατότητα εφαρμογής σε δεδομένα με έντονη ετερογένεια, γεγονός που τους καθιστά ιδιαίτερα κατάλληλους για εφαρμογές σε ιατρικά δεδομένα.

Στην παρούσα εργασία, η ανάλυση συστάδων αξιοποιείται για την ομαδοποίηση ασθενών με καρωτιδική νόσο, με στόχο την αναγνώριση μορφολογικών προτύπων που σχετίζονται με διαφορετικά επίπεδα μορφολογικού κινδύνου και τη βαθύτερη κατανόηση της δομής των δεδομένων. Αρχικά, εφαρμόζεται ο αλγόριθμος K-Modes στα κατηγορικά χαρακτηριστικά της αθηρωματικής πλάκας. Στη συνέχεια, εφαρμόζεται και ο K-Prototypes, ώστε να αξιοποιηθούν συνδυαστικά τόσο τα κατηγορικά, όσο και τα αριθμητικά χαρακτηριστικά των ασθενών. Ως εκ τούτου, καθίσταται δυνατή η σύγκριση μίας καθαρά κατηγορικής προσέγγισης clustering με μία προσέγγιση μικτού τύπου δεδομένων.

Η βασική έννοια της ανάλυσης συστάδων και της ομαδοποίησης δεδομένων με βάση την ομοιότητα των χαρακτηριστικών τους απεικονίζεται στο Σχήμα 2.3.1 .



Σχήμα 2.3.1 Γραφική απεικόνιση της διαδικασίας ανάλυσης συστάδων, όπου τα δεδομένα ομαδοποιούνται σε διακριτές περιοχές του χώρου χαρακτηριστικών.

2.4 Αλγόριθμος K-Modes

Ο αλγόριθμος K-Modes αποτελεί επέκταση του αλγορίθμου K-Means, προσαρμοσμένη ειδικά για την ανάλυση δεδομένων κατηγορικής μορφής [9]. Σε αντίθεση με τον K-Means, ο οποίος βασίζεται στον υπολογισμό αριθμητικών μέσων τιμών και στη χρήση της ευκλείδειας απόστασης, ο K-Modes αντικαθιστά τα στοιχεία αυτά με την επικρατούσα τιμή (mode) και μία κατάλληλη μετρική ασυμφωνίας (matching dissimilarity), γεγονός που τον καθιστά κατάλληλο για κατηγορικά δεδομένα.

Η βασική ιδέα του αλγορίθμου είναι η ομαδοποίηση των δεδομένων σε k συστάδες, έτσι ώστε οι παρατηρήσεις που ανήκουν στην ίδια συστάδα να παρουσιάζουν υψηλό βαθμό ομοιότητας ως προς τις κατηγορικές τους τιμές. Η ομοιότητα αυτή δεν εκφράζεται μέσω γεωμετρικής απόστασης, αλλά μέσω του αριθμού των ασυμφωνιών μεταξύ των χαρακτηριστικών.

$$d(X, Y) = \sum_{i=1}^m \delta(x_i, y_i)$$

όπου:

- m είναι ο αριθμός των χαρακτηριστικών
- $\begin{cases} \delta(x_i, y_i) = 0, \text{ αν } x_i = y_i, \\ \delta(x_i, y_i) = 1, \text{ αν } x_i \neq y_i \end{cases}$

Η συνάρτηση αυτή είναι γνωστή ως *matching dissimilarity* και αποτελεί μία απλή, αλλά αποτελεσματική μετρική για κατηγορικά δεδομένα.

Ο στόχος του αλγορίθμου είναι η ελαχιστοποίηση της συνολικής ασυμφωνίας εντός των συστάδων, δηλαδή της συνάρτησης κόστους:

$$J = \sum_{j=1}^k \sum_{X \in C_j} d(X, \mu_j)$$

όπου:

- C_j είναι η j -οστή συστάδα
- μ_j είναι το mode της συστάδας, δηλαδή το διάνυσμα επικρατουσών τιμών

Η συνάρτηση αυτή εκφράζει τον συνολικό βαθμό απόκλισης των παρατηρήσεων από τα αντίστοιχα modes των συστάδων και ελαχιστοποιείται επαναληπτικά κατά τη διαδικασία ομαδοποίησης.

Η λειτουργία του K-Modes βασίζεται σε μία επαναληπτική διαδικασία δύο βασικών βημάτων:

1. Βήμα ανάθεσης (Assignment step): Κάθε παρατήρηση ανατίθεται στη συστάδα της οποίας το mode παρουσιάζει τη μικρότερη ασυμφωνία με την παρατήρηση.
2. Βήμα ενημέρωσης (Update step): Για κάθε συστάδα υπολογίζεται εκ νέου το mode, επιλέγοντας για κάθε χαρακτηριστικό την πιο συχνά εμφανιζόμενη κατηγορία μεταξύ των παρατηρήσεων της συστάδας.

Η διαδικασία επαναλαμβάνεται έως ότου επιτευχθεί σύγκλιση, δηλαδή όταν δεν μεταβάλλονται πλέον οι αναθέσεις των παρατηρήσεων ή όταν η μεταβολή της συνάρτησης κόστους γίνει αμελητέα.

Ένα σημαντικό χαρακτηριστικό του K-Modes είναι ότι, όπως και ο K-Means, απαιτεί τον εκ των προτέρων καθορισμό του αριθμού των συστάδων k . Για τον λόγο αυτό, είναι απαραίτητη η χρήση τεχνικών επιλογής του κατάλληλου αριθμού clusters, όπως η μέθοδος «elbow» και οι δείκτες αξιολόγησης συστάδων.

Ο αλγόριθμος K-Modes είναι ιδιαίτερα κατάλληλος για την ανάλυση ιατρικών δεδομένων κατηγορικής φύσης, καθώς επιτρέπει την ομαδοποίηση των παρατηρήσεων, χωρίς την ανάγκη μετατροπής των μεταβλητών σε αριθμητική μορφή. Με τον τρόπο αυτό αποφεύγεται η απώλεια πληροφορίας ή η εισαγωγή τεχνητών σχέσεων μεταξύ των κατηγοριών.

Στο πλαίσιο της παρούσας εργασίας, ο K-Modes εφαρμόζεται σε δεδομένα ασθενών με καρωτιδική νόσο, τα οποία περιλαμβάνουν κατηγορικές μεταβλητές, όπως ο βαθμός στένωσης (STENOSIS), η ηχογένεια της πλάκας (GSM), η επιφάνεια της πλάκας (PLAREA), η παρουσία Discrete White Areas (DWA), η παρουσία Juxtaluminal Black Area (JBA) και το ιστορικό παροδικών ισχαιμικών επεισοδίων (CTIASTR). Η εφαρμογή του αλγορίθμου επιτρέπει την ανάδειξη υποκρυπτόμενων ομάδων ασθενών με παρόμοια μορφολογικά χαρακτηριστικά, συμβάλλοντας στην καλύτερη κατανόηση της ετερογένειας της καρωτιδικής νόσου.

Ας σημειωθεί ότι, ο αλγόριθμος παρουσιάζει ευαισθησία στην αρχικοποίηση των συστάδων και ενδέχεται να συγκλίνει σε τοπικά βέλτιστες λύσεις. Για τον λόγο αυτό, συνήθως απαιτείται η εκτέλεση πολλαπλών επαναλήψεων ή η χρήση κατάλληλων στρατηγικών αρχικοποίησης, ώστε να βελτιώνεται η σταθερότητα και η αξιοπιστία των αποτελεσμάτων.

2.5 Αλγόριθμος K-Prototypes

Ο αλγόριθμος K-Prototypes αποτελεί επέκταση των αλγορίθμων K-Means και K-Modes και έχει σχεδιαστεί για την ομαδοποίηση δεδομένων μικτού τύπου, δηλαδή δεδομένων που περιλαμβάνουν τόσο αριθμητικές, όσο και κατηγορικές μεταβλητές [15]. Σε αντίθεση με τον K-Means, ο οποίος εφαρμόζεται αποκλειστικά σε αριθμητικά δεδομένα και τον K-Modes, ο οποίος χρησιμοποιείται μόνο για κατηγορικές μεταβλητές, ο K-Prototypes συνδυάζει στοιχεία και των δύο προσεγγίσεων, επιτρέποντας την ταυτόχρονη αξιοποίηση διαφορετικών τύπων δεδομένων.

Η βασική ιδέα του αλγορίθμου είναι η ομαδοποίηση των παρατηρήσεων σε k συστάδες, έτσι ώστε οι ασθενείς που ανήκουν στην ίδια συστάδα να παρουσιάζουν υψηλό βαθμό ομοιότητας τόσο ως προς τα αριθμητικά, όσο και ως προς τα κατηγορικά χαρακτηριστικά τους.

Πιο συγκεκριμένα, η συνολική απόσταση μεταξύ δύο παρατηρήσεων X και Y ορίζεται ως:

$$d(X, Y) = \sum_{i=1}^q (x_i - y_i)^2 + \gamma \sum_{i=q+1}^m \delta(x_i, y_i)$$

όπου:

- q είναι ο αριθμός των αριθμητικών χαρακτηριστικών,
- $m - q$ είναι ο αριθμός των κατηγορικών χαρακτηριστικών,
- $\delta(x_i, y_i)$ είναι η συνάρτηση matching dissimilarity:

$$\delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases}$$

- γ είναι παράμετρος στάθμισης, η οποία καθορίζει τη συνεισφορά των κατηγορικών μεταβλητών στη συνολική απόσταση.

Η συνάρτηση αυτή επιτρέπει στον αλγόριθμο να συνδυάζει την πληροφορία από αριθμητικά και κατηγορικά δεδομένα σε ένα ενιαίο μέτρο ομοιότητας.

Ο στόχος του αλγορίθμου είναι η ελαχιστοποίηση της συνολικής απόστασης εντός των συστάδων, δηλαδή της συνάρτησης κόστους:

$$J = \sum_{j=1}^k \sum_{X \in C_j} d(X, \mu_j)$$

όπου:

- C_j είναι η j -οστή συστάδα,
- X είναι το διάνυσμα των χαρακτηριστικών μιας παρατήρησης (δηλαδή του εκάστοτε ασθενή),
- μ_j αντιστοιχεί στο πρωτότυπο (prototype) της συστάδας, το οποίο περιλαμβάνει τις μέσες τιμές των αριθμητικών χαρακτηριστικών και τις επικρατούσες τιμές των κατηγορικών χαρακτηριστικών.

Η λειτουργία του K-Prototypes βασίζεται σε μία επαναληπτική διαδικασία δύο βασικών βημάτων:

1. Ανάθεση (Assignment step): Κάθε παρατήρηση ανατίθεται στη συστάδα με τη μικρότερη συνολική απόσταση, λαμβάνοντας υπόψη τόσο τα αριθμητικά όσο και τα κατηγορικά χαρακτηριστικά.
2. Ενημέρωση (Update step): Για κάθε συστάδα, ενημερώνονται:
 - οι μέσες τιμές των αριθμητικών χαρακτηριστικών και
 - οι επικρατούσες τιμές (modes) των κατηγορικών χαρακτηριστικών.

Η διαδικασία επαναλαμβάνεται έως ότου επιτευχθεί σύγκλιση, δηλαδή έως ότου οι αναθέσεις των παρατηρήσεων στις συστάδες παραμείνουν σταθερές ή η μεταβολή της συνάρτησης κόστους γίνει αμελητέα.

Όπως και οι αλγόριθμοι K-Means και K-Modes, ο K-Prototypes απαιτεί τον εκ των προτέρων καθορισμό του αριθμού των συστάδων k . Για τον λόγο αυτό, είναι απαραίτητη η χρήση τεχνικών επιλογής του κατάλληλου αριθμού clusters, καθώς και μετρικών αξιολόγησης της ποιότητας της ομαδοποίησης.

Ο αλγόριθμος K-Prototypes είναι ιδιαίτερα κατάλληλος για δεδομένα που περιλαμβάνουν ταυτόχρονα κατηγορικές και συνεχείς αριθμητικές μεταβλητές, καθώς επιτρέπει την αξιοποίηση της πληροφορίας και των δύο τύπων χαρακτηριστικών, χωρίς να απαιτείται πλήρης μετασχηματισμός τους σε ενιαία μορφή.

Στο πλαίσιο της παρούσας εργασίας, ο K-Prototypes εφαρμόζεται στα υποσύνολα των συμπτωματικών και ασυμπτωματικών ασθενών, αξιοποιώντας τόσο κατηγορικές μεταβλητές όσο και τις αντίστοιχες συνεχείς αριθμητικές τιμές τους. Συγκεκριμένα, χρησιμοποιούνται αριθμητικά χαρακτηριστικά όπως ο βαθμός στένωσης (STENOSIS), το Gray Scale Median (GSM), η επιφάνεια της πλάκας (PLAREA), η τιμή JBA (Juxtaluminal Black Area) και η ηλικία των ασθενών (age). Πριν από την εφαρμογή του αλγορίθμου, οι αριθμητικές μεταβλητές κανονικοποιούνται με χρήση της μεθόδου RobustScaler, ώστε να περιοριστεί η επίδραση ακραίων τιμών και διαφορών κλίμακας μεταξύ των μεταβλητών.

Η χρήση του K-Prototypes επιτρέπει τη σύγκριση των αποτελεσμάτων με εκείνα του K-Modes και προσφέρει μία πιο ολοκληρωμένη εικόνα της δομής των δεδομένων, καθώς λαμβάνει υπόψη όχι μόνο τις κατηγοριοποιημένες μεταβλητές αλλά και την αρχική αριθμητική πληροφορία.

Παράλληλα, η συνδυαστική αξιοποίηση αριθμητικών και κατηγορικών χαρακτηριστικών συμβάλλει στην καλύτερη αποτύπωση της μορφολογικής ετερογένειας της καρωτιδικής νόσου επιτρέποντας την ανάδειξη περισσότερο συνεκτικών και διαχωρίσιμων μορφολογικών προτύπων ασθενών.

Ωστόσο, όπως και ο K-Modes, ο αλγόριθμος παρουσιάζει ευαισθησία στην αρχικοποίηση των συστάδων και ενδέχεται να συγκλίνει σε τοπικά βέλτιστες λύσεις. Για τον λόγο αυτό, απαιτείται συνήθως η εκτέλεση πολλαπλών επαναλήψεων και η χρήση κατάλληλων στρατηγικών αρχικοποίησης, ώστε να βελτιώνεται η σταθερότητα και η αξιοπιστία των αποτελεσμάτων.

2.6 Μετρικές αξιολόγησης

Η αξιολόγηση της ποιότητας των συστάδων αποτελεί ιδιαίτερα σημαντικό στάδιο στην ανάλυση συστάδων, μιας και επιτρέπει την εκτίμηση του κατά πόσο τα δεδομένα έχουν ομαδοποιηθεί με ουσιαστικό και συνεκτικό τρόπο. Μέσω κατάλληλων μετρικών αξιολόγησης, μπορεί να εξεταστεί τόσο η συνοχή των συστάδων, όσο και ο βαθμός διαχωρισμού μεταξύ τους, συμβάλλοντας στην αξιολόγηση και σύγκριση διαφορετικών αποτελεσμάτων clustering.

Οι μετρικές αξιολόγησης που χρησιμοποιούνται στην ανάλυση συστάδων μπορούν γενικά να διακριθούν σε δύο βασικές κατηγορίες:

- μετρικές ποιότητας του διαχωρισμού των συστάδων και
- μετρικές σταθερότητας των αποτελεσμάτων.

Στην παρούσα εργασία χρησιμοποιούνται οι δείκτες Silhouette, Davies–Bouldin και Calinski–Harabasz, για την αξιολόγηση της ποιότητας της ομαδοποίησης, όπως και ο δείκτης Adjusted Rand Index (ARI), για την αξιολόγηση της σταθερότητας των αποτελεσμάτων.

2.6.1 Δείκτης Silhouette

Μία από τις πιο διαδεδομένες μετρικές αξιολόγησης είναι ο δείκτης Silhouette [11]. Ο δείκτης αυτός λαμβάνει τιμές από -1 έως 1 και εκτιμά σε ποιο βαθμό κάθε παρατήρηση μοιάζει περισσότερο με τις παρατηρήσεις της δικής της συστάδας σε σχέση με τις παρατηρήσεις άλλων συστάδων. Υψηλές θετικές τιμές υποδηλώνουν ότι οι συστάδες είναι καλά διαχωρισμένες και εσωτερικά συνεκτικές, ενώ χαμηλές ή αρνητικές τιμές υποδηλώνουν πιθανή επικάλυψη μεταξύ των ομάδων.

Η τιμή του δείκτη για μία παρατήρηση i ορίζεται ως:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

όπου:

- $a(i)$ είναι η μέση απόσταση της παρατήρησης i από όλα τα σημεία της ίδιας συστάδας,
- $b(i)$ είναι η μικρότερη μέση απόσταση της παρατήρησης i από διαφορετική συστάδα.

Ο συνολικός δείκτης Silhouette προκύπτει από τον μέσο όρο των τιμών $s(i)$ για όλες τις παρατηρήσεις του συνόλου δεδομένων.

2.6.2 Δείκτης Davies–Bouldin

Ο δείκτης Davies–Bouldin (DB) [12] βασίζεται στη σύγκριση της διασποράς εντός των συστάδων με την απόσταση μεταξύ των συστάδων. Ο δείκτης αυτός εκτιμά κατά πόσο οι συστάδες είναι ταυτόχρονα συμπαγείς και καλά διαχωρισμένες.

Ο δείκτης υπολογίζεται ως:

$$DB = \frac{1}{k} \sum_{j=1}^k \max_{l \neq j} \left(\frac{S_j + S_l}{M_{j,l}} \right)$$

όπου:

- S_j και S_l αντιπροσωπεύουν την εσωτερική διασπορά των συστάδων j και l ,
- $M_{j,l}$ είναι η απόσταση μεταξύ των κεντροειδών των δύο συστάδων,
- k είναι ο συνολικός αριθμός συστάδων.

Χαμηλότερες τιμές του δείκτη υποδηλώνουν καλύτερο διαχωρισμό και μεγαλύτερη συνοχή των ομάδων.

2.6.3 Δείκτης Calinski–Harabasz

Ο δείκτης Calinski–Harabasz (CH) [13], γνωστός και ως Variance Ratio Criterion, αξιολογεί τη σχέση μεταξύ της διασποράς μεταξύ των συστάδων και της διασποράς εντός των συστάδων.

Ο δείκτης υπολογίζεται ως:

$$CH = \frac{SS_B}{SS_W} \times \frac{N - k}{k - 1}$$

όπου:

- SS_B είναι η διασπορά μεταξύ των συστάδων,
- SS_W είναι η διασπορά εντός των συστάδων,
- N είναι ο συνολικός αριθμός παρατηρήσεων,
- k είναι ο αριθμός των συστάδων.

Υψηλότερες τιμές του δείκτη υποδηλώνουν καλύτερη οργάνωση των δεδομένων σε διακριτές και καλά διαχωρισμένες ομάδες.

2.6.4 Δείκτης Adjusted Rand Index (ARI)

Πέρα από την ποιότητα του διαχωρισμού των συστάδων, ιδιαίτερα σημαντική θεωρείται και η σταθερότητα των αποτελεσμάτων clustering. Για τον λόγο αυτό, στην παρούσα μελέτη πραγματοποιούνται πολλαπλές εκτελέσεις των αλγορίθμων και αξιολογείται η συμφωνία μεταξύ των διαφορετικών αναθέσεων συστάδων μέσω του δείκτη Adjusted Rand Index (ARI).

Ο δείκτης ARI εκτιμά τον βαθμό ομοιότητας μεταξύ δύο διαφορετικών ομαδοποιήσεων, λαμβάνοντας υπόψη και τη συμφωνία που θα μπορούσε να προκύψει τυχαία.

Ο μαθηματικός ορισμός του δείκτη είναι:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}$$

όπου:

- n_{ij} είναι ο αριθμός παρατηρήσεων που ανήκουν ταυτόχρονα στις συστάδες u_i και v_j ,
- a_i και b_j αντιστοιχούν στα αθροίσματα γραμμών και στηλών του πίνακα συνάφειας,
- n είναι ο συνολικός αριθμός παρατηρήσεων.

Τιμές του δείκτη κοντά στο 1 υποδηλώνουν υψηλή συμφωνία μεταξύ των διαφορετικών ομαδοποιήσεων, ενώ τιμές κοντά στο 0 αντιστοιχούν σε τυχαία ομαδοποίηση.

2.6.5 Μεθοδολογική προσέγγιση αξιολόγησης στην παρούσα εργασία

Στην παρούσα εργασία, ο δείκτης Silhouette υπολογίζεται με χρήση της απόστασης Hamming, ώστε να είναι συμβατός με τη φύση των κατηγορικών δεδομένων και ιδιαίτερα με την εφαρμογή του αλγορίθμου K-Modes.

Αντίθετα, οι δείκτες Davies–Bouldin (DB) και Calinski–Harabasz (CH) υπολογίζονται μετά από κατάλληλη αριθμητική κωδικοποίηση των κατηγορικών μεταβλητών, αποκλειστικά για σκοπούς αξιολόγησης και όχι για την εκπαίδευση των αλγορίθμων clustering.

Για τον αλγόριθμο K-Prototypes, οι μετρικές αξιολόγησης υπολογίζονται μετά από κατάλληλη αναπαράσταση των δεδομένων μικτού τύπου και κανονικοποίηση των αριθμητικών χαρακτηριστικών, καθώς ο αλγόριθμος αξιοποιεί συνδυαστικά αριθμητικές και κατηγορικές μεταβλητές.

Επιπλέον, για την εκτίμηση της σταθερότητας των αποτελεσμάτων πραγματοποιούνται πολλαπλές εκτελέσεις των αλγορίθμων clustering με διαφορετικές αρχικοποιήσεις. Η συμφωνία μεταξύ των διαφορετικών ομαδοποιήσεων αξιολογείται μέσω του δείκτη Adjusted Rand Index.

Η συνδυαστική αξιοποίηση των παραπάνω μετρικών επιτρέπει μία πιο ολοκληρωμένη αξιολόγηση τόσο της ποιότητας, όσο και της σταθερότητας της ομαδοποίησης, συμβάλλοντας στην ασφαλέστερη και πιο αξιόπιστη ερμηνεία των αποτελεσμάτων της ανάλυσης συστάδων.

2.7 Εξηγήσιμη Τεχνητή Νοημοσύνη (Explainable Artificial Intelligence – XAI)

Η συνεχώς αυξανόμενη αξιοποίηση τεχνικών μηχανικής μάθησης στην ιατρική έχει οδηγήσει στην ανάπτυξη μοντέλων υψηλής υπολογιστικής απόδοσης, τα οποία όμως συχνά χαρακτηρίζονται από περιορισμένη ερμηνευσιμότητα. Πολλά σύγχρονα μοντέλα, όπως τα νευρωνικά δίκτυα και οι μέθοδοι συνόλου (ensemble methods), λειτουργούν ως συστήματα «μαύρου κουτιού» (black-box), καθώς παρέχουν αποτελέσματα χωρίς να είναι εύκολα κατανοητός ο τρόπος με τον οποίο καταλήγουν στις τελικές αποφάσεις.

Η περιορισμένη διαφάνεια τέτοιων μοντέλων αποτελεί σημαντική πρόκληση στον ιατρικό τομέα, όπου η αξιοπιστία και η δυνατότητα ερμηνείας των αποτελεσμάτων είναι ιδιαίτερα σημαντικές. Οι κλινικοί ιατροί χρειάζονται σαφείς και τεκμηριωμένες εξηγήσεις για τους παράγοντες που επηρεάζουν μία πρόβλεψη ή μία απόφαση, ώστε να μπορούν να αξιολογούν την εγκυρότητά της και να την ενσωματώνουν με ασφάλεια στη διαδικασία λήψης κλινικών αποφάσεων.

Ως εκ τούτου, τα τελευταία χρόνια έχει αναπτυχθεί σημαντικά το πεδίο της Εξηγήσιμης Τεχνητής Νοημοσύνης (Explainable Artificial Intelligence – XAI), το οποίο στοχεύει στη βελτίωση της διαφάνειας, της κατανόησης και της ερμηνευσιμότητας των μοντέλων μηχανικής μάθησης. Οι τεχνικές XAI επιτρέπουν την αναγνώριση των χαρακτηριστικών που επηρεάζουν περισσότερο τις αποφάσεις ενός μοντέλου και συμβάλλουν στην καλύτερη κατανόηση της σχέσης μεταξύ των δεδομένων εισόδου και των αποτελεσμάτων που παράγονται.

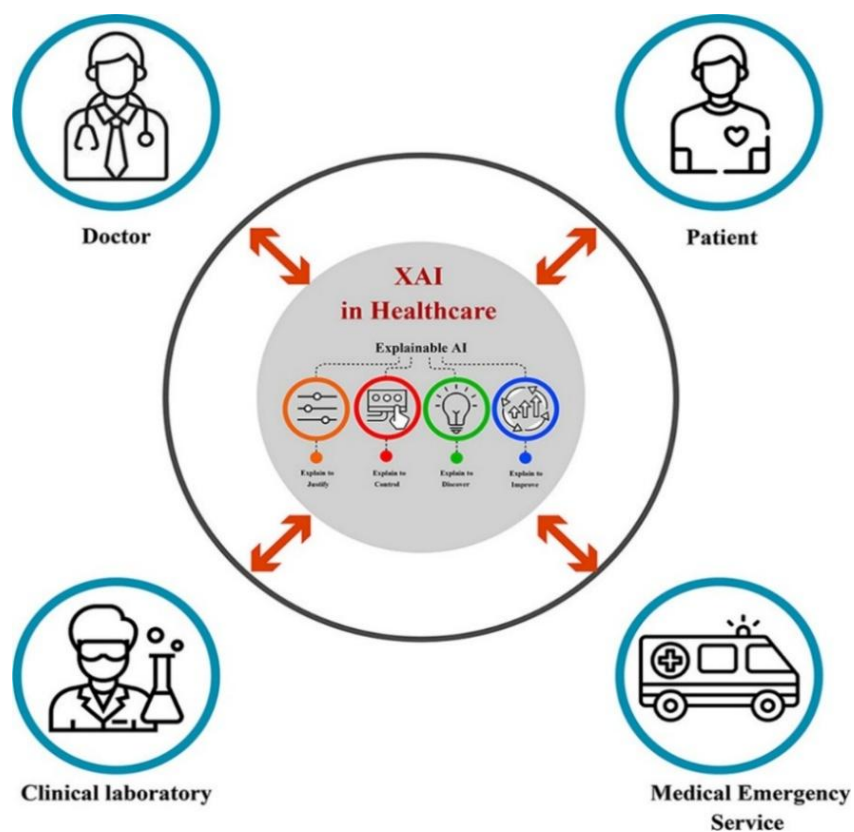
Οι προσεγγίσεις XAI μπορούν να εφαρμοστούν είτε σε συνολικό επίπεδο (global interpretability), παρέχοντας μία γενική εικόνα της λειτουργίας ενός μοντέλου, είτε σε επίπεδο μεμονωμένων προβλέψεων (local interpretability), εξηγώντας τον τρόπο με τον οποίο προέκυψε μία συγκεκριμένη απόφαση. Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική στην ιατρική, όπου η ερμηνεία των αποτελεσμάτων συνδέεται άμεσα με ζητήματα αξιοπιστίας, ασφάλειας και ηθικής ευθύνης.

Η σημασία της εξηγήσιμης τεχνητής νοημοσύνης στην ιατρική έχει αναδειχθεί και από σύγχρονες ερευνητικές μελέτες. Συγκεκριμένα, η εργασία των Prentzas et al. [2] έδειξε ότι η ενσωμάτωση τεχνικών XAI σε μοντέλα μηχανικής μάθησης που αξιοποιούν υπερηχογραφικά και κλινικά χαρακτηριστικά μπορεί να διατηρήσει υψηλή προγνωστική απόδοση, ενώ παράλληλα βελτιώνει σημαντικά την κατανόηση των αποτελεσμάτων από τους κλινικούς ιατρούς.

Ταυτόχρονα, οι Explainable AI τεχνικές συμβάλλουν στον εντοπισμό πιθανών σφαλμάτων ή μεροληψιών των μοντέλων, καθώς και στην καλύτερη κατανόηση της σχέσης μεταξύ των μορφολογικών χαρακτηριστικών της αθηρωματικής πλάκας και του κινδύνου εμφάνισης αγγειακών συμβάντων. Με τον τρόπο αυτό, τα αποτελέσματα των μοντέλων καθίστανται περισσότερο αξιόπιστα, διαφανή και κλινικά αξιοποιήσιμα.

Στο πλαίσιο της παρούσας εργασίας, η ερμηνευσιμότητα αποτελεί βασικό στοιχείο της ανάλυσης, αφού η μελέτη δεν περιορίζεται μόνο στην ομαδοποίηση των ασθενών, αλλά επεκτείνεται και στην κατανόηση των χαρακτηριστικών που διαμορφώνουν τα μορφολογικά πρότυπα των συστάδων. Για τον σκοπό αυτό, αξιοποιούνται τεχνικές Explainable Artificial Intelligence και ειδικότερα η μέθοδος TE2Rules, η οποία παρουσιάζεται αναλυτικά στην επόμενη ενότητα.

Στο Σχήμα 2.7.1 παρουσιάζεται σχηματικά ο ρόλος της εξηγήσιμης τεχνητής νοημοσύνης στον τομέα της υγείας και η αλληλεπίδρασή της με τους βασικούς εμπλεκόμενους φορείς.



Σχήμα 2.7.1: Ρόλος της εξηγήσιμης τεχνητής νοημοσύνης (XAI) στον τομέα της υγείας και η αλληλεπίδρασή της με τους βασικούς εμπλεκόμενους φορείς.

2.8 TE2Rules

Η μέθοδος TE2Rules αποτελεί μία τεχνική ερμηνευσιμότητας (Explainable Artificial Intelligence – XAI), η οποία στοχεύει στην εξαγωγή κατανοητών και ερμηνεύσιμων κανόνων από μοντέλα μηχανικής μάθησης [10]. Η βασική φιλοσοφία της προσέγγισης είναι η μετατροπή πολύπλοκων μοντέλων, όπως τα σύνολα δέντρων αποφάσεων (tree ensembles), σε ένα σύνολο απλών λογικών κανόνων που μπορούν να κατανοηθούν εύκολα από τον άνθρωπο.

Οι κανόνες που παράγονται έχουν συνήθως τη μορφή συνθηκών τύπου «if-then» και περιγράφουν συνδυασμούς χαρακτηριστικών που οδηγούν σε μία συγκεκριμένη απόφαση ή κατηγορία. Με τον τρόπο αυτό καθίσταται δυνατή η καλύτερη κατανόηση των παραγόντων που επηρεάζουν τη συμπεριφορά του μοντέλου και η ερμηνεία των αποτελεσμάτων με περισσότερο διαφανή και κατανοητό τρόπο.

Η TE2Rules επιχειρεί ουσιαστικά να γεφυρώσει το χάσμα μεταξύ υψηλής υπολογιστικής απόδοσης και ερμηνευσιμότητας. Παρόλο που, πολλά σύγχρονα μοντέλα μηχανικής μάθησης επιτυγχάνουν ιδιαίτερα υψηλή ακρίβεια, συχνά χαρακτηρίζονται από περιορισμένη διαφάνεια, γεγονός που δυσκολεύει την κλινική αξιοποίησή τους. Η συγκεκριμένη μέθοδος επιτρέπει την παραγωγή απλών και εύκολα αναγνώσιμων κανόνων, χωρίς να χάνεται σημαντικό μέρος της πληροφορίας που περιέχεται στα δεδομένα.

Στο πλαίσιο της ανάλυσης συστάδων, η TE2Rules μπορεί να αξιοποιηθεί για την ερμηνεία των clusters, μέσω της εξαγωγής κανόνων που περιγράφουν τα χαρακτηριστικά κάθε ομάδας. Οι κανόνες αυτοί βασίζονται στις αναθέσεις των παρατηρήσεων στις συστάδες και επιτρέπουν την καλύτερη κατανόηση των μορφολογικών προτύπων που προκύπτουν από τα δεδομένα, καθώς και τη σύνδεσή τους με συγκεκριμένα κλινικά χαρακτηριστικά.

Ένα από τα σημαντικότερα πλεονεκτήματα της μεθόδου είναι ότι, οι κανόνες που παράγονται είναι εύκολα αναγνώσιμοι και μπορούν να αξιολογηθούν άμεσα από κλινικούς ιατρούς ή ερευνητές. Έτσι, η TE2Rules δεν περιορίζεται μόνο στην περιγραφή των αποτελεσμάτων ενός μοντέλου, αλλά συμβάλλει και στην καλύτερη κατανόηση των σχέσεων μεταξύ των χαρακτηριστικών των ασθενών και των μορφολογικών προτύπων που αναδεικνύονται.

Στην παρούσα εργασία, οι αναθέσεις των ασθενών στις συστάδες χρησιμοποιούνται ως μεταβλητή-στόχος, ενώ τα μορφολογικά και κλινικά χαρακτηριστικά αξιοποιούνται ως μεταβλητές εισόδου. Μέσω της εφαρμογής της TE2Rules εξάγονται ερμηνεύσιμοι κανόνες που περιγράφουν τα χαρακτηριστικά κάθε συστάδας και συμβάλλουν στην κατανόηση των μορφολογικών διαφοροποιήσεων μεταξύ των ομάδων ασθενών.

Η αξιοποίηση της συγκεκριμένης μεθόδου ενισχύει σημαντικά τη διαφάνεια της ανάλυσης και επιτρέπει τη σύνδεση των υπολογιστικών αποτελεσμάτων με κλινικά ερμηνεύσιμα μορφολογικά πρότυπα. Με τον τρόπο αυτό, διευκολύνεται η καλύτερη αξιοποίηση των τεχνικών μηχανικής μάθησης στον τομέα της καρωτιδικής νόσου και ενισχύεται η δυνατότητα μελλοντικής ενσωμάτωσης τέτοιων προσεγγίσεων στην κλινική πράξη.

Κεφάλαιο 3: Μεθοδολογία

3.1	Περιγραφή δεδομένων	38
3.2	Διάγραμμα ροής μεθοδολογίας	41
3.3	Προεπεξεργασία δεδομένων	44
3.4	Διαχωρισμός δεδομένων	46
3.5	Μέθοδος Clustering (K-Modes & K-Prototypes)	47
3.6	Επιλογή αριθμού συστάδων	50
3.7	Μετρικές Αξιολόγησης	53
3.8	Μέθοδος TE2Rules	55

3.1 Περιγραφή του συνόλου δεδομένων

Η παρούσα μελέτη βασίστηκε σε σύνολο δεδομένων που περιλαμβάνει 996 ασθενείς με καρωτιδική νόσο. Τα δεδομένα προήλθαν από δύο βασικά αρχεία, τα *ACSRs_JBA_Categorical* και *acsrJsba_EventsLastFuAge*, τα οποία περιέχουν πληροφορίες σχετικά με τα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας, καθώς και δεδομένα που αφορούν την εμφάνιση ισχαιμικών επεισοδίων και τον χρόνο παρακολούθησης των ασθενών.

Οι μεταβλητές του συνόλου δεδομένων σχετίζονται κυρίως με μορφολογικά και υπερηχογραφικά χαρακτηριστικά της αθηρωματικής πλάκας, όπως ο βαθμός στένωσης της καρωτίδας, η ηχογένεια της πλάκας, η επιφάνειά της και ειδικοί δείκτες μορφολογικού κινδύνου, όπως η παρουσία *Discrete White Areas (DWA)* και *Juxtaluminal Black Area (JBA)*. Τα χαρακτηριστικά αυτά αποτέλεσαν τη βασική πληροφορία για τη διαδικασία ομαδοποίησης των ασθενών.

Επιπλέον, το σύνολο δεδομένων περιλαμβάνει πρόσθετες κλινικές και δημογραφικές μεταβλητές, όπως η κλινική κατάσταση των ασθενών (*PATSTAT*), η καταγραφή ισχαιμικών επεισοδίων (*EVENTS*), ο χρόνος παρακολούθησης (*LASTFU*) και η ηλικία (*AGE*). Επισημαίνεται ότι η μεταβλητή *AGE* αξιοποιήθηκε ως αριθμητικό χαρακτηριστικό εισόδου αποκλειστικά κατά την εφαρμογή του αλγορίθμου *K-Prototypes*, προκειμένου να ενσωματωθεί η ηλικιακή πληροφορία στη διαδικασία ομαδοποίησης. Αντίθετα, οι μεταβλητές *PATSTAT*, *EVENTS* και *LASTFU* δεν συμμετείχαν ως χαρακτηριστικά εισόδου σε κανέναν από τους δύο αλγορίθμους *clustering*, αλλά χρησιμοποιήθηκαν μεταγενέστερα, τόσο για τον διαχωρισμό του πληθυσμού σε επιμέρους υποσύνολα μελέτης όσο και για την οπτικοποίηση, την αξιολόγηση και την κλινική ερμηνεία των παραγόμενων συστάδων.

Για την εφαρμογή του αλγορίθμου K-Prototypes χρησιμοποιήθηκε η συνεχής αριθμητική μεταβλητή της ηλικίας (AGE) σε συνδυασμό με τις έξι κατηγορικές μεταβλητές του συνόλου δεδομένων, επιτρέποντας την ταυτόχρονη αξιοποίηση τόσο ποιοτικών όσο και ποσοτικών χαρακτηριστικών. Πριν από την εκτέλεση του αλγορίθμου, η αριθμητική μεταβλητή AGE κανονικοποιήθηκε με τη χρήση της μεθόδου RobustScaler, με στόχο τον περιορισμό της επίδρασης τυχόν ακραίων τιμών και την αποφυγή σφαλμάτων λόγω διαφοράς κλίμακας με τα υπόλοιπα χαρακτηριστικά. Η συγκεκριμένη δομή επιτρέπει τη διεξοδική διερεύνηση της μορφολογικής ετερογένειας των ασθενών και την αναγνώριση διακριτών προτύπων που σχετίζονται με την καρωτιδική νόσο.

Οι μεταβλητές του συνόλου δεδομένων οργανώθηκαν με βάση τον λειτουργικό τους ρόλο στη μεθοδολογία της μελέτης σε τρεις βασικές κατηγορίες, οι οποίες παρουσιάζονται αναλυτικά στους Πίνακες 3.1.1, 3.1.2 και 3.1.3.

Μεταβλητή	Περιγραφή
STN	Ποσοστό στένωσης της καρωτίδας σε αριθμητική μορφή.
GSMN	Αριθμητική τιμή ηχογένειας της πλάκας (<i>Gray Scale Median</i>).
PLAREAN	Αριθμητική τιμή επιφάνειας της αθηρωματικής πλάκας.
DWAN	Δυαδική αριθμητική μεταβλητή για την παρουσία/απουσία <i>Discrete White Areas</i> (0 ή 1).
CTIASTRN	Δυαδική αριθμητική μεταβλητή για το ιστορικό αντίπλευρου <i>TIA</i> ή εγκεφαλικού.
JBAN	Αριθμητική τιμή του δείκτη <i>Juxtaluminal Black Area</i> .
AGE	Ηλικία ασθενούς σε έτη (η οποία αποτελεί και την αριθμητική είσοδο στον <i>K-Prototypes</i>).

Πίνακας 3.1.1 Αριθμητικές μεταβλητές του συνόλου δεδομένων

Μεταβλητή	Περιγραφή
STENOSIS	Κατηγορία βαθμού στένωσης (low, medium, high)
GSM	Κατηγορία ηχογένειας της πλάκας
PLAREA	Κατηγορία επιφάνειας της πλάκας
DWA	Παρουσία ή απουσία Discrete White Areas (present / absent)
CTIASTR	Παρουσία ή απουσία ιστορικού αντίπλευρου ΤΙΑ ή εγκεφαλικού επεισοδίου
JBA	Κατηγορία Juxtaluminal Black Area (low, medium, high, very high)

Πίνακας 3.1.2 Κατηγορικές μεταβλητές που χρησιμοποιήθηκαν στην ανάλυση συστάδων

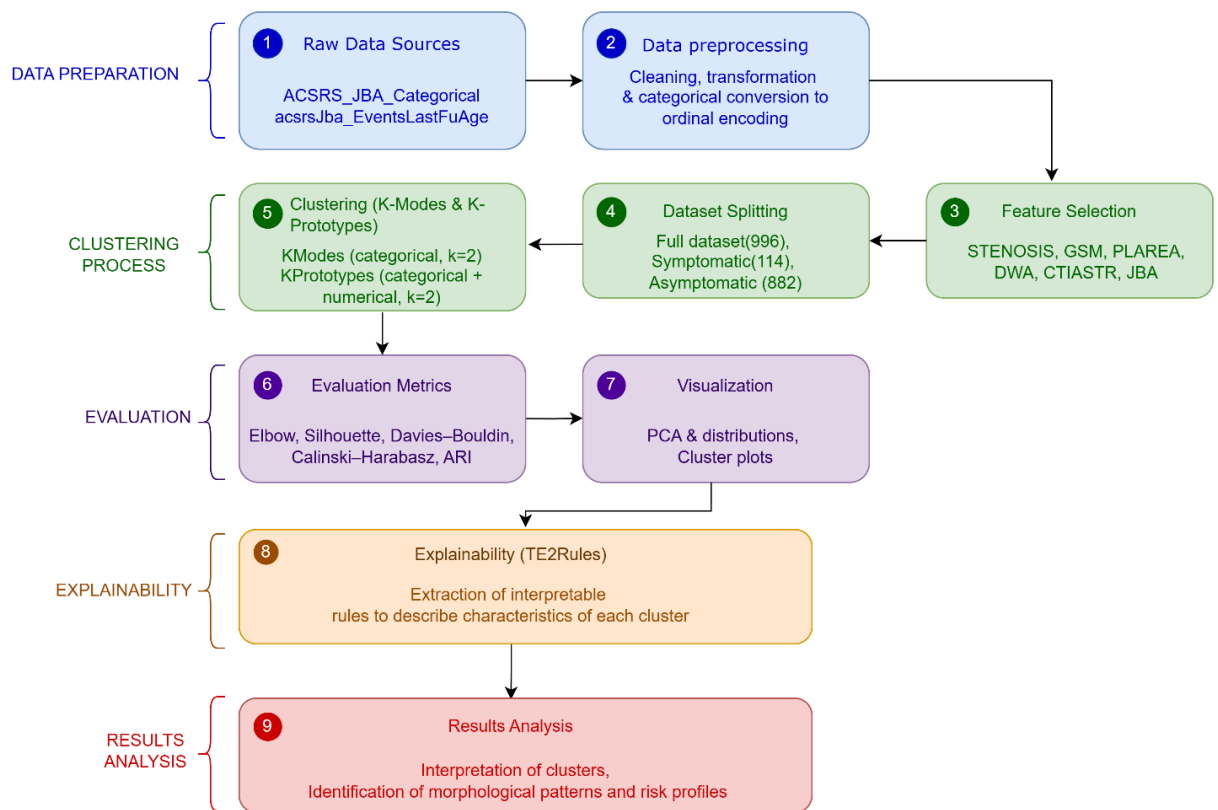
Μεταβλητή	Περιγραφή
PATSTAT	Κλινική κατηγοριοποίηση ασθενών ως συμπτωματικών (sympt) ή ασυμπτωματικών (asympt)
EVENTS	Αριθμός ισχαιμικών επεισοδίων
LASTFU	Χρόνος παρακολούθησης (σε εξάμηνα)
CUT72	Κατηγορικός διαχωρισμός lastfu σε ≤ 3 έτη και > 3 έτη — χρησιμοποιείται αποκλειστικά για οπτικοποίηση στα PCA plots και όχι ως χαρακτηριστικό εισόδου στους αλγορίθμους clustering
AGECL	Κατηγορία ηλικίας

Πίνακας 3.1.3 Μεταβλητές που χρησιμοποιήθηκαν για την αξιολόγηση και ερμηνεία των συστάδων.

3.2 Διάγραμμα ροής μεθοδολογίας

Στην παρούσα ενότητα παρουσιάζεται το διάγραμμα ροής της μεθοδολογίας που ακολουθήθηκε για την επεξεργασία και ανάλυση των δεδομένων. Το διάγραμμα συνοψίζει τα βασικά στάδια της διαδικασίας, από τη συλλογή και προεπεξεργασία των δεδομένων έως την εφαρμογή των αλγορίθμων ομαδοποίησης και την ερμηνεία των αποτελεσμάτων.

Η διαδικασία της ανάλυσης αποτελείται από διαδοχικά στάδια, τα οποία αποτυπώνονται συνοπτικά στο Σχήμα 3.2.1 και επεξηγούνται αναλυτικά στη συνέχεια.



Σχήμα 3.2.1 Διάγραμμα ροής των διαδοχικών σταδίων της προτεινόμενης μεθοδολογίας (Data Preparation, Clustering Process, Evaluation, Explainability, Results Analysis).

Αρχικά πραγματοποιήθηκε η συλλογή των δεδομένων από τις δύο βασικές πηγές, τα αρχεία ACSRS_JBA_Categorical και acsrsJba_EventsLastFuAge, τα οποία περιλαμβάνουν μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας, καθώς και πληροφορίες σχετικά με την εμφάνιση ισχαιμικών επεισοδίων και τον χρόνο παρακολούθησης των ασθενών. Το αρχείο ACSRS_JBA_Categorical περιλαμβάνει κυρίως τις κατηγορικές μεταβλητές που χρησιμοποιήθηκαν στον αλγόριθμο K-Modes, ενώ το αρχείο acsrsJba_EventsLastFuAge περιλαμβάνει επιπλέον τις αντίστοιχες αριθμητικές μορφές των μεταβλητών, οι οποίες αξιοποιήθηκαν για τις ανάγκες της στατιστικής ανάλυσης και της οπτικοποίησης, καθώς και τη συνεχή μεταβλητή της ηλικίας.

Στη συνέχεια, ακολούθησε το στάδιο της προεπεξεργασίας των δεδομένων, το οποίο περιλάμβανε καθαρισμό, έλεγχο συνέπειας και κατάλληλο μετασχηματισμό των μεταβλητών σύμφωνα με τις απαιτήσεις των αλγορίθμων clustering. Για τον K-Modes χρησιμοποιήθηκαν αποκλειστικά κατηγορικές μεταβλητές, ενώ για τον K-Prototypes αξιοποιήθηκαν τόσο κατηγορικά, όσο και αριθμητικά χαρακτηριστικά. Οι αριθμητικές μεταβλητές κανονικοποιήθηκαν με χρήση της μεθόδου RobustScaler, ώστε να περιοριστεί η επίδραση ακραίων τιμών και διαφορών κλίμακας μεταξύ των χαρακτηριστικών.

Έπειτα, πραγματοποιήθηκε η επιλογή των βασικών χαρακτηριστικών που χρησιμοποιήθηκαν στην ανάλυση. Συγκεκριμένα, για τον K-Modes χρησιμοποιήθηκαν οι κατηγορικές μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA, ενώ για τον αλγόριθμο K-Prototypes αξιοποιήθηκαν οι ίδιες έξι κατηγορικές μεταβλητές με την προσθήκη της συνεχούς μεταβλητής AGE.

Κατόπιν, το σύνολο δεδομένων διαχωρίστηκε σε δύο υποομάδες με βάση την κλινική κατάσταση των ασθενών: συμπτωματικούς (n=114) και ασυμπτωματικούς ασθενείς (n=882). Η ανάλυση συστάδων εφαρμόστηκε ξεχωριστά σε κάθε υποομάδα, με στόχο τη διερεύνηση πιθανών διαφορών στα μορφολογικά πρότυπα μεταξύ των δύο ομάδων.

Παράλληλα, η μεταβλητή LASTFU, η οποία εκφράζει τον χρόνο παρακολούθησης σε εξάμηνα, χρησιμοποιήθηκε αποκλειστικά για σκοπούς οπτικοποίησης στα PCA plots. Συγκεκριμένα, οι ασθενείς διακρίθηκαν γραφικά σε δύο χρονικές κατηγορίες, με τιμή $LASTFU \leq 6$ να αντιστοιχεί σε χρόνο παρακολούθησης έως 3 έτη και τιμή $LASTFU > 6$ σε χρόνο παρακολούθησης μεγαλύτερο των 3 ετών. Διευκρινίζεται ότι η μεταβλητή αυτή δεν χρησιμοποιήθηκε ως χαρακτηριστικό εισόδου στους αλγορίθμους clustering.

Για την ανάλυση συστάδων εφαρμόστηκαν δύο αλγόριθμοι: ο K-Modes, ο οποίος είναι κατάλληλος για καθαρά κατηγορικά δεδομένα και ο K-Prototypes, ο οποίος αξιοποιεί συνδυαστικά κατηγορικές και αριθμητικές μεταβλητές. Ο αριθμός των συστάδων προσδιορίστηκε με χρήση της μεθόδου elbow σε συνδυασμό με δείκτες αξιολόγησης clustering, οδηγώντας στην επιλογή $k = 2$ και για τους δύο αλγορίθμους.

Η αξιολόγηση της ποιότητας των συστάδων πραγματοποιήθηκε με χρήση των δεικτών Silhouette, Davies–Bouldin και Calinski–Harabasz, ενώ παράλληλα αξιοποιήθηκε και η σταθερότητα των αποτελεσμάτων μέσω του δείκτη Adjusted Rand Index (ARI). Για την σαφέστερη κατανόηση της δομής των δεδομένων εφαρμόστηκαν επίσης τεχνικές οπτικοποίησης, όπως η ανάλυση κύριων συνιστωσών (Principal Component Analysis – PCA) και διαγράμματα κατανομών χαρακτηριστικών.

Ακολούθως, εφαρμόστηκε η μέθοδος ερμηνευσιμότητας TE2Rules, με στόχο την εξαγωγή κατανοητών κανόνων που περιγράφουν τα χαρακτηριστικά κάθε συστάδας. Η προσέγγιση που ακολουθήθηκε βασίστηκε στη λογική “cluster vs. rest”, όπου κάθε συστάδα συγκρίνεται με το σύνολο των υπόλοιπων παρατηρήσεων.

Τέλος, τα αποτελέσματα της ανάλυσης ερμηνεύτηκαν με στόχο την κατανόηση των διαφορών μεταξύ των συστάδων και την ανάδειξη διακριτών μορφολογικών προτύπων που σχετίζονται με διαφορετικά επίπεδα μορφολογικού κινδύνου.

3.3 Προεπεξεργασία δεδομένων

Πριν από την εφαρμογή της ανάλυσης συστάδων, πραγματοποιήθηκε κατάλληλη προεπεξεργασία των δεδομένων, με στόχο τη διασφάλιση της συμβατότητάς τους με τους αλγόριθμους K-Modes και K-Prototypes και τη βελτίωση της αξιοπιστίας της διαδικασίας ομαδοποίησης.

Αρχικά έγινε επιλογή των βασικών χαρακτηριστικών που χρησιμοποιήθηκαν στην ανάλυση. Αναλυτικότερα, επιλέχθηκαν οι μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA, οι οποίες περιγράφουν μορφολογικά και υπερηχογραφικά χαρακτηριστικά της αθηρωματικής πλάκας και σχετίζονται άμεσα με τον μορφολογικό κίνδυνο της καρωτιδικής νόσου.

Οι μεταβλητές αυτές ήταν ήδη διαθέσιμες σε κατηγορική μορφή στο αρχικό σύνολο δεδομένων. Για τον λόγο αυτό δεν απαιτήθηκε επιπλέον διαδικασία κατηγοριοποίησης (discretization). Ωστόσο, πραγματοποιήθηκε μετατροπή τους σε μορφή συμβολοσειράς (string), ώστε να είναι πλήρως συμβατές με τον αλγόριθμο K-Modes, ο οποίος βασίζεται στη σύγκριση κατηγορικών τιμών μέσω μετρικών ασυμφωνίας.

Σε μετέπειτα στάδιο πραγματοποιήθηκε έλεγχος ποιότητας και συνέπειας των δεδομένων. Οι εγγραφές που περιείχαν ελλείπουσες τιμές (missing values) στις βασικές μεταβλητές της ανάλυσης απομακρύνθηκαν, προκειμένου να διασφαλιστεί η σταθερότητα και η αξιοπιστία της διαδικασίας clustering. Συγχρόνως, αφαιρέθηκαν μεταβλητές που δεν συνεισέφεραν άμεσα στην ομαδοποίηση, όπως αναγνωριστικά πεδία (π.χ. caseID), προκειμένου να αποφευχθεί η εισαγωγή θορύβου στο μοντέλο.

Επιπροσθέτως, εξετάστηκε η δυνατότητα αξιοποίησης των αριθμητικών μεταβλητών μέσω διαφορετικών σχημάτων κατηγοριοποίησης για χρήση στον αλγόριθμο K-Modes. Συγκεκριμένα, εξετάστηκαν οι μεταβλητές STN, GSMN, PLAREAN, DWAN, CTIASTRN και JBAN. Παρόλα αυτά, η προσέγγιση αυτή δεν οδήγησε σε ικανοποιητικά αποτελέσματα, καθώς οι συστάδες που προέκυψαν παρουσίασαν χαμηλότερη συνοχή και μειωμένη διακριτότητα. Για τον λόγο αυτό, αποφασίστηκε η χρήση αποκλειστικά των αρχικών κατηγορικών μεταβλητών στον K-Modes.

Αντιθέτως, για την εφαρμογή του αλγορίθμου K-Prototypes, οι έξι κατηγορικές μεταβλητές συνδυάστηκαν με τη συνεχή αριθμητική μεταβλητή της ηλικίας των ασθενών (AGE). Οι αριθμητικές μεταβλητές χρησιμοποιήθηκαν στην αρχική τους μορφή, αφού προηγουμένως υποβλήθηκαν σε κανονικοποίηση με χρήση της μεθόδου RobustScaler. Η συγκεκριμένη μέθοδος επιλέχθηκε λόγω της ανθεκτικότητάς της απέναντι σε ακραίες τιμές (outliers), αφού βασίζεται στη διάμεσο και στο ενδοτεταρτημοριακό εύρος (Interquartile Range – IQR), επιτρέποντας πιο αξιόπιστη κανονικοποίηση των κλινικών δεδομένων, χωρίς σημαντική αλλοίωση της κατανομής τους.

Η διαδικασία αυτή επέτρεψε την κατάλληλη προσαρμογή των δεδομένων στις απαιτήσεις των δύο αλγορίθμων clustering και συνέβαλε ουσιαστικά στη βελτίωση της ποιότητας και της αξιοπιστίας των αποτελεσμάτων της ανάλυσης.

3.4 Διαχωρισμός δεδομένων

Μετά την ολοκλήρωση της προεπεξεργασίας και της επιλογής των χαρακτηριστικών, το σύνολο δεδομένων διαχωρίστηκε σε επιμέρους υποσύνολα, με στόχο τη διερεύνηση της δομής των δεδομένων υπό διαφορετικές κλινικές συνθήκες και την καλύτερη κατανόηση των προτύπων που σχετίζονται με την καρωτιδική νόσο.

Αρχικά, η ανάλυση πραγματοποιήθηκε στο σύνολο των δεδομένων, το οποίο περιλαμβάνει όλους τους ασθενείς της μελέτης ($n = 996$), παρέχοντας μία συνολική εικόνα της κατανομής των χαρακτηριστικών και των προτύπων που εμφανίζονται στον πληθυσμό.

Στη συνέχεια, το σύνολο δεδομένων διαχωρίστηκε με βάση την κλινική κατάσταση των ασθενών, δημιουργώντας δύο βασικές υποομάδες:

- Συμπτωματικοί ασθενείς ($n = 114$)
- Ασυμπτωματικοί ασθενείς ($n = 882$)

Στο πλαίσιο της παρούσας μελέτης, ως συμπτωματικοί θεωρούνται οι ασθενείς που παρουσίασαν τουλάχιστον ένα ισχαιμικό επεισόδιο κατά την περίοδο παρακολούθησης. Κατά αντίστοιχα ως ασυμπτωματικοί θεωρούνται όσοι δεν εμφάνισαν αντίστοιχο επεισόδιο. Ο διαχωρισμός αυτός κρίνεται ιδιαίτερα καθοριστικός, καθώς οι δύο ομάδες ενδέχεται να παρουσιάζουν διαφορετικά χαρακτηριστικά της αθηρωματικής πλάκας και διαφορετικό επίπεδο μορφολογικού κινδύνου.

Η ανάλυση συστάδων εφαρμόστηκε τόσο στο σε όλο το σύνολο δεδομένων, όσο και ξεχωριστά σε κάθε υποομάδα ασθενών, με στόχο τη διερεύνηση πιθανών διαφοροποιήσεων στη δομή των συστάδων και στα πρότυπα που προκύπτουν από τα δεδομένα.

Η προσέγγιση αυτή επιτρέπει τη συγκριτική μελέτη των διαφορετικών ομάδων ασθενών και συμβάλλει στην καλύτερη κατανόηση της ετερογένειας που χαρακτηρίζει την καρωτιδική νόσο. Στο επόμενο στάδιο παρουσιάζεται η εφαρμογή των αλγορίθμων K-Modes και K-Prototypes για την ομαδοποίηση των δεδομένων και την ανάδειξη των αντίστοιχων συστάδων.

3.5 Μέθοδος clustering (K-Modes & K-Prototypes)

Μετά τον διαχωρισμό των δεδομένων σε επιμέρους υποομάδες, εφαρμόστηκε ανάλυση συστάδων με χρήση δύο αλγορίθμων: του *K-Modes*, για τα κατηγορικά χαρακτηριστικά, και του *K-Prototypes*, για την ταυτόχρονη αξιοποίηση κατηγορικών και αριθμητικών μεταβλητών. Οι δύο προαναφερθέντες αλγόριθμοι εφαρμόστηκαν στο σύνολο των δεδομένων και στις επιμέρους υποομάδες ασθενών, με πολλαπλές αρχικοποιήσεις (*multiple initializations*), για να μειωθεί η πιθανότητα σύγκλισης σε τοπικά βέλτιστες λύσεις.

Η ανάλυση συστάδων μπορεί να πραγματοποιηθεί μέσω διαφορετικών αλγορίθμων, όπως ο *K-Means*, η ιεραρχική ομαδοποίηση (*hierarchical clustering*) και ο αλγόριθμος *DBSCAN*. Ωστόσο, πολλές από αυτές τις μεθόδους είναι σχεδιασμένες κυρίως για αριθμητικά δεδομένα και παρουσιάζουν σημαντικούς περιορισμούς όταν εφαρμόζονται σε κατηγορικές μεταβλητές.

Λόγου χάρη, ο αλγόριθμος *K-Means*, βασίζεται στην ευκλείδεια απόσταση και στον υπολογισμό μέσων τιμών, έννοιες που δεν έχουν φυσική ερμηνεία όταν εφαρμόζονται σε κατηγορικά δεδομένα. Αντίστοιχα, μέθοδοι βασισμένες στην πυκνότητα, όπως ο *DBSCAN*, απαιτούν τον ορισμό μέτρου απόστασης, το οποίο δεν μπορεί να αποδώσει ικανοποιητικά τις σχέσεις μεταξύ μη αριθμητικών χαρακτηριστικών. Για τον λόγο αυτό, οι συγκεκριμένες προσεγγίσεις δεν κρίθηκαν κατάλληλες για την παρούσα ανάλυση.

Λαμβάνοντας υπόψη ότι, το σύνολο δεδομένων περιλαμβάνει κυρίως κατηγορικά μορφολογικά χαρακτηριστικά, όπως και αριθμητικές μεταβλητές που περιέχουν σημαντική πληροφορία, επιλέχθηκαν οι αλγόριθμοι *K-Modes* και *K-Prototypes*. Ο *K-Modes* αποτελεί προσαρμογή του *K-Means* για κατηγορικά δεδομένα, αντικαθιστώντας τα κεντροειδή με τις επικρατούσες τιμές (*modes*).

Ο αλγόριθμος *K-Prototypes* επεκτείνει την προσέγγιση αυτή σε δεδομένα μικτού τύπου, συνδυάζοντας τη μέτρηση ασυμφωνίας (*matching dissimilarity*) για τα κατηγορικά χαρακτηριστικά και την ευκλείδεια απόσταση για τα αριθμητικά δεδομένα. Έτσι, οι αριθμητικές μεταβλητές υποβλήθηκαν σε κανονικοποίηση πριν από την εφαρμογή του αλγορίθμου, προκειμένου να εξισορροπηθεί η συνεισφορά τους στη συνολική απόσταση.

Οι δύο αλγόριθμοι λειτουργούν επαναληπτικά, μέσω δύο βασικών βημάτων:

- Βήμα ανάθεσης (*assignment step*): Κάθε παρατήρηση ανατίθεται στη συστάδα με τη μικρότερη απόσταση, με βάση τη *matching dissimilarity* στον *K-Modes* ή τον συνδυασμό *matching dissimilarity* και ευκλείδειας απόστασης στον *K-Prototypes*.
- Βήμα ενημέρωσης (*update step*): Ενημερώνονται τα κέντρα των συστάδων. Στον *K-Modes* υπολογίζονται εκ νέου οι επικρατούσες τιμές (*modes*) των κατηγορικών χαρακτηριστικών για κάθε ομάδα, ενώ στον *K-Prototypes* ενημερώνονται τόσο οι επικρατούσες τιμές των κατηγορικών χαρακτηριστικών όσο και οι μέσες τιμές (*means*) των αριθμητικών μεταβλητών.

Η διαδικασία αυτή επαναλαμβάνεται μέχρι την επίτευξη σύγκλισης, δηλαδή όταν οι αναθέσεις των παρατηρήσεων στις συστάδες παραμένουν σταθερές μεταξύ διαδοχικών επαναλήψεων ή όταν η μεταβολή της συνολικής συνάρτησης κόστους γίνει αμελητέα.

Διευκρινίζεται ότι, οι αλγόριθμοι εφαρμόστηκαν σε κάθε υποσύνολο δεδομένων που προέκυψε από τον διαχωρισμό των ασθενών, με στόχο την ανάδειξη μορφολογικών προτύπων και τη διερεύνηση πιθανών διαφοροποιήσεων μεταξύ των ομάδων. Η επιλογή του *K-Modes* κρίθηκε ιδιαίτερα κατάλληλη για τα κατηγορικά δεδομένα της μελέτης, διότι επιτρέπει την αξιόπιστη ομαδοποίηση, χωρίς την ανάγκη μετασχηματισμού των μεταβλητών σε αριθμητική μορφή. Παράλληλα, ο *K-Prototypes* επιτρέπει την ενσωμάτωση και της αριθμητικής πληροφορίας, προσφέροντας μία πιο ολοκληρωμένη εικόνα της δομής των δεδομένων.

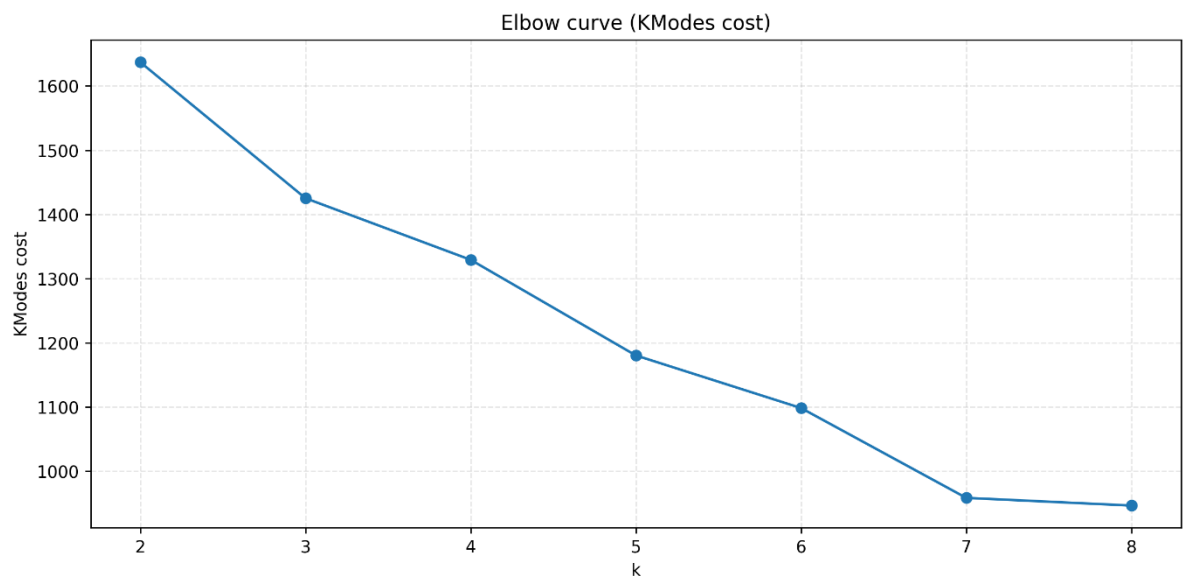
Για την εφαρμογή των αλγορίθμων απαιτείται ο προκαθορισμός του αριθμού των συστάδων (k). Η επιλογή της κατάλληλης τιμής του k πραγματοποιήθηκε μέσω της συνδυαστικής χρήσης της μεθόδου «elbow» και δεικτών αξιολόγησης clustering, όπως παρουσιάζεται αναλυτικά στην επόμενη ενότητα.

3.6 Επιλογή αριθμού συστάδων

Η επιλογή του κατάλληλου αριθμού συστάδων (k) αποτελεί ιδιαίτερα σημαντικό στάδιο στην ανάλυση συστάδων, καθώς επηρεάζει άμεσα τη δομή, τη σταθερότητα και την ερμηνευσιμότητα των αποτελεσμάτων. Στην παρούσα εργασία, η επιλογή της τιμής του k βασίστηκε σε συνδυασμό εμπειρικών, ποσοτικών και κλινικών κριτηρίων, με στόχο την εξασφάλιση αξιόπιστης και κλινικά ερμηνεύσιμης ομαδοποίησης.

Αρχικά, χρησιμοποιήθηκε η μέθοδος του αγκώνα (elbow method), η οποία βασίζεται στη μεταβολή του κόστους των αλγορίθμων K-Modes και K-Prototypes σε συνάρτηση με τον αριθμό των συστάδων. Η μέθοδος αυτή επιτρέπει τον εντοπισμό ενός σημείου καμπής (elbow), πέρα από το οποίο η αύξηση του αριθμού των συστάδων οδηγεί σε ολοένα μικρότερη βελτίωση της συνάρτησης κόστους του μοντέλου.

Όπως παρουσιάζεται στο Σχήμα 3.6.1, η καμπύλη elbow εμφανίζει τη σημαντικότερη μείωση του κόστους κατά τη μετάβαση από $k=2$ στο $k=3$, χωρίς ωστόσο να παρατηρείται ένα μοναδικό και απόλυτα σαφές σημείο καμπής. Παρ' όλα αυτά, η μεγαλύτερη σχετική μεταβολή εντοπίζεται στις μικρές τιμές του k , ενώ μετά το $k=2$ η μείωση του κόστους εξελίσσεται με σαφώς πιο σταδιακό και ομαλό ρυθμό. Το εύρημα αυτό υποδηλώνει ότι η αύξηση του αριθμού των συστάδων πέρα από δύο ομάδες δεν συνοδεύεται από αντίστοιχα σημαντική βελτίωση της ομαδοποίησης.

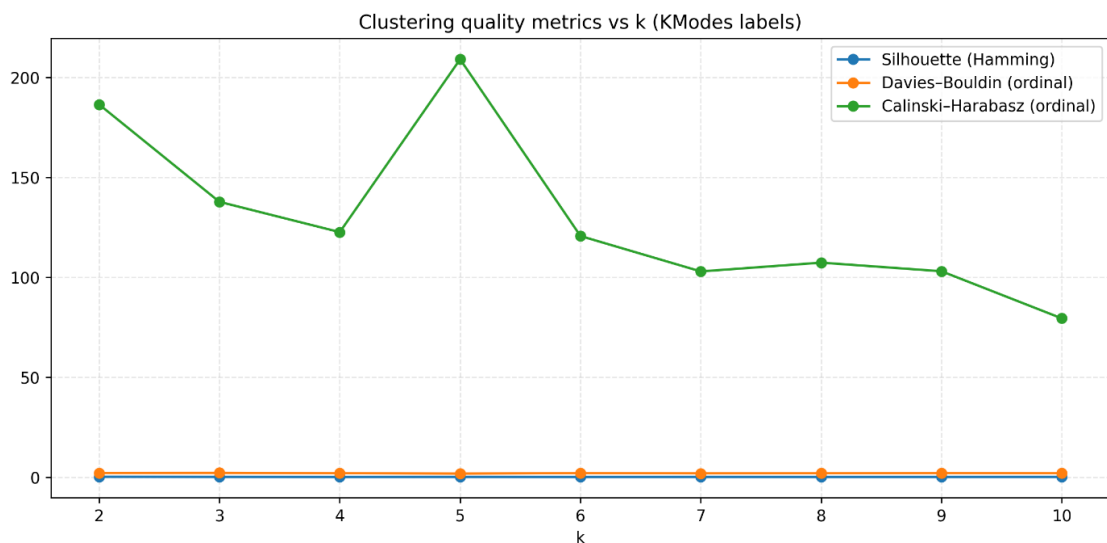


Σχήμα 3.6.1 Καμπύλη elbow για την επιλογή του αριθμού k

Ταυτόχρονα, η ποιότητα της ομαδοποίησης αξιολογήθηκε ποσοτικά με χρήση εσωτερικών δεικτών clustering, όπως οι δείκτες Silhouette, Davies–Bouldin και Calinski–Harabasz, οι οποίοι εκτιμούν τη συνοχή στο εσωτερικό των ομάδων και του διαχωρισμού μεταξύ τους. Επιπλέον, υπολογίστηκε ο δείκτης Adjusted Rand Index (ARI) για την αξιολόγηση της σταθερότητας των αποτελεσμάτων έπειτα από πολλαπλές εκτελέσεις των αλγορίθμων.

Στο Σχήμα 3.6.2 παρουσιάζονται οι τιμές των δεικτών αξιολόγησης για διαφορετικές τιμές του k . Από την ανάλυση των μετρικών, προκύπτει ότι η επιλογή $k=2$ εξασφαλίζει την καλύτερη συνολική ισορροπία μεταξύ συνοχής, διαχωρισμού και σταθερότητας. Συγκεκριμένα, για $k=2$ παρατηρείται η υψηλότερη τιμή του δείκτη Silhouette (Hamming) και η χαμηλότερη τιμή του δείκτη Davies–Bouldin (ordinal), γεγονός που υποδηλώνει μεγαλύτερη εσωτερική ομοιογένεια και καλύτερο διαχωρισμό μεταξύ των συστάδων.

Αν και ο δείκτης Calinski–Harabasz (ordinal) εμφανίζει τη μέγιστη τιμή του στο $k=5$, η επιλογή μικρότερου αριθμού συστάδων κρίθηκε καταλληλότερη, καθώς το $k=2$ συνοδεύεται από ισχυρή τοπική τιμή του δείκτη, ενώ ταυτόχρονα διασφαλίζει μεγαλύτερη απλότητα, σταθερότητα και κλινική ερμηνευσιμότητα των αποτελεσμάτων. Επιπρόσθετα, για $k=2$ ο δείκτης σταθερότητας ARI έλαβε τιμή ίση με 1.0, γεγονός που υποδηλώνει απόλυτη σταθερότητα και αναπαραγωγή των αποτελεσμάτων μεταξύ επαναλαμβανόμενων εκτελέσεων.



Σχήμα 3.6.2 Μετρικές ποιότητας clustering ως συνάρτηση του k .

Ο συνδυασμός της μεθόδου elbow και των εσωτερικών δεικτών αξιολόγησης επέτρεψε την αξιόπιστη εκτίμηση της καταλληλότερης τιμής του k , λαμβάνοντας εξίσου υπόψη τη δομή των δεδομένων και τη σταθερότητα της ομαδοποίησης.

Με βάση τα παραπάνω, επιλέχθηκε η τιμή $k=2$ ως η πλέον κατάλληλη για την παρούσα ανάλυση, καθώς προσφέρει την καλύτερη ισορροπία μεταξύ απλότητας, σταθερότητας και κλινικής ερμηνευσιμότητας των αποτελεσμάτων. Η επιλογή αυτή διατηρήθηκε σταθερή σε όλα τα υποσύνολα δεδομένων και στους δύο αλγορίθμους, με στόχο να διασφαλιστεί η δυνατότητα άμεσης και αξιόπιστης συγκριτικής ανάλυσης μεταξύ των διαφορετικών ομάδων ασθενών.

3.7 Μετρικές αξιολόγησης

Η αξιολόγηση της ποιότητας της ανάλυσης συστάδων πραγματοποιήθηκε με χρήση εσωτερικών δεικτών clustering και με ανάλυση της σταθερότητας των αποτελεσμάτων. Οι μετρικές αυτές χρησιμοποιήθηκαν με στόχο την εκτίμηση της συνοχής των συστάδων, του βαθμού διαχωρισμού μεταξύ τους και της σταθερότητας της ομαδοποίησης, σε επαναλαμβανόμενες εκτελέσεις των αλγορίθμων.

Για την αξιολόγηση του αλγορίθμου K-Modes υπολογίστηκε αρχικά ο δείκτης Silhouette με χρήση απόστασης Hamming, ώστε η διαδικασία αξιολόγησης να είναι συμβατή με τη φύση των κατηγορικών δεδομένων. Ο συγκεκριμένος δείκτης εκτιμά κατά πόσο κάθε παρατήρηση παρουσιάζει μεγαλύτερη ομοιότητα με τη συστάδα στην οποία ανήκει, σε σχέση με τις υπόλοιπες συστάδες.

Παράλληλα, υπολογίστηκαν και οι δείκτες Davies–Bouldin και Calinski–Harabasz. Για τον υπολογισμό τους πραγματοποιήθηκε μετατροπή των κατηγορικών μεταβλητών σε αριθμητική μορφή μέσω ordinal encoding, αποκλειστικά για σκοπούς αξιολόγησης και όχι για την εφαρμογή του αλγορίθμου K-Modes. Οι δείκτες αυτοί χρησιμοποιήθηκαν για την εκτίμηση της συνοχής στο εσωτερικό των συστάδων και του βαθμού διαχωρισμού μεταξύ διαφορετικών ομάδων.

Για τον αλγόριθμο K-Prototypes, οι δείκτες Silhouette, Davies–Bouldin και Calinski–Harabasz υπολογίστηκαν με βάση τη μικτή αναπαράσταση κατηγορικών και κανονικοποιημένων αριθμητικών χαρακτηριστικών, εφόσον ο αλγόριθμος αξιοποιεί ταυτόχρονα και τους δύο τύπους δεδομένων. Ως εκ τούτου, οι απόλυτες τιμές των μετρικών μεταξύ K-Modes και K-Prototypes δεν θεωρούνται άμεσα συγκρίσιμες, διότι βασίζονται σε διαφορετικές προσεγγίσεις υπολογισμού της απόστασης και της ομοιότητας μεταξύ των παρατηρήσεων.

Επιπλέον, για τον K-Modes υπολογίστηκε και η μέση απόσταση Hamming εντός των συστάδων, η οποία χρησιμοποιήθηκε ως συμπληρωματική ένδειξη του βαθμού ομοιογένειας των ομάδων. Μικρότερες τιμές της μέσης απόστασης υποδηλώνουν μεγαλύτερη ομοιότητα μεταξύ των παρατηρήσεων που ανήκουν στην ίδια συστάδα.

Πέρα από την αξιολόγηση της ποιότητας της ομαδοποίησης, πραγματοποιήθηκε και ανάλυση σταθερότητας των αποτελεσμάτων για τους αλγορίθμους K-Modes και K-Prototypes. Για τον σκοπό αυτό, κάθε αλγόριθμος εκτελέστηκε πολλαπλές φορές με διαφορετικές αρχικοποιήσεις και οι αναθέσεις των παρατηρήσεων στις συστάδες συγκρίθηκαν μεταξύ των εκτελέσεων μέσω του δείκτη Adjusted Rand Index (ARI). Η διαδικασία αυτή επιτρέπει την εκτίμηση της σταθερότητας των συστάδων και της ευαισθησίας των αλγορίθμων σε διαφορετικές αρχικοποιήσεις.

Οι παραπάνω μετρικές υπολογίστηκαν αφενός για το συνολικό σύνολο δεδομένων, αφετέρου και για τα επιμέρους υποσύνολα ασθενών (συμπτωματικούς και ασυμπτωματικούς), προσφέροντας μια ολοκληρωμένη και αντικειμενική εικόνα της ποιότητας της ανάλυσης. Τα ποσοτικά αποτελέσματα των μετρικών αυτών παρουσιάζονται αναλυτικά στο Κεφάλαιο 4, ενώ η περαιτέρω αξιολόγηση και η κλινική τους ερμηνεία αναπτύσσονται στο Κεφάλαιο 5.

3.8 Μέθοδος TE2Rules

Για την ερμηνεία των αποτελεσμάτων της ανάλυσης συστάδων χρησιμοποιήθηκε η μέθοδος TE2Rules, η οποία αποτελεί τεχνική εξηγήσιμης τεχνητής νοημοσύνης (Explainable Artificial Intelligence – XAI) με στόχο την εξαγωγή ερμηνεύσιμων κανόνων από μοντέλα μηχανικής μάθησης [10]. Σκοπός της εφαρμογής της, στην παρούσα μελέτη, ήταν η μετατροπή των αποτελεσμάτων της ομαδοποίησης σε απλούς και κατανοητούς λογικούς κανόνες, ώστε να καταστεί δυνατή η σαφέστερη περιγραφή των μορφολογικών χαρακτηριστικών που διαφοροποιούν κάθε παραγόμενη συστάδα.

Στο πλαίσιο της παρούσας εργασίας, η μέθοδος TE2Rules εφαρμόστηκε στα αποτελέσματα που προέκυψαν από τον αλγόριθμο K-Modes και από τον αλγόριθμο K-Prototypes. Για κάθε ασθενή αξιοποιήθηκαν τα μορφολογικά χαρακτηριστικά που συμμετείχαν στη διαδικασία clustering, σε συνδυασμό με την ετικέτα συστάδας που αποδόθηκε από τον αντίστοιχο αλγόριθμο ομαδοποίησης. Για σκοπούς σύγκρισης και ομοιογένειας, ως μεταβλητές εισόδου στη μέθοδο TE2Rules χρησιμοποιήθηκαν αποκλειστικά οι κατηγορικές μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA, τόσο για τα αποτελέσματα του K-Modes, όσο και για εκείνα του K-Prototypes. Με την προσέγγιση αυτή κατέστη δυνατή η άμεση σύγκριση των κανόνων που προέκυψαν από τις δύο διαφορετικές μεθοδολογίες clustering.

Πριν από την εφαρμογή της μεθόδου, οι κατηγορικές τιμές μετατράπηκαν σε αριθμητική μορφή, μέσω χειροκίνητης αντιστοίχισης κατηγοριών. Συγκεκριμένα, οι διαβαθμισμένες κατηγορίες όπως low, medium και high αντιστοιχίστηκαν σε διαδοχικές ακέραιες αριθμητικές τιμές, ενώ οι δυαδικές κατηγορίες absent και present αντιστοιχίστηκαν στις τιμές 0 και 1 αντίστοιχα. Ως μεταβλητή-στόχος χρησιμοποιήθηκε η ετικέτα της συστάδας (Cluster), η οποία είχε προκύψει από τη διαδικασία clustering.

Για την εξαγωγή των κανόνων εφαρμόστηκε η προσέγγιση «μία έναντι όλων» (one-vs-rest). Πιο συγκεκριμένα, για κάθε συστάδα δημιουργήθηκε ξεχωριστό δυαδικό πρόβλημα ταξινόμησης, στο οποίο η εκάστοτε συστάδα θεωρήθηκε ως θετική κλάση, ενώ το σύνολο των υπόλοιπων παρατηρήσεων αποτέλεσε την αρνητική κλάση. Στη συνέχεια, εκπαιδεύτηκε μοντέλο Random Forest, πάνω στο οποίο εφαρμόστηκε η μέθοδος TE2Rules για την εξαγωγή κανόνων της μορφής «if-then».

Οι κανόνες που προέκυψαν περιγράφουν τους συνδυασμούς μορφολογικών χαρακτηριστικών που σχετίζονται με την ένταξη των ασθενών σε συγκεκριμένες συστάδες. Μέσω των κανόνων αυτών κατέστη δυνατή η βαθύτερη κατανόηση των διαφορών μεταξύ των ομάδων ασθενών και η ανάδειξη μορφολογικών προτύπων που σχετίζονται με διαφορετικά επίπεδα μορφολογικού κινδύνου.

Η αξιοποίηση της μεθόδου TE2Rules συνέβαλε ουσιαστικά στη βελτίωση της ερμηνευσιμότητας της ανάλυσης, επιτρέποντας τη σύνδεση των αποτελεσμάτων του clustering με κλινικά κατανοητούς και εύκολα ερμηνεύσιμους κανόνες. Οι κανόνες που εξήχθησαν για κάθε υποομάδα ασθενών παρουσιάζονται αναλυτικά στο Κεφάλαιο 4, ενώ η κλινική τους αξιολόγηση αναπτύσσεται στο Κεφάλαιο 5.

Κεφάλαιο 4: Αποτελέσματα

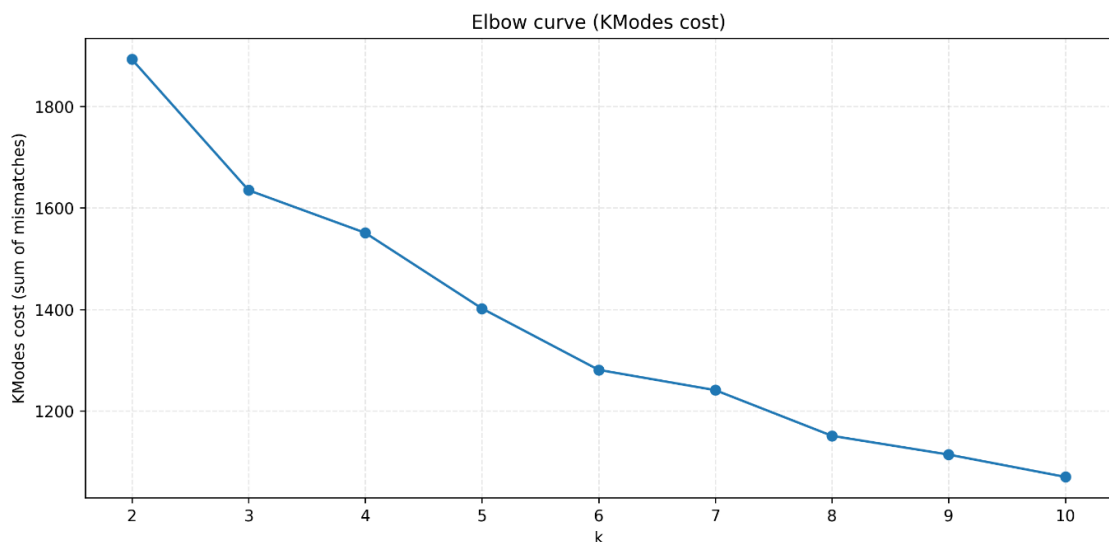
4.1	Σύνολο ασθενών	51
4.1.1	Αποτελέσματα clustering - Αλγόριθμος K-Modes	51
4.1.2	Αποτελέσματα κανόνων με TE2Rules - Αλγόριθμος K-Modes	66
4.1.3	Αποτελέσματα clustering – Αλγόριθμος K-Prototypes	67
4.1.4	Αποτελέσματα κανόνων με TE2Rules – Αλγόριθμος K-Prototypes	75
4.2	Σύνολο συμπτωματικών ασθενών	76
4.2.1	Αποτελέσματα clustering - Αλγόριθμος K-Modes	76
4.2.2	Αποτελέσματα κανόνων με TE2Rules - Αλγόριθμος K-Modes	84
4.2.3	Αποτελέσματα clustering – Αλγόριθμος K-Prototypes	86
4.2.4	Αποτελέσματα κανόνων με TE2Rules – Αλγόριθμος K-Prototypes	95
4.3	Σύνολο ασυμπτωματικών ασθενών	97
4.3.1	Αποτελέσματα clustering - Αλγόριθμος K-Modes	97
4.3.2	Αποτελέσματα κανόνων με TE2Rules - Αλγόριθμος K-Modes	105
4.3.3	Αποτελέσματα clustering – Αλγόριθμος K-Prototypes	106
4.3.4	Αποτελέσματα κανόνων με TE2Rules – Αλγόριθμος K-Prototypes	115

4.1 Συνολικό σύνολο ασθενών

4.1.1 Αποτελέσματα ομαδοποίησης στο σύνολο των ασθενών με τον αλγόριθμο K-Modes

Η ανάλυση συστάδων στο σύνολο των δεδομένων πραγματοποιήθηκε με χρήση του αλγορίθμου K-Modes, με στόχο την ομαδοποίηση ασθενών με παρόμοια μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας.

Για την επιλογή του κατάλληλου αριθμού συστάδων εφαρμόστηκε η μέθοδος elbow. Όπως παρουσιάζεται στο Σχήμα 4.1.1, η μείωση του κόστους είναι πιο έντονη για μικρότερες τιμές του k και γίνεται πιο σταδιακή για μεγαλύτερες τιμές.



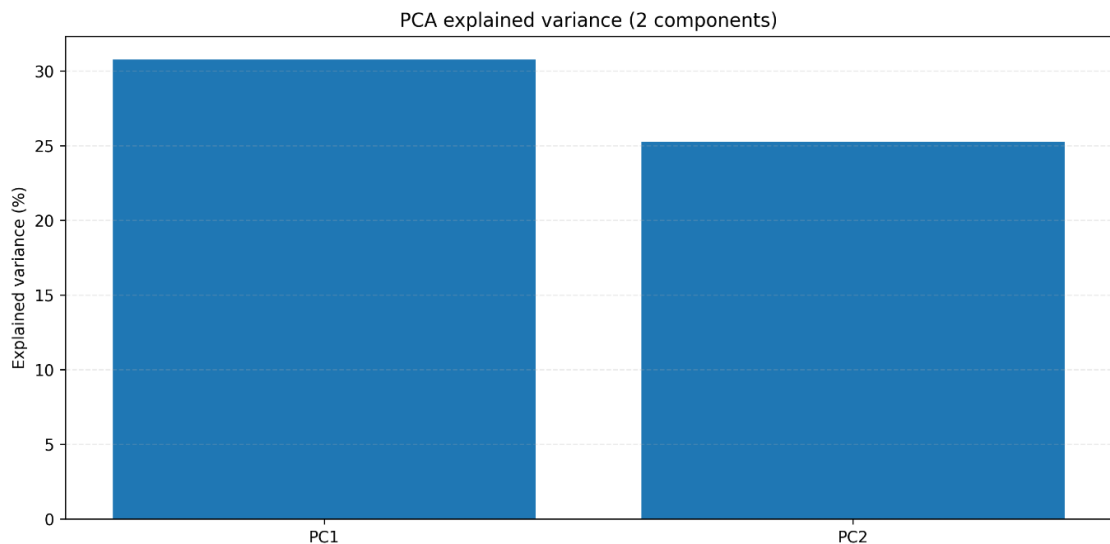
Σχήμα 4.1.1 Καμπύλη elbow για τον αλγόριθμο K-Modes στο σύνολο ασθενών

Η επιλογή του $k = 2$ υποστηρίζεται και από τις μετρικές αξιολόγησης που παρουσιάζονται στο Σχήμα 4.1.2. Οι τιμές των δεικτών Silhouette, Davies–Bouldin και Calinski–Harabasz υποδεικνύουν ικανοποιητική ποιότητα ομαδοποίησης για δύο συστάδες. Ταυτόχρονα, η ανάλυση σταθερότητας έδειξε υψηλές τιμές ARI σε επαναλαμβανόμενες εκτελέσεις.



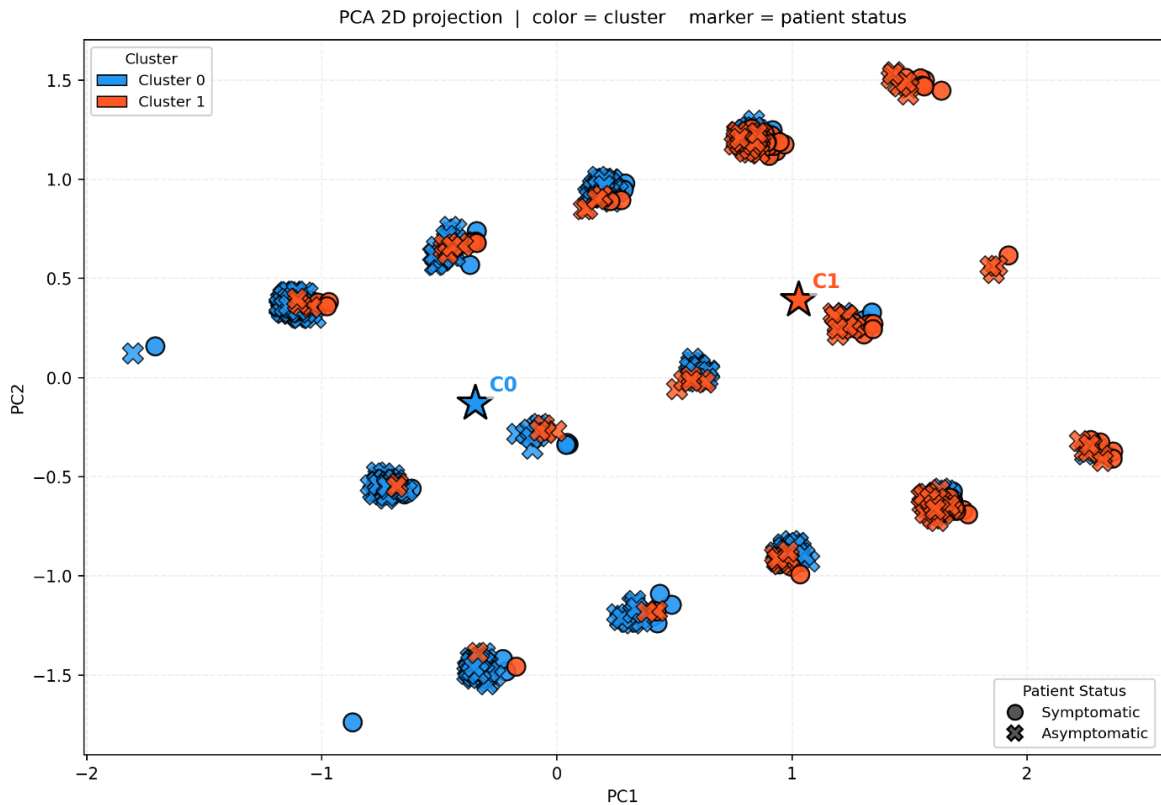
Σχήμα 4.1.2 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο ασθενών K-Modes.

Η οπτική διερεύνηση της δομής των δεδομένων πραγματοποιήθηκε μέσω Ανάλυσης Κύριων Συνιστωσών (PCA). Όπως παρουσιάζεται στο Σχήμα 4.1.3, οι δύο πρώτες κύριες συνιστώσες εξηγούν συνολικά περίπου 56% της συνολικής διακύμανσης των δεδομένων.



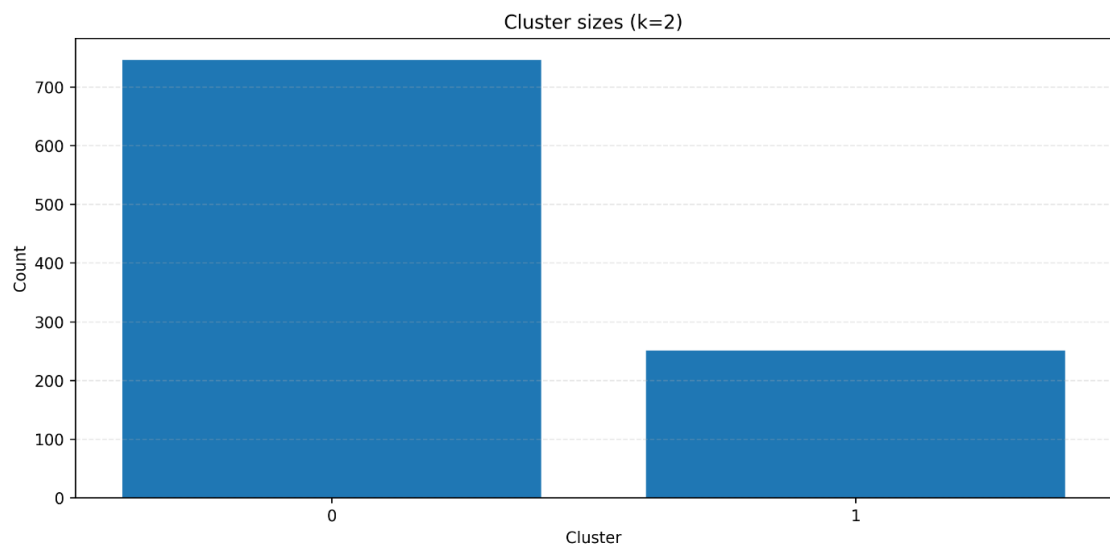
Σχήμα 4.1.3 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο ασθενών K-Modes.

Η προβολή των δεδομένων στον χώρο των δύο πρώτων κύριων συνιστωσών παρουσιάζεται στο Σχήμα 4.1.4. Παρατηρείται μερικός διαχωρισμός μεταξύ των δύο συστάδων, καθώς και επικάλυψη σε ορισμένες περιοχές του χώρου προβολής. Στο ίδιο διάγραμμα αποτυπώνεται και η κατανομή συμπτωματικών και ασυμπτωματικών ασθενών ανά συστάδα.



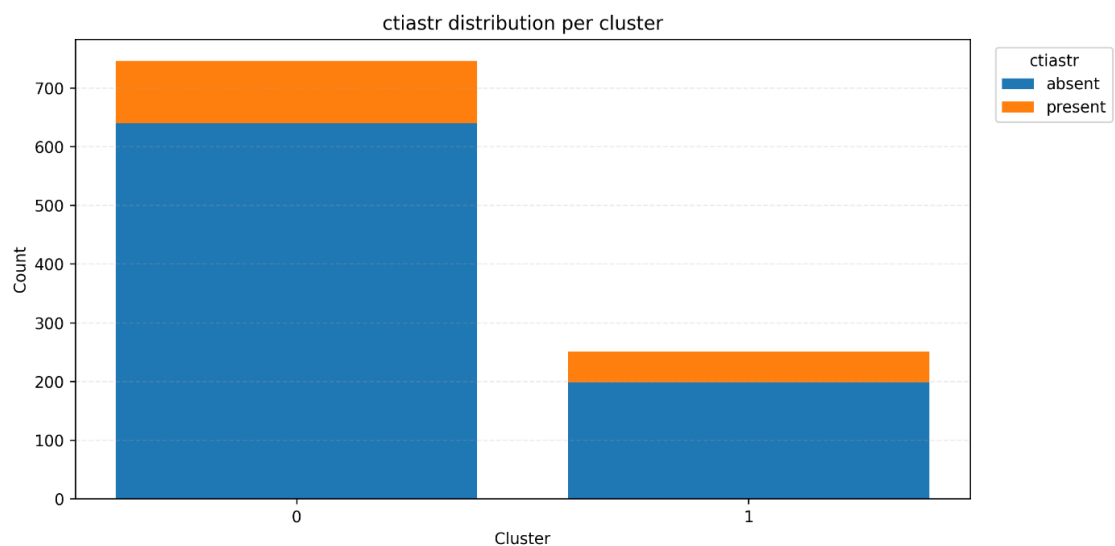
Σχήμα 4.1.4 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο ασθενών K-Modes

Η κατανομή των ασθενών ανά συστάδα παρουσιάζεται στο σχήμα 4.1.5. Από την γραφική παράσταση παρατηρούμε ότι το cluster 0 είναι πολύ μεγαλύτερο σε μέγεθος με 748 ασθενείς ενώ, το cluster 1 έχει μόνο 252 ασθενείς.

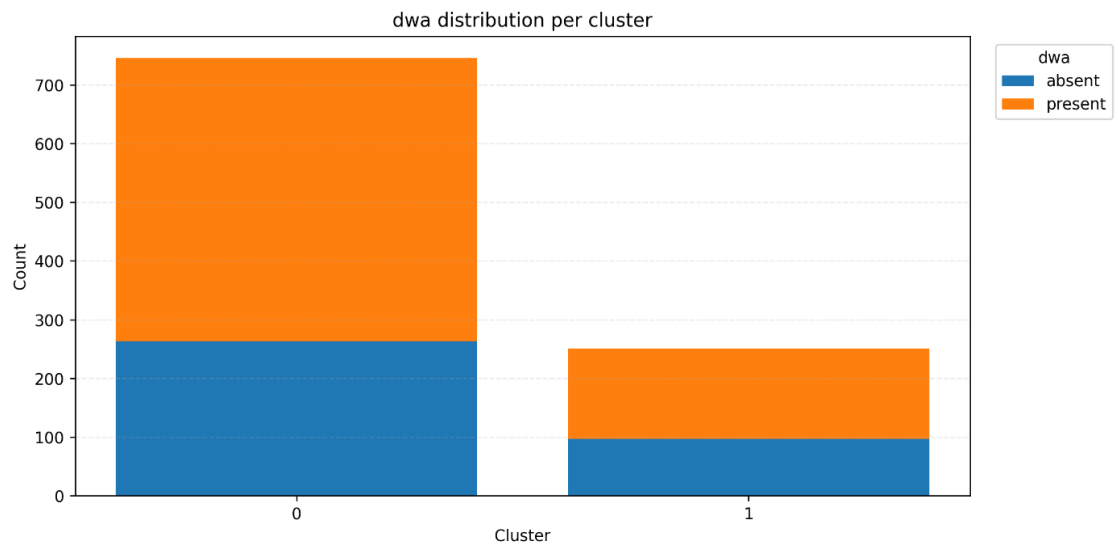


Σχήμα 4.1.5 Κατανομή των ασθενών ανά συστάδα στο σύνολο ασθενών K-Modes

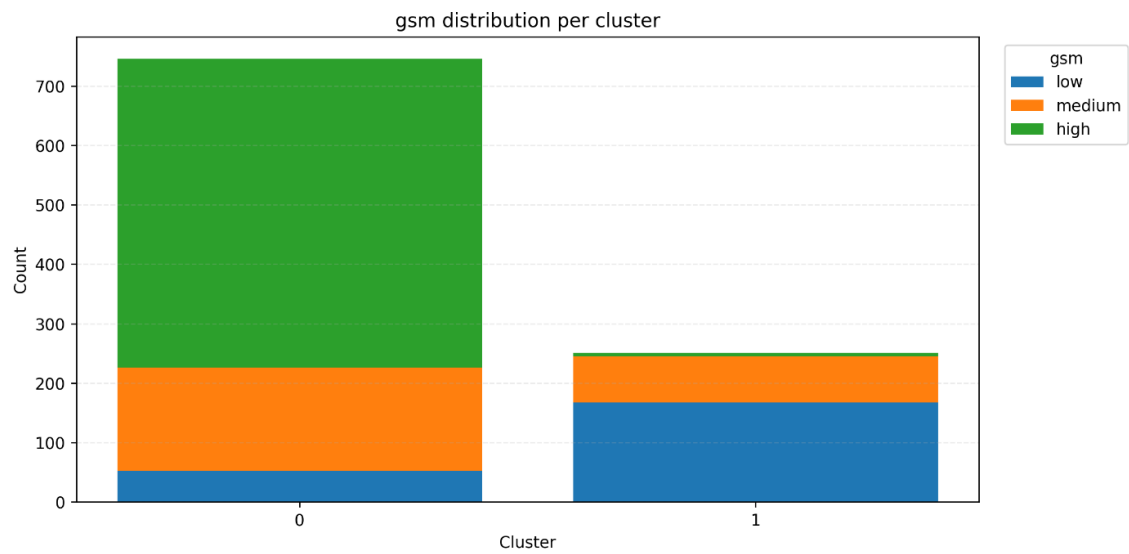
Στα Σχήματα 4.1.6 - 4.1.11, που ακολουθούν πιο κάτω, παρουσιάζονται οι κατανομές των βασικών μορφολογικών χαρακτηριστικών ανά συστάδα. Παρατηρούνται διαφοροποιήσεις μεταξύ των clusters ως προς τις μεταβλητές CTIASTR, DWA, GSM, JBA, PLAREA και STENOSIS.



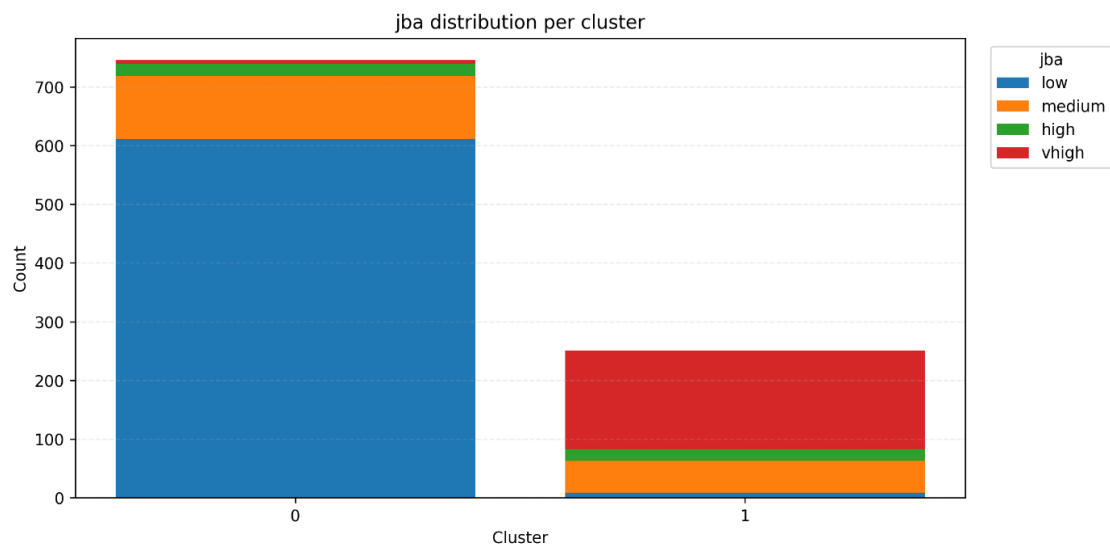
Σχήμα 4.1.6 Κατανομή CTIASTR ανά συστάδα στο σύνολο ασθενών K-Modes



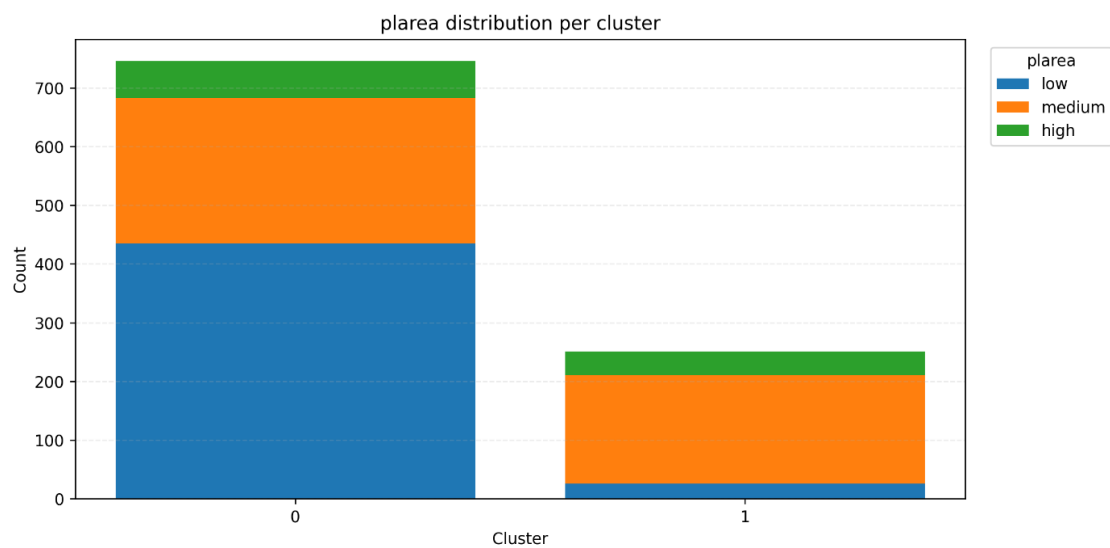
Σχήμα 4.1.7 Κατανομή DWA ανά συστάδα στο σύνολο ασθενών K-Modes



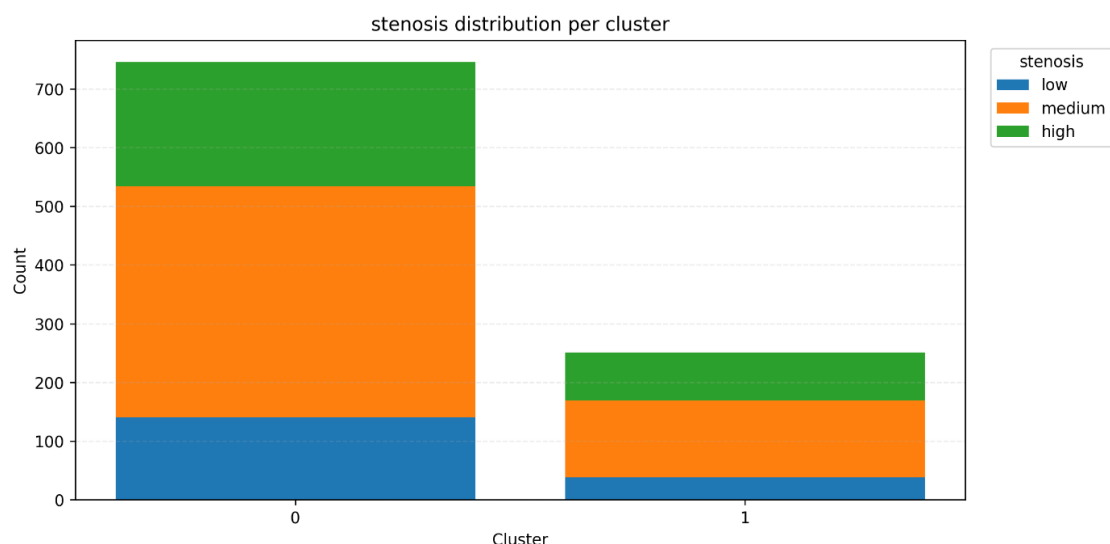
Σχήμα 4.1.8 Κατανομή GSM ανά συστάδα στο σύνολο ασθενών K-Modes



Σχήμα 4.1.9 Κατανομή JBA ανά συστάδα στο σύνολο ασθενών K-Modes



Σχήμα 4.1.10 Κατανομή PLAREA ανά συστάδα στο σύνολο ασθενών K-Modes



Σχήμα 4.1.11 Κατανομή STENOSIS ανά συστάδα στο σύνολο ασθενών K-Modes

Τέλος, η ποσοτική αξιολόγηση της ποιότητας της ομαδοποίησης παρουσιάζεται στους Πίνακες 4.1.1 και 4.1.2. Για το συνολικό σύνολο δεδομένων ($n=996$, $k=2$), ο δείκτης Silhouette με απόσταση Hamming 0.45, έλαβε τιμή 0.30, υποδηλώνοντας μέτριο βαθμό διαχωρισμού μεταξύ των συστάδων. Ο δείκτης Davies–Bouldin έλαβε τιμή 1.968 και ο δείκτης Calinski–Harabasz ανήλθε στο 187.7. Για την αξιολόγηση της σταθερότητας, ο αλγόριθμος εκτελέστηκε 20 φορές με μέσο ARI 0.69, ελάχιστο 0.29 και μέγιστο 1.00.

Τα αριθμητικά αποτελέσματα των μετρικών παρουσιάζονται στους Πίνακες 4.1.1 και 4.1.2.

Group	n	k	Silhouette	Davies - Bouldin	Calinski - Harabasz	Avg. Hamming Distance
Full dataset	996	2	0.30	1.97	187.7	0.45

Πίνακας 4.1.1 Εσωτερικές μετρικές ομαδοποίησης K -modes στο σύνολο ασθενών

Group	Runs	ARI (Mean)	ARI (Min)	ARI (Max)
Full dataset	20	0.69	0.29	1.00

Πίνακας 4.1.2 Σταθερότητα της ομαδοποίησης K-modes στο σύνολο ασθενών

4.1.2 Αποτελέσματα κανόνων TE2Rules στο σύνολο των ασθενών με τον αλγόριθμο K-Modes

Η μέθοδος TE2Rules εφαρμόστηκε στις συστάδες που προέκυψαν από την ανάλυση K-Modes στο σύνολο των ασθενών, με $k=2$. Στόχος ήταν η εξαγωγή ερμηνεύσιμων κανόνων που περιγράφουν τα βασικά χαρακτηριστικά κάθε συστάδας.

Ως μεταβλητή-στόχος χρησιμοποιήθηκε η μεταβλητή Cluster. Ως μεταβλητές εισόδου χρησιμοποιήθηκαν οι κατηγορικές μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA. Για κάθε συστάδα εφαρμόστηκε προσέγγιση one-vs-rest, ώστε να εντοπιστούν οι κανόνες που διαφοροποιούν τη συγκεκριμένη συστάδα από την υπόλοιπη ομάδα ασθενών.

Οι κανόνες που προέκυψαν παρουσιάζονται στον Πίνακα 4.1.3.

Cluster	Κανόνες TE2Rules
Cluster 0	JBA = low
	JBA $\leq$ high AND PLAREA = low
	GSM = high
	JBA $\leq$ high AND PLAREA = high
	CTIASTR = absent AND PLAREA = low
Cluster 1	JBA = very high
	GSM $\leq$ medium AND JBA = medium ή high AND PLAREA = medium
	GSM = low AND PLAREA $\geq$ medium

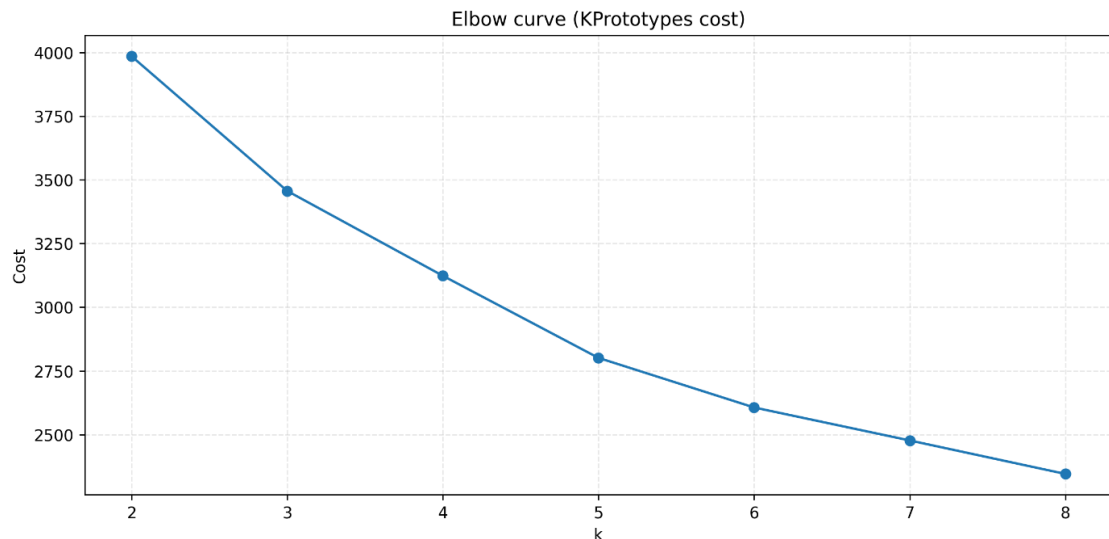
Πίνακας 4.1.3 Αντιπροσωπευτικοί κανόνες TE2Rules στο σύνολο ασθενών K-Modes

Από τους κανόνες που προέκυψαν παρατηρούμε ότι, το Cluster 0 περιγράφεται κυρίως από χαμηλότερες τιμές JBA, υψηλότερες τιμές GSM και χαμηλότερες τιμές PLAREA. Αντίθετα, το Cluster 1 περιγράφεται κυρίως από πολύ υψηλή τιμή JBA, καθώς και από συνδυασμούς χαμηλού ή μέτριου GSM με μέτρια ή υψηλότερη PLAREA.

4.1.3 Αποτελέσματα ομαδοποίησης στο σύνολο ασθενών με τον αλγόριθμο K-Prototypes

Η ανάλυση συστάδων στο σύνολο των ασθενών πραγματοποιήθηκε με χρήση του αλγορίθμου K-Prototypes, με στόχο την ομαδοποίηση ασθενών αξιοποιώντας τόσο αριθμητικά, όσο και κατηγορικά χαρακτηριστικά της αθηρωματικής πλάκας.

Για την επιλογή του κατάλληλου αριθμού συστάδων εφαρμόστηκε η μέθοδος elbow, με αξιολόγηση του κόστους clustering για διαφορετικές τιμές του k . Όπως παρουσιάζεται στο Σχήμα 4.1.12, παρατηρείται σημαντική μείωση του κόστους από $k=2$ έως $k=4$, ενώ για μεγαλύτερες τιμές του k η μείωση γίνεται περισσότερο σταδιακή. Το σημείο καμπής εντοπίζεται κοντά στο $k=2$, γεγονός που υποστηρίζει την επιλογή δύο clusters για την περιγραφή της δομής των δεδομένων.



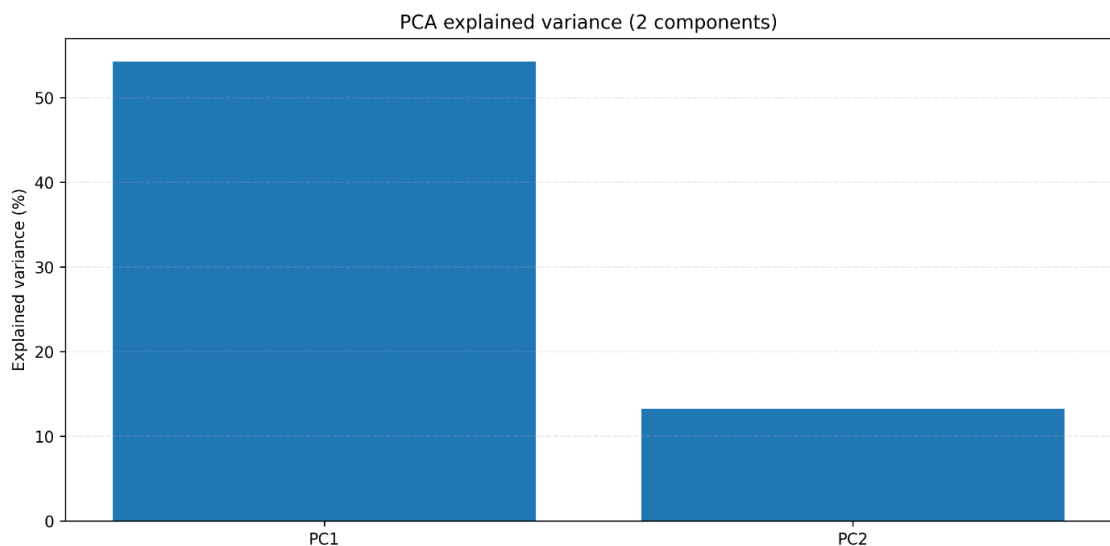
Σχήμα 4.1.12 Καμπύλη elbow για τον αλγόριθμο K-Prototypes στο σύνολο ασθενών

Παράλληλα, αξιολογήθηκαν οι δείκτες ποιότητας clustering Silhouette, Davies–Bouldin και Calinski–Harabasz για διαφορετικές τιμές του k . Όπως φαίνεται στο Σχήμα 4.1.13, η τιμή του δείκτη Silhouette παρουσιάζει την υψηλότερη τιμή για $k=2$, ενώ ο δείκτης Davies–Bouldin εμφανίζει τη χαμηλότερη τιμή στην ίδια περιοχή. Επιπλέον, ο δείκτης Calinski–Harabasz λαμβάνει επίσης τη μέγιστη τιμή του για $k=2$, υποδεικνύοντας καλύτερο διαχωρισμό μεταξύ των clusters.



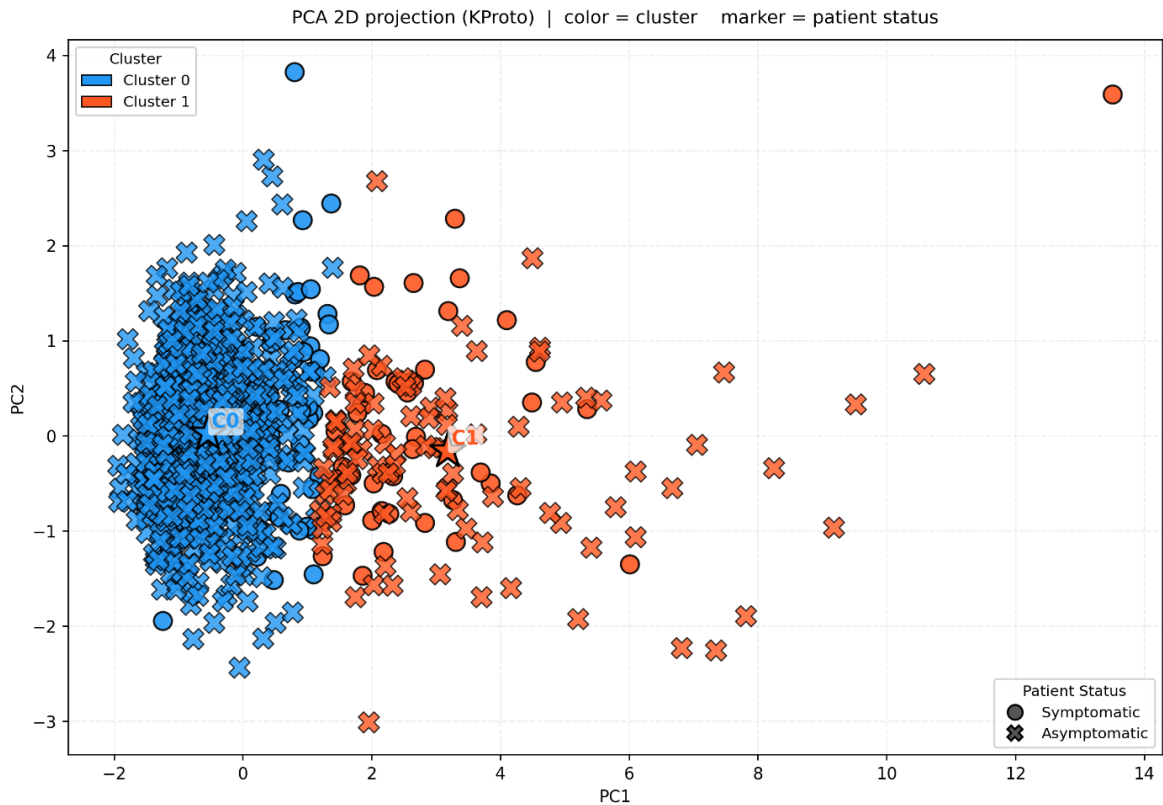
Σχήμα 4.1.13 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο ασθενών K-Prototypes.

Το Σχήμα 4.1.14 παρουσιάζει το ποσοστό διακύμανσης που εξηγείται από τις δύο πρώτες κύριες συνιστώσες της PCA. Η πρώτη κύρια συνιστώσα (PC1) εξηγεί το μεγαλύτερο μέρος της συνολικής διακύμανσης των δεδομένων, ενώ η δεύτερη συνιστώσα (PC2) συνεισφέρει σημαντικά λιγότερο.



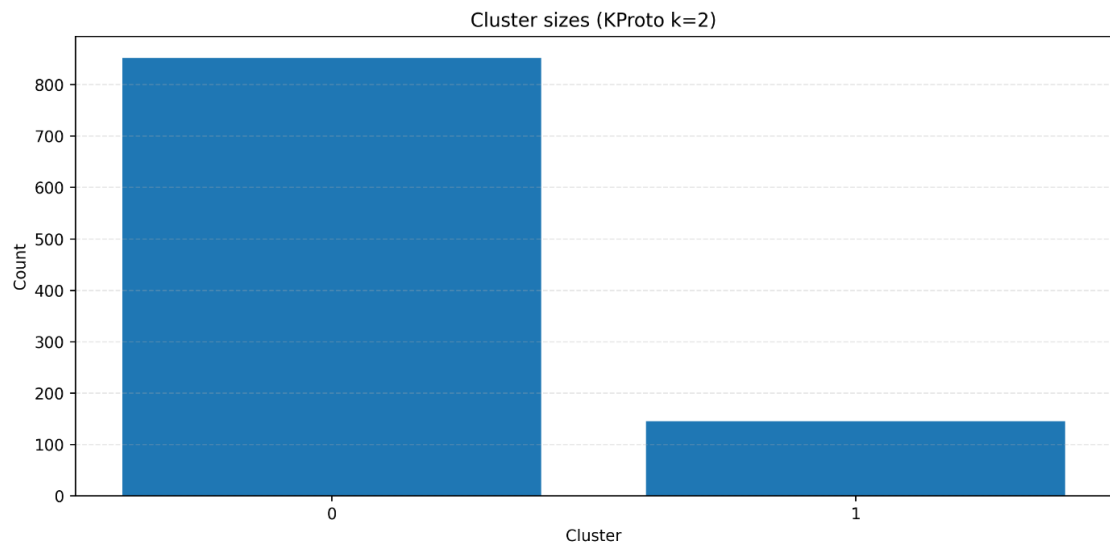
Σχήμα 4.1.14 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο ασθενών K-Prototypes.

Η προβολή των ασθενών στις δύο πρώτες κύριες συνιστώσες παρουσιάζεται στο Σχήμα 4.1.15. Παρατηρείται ικανοποιητικός διαχωρισμός των δύο clusters κυρίως κατά μήκος της πρώτης κύριας συνιστώσας, με το Cluster 0 να συγκεντρώνει τη μεγαλύτερη πλειονότητα των ασθενών και το Cluster 1 να αντιστοιχεί σε μικρότερη αλλά περισσότερο διακριτή ομάδα ασθενών.



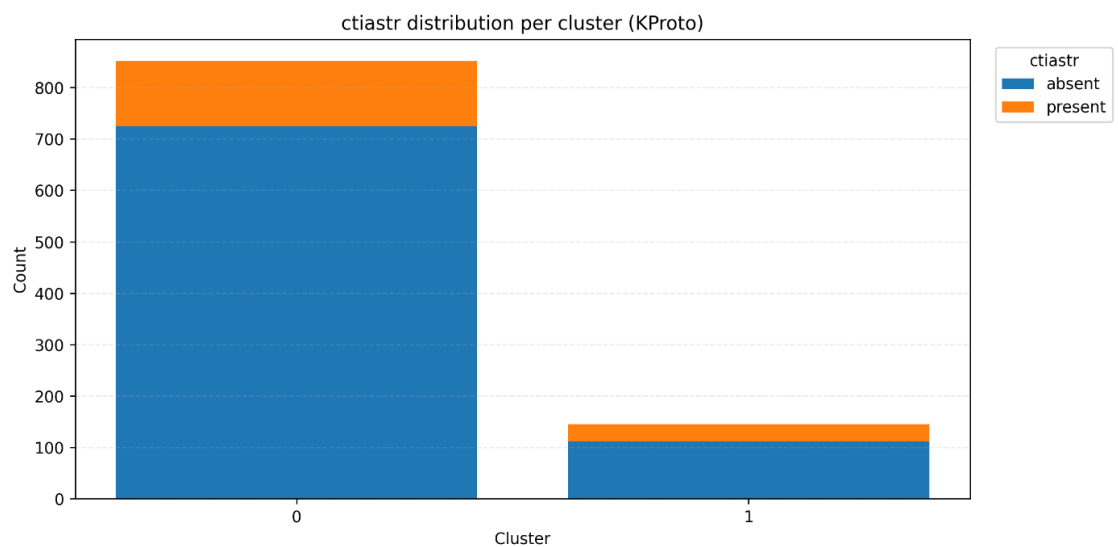
Σχήμα 4.1.15 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο ασθενών K-Prototypes

Η κατανομή των ασθενών ανά cluster παρουσιάζεται στο Σχήμα 4.1.16. Από την γραφική παράσταση παρατηρείται ότι, στην εκτέλεση του αλγορίθμου K-Prototypes, το Cluster 0 υπερτερεί αριθμητικά από το Cluster 1 το οποίο περιλαμβάνει σημαντικά μικρότερο αριθμό ασθενών.

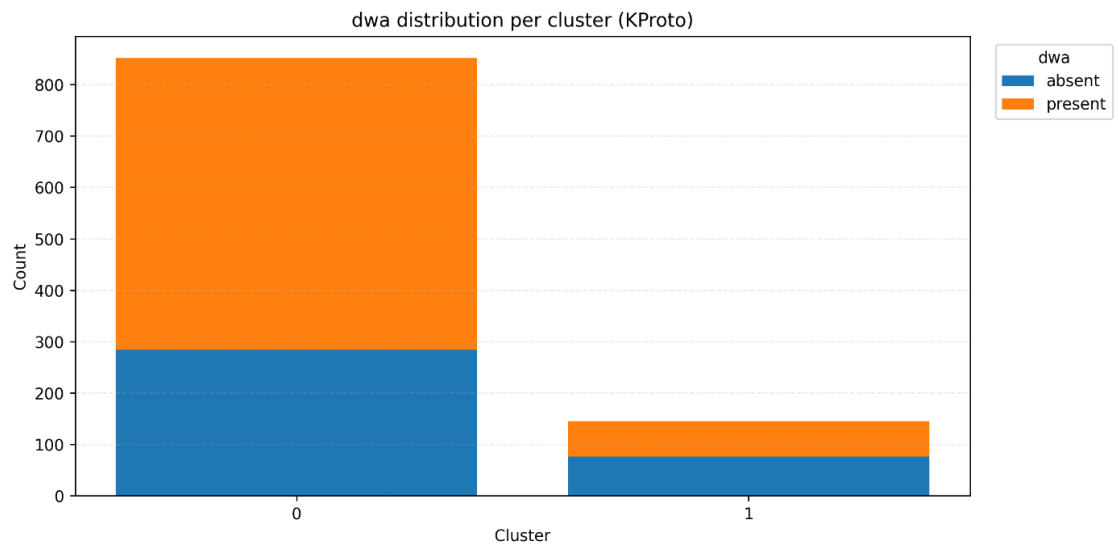


Σχήμα 4.1.16 Κατανομή των ασθενών ανά συστάδα στο σύνολο ασθενών K-Prototypes

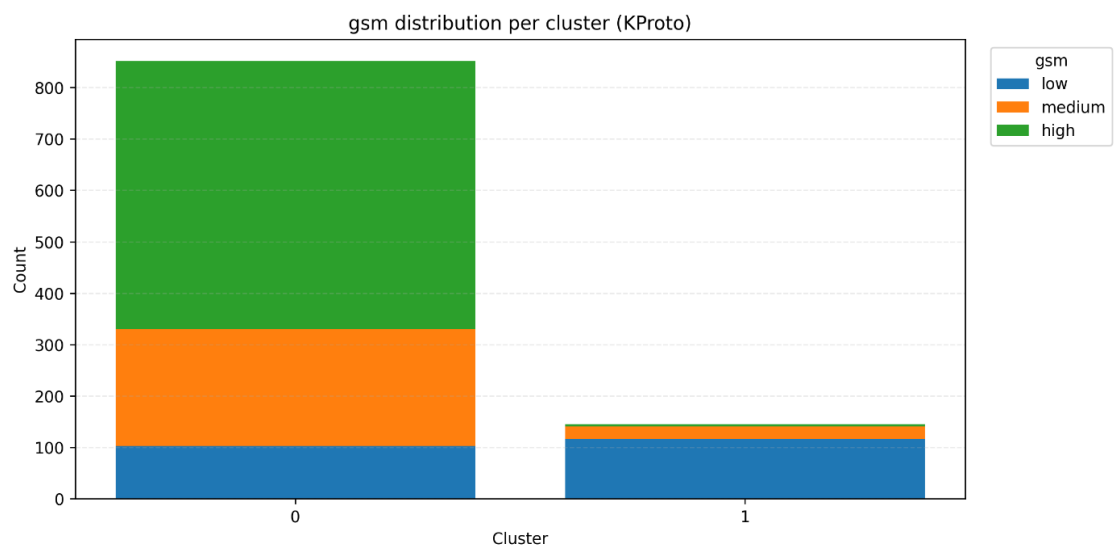
Η ανάλυση της κατανομής των χαρακτηριστικών ανά cluster παρουσιάζεται στα Σχήματα 4.1.17 - 4.1.22. Τα διαγράμματα απεικονίζουν την κατανομή των μεταβλητών CTIASTR, DWA, GSM, JBA, PLAREA και STENOSIS για κάθε cluster που προέκυψε από τον αλγόριθμο K-Prototypes.



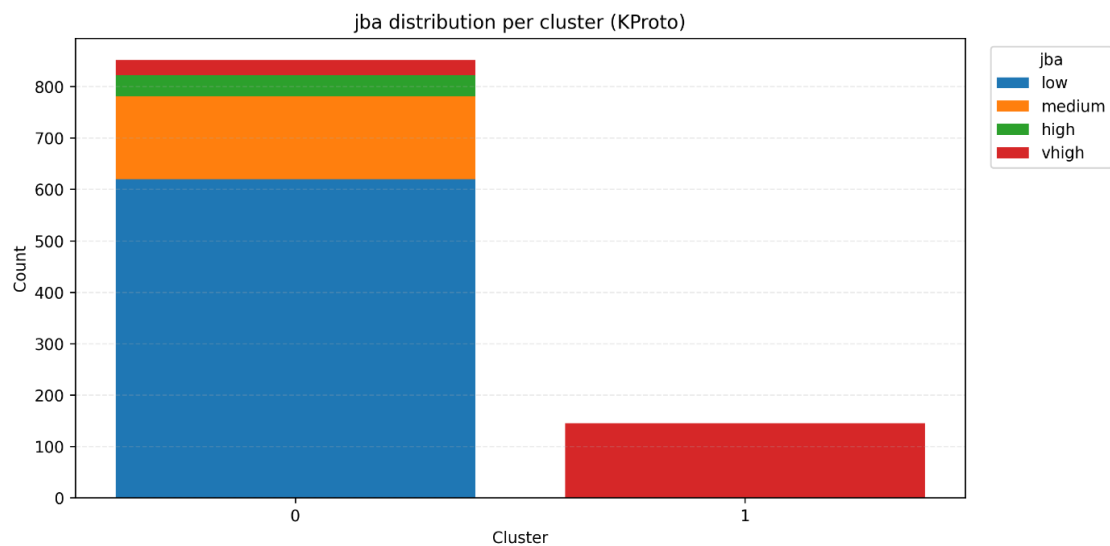
Σχήμα 4.1.17 Κατανομή CTIASTR ανά συστάδα στο σύνολο των ασθενών K-Prototypes.



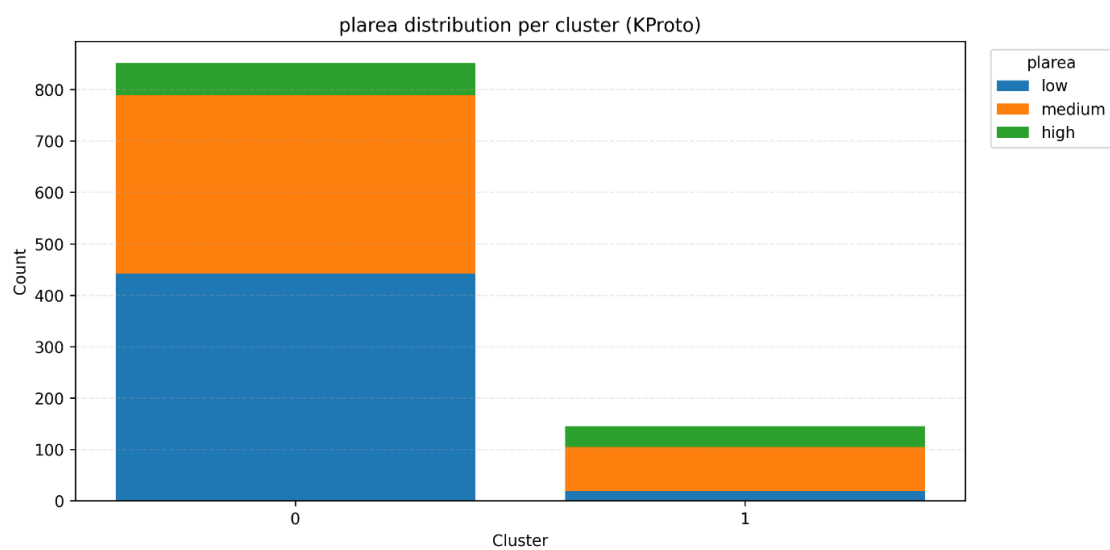
Σχήμα 4.1.18 Κατανομή DWA ανά συστάδα στο σύνολο των ασθενών K-Prototypes.



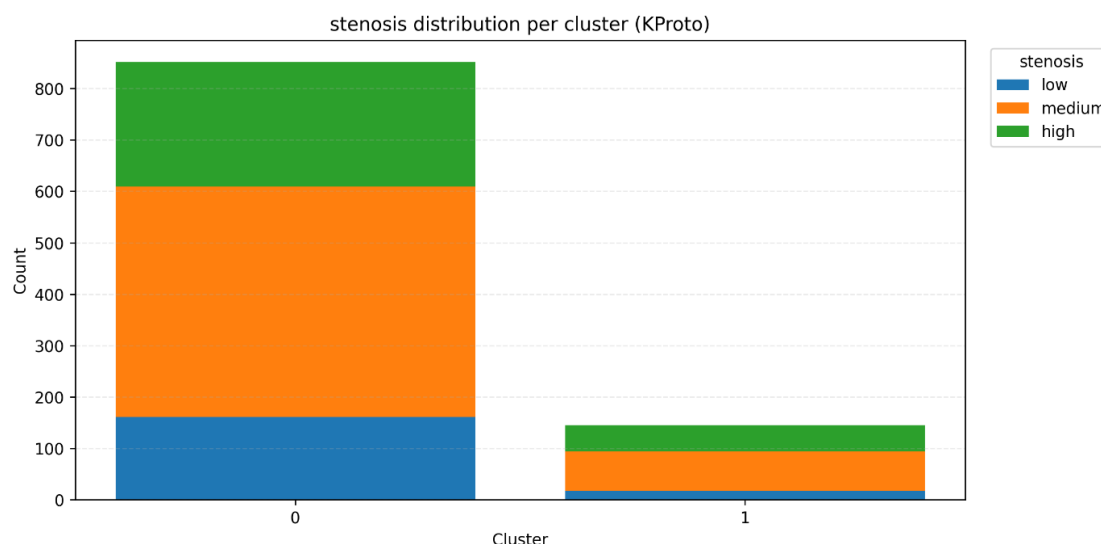
Σχήμα 4.1.19 Κατανομή GSM ανά συστάδα στο σύνολο των ασθενών K-Prototypes.



Σχήμα 4.1.20 Κατανομή JBA ανά συστάδα στο σύνολο των ασθενών K-Prototypes.



Σχήμα 4.1.21 Κατανομή PLAREA ανά συστάδα στο σύνολο των ασθενών K-Prototypes.



Σχήμα 4.1.22 Κατανομή STENOSIS ανά συστάδα στο σύνολο των ασθενών K-Prototypes.

Η αξιολόγηση της ποιότητας του clustering πραγματοποιήθηκε με χρήση των δεικτών Silhouette, Davies–Bouldin και Calinski–Harabasz. Για το συνολικό σύνολο ασθενών και για $k=2$, ο δείκτης Silhouette έλαβε τιμή 0.46, ο δείκτης Davies–Bouldin τιμή 1.00 και ο δείκτης Calinski–Harabasz τιμή 535.87. Οι τιμές αυτές υποδεικνύουν ικανοποιητικό διαχωρισμό μεταξύ των clusters και σχετικά καλή συνοχή εντός των συστάδων.

Παράλληλα, αξιολογήθηκε η σταθερότητα του μοντέλου μέσω πολλαπλών επαναλήψεων της διαδικασίας clustering ($n=20$), με χρήση του δείκτη Adjusted Rand Index (ARI). Η μέση τιμή του Pairwise ARI ήταν 0.999, με ελάχιστη τιμή 0.995 και μέγιστη τιμή 1.000, γεγονός που δείχνει πολύ υψηλή σταθερότητα και επανάληψη των αποτελεσμάτων του αλγορίθμου K-Prototypes στο συγκεκριμένο σύνολο δεδομένων, σε αντίθεση με αυτή που παρατηρήθηκε στον αλγόριθμο K-Modes.

Τα αριθμητικά αποτελέσματα των μετρικών παρουσιάζονται στους Πίνακες 4.1.4 και 4.1.5.

Group	n	k	Silhouette	Davies -Bouldin	Calinski - Harabasz
Full dataset	996	2	0.46	1.00	535.87

Πίνακας 4.1.4 Εσωτερικές μετρικές ομαδοποίησης K -Prototypes στο σύνολο ασθενών

Group	Runs	ARI (Mean)	ARI (Min)	ARI (Max)
Full dataset	20	0.99	0.99	1.00

Πίνακας 4.1.5 Σταθερότητα της ομαδοποίησης K-Prototypes στο σύνολο ασθενών

4.1.4 Αποτελέσματα κανόνων TE2Rules στο σύνολο ασθενών με τον αλγόριθμο K-Prototypes

Η μέθοδος TE2Rules εφαρμόστηκε στις συστάδες που προέκυψαν από την ανάλυση K-Prototypes στο σύνολο των ασθενών, με $k=2$. Στόχος ήταν η εξαγωγή ερμηνεύσιμων κανόνων που περιγράφουν τα βασικά χαρακτηριστικά κάθε συστάδας.

Όπως και στον αλγόριθμο K-Modes, έτσι και εδώ χρησιμοποιήθηκε ως μεταβλητή-στόχος η μεταβλητή Cluster. Ως μεταβλητές εισόδου χρησιμοποιήθηκαν οι κατηγορικές μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA. Για κάθε συστάδα εφαρμόστηκε προσέγγιση one-vs-rest, ώστε να εντοπιστούν οι κανόνες που διαφοροποιούν τη συγκεκριμένη συστάδα από την υπόλοιπη ομάδα ασθενών.

Οι κανόνες που προέκυψαν παρουσιάζονται στον Πίνακα 4.1.6.

Cluster	Κανόνες TE2Rules
Cluster 0	$JBA \leq \text{high}$
	$CTIASTR = \text{absent AND DWA} = \text{present AND PLAREA} \leq \text{medium}$
	$PLAREA = \text{low}$
	$GSM \geq \text{medium}$
Cluster 1	$DWA = \text{absent AND JBA} = \text{very high}$
	$GSM \leq \text{medium AND JBA} = \text{very high AND PLAREA} = \text{high}$
	$GSM = \text{low AND JBA} = \text{very high}$
	$JBA = \text{very high AND PLAREA} = \text{high}$

Πίνακας 4.1.6 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο ασθενών K-Prototypes

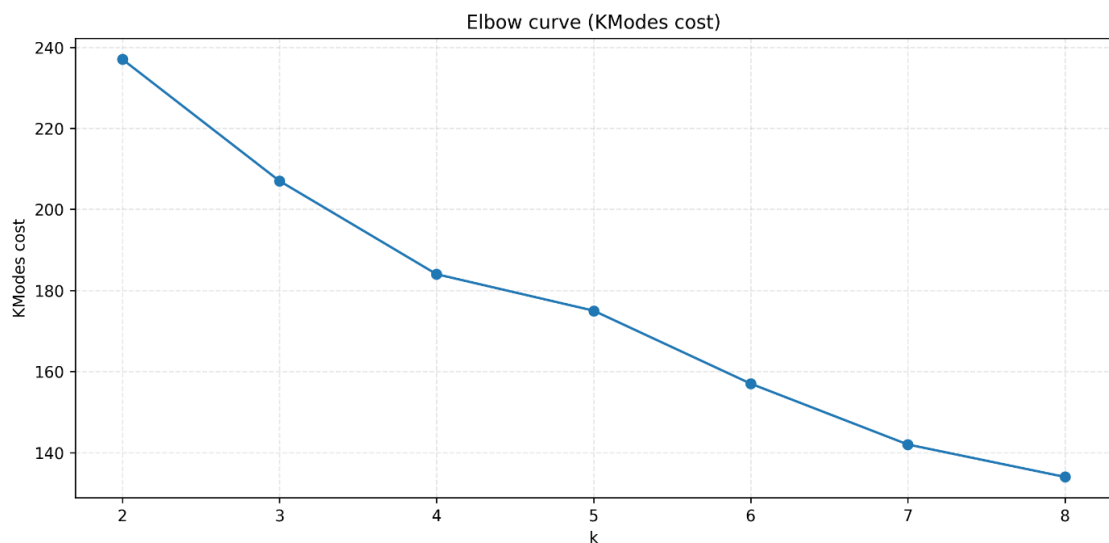
Οι κανόνες δείχνουν ότι το Cluster 0 περιγράφεται κυρίως από χαμηλότερες τιμές JBA, χαμηλότερες τιμές PLAREA και υψηλότερες τιμές GSM. Αντίθετα, το Cluster 1 περιγράφεται κυρίως από πολύ υψηλές τιμές JBA, καθώς και από συνδυασμούς χαμηλού GSM με υψηλότερες τιμές PLAREA και STENOSIS.

4.2 Σύνολο συμπτωματικών ασθενών

4.2.1 Αποτελέσματα ομαδοποίησης στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K-Modes

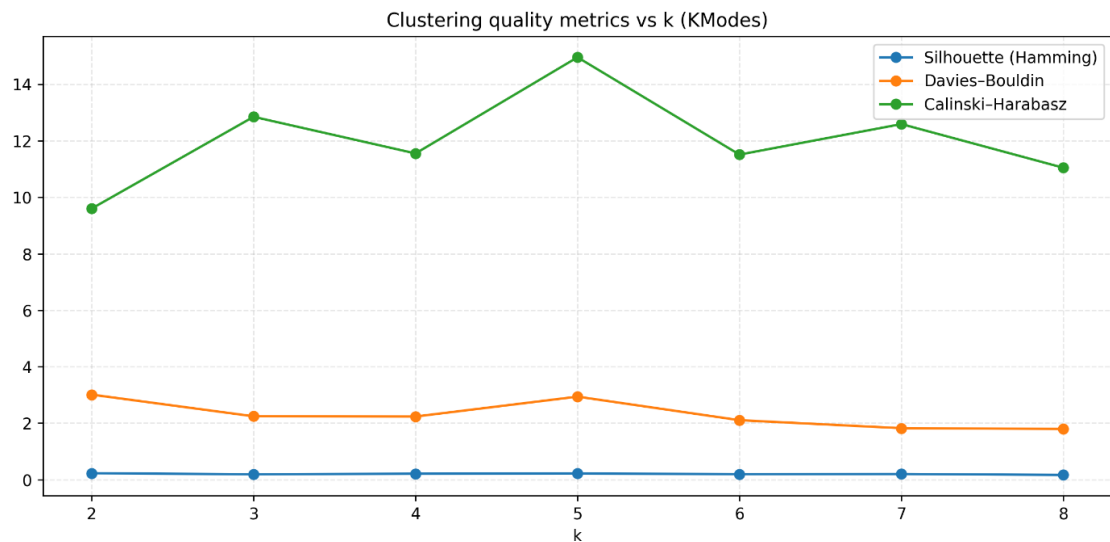
Η ανάλυση συστάδων εφαρμόστηκε στο σύνολο των συμπτωματικών ασθενών με χρήση του αλγορίθμου K-Modes, με στόχο την αναγνώριση υποομάδων ασθενών με παρόμοια μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας.

Για την επιλογή του κατάλληλου αριθμού συστάδων εφαρμόστηκε η μέθοδος elbow, όπως και στις προηγούμενες εκτελέσεις στο σύνολο των ασθενών. Όπως φαίνεται στο Σχήμα 4.2.1, παρατηρείται σημαντική μείωση του κόστους από $k=2$ έως $k=4$, ενώ για μεγαλύτερες τιμές του k η μείωση γίνεται πιο σταδιακή. Με βάση το σημείο καμπής της καμπύλης, όπως και τη συνολική συμπεριφορά των δεικτών αξιολόγησης, τελικά ο αριθμός συστάδων που επιλέχθηκε είναι $k=2$.



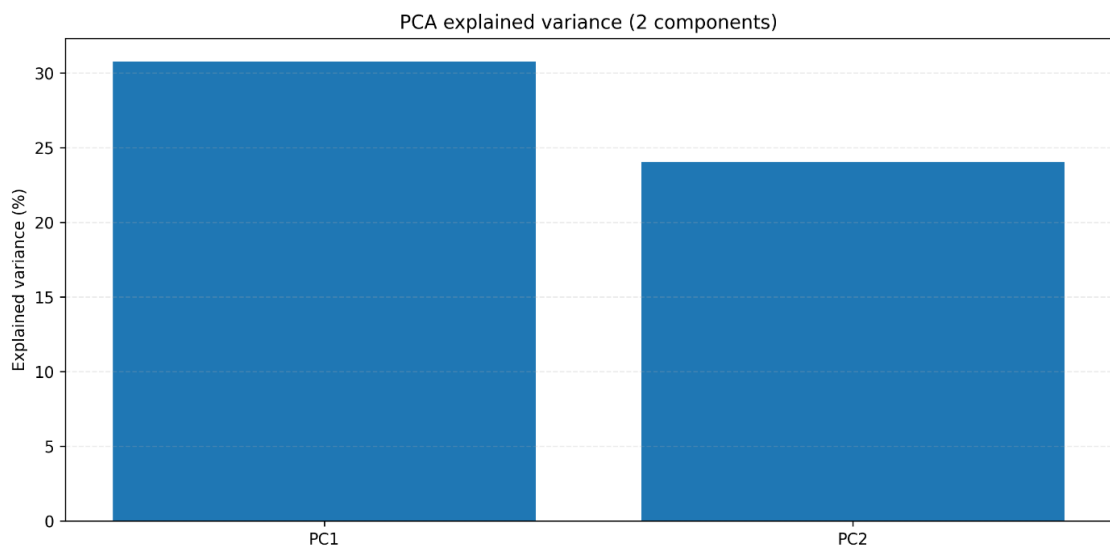
Σχήμα 4.2.1 Καμπύλη elbow για τον αλγόριθμο K-Modes στο σύνολο των συμπτωματικών ασθενών

Η αξιολόγηση της ποιότητας των συστάδων πραγματοποιήθηκε με χρήση των δεικτών Silhouette, Davies–Bouldin και Calinski–Harabasz. Όπως παρουσιάζεται στο Σχήμα 4.2.2, οι τιμές των δεικτών παραμένουν σχετικά σταθερές για διαφορετικές τιμές του k , χωρίς να παρατηρείται ουσιαστική βελτίωση για μεγαλύτερο αριθμό clusters. Η επιλογή του $k=2$ κρίθηκε επαρκής, καθώς παρέχει ικανοποιητικό διαχωρισμό των δεδομένων με απλούστερη και πιο ερμηνεύσιμη δομή.



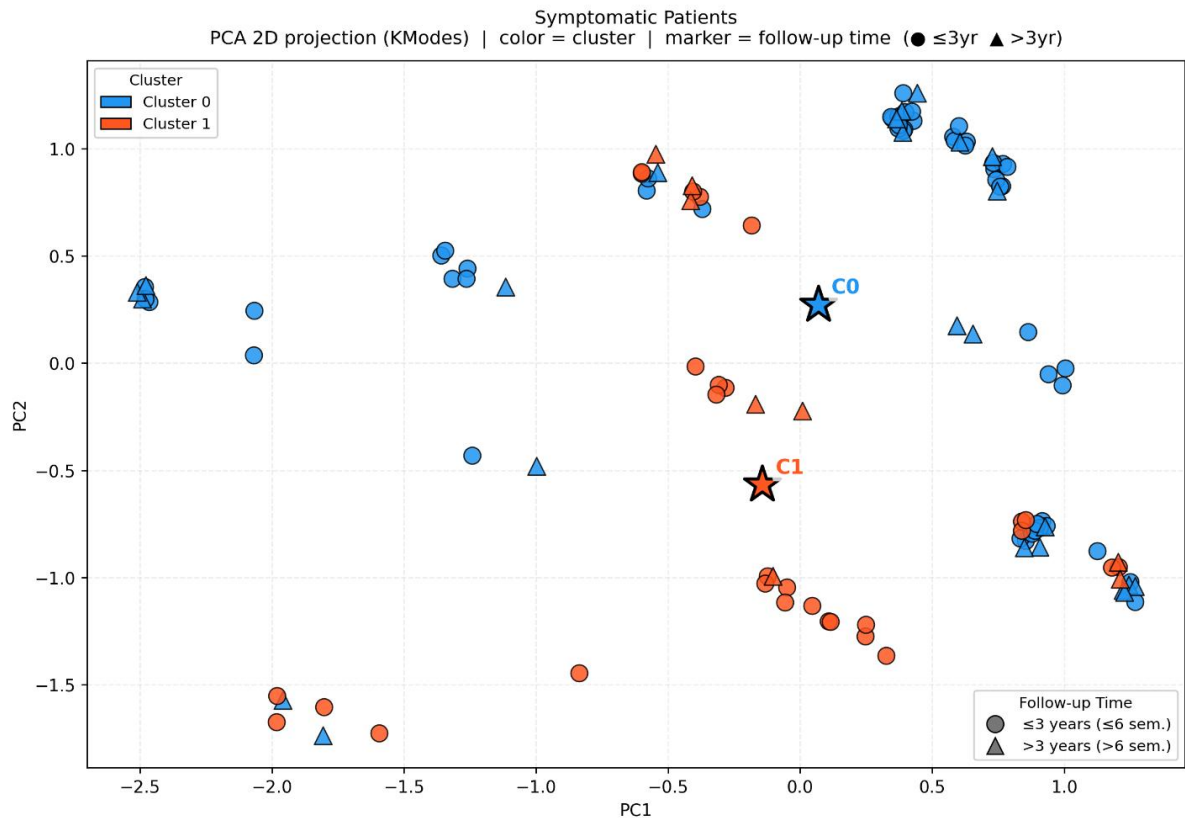
Σχήμα 4.2.2 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των συμπτωματικών ασθενών K-Modes.

Για την οπτικοποίηση των δεδομένων εφαρμόστηκε Ανάλυση Κύριων Συνιστωσών (PCA) σε δύο διαστάσεις. Οι δύο πρώτες κύριες συνιστώσες εξηγούν συνολικά περίπου το 55% της συνολικής διακύμανσης των δεδομένων, με την πρώτη συνιστώσα να εξηγεί περίπου 31% και τη δεύτερη περίπου 24% της διακύμανσης.



Σχήμα 4.2.3 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των συμπτωματικών ασθενών K-Modes.

Η προβολή των ασθενών στον δισδιάστατο χώρο PCA παρουσιάζεται στο Σχήμα 4.2.4. Παρατηρείται μερικός διαχωρισμός μεταξύ των δύο clusters, με ορισμένες περιοχές επικάλυψης, γεγονός αναμενόμενο λόγω της πολυπλοκότητας και της ετερογένειας των κλινικών δεδομένων. Παράλληλα, δεν διακρίνεται σαφής διαφοροποίηση των clusters ως προς τον χρόνο εμφάνισης συμβάντος (≤ 3 έτη και >3 έτη).



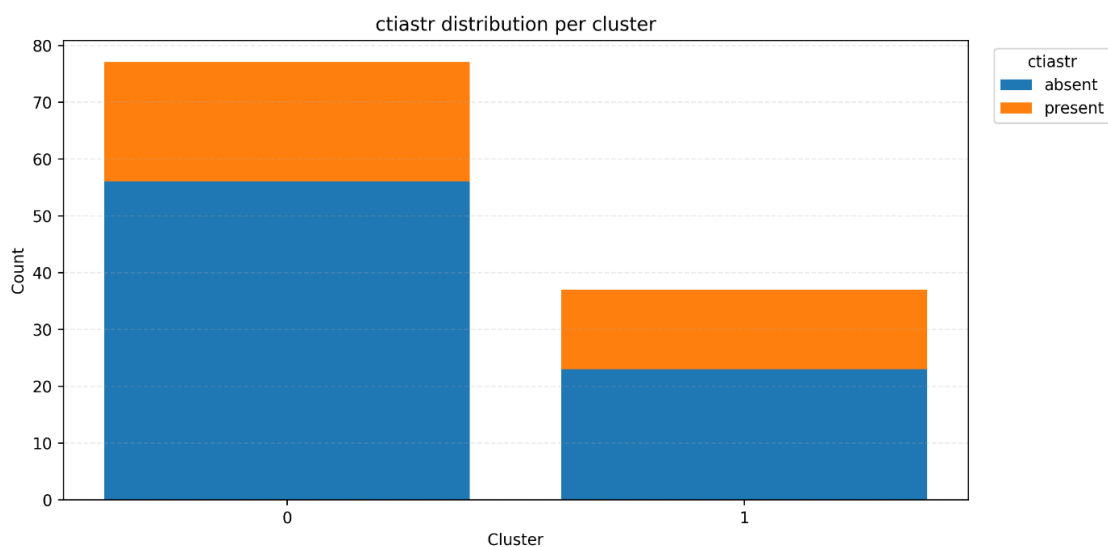
Σχήμα 4.2.4 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των συμπτωματικών ασθενών K-Modes.

Η κατανομή των ασθενών ανά cluster παρουσιάζεται στο Σχήμα 4.2.5. Όπως φαίνεται, το Cluster 0 αποτελεί τη μεγαλύτερη ομάδα με 77 ασθενείς, ενώ το Cluster 1 περιλαμβάνει 37 ασθενείς.

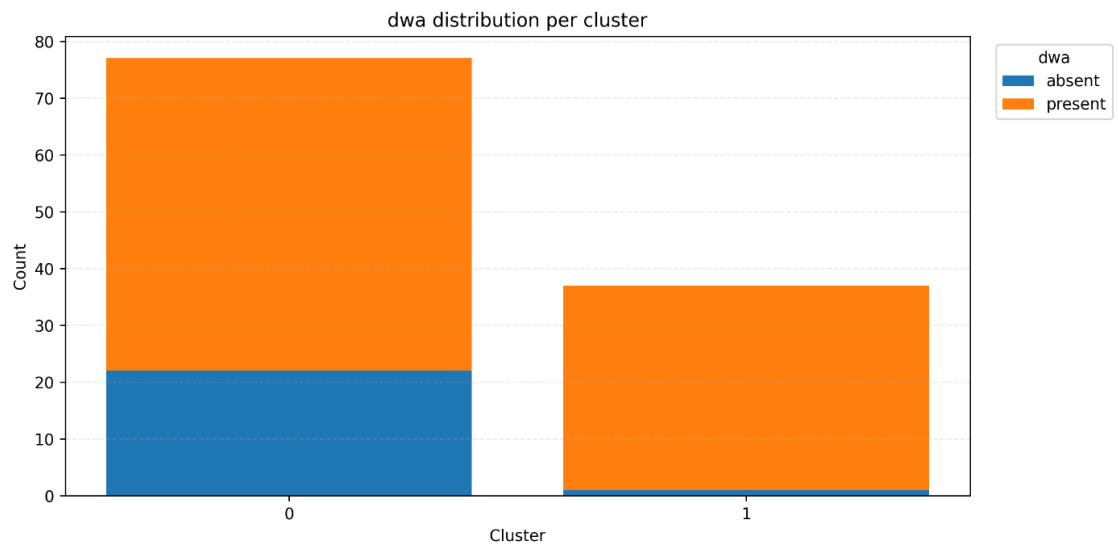


Σχήμα 4.2.5 Κατανομή των ασθενών ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.

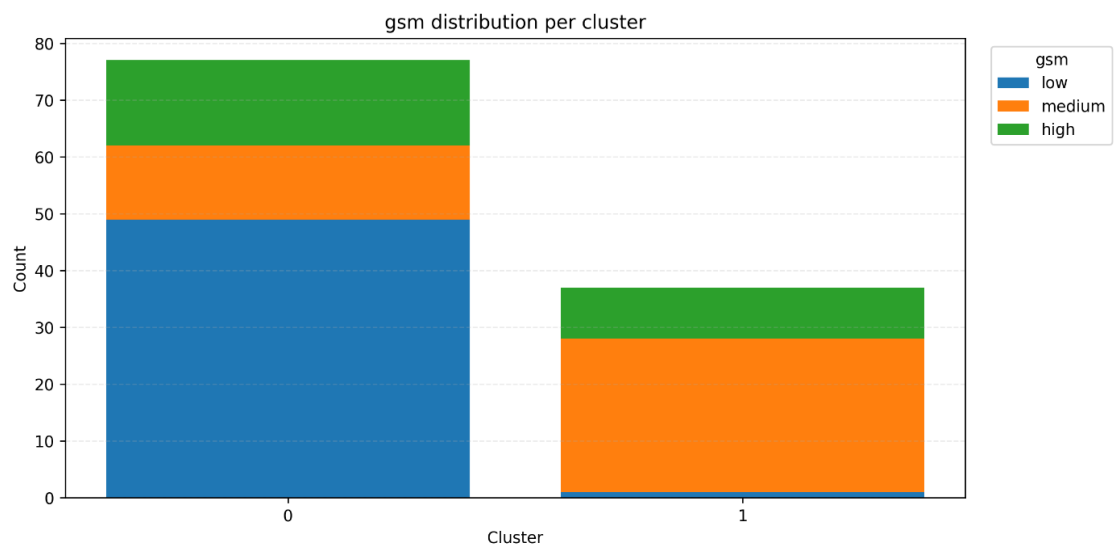
Η ανάλυση της κατανομής των χαρακτηριστικών ανά cluster παρουσιάζεται στα Σχήματα 4.2.6 - 4.2.11. Τα διαγράμματα απεικονίζουν την κατανομή των μεταβλητών CTIASTR, DWA, GSM, JBA, PLAREA και STENOSIS για κάθε cluster που προέκυψε από τον αλγόριθμο K-Modes.



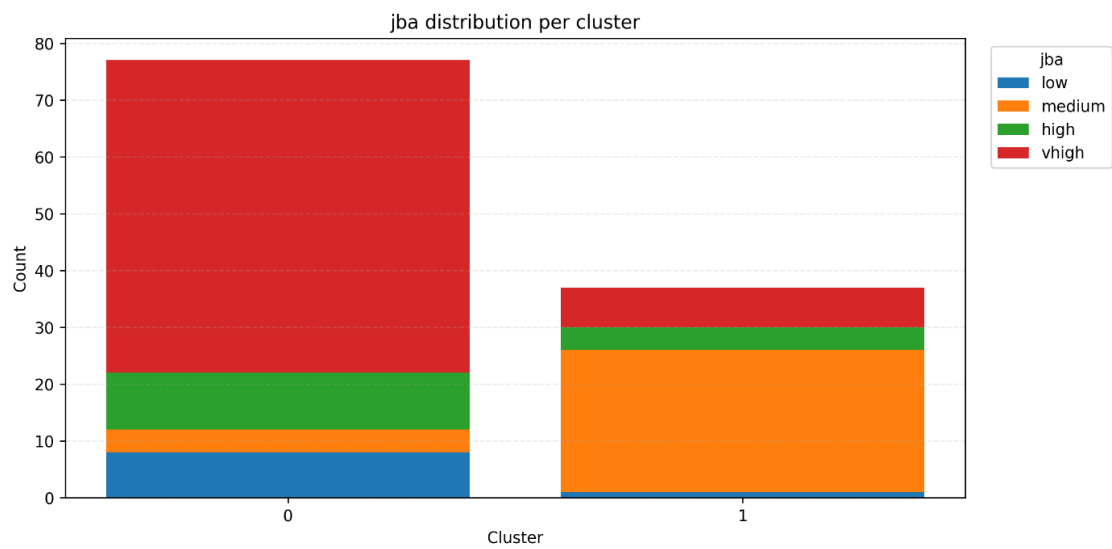
Σχήμα 4.2.6 Κατανομή CTIASTR ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.



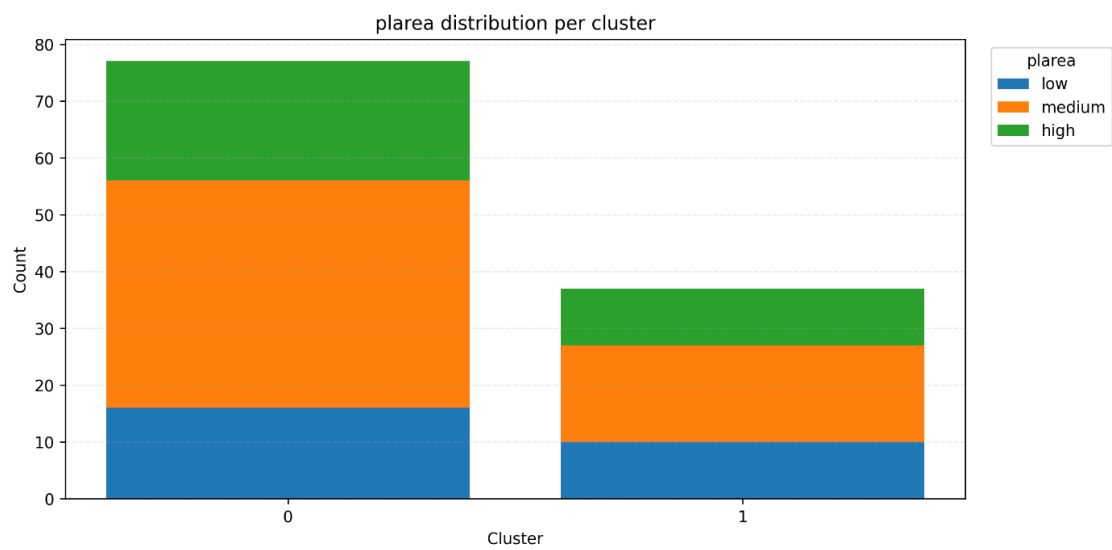
Σχήμα 4.2.7 Κατανομή DWA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.



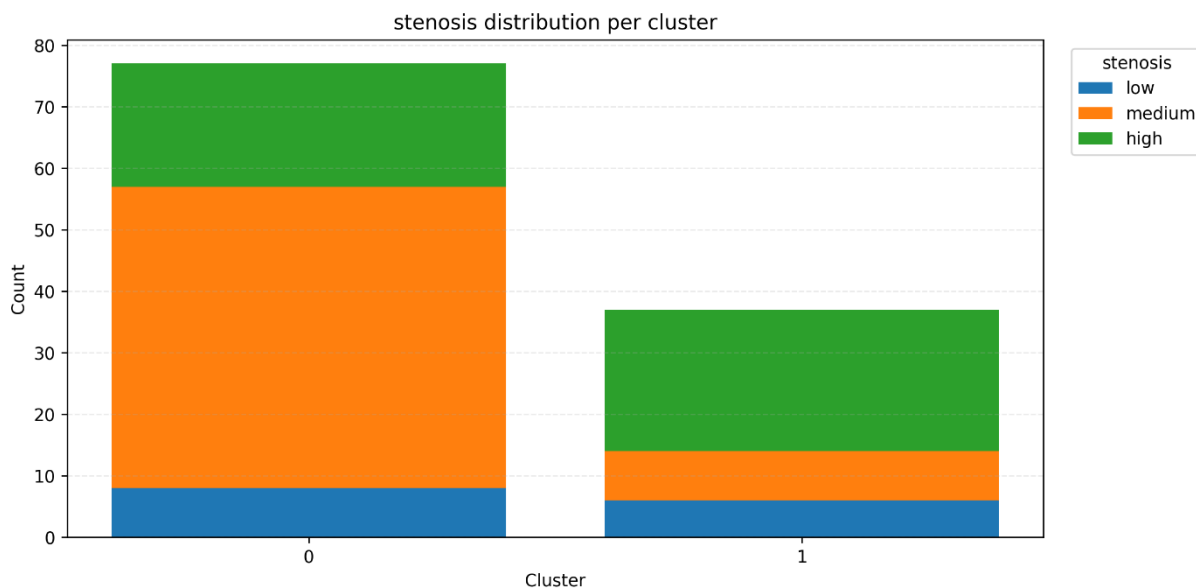
Σχήμα 4.2.8 Κατανομή GSM ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes



Σχήμα 4.2.9 Κατανομή JBA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.



Σχήμα 4.2.10 Κατανομή PLAREA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.



Σχήμα 4.2.11 Κατανομή STENOSIS ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Modes.

Η τελική αξιολόγηση του μοντέλου έδωσε τιμή Silhouette 0.21 με μέση απόσταση Hamming 0.47, η μετρική του Davies–Bouldin 3.29 και η τιμή Calinski–Harabasz 8.28. Οι μετρικές αυτές μας υποδεικνύουν ότι η ανάλυση με τον αλγόριθμο K-Modes έδωσε μέτριο βαθμό συνοχής μεταξύ των ασθενών κάθε cluster.

Επιπλέον, η ανάλυση σταθερότητας του αλγορίθμου μέσω πολλαπλών επαναλήψεων έδωσε μέση τιμή Adjusted Rand Index (ARI) 0.798, με ελάχιστη τιμή 0.372 και μέγιστη τιμή 1.000, γεγονός που δείχνει ικανοποιητική σταθερότητα της ομαδοποίησης.

Τα αριθμητικά αποτελέσματα των μετρικών παρουσιάζονται στους Πίνακες 4.2.1 και 4.2.2.

Group	n	k	Silhouette	Davies - Bouldin	Calinski - Harabasz	Avg within Hamming Distance
Symptomatic dataset	114	2	0.21	3.29	8.28	0.47

Πίνακας 4.2.1 Εσωτερικές μετρικές ομαδοποίησης K -Modes στο σύνολο των συμπτωματικών ασθενών

Group	Runs	ARI (Mean)	ARI (Min)	ARI (Max)
Symptomatic dataset	20	0.79	0.37	1.00

Πίνακας 4.2.2 Σταθερότητα της ομαδοποίησης K-Modes στο σύνολο των συμπτωματικών ασθενών

4.2.2 Αποτελέσματα κανόνων TE2Rules στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K-Modes

Η μέθοδος TE2Rules εφαρμόστηκε στις συστάδες που προέκυψαν από την ανάλυση K-Modes στο σύνολο των συμπτωματικών ασθενών, με $k=2$, με στόχο την εξαγωγή ερμηνεύσιμων κανόνων που περιγράφουν τα βασικά χαρακτηριστικά κάθε συστάδας.

Η μεταβλητή-στόχος που χρησιμοποιήθηκε και σε αυτό το μοντέλο, ήταν η μεταβλητή Cluster και ως μεταβλητές εισόδου οι κατηγορικές μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA. Για κάθε συστάδα ακολουθήθηκε η ίδια προσέγγιση one-vs-rest, όπως και στα υπόλοιπα μοντέλα, ώστε να εντοπιστούν οι κανόνες που διαφοροποιούν τη συγκεκριμένη συστάδα από την υπόλοιπη ομάδα ασθενών.

Οι κανόνες που προέκυψαν παρουσιάζονται στον Πίνακα 4.2.3.

Cluster	Κανόνες TE2Rules
Cluster 0	GSM = low
	JBA $\geq$ high AND STENOSIS $\leq$ medium
	JBA = low AND STENOSIS $\leq$ medium
	GSM = high AND JBA = very high
	GSM = high AND JBA = medium ή high AND STENOSIS = medium
Cluster 1	GSM = medium ή high AND JBA = medium ή high
	GSM = medium AND STENOSIS = high
	JBA $\leq$ high AND STENOSIS = high
	GSM $\geq$ medium AND JBA $\leq$ very high AND STENOSIS = high

Πίνακας 4.2.3 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των συμπτωματικών ασθενών K-Modes.

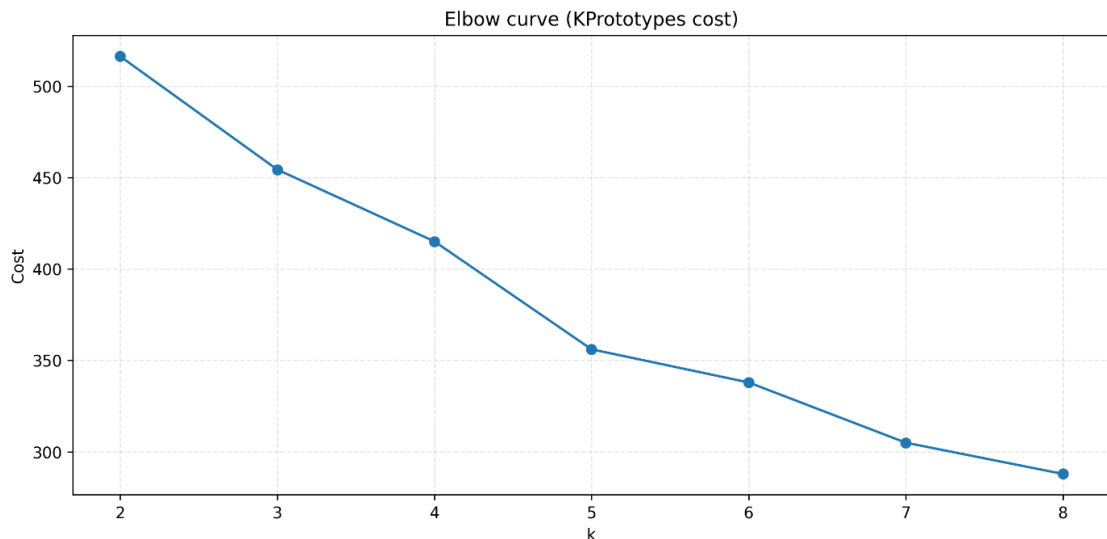
Οι κανόνες δείχνουν ότι το Cluster 0 περιγράφεται κυρίως από χαμηλές τιμές GSM, καθώς και από συνδυασμούς πολύ υψηλών τιμών JBA με μέτριες τιμές STENOSIS. Παράλληλα, εμφανίζονται και κανόνες που συνδέουν υψηλές τιμές GSM με υψηλές ή πολύ υψηλές τιμές JBA.

Αντίθετα, το Cluster 1 χαρακτηρίζεται κυρίως από υψηλότερες τιμές STENOSIS σε συνδυασμό με μέτριες ή υψηλότερες τιμές GSM και JBA. Οι κανόνες της συγκεκριμένης συστάδας δείχνουν ότι η σοβαρότητα της στένωσης αποτελεί βασικό στοιχείο διαφοροποίησης του cluster.

4.2.3 Αποτελέσματα ομαδοποίησης στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K – Prototypes

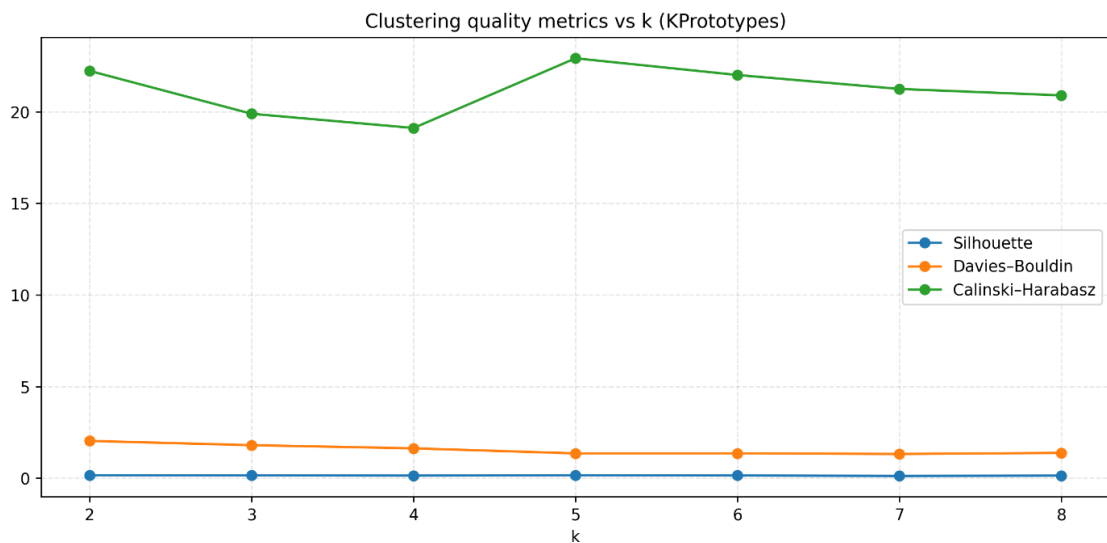
Η ανάλυση συστάδων στο σύνολο των συμπτωματικών ασθενών πραγματοποιήθηκε και με χρήση του αλγορίθμου K-Prototypes, με στόχο την ομαδοποίηση ασθενών αξιοποιώντας τόσο αριθμητικά, όσο και κατηγορικά χαρακτηριστικά της αθηρωματικής πλάκας.

Για την επιλογή του κατάλληλου αριθμού συστάδων εφαρμόστηκε η μέθοδος elbow, με αξιολόγηση του κόστους clustering για διαφορετικές τιμές του k. Όπως παρουσιάζεται στο Σχήμα 4.2.12, παρατηρείται σημαντική μείωση του κόστους από k=2 έως k=5, ενώ για μεγαλύτερες τιμές του k η μείωση γίνεται περισσότερο σταδιακή. Το σημείο καμπής εντοπίζεται κοντά στο k=2, γεγονός που υποστηρίζει την επιλογή των δύο clusters.



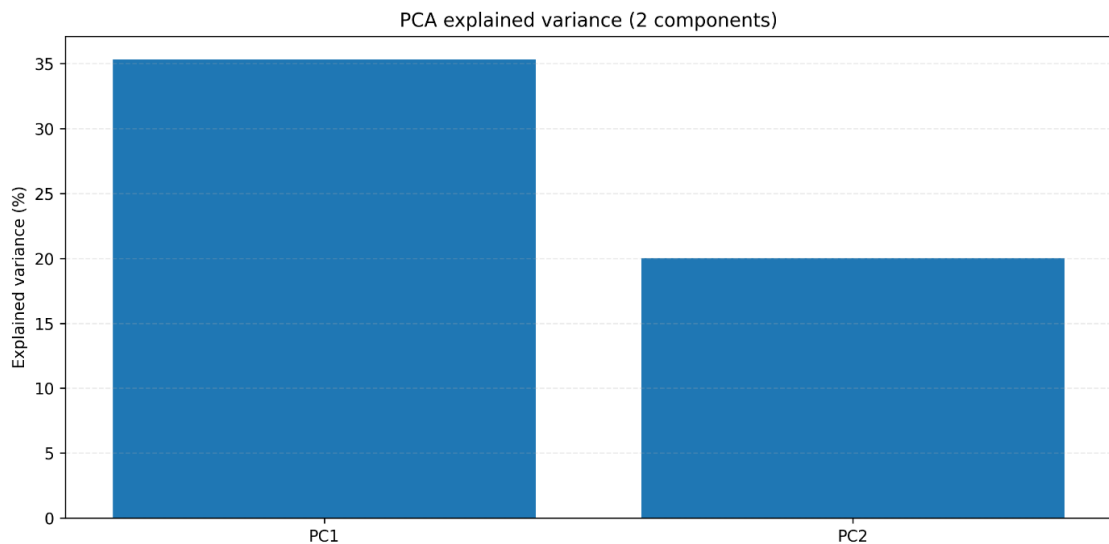
Σχήμα 4.2.12 Καμπύλη elbow για τον αλγόριθμο K-Prototypes στο σύνολο των συμπτωματικών ασθενών

Παράλληλα, αξιολογήθηκαν οι δείκτες ποιότητας clustering Silhouette, Davies–Bouldin και Calinski–Harabasz για διαφορετικές τιμές του k . Όπως φαίνεται στο Σχήμα 4.2.13, ο δείκτης Silhouette παρουσιάζει σχετικά σταθερές τιμές για όλες τις τιμές του k , ενώ ο δείκτης Davies–Bouldin εμφανίζει τη χαμηλότερη τιμή κοντά στο $k=5$. Επιπλέον, ο δείκτης Calinski–Harabasz λαμβάνει τη μέγιστη τιμή του επίσης κοντά στο $k=5$. Ωστόσο, λόγω του ότι το σύνολο των συμπτωματικών ασθενών είναι πολύ μικρό ($n=114$) και για λόγους απλούστερης ερμηνείας και καλύτερης σύγκρισης με τα αποτελέσματα του K-Modes, επιλέχθηκε τελικά η λύση με δύο clusters.



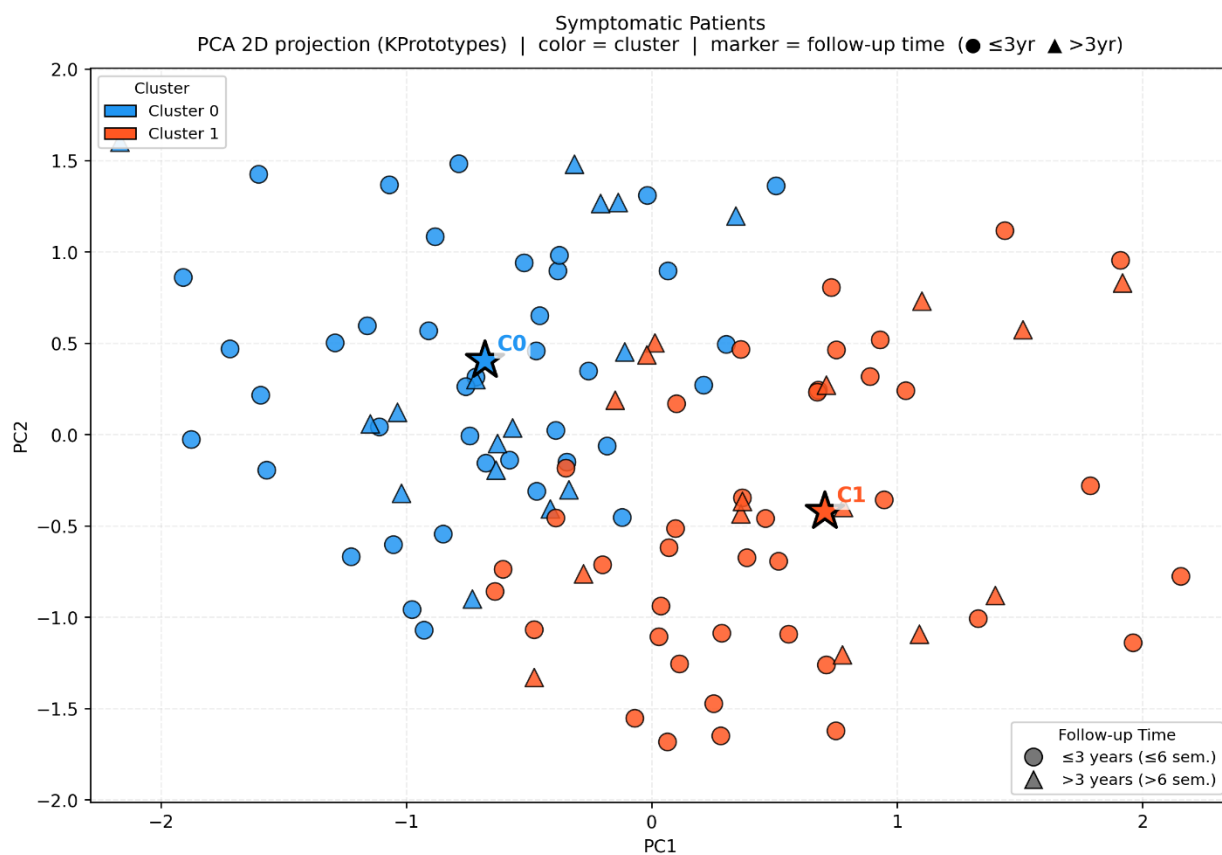
Σχήμα 4.2.13 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των συμπτωματικών ασθενών K-Prototypes.

Το Σχήμα 4.2.14 παρουσιάζει το ποσοστό διακύμανσης που εξηγείται από τις δύο πρώτες κύριες συνιστώσες της PCA. Η πρώτη κύρια συνιστώσα (PC1) εξηγεί περίπου το 35% της συνολικής διακύμανσης των δεδομένων, ενώ η δεύτερη κύρια συνιστώσα (PC2) εξηγεί περίπου το 20%.



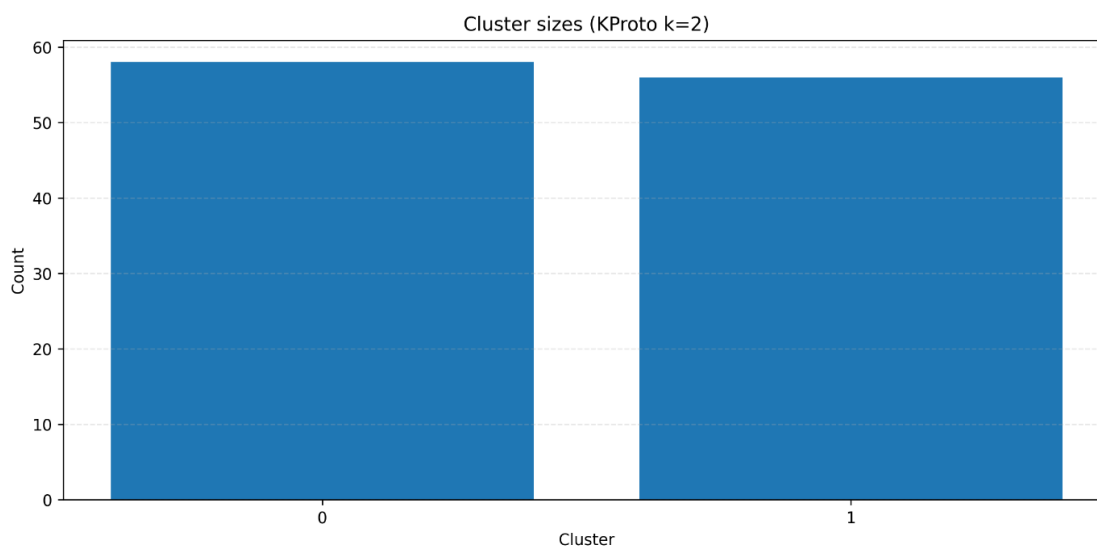
Σχήμα 4.2.14 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των συμπτωματικών ασθενών K-Prototypes.

Η προβολή των ασθενών στις δύο πρώτες κύριες συνιστώσες παρουσιάζεται στο Σχήμα 4.2.15. Παρατηρείται μερικός διαχωρισμός των δύο clusters κυρίως κατά μήκος της πρώτης κύριας συνιστώσας, με το Cluster 0 να συγκεντρώνεται κυρίως σε αρνητικές τιμές του άξονα PC1 και το Cluster 1 σε θετικές τιμές. Παράλληλα, παρατηρείται αλληλοεπικάλυψη μεταξύ των δύο ομάδων, γεγονός που αντανακλά την πολυπλοκότητα και την ετερογένεια των χαρακτηριστικών των συμπτωματικών ασθενών.



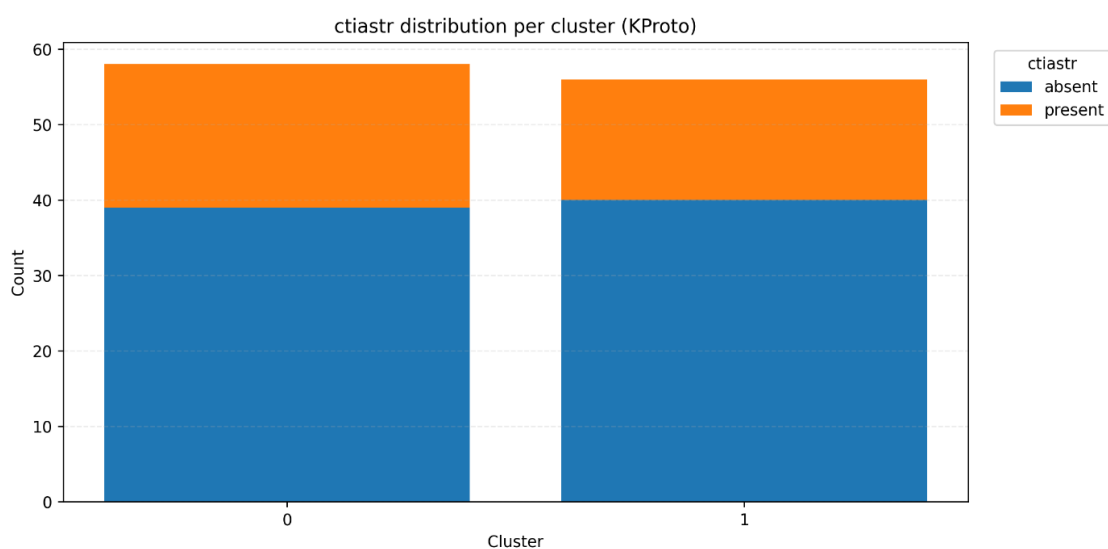
Σχήμα 4.2.15 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των συμπτωματικών ασθενών K-Prototypes

Η κατανομή των ασθενών ανά cluster παρουσιάζεται στο Σχήμα 4.2.16. Τα δύο clusters εμφανίζουν πολύ παρόμοιο αριθμό ασθενών, με το Cluster 0 να περιλαμβάνει 58 ασθενείς και το Cluster 1 να περιλαμβάνει 56 ασθενείς.

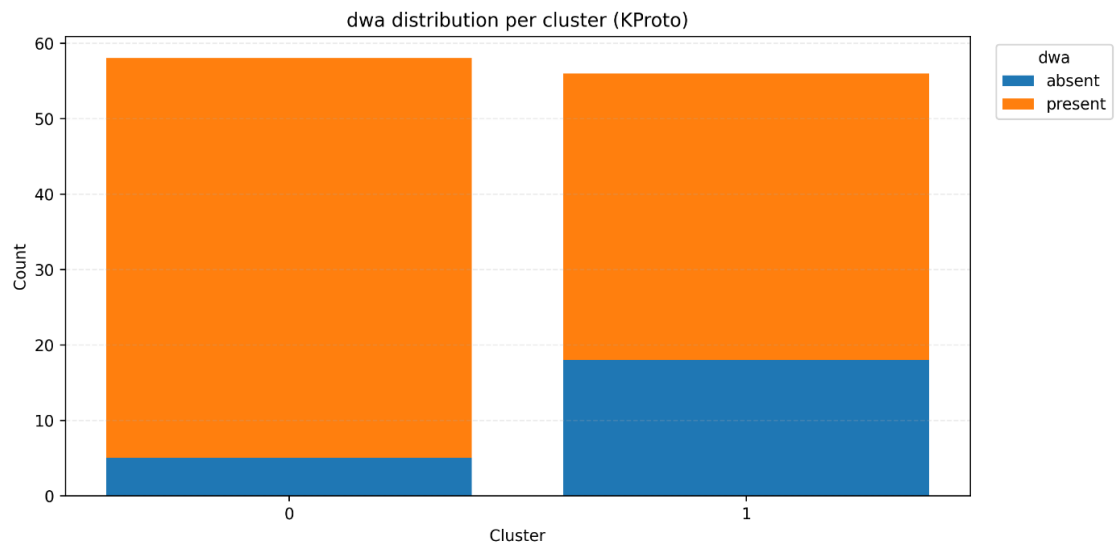


Σχήμα 4.2.16 Κατανομή των ασθενών ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes

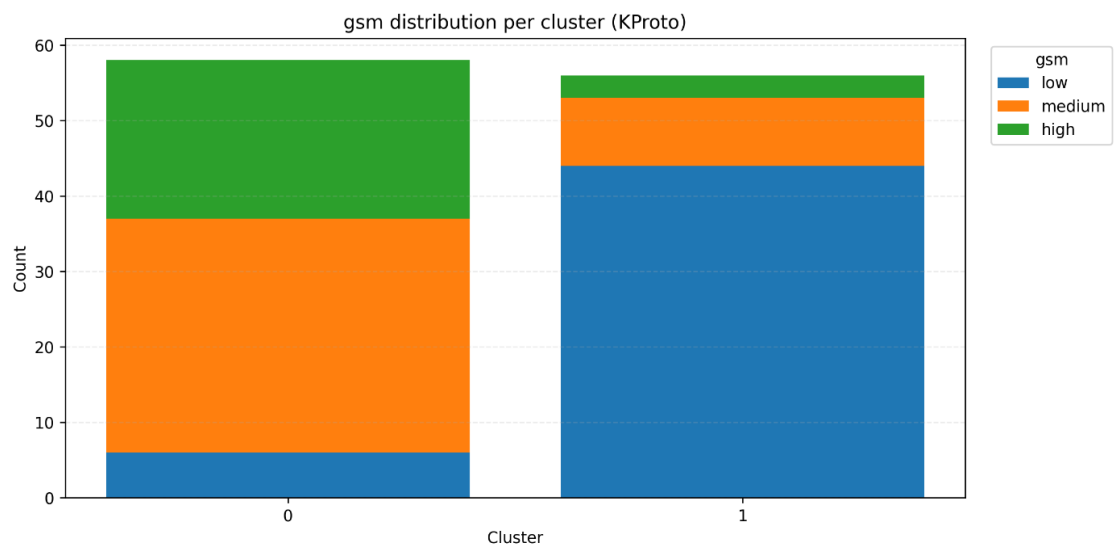
Η ανάλυση της κατανομής των χαρακτηριστικών ανά cluster στο σύνολο των συμπτωματικών ασθενών παρουσιάζεται στα Σχήματα 4.2.17 - 4.2.22. Τα διαγράμματα απεικονίζουν την κατανομή των μεταβλητών CTIASTR, DWA, GSM, JBA, PLAREA και STENOSIS για κάθε cluster που προέκυψε από τον αλγόριθμο K-Prototypes.



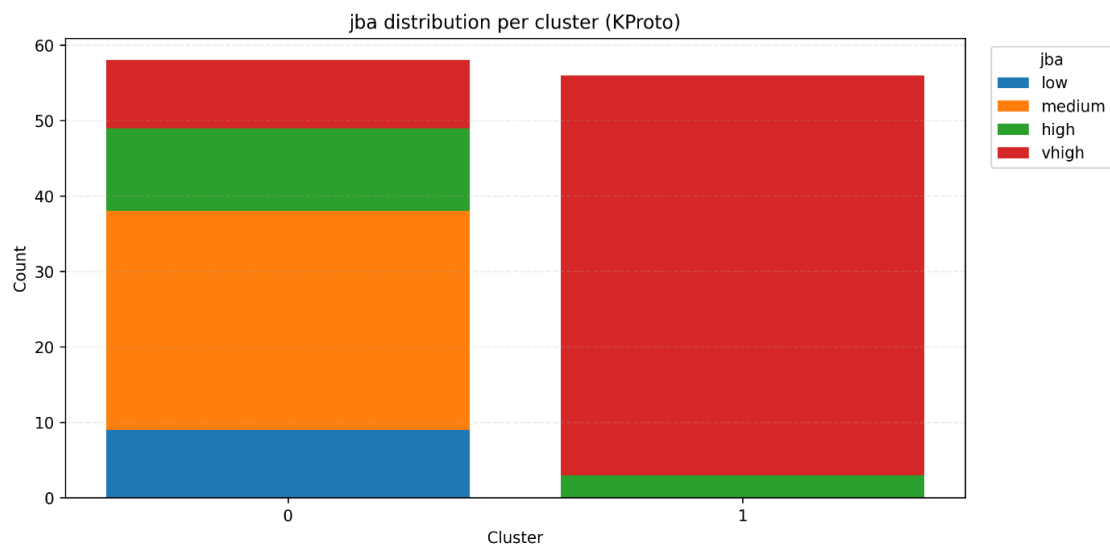
Σχήμα 4.2.17 Κατανομή CTIASTR ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes



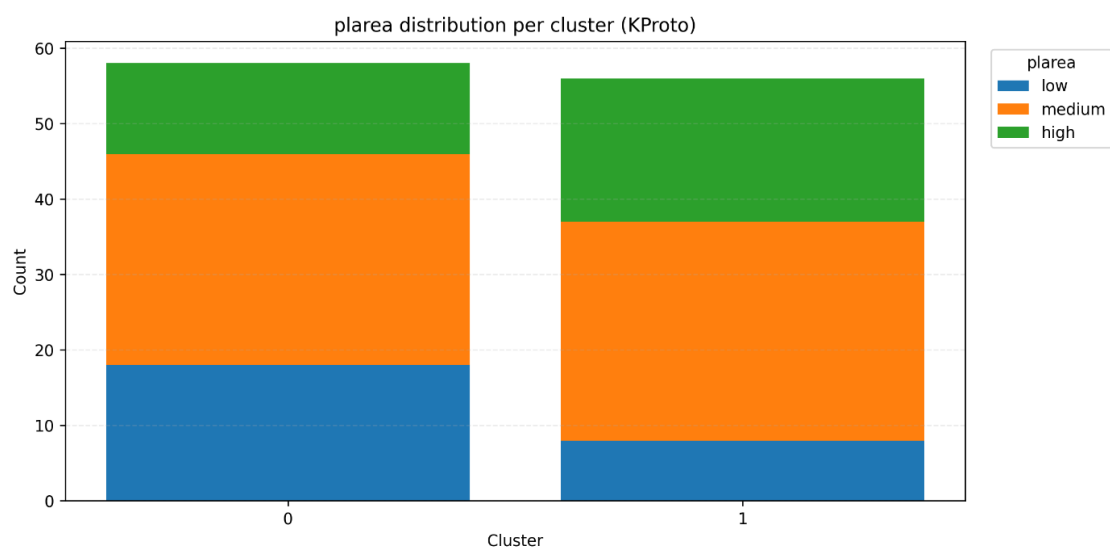
Σχήμα 4.2.18 Κατανομή DWA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.



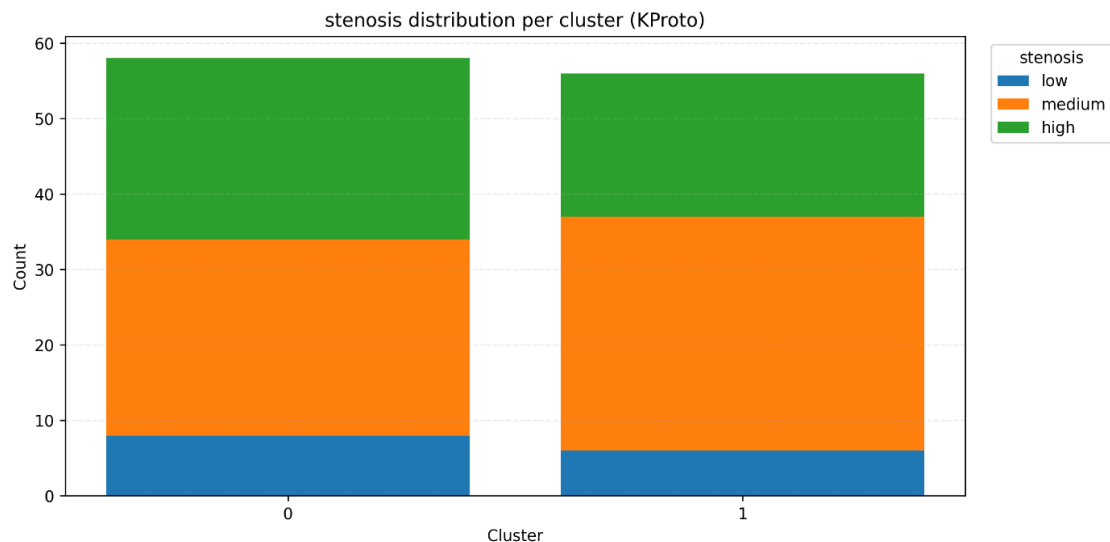
Σχήμα 4.2.19 Κατανομή GSM ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.



Σχήμα 4.2.20 Κατανομή JBA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.



Σχήμα 4.2.21 Κατανομή PLAREA ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.



Σχήμα 4.2.22 Κατανομή STENOSIS ανά συστάδα στο σύνολο των συμπτωματικών ασθενών K-Prototypes.

Η αξιολόγηση της ποιότητας του clustering πραγματοποιήθηκε με χρήση των δεικτών Silhouette, Davies–Bouldin και Calinski–Harabasz. Για το σύνολο των συμπτωματικών ασθενών και για $k=2$, ο δείκτης Silhouette έλαβε τιμή 0.15, ο δείκτης Davies–Bouldin τιμή 2.079 και ο δείκτης Calinski–Harabasz τιμή 22.23. Οι τιμές αυτές υποδεικνύουν μέτριο διαχωρισμό μεταξύ των clusters, κάτι που θεωρείται αναμενόμενο λόγω της αυξημένης ετερογένειας των συμπτωματικών ασθενών.

Παράλληλα, αξιολογήθηκε η σταθερότητα του μοντέλου μέσω πολλαπλών επαναλήψεων της διαδικασίας clustering ($n=20$), με χρήση του δείκτη Adjusted Rand Index (ARI). Η μέση τιμή του Pairwise ARI ήταν 0.977, με ελάχιστη τιμή 0.911 και μέγιστη τιμή 1.000, γεγονός που δείχνει πολύ υψηλή σταθερότητα και επανάληψη των αποτελεσμάτων του αλγορίθμου K-Prototypes στο συγκεκριμένο σύνολο δεδομένων.

Τα αριθμητικά αποτελέσματα των μετρικών παρουσιάζονται στους Πίνακες 4.2.4 και 4.2.5.

Group	n	k	Silhouette	Davies -Bouldin	Calinski - Harabasz
Symptomatic dataset	114	2	0.15	2.03	22.23

Πίνακας 4.2.4 Εσωτερικές μετρικές ομαδοποίησης K -Prototypes στο σύνολο των συμπτωματικών ασθενών

Group	Runs	ARI (Mean)	ARI (Min)	ARI (Max)
Symptomatic dataset	20	0.99	0.93	1.00

Πίνακας 4.2.5 Σταθερότητα της ομαδοποίησης K-Prototypes στο σύνολο των συμπτωματικών ασθενών

4.2.4 Αποτελέσματα κανόνων TE2Rules στο σύνολο των συμπτωματικών ασθενών με τον αλγόριθμο K-Prototypes

Η μέθοδος TE2Rules εφαρμόστηκε στις συστάδες που προέκυψαν από την ανάλυση K-Prototypes στο σύνολο των συμπτωματικών ασθενών, με $k=2$. Στόχος ήταν η εξαγωγή ερμηνεύσιμων κανόνων που περιγράφουν τα βασικά χαρακτηριστικά κάθε συστάδας.

Όπως και στα προηγούμενα, η μεταβλητή-στόχος που χρησιμοποιήθηκε η μεταβλητή Cluster. Ως μεταβλητές εισόδου χρησιμοποιήθηκαν οι μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA. Για κάθε συστάδα εφαρμόστηκε προσέγγιση one-vs-rest, ώστε να εντοπιστούν οι κανόνες που διαφοροποιούν τη συγκεκριμένη συστάδα από την υπόλοιπη ομάδα ασθενών.

Οι κανόνες που προέκυψαν παρουσιάζονται στον Πίνακα 4.2.6.

Cluster	Κανόνες TE2Rules
Cluster 0	$GSM \geq \text{medium AND } JBA \leq \text{high}$
	$PLAREA \leq \text{medium AND } GSM \geq \text{medium}$
	$JBA \leq \text{medium}$
	$JBA \leq \text{high AND STENOSIS} = \text{high}$
	$CTIASTR = \text{absent AND } GSM = \text{high AND } JBA = \text{very high AND STENOSIS} = \text{high}$
	$CTIASTR = \text{present AND } DWA = \text{present AND } GSM \geq \text{medium AND } PLAREA = \text{high AND STENOSIS} \leq \text{medium}$
Cluster 1	$GSM = \text{low AND } JBA \geq \text{high}$
	$CTIASTR = \text{absent AND } GSM \leq \text{medium AND } JBA = \text{very high AND } PLAREA = \text{high}$
	$DWA = \text{absent AND } JBA = \text{very high}$
	$CTIASTR = \text{present AND } JBA = \text{very high AND } PLAREA = \text{high AND STENOSIS} = \text{high}$

Πίνακας 4.2.6 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των συμπτωματικών ασθενών με K-Prototypes.

Οι κανόνες δείχνουν ότι το Cluster 0 χαρακτηρίζεται κυρίως από μέτριες ή υψηλές τιμές GSM σε συνδυασμό με χαμηλότερες τιμές JBA, καθώς και από μικρότερες τιμές PLAREA. Παράλληλα, εμφανίζονται κανόνες που συνδέουν υψηλές τιμές STENOSIS με χαμηλότερες τιμές JBA, ενώ σε ορισμένες περιπτώσεις παρατηρείται συνδυασμός παρουσίας DWA και υψηλότερων τιμών GSM.

Αντίθετα, το Cluster 1 χαρακτηρίζεται κυρίως από χαμηλές τιμές GSM και πολύ υψηλές τιμές JBA. Επιπλέον, αρκετοί κανόνες της συγκεκριμένης συστάδας περιλαμβάνουν αυξημένες τιμές PLAREA και υψηλές τιμές STENOSIS, υποδεικνύοντας μορφολογικά χαρακτηριστικά που σχετίζονται με περισσότερο ευάλωτες αθηρωματικές πλάκες.

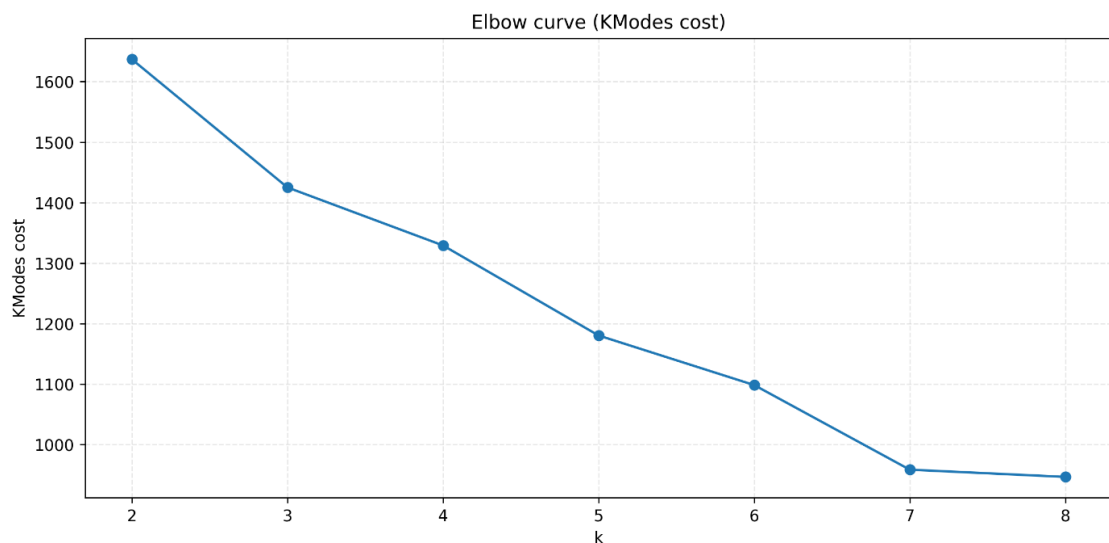
Συνολικά, οι κανόνες TE2Rules παρέχουν μια πιο ερμηνεύσιμη περιγραφή των μορφολογικών χαρακτηριστικών που διαφοροποιούν τις δύο συστάδες των συμπτωματικών ασθενών και συμβάλλουν στην καλύτερη κατανόηση των προτύπων που εντοπίστηκαν μέσω της ομαδοποίησης με τον αλγόριθμο K-Prototypes.

4.3 Σύνολο ασυμπτωματικών ασθενών

Αποτελέσματα ομαδοποίησης στο σύνολο των ασυμπτωματικών ασθενών με τον αλγόριθμο K-Modes

Η ανάλυση συστάδων στο σύνολο των ασυμπτωματικών ασθενών πραγματοποιήθηκε αρχικά, με χρήση του αλγορίθμου K-Modes, με στόχο την αναγνώριση ομάδων ασθενών με παρόμοια μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας.

Για την επιλογή του κατάλληλου αριθμού συστάδων εφαρμόστηκε, όπως και στα προηγούμενα μοντέλα, η μέθοδος elbow, με αξιολόγηση του κόστους clustering για διαφορετικές τιμές του k . Όπως παρουσιάζεται στο Σχήμα 4.3.1, παρατηρείται σταδιακή μείωση του κόστους όσο αυξάνεται ο αριθμός των clusters, με το σημείο καμπής να εντοπίζεται κοντά στο $k=2$, γεγονός που υποστηρίζει την επιλογή δύο συστάδων για την περιγραφή της δομής των δεδομένων.



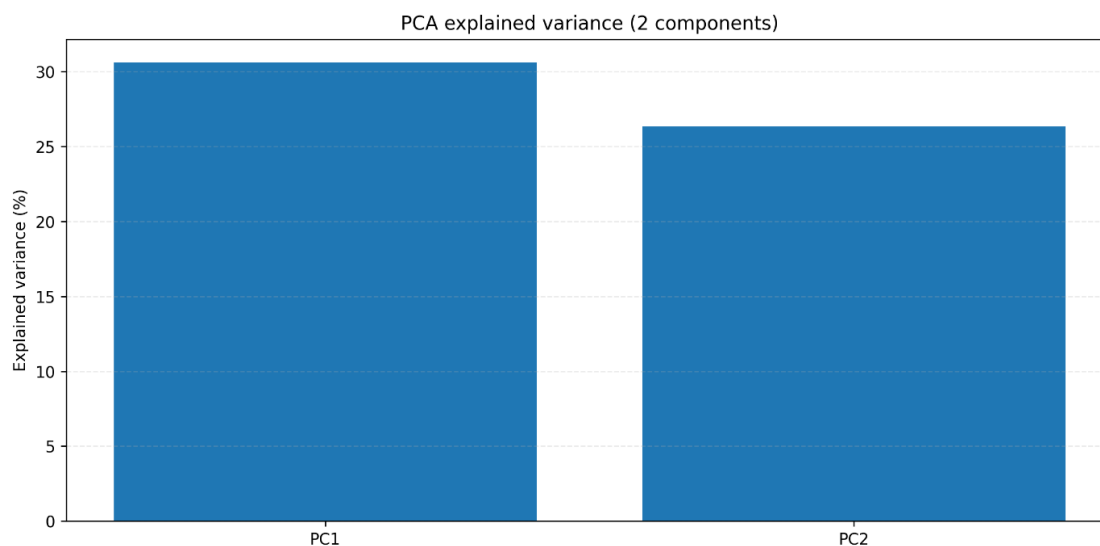
Σχήμα 4.3.1 Καμπύλη elbow για τον αλγόριθμο K-Modes στο σύνολο των ασυμπτωματικών ασθενών

Παράλληλα, αξιολογήθηκαν οι δείκτες ποιότητας clustering Silhouette, Davies–Bouldin και Calinski–Harabasz για διαφορετικές τιμές του k . Όπως φαίνεται στο Σχήμα 4.3.2, οι τιμές των μετρικών παρουσιάζουν μικρές διαφοροποιήσεις μεταξύ των διαφορετικών αριθμών συστάδων, όμως, η επιλογή του $k=2$ θεωρήθηκε κατάλληλη για την περιγραφή των δεδομένων.



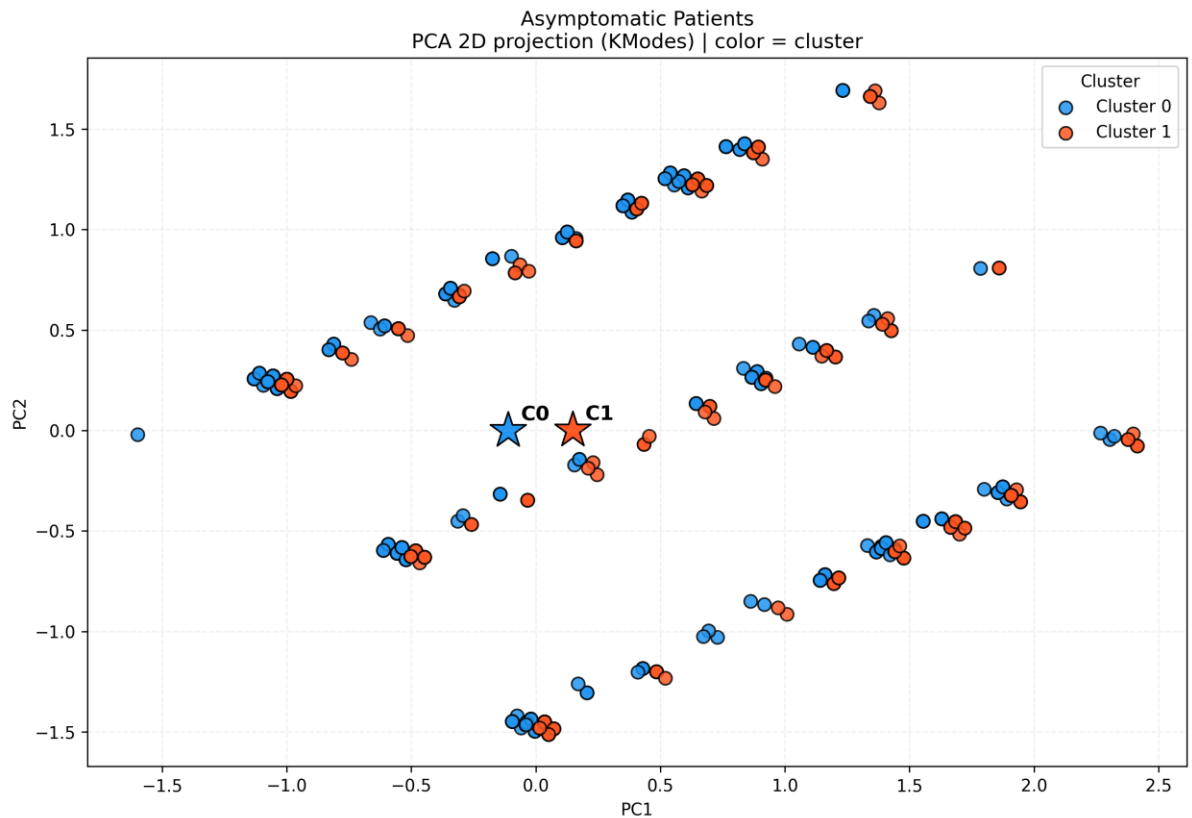
Σχήμα 4.3.2 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των ασυμπτωματικών ασθενών K-Modes

Το Σχήμα 4.3.3 παρουσιάζει το ποσοστό διακύμανσης που εξηγείται από τις δύο πρώτες κύριες συνιστώσες της PCA. Η πρώτη κύρια συνιστώσα (PC1) εξηγεί μεγαλύτερο ποσοστό της συνολικής διακύμανσης σε σχέση με τη δεύτερη κύρια συνιστώσα (PC2).



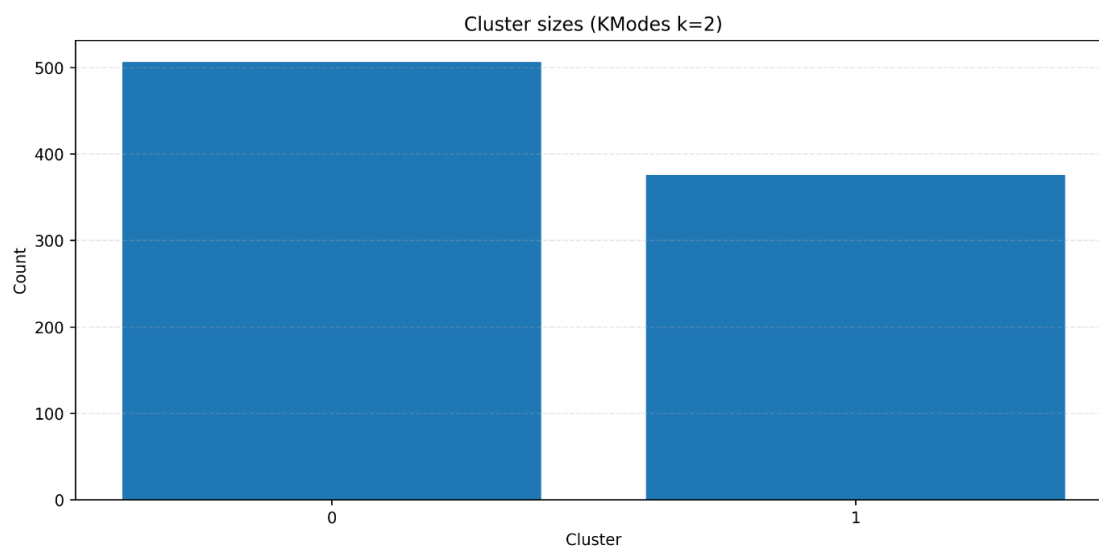
Σχήμα 4.3.3 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των ασυμπτωματικών ασθενών K-Modes

Η προβολή των ασθενών στις δύο πρώτες κύριες συνιστώσες παρουσιάζεται στο Σχήμα 4.3.4. Παρατηρείται μερικός διαχωρισμός των δύο clusters στον χώρο των κύριων συνιστωσών. Παράλληλα, εμφανίζεται επικάλυψη μεταξύ ορισμένων περιπτώσεων των δύο συστάδων.



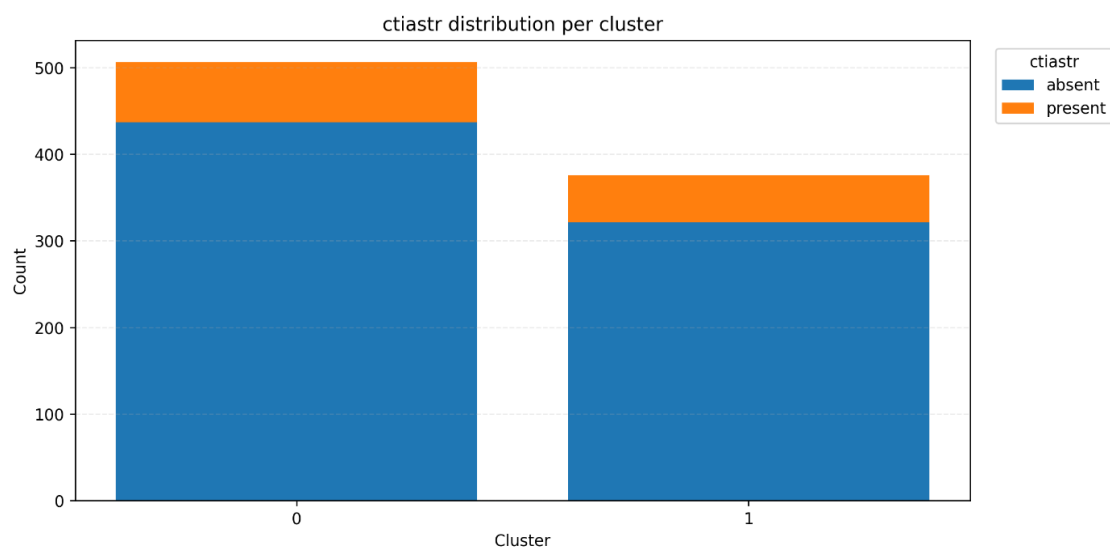
Σχήμα 4.3.4 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των ασυμπτωματικών ασθενών K-Modes

Η κατανομή των ασθενών ανά cluster παρουσιάζεται στο Σχήμα 4.3.5. Όπως φαίνεται και στην γραφική παράσταση, το Cluster 0 περιλαμβάνει μεγαλύτερο αριθμό ασθενών συγκριτικά με το Cluster 1.

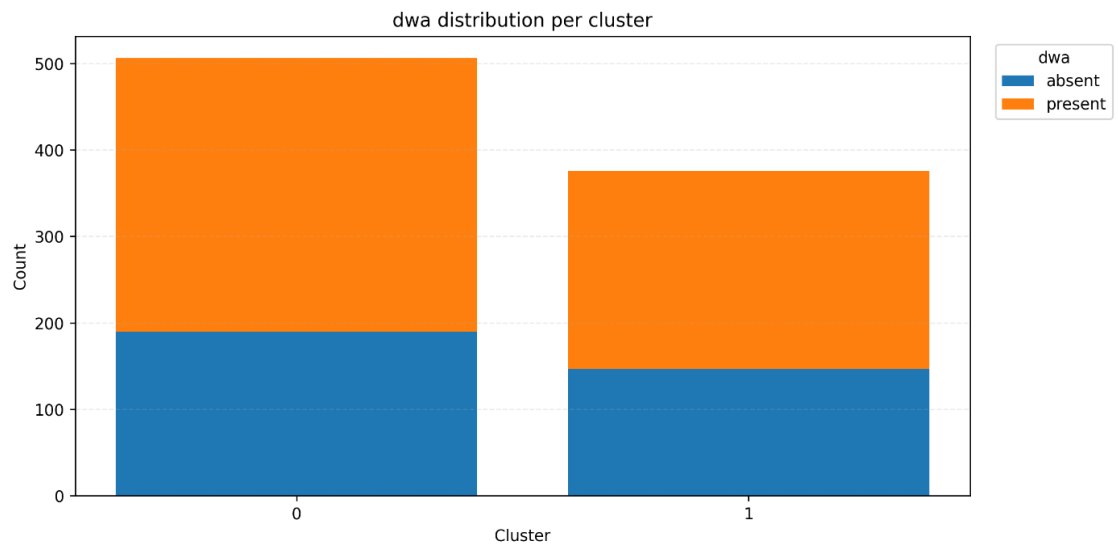


Σχήμα 4.3.5 Κατανομή των ασθενών ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes

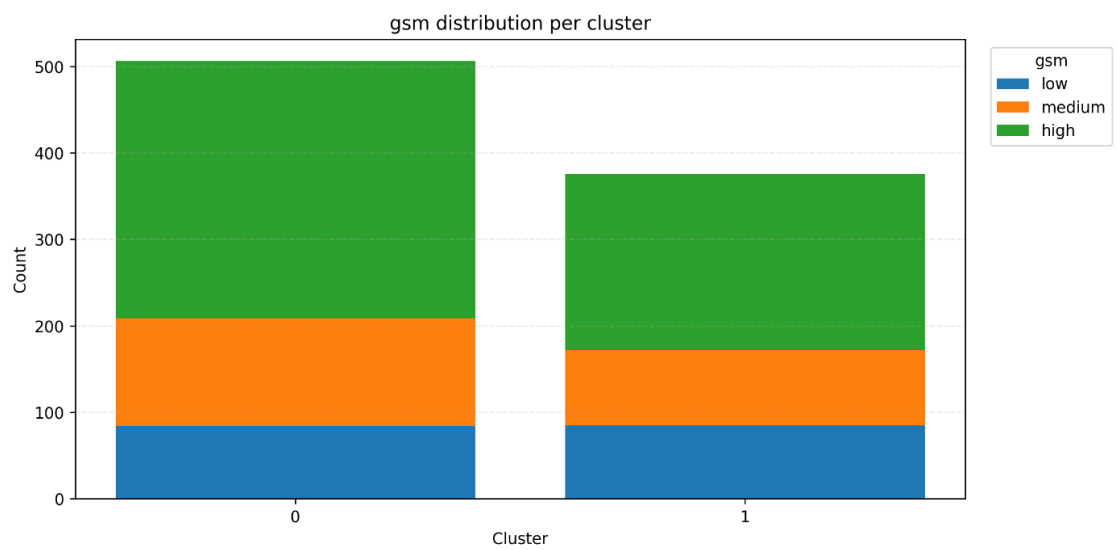
Η ανάλυση της κατανομής των χαρακτηριστικών ανά cluster παρουσιάζεται στα Σχήματα 4.3.6 - 4.3.11. Τα διαγράμματα απεικονίζουν την κατανομή των μεταβλητών CTIASTR, DWA, GSM, JBA, PLAREA και STENOSIS για κάθε cluster που προέκυψε από τον αλγόριθμο K-Modes.



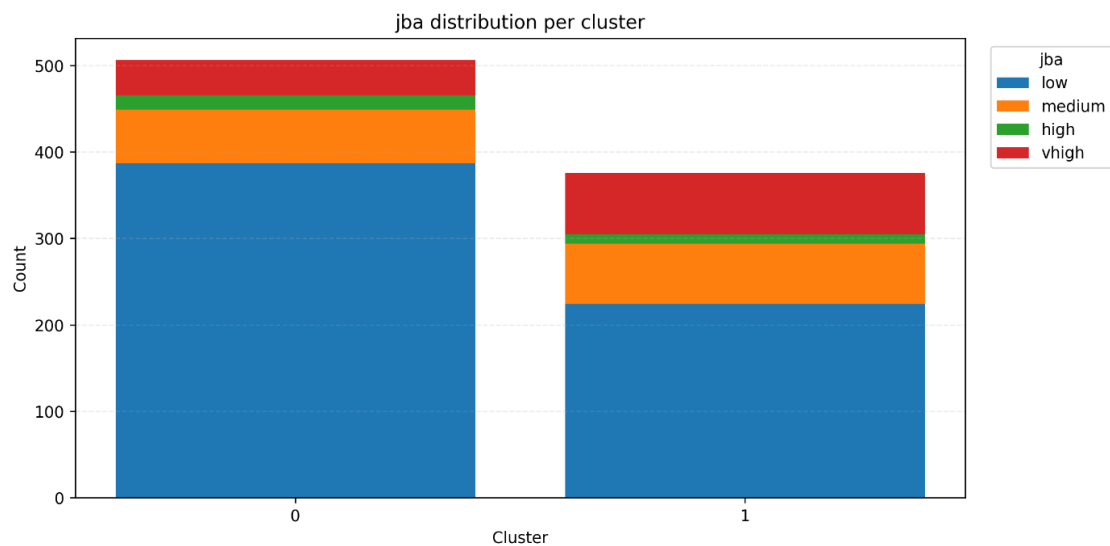
Σχήμα 4.3.6 Κατανομή CTIASTR ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes



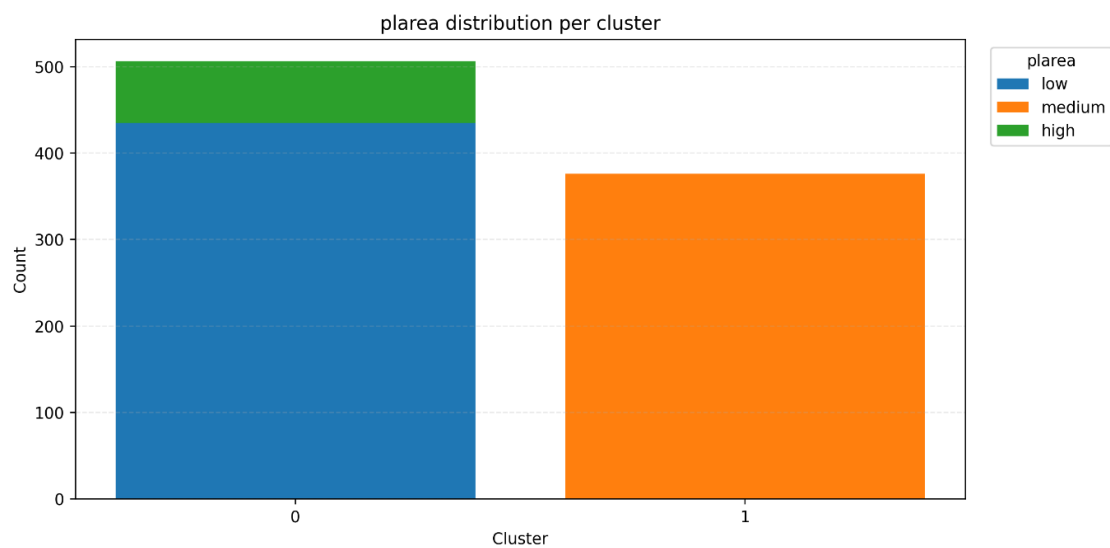
Σχήμα 4.3.7 Κατανομή DWA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes



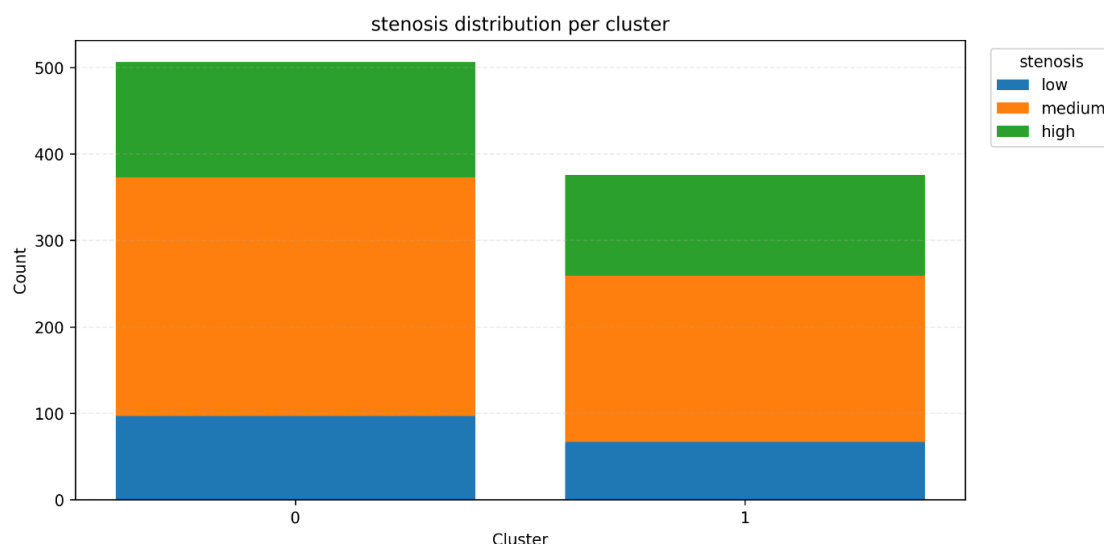
Σχήμα 4.3.8 Κατανομή GSM ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes



Σχήμα 4.3.9 Κατανομή JBA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes



Σχήμα 4.3.10 Κατανομή PLAREA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes



Σχήμα 4.3.11 Κατανομή STENOSIS ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Modes

Η αξιολόγηση της ποιότητας του clustering για το σύνολο των ασυμπτωματικών ασθενών για $k=2$, πραγματοποιήθηκε και εδώ με τον δείκτη Silhouette να λαμβάνει τιμή 0.27, με μέση απόσταση Hamming εντός των συστάδων 0.42, ο δείκτης Davies–Bouldin έλαβε τιμή 2.56 και ο δείκτης Calinski–Harabasz τιμή 124.04. Οι μετρικές αξιολόγησης υποδεικνύουν μέτριο βαθμό συνοχής μεταξύ των ασθενών κάθε συστάδας.

Συγχρόνως, αξιολογήθηκε η σταθερότητα του μοντέλου μέσω πολλαπλών επαναλήψεων της διαδικασίας clustering ($n=20$), με χρήση του δείκτη Adjusted Rand Index (ARI). Η μέση τιμή του Pairwise ARI ήταν 0.46, με ελάχιστη τιμή 0.05 και μέγιστη τιμή 1.00. Τα αποτελέσματα αυτά δείχνουν μεταβλητή σταθερότητα της ομαδοποίησης μεταξύ των επαναλαμβανόμενων εκτελέσεων, γεγονός που υποδηλώνει ότι η δομή των συστάδων στο σύνολο των ασυμπτωματικών ασθενών δεν είναι πλήρως σταθερή για όλες τις αρχικοποιήσεις του αλγορίθμου.

Τα αριθμητικά αποτελέσματα των μετρικών παρουσιάζονται στους Πίνακες 4.3.1 και 4.3.2.

Group	n	k	Silhouette	Davies - Bouldin	Calinski - Harabasz	Avg within Hamming Distance
Asymptomatic dataset	882	2	0.27	2.56	124.04	0.42

Πίνακας 4.3.1 Εσωτερικές μετρικές ομαδοποίησης K-Modes στο σύνολο των ασυμπτωματικών ασθενών

Group	Runs	ARI (Mean)	ARI (Min)	ARI (Max)
Asymptomatic dataset	20	0.46	0.05	1.00

Πίνακας 4.3.2 Σταθερότητα της ομαδοποίησης K-Modes στο σύνολο των ασυμπτωματικών ασθενών

4.3.1 Αποτελέσματα κανόνων με TE2Rules στο σύνολο των ασυμπτωματικών ασθενών με K-Modes

Η μέθοδος TE2Rules εφαρμόστηκε στις συστάδες που προέκυψαν από την ανάλυση K-Modes στο σύνολο των ασυμπτωματικών ασθενών, με $k=2$. Στόχος ήταν η εξαγωγή ερμηνεύσιμων κανόνων που περιγράφουν τα βασικά χαρακτηριστικά κάθε συστάδας.

Όπως και στα προηγούμενα μοντέλα, ως μεταβλητή-στόχος χρησιμοποιήθηκε η μεταβλητή Cluster. Ως μεταβλητές εισόδου χρησιμοποιήθηκαν οι κατηγορικές μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA. Για κάθε συστάδα εφαρμόστηκε προσέγγιση one-vs-rest, ώστε να εντοπιστούν οι κανόνες που διαφοροποιούν τη συγκεκριμένη συστάδα από την υπόλοιπη ομάδα ασθενών.

Οι κανόνες που προέκυψαν παρουσιάζονται στον Πίνακα 4.3.3.

Cluster	Κανόνες TE2Rules
Cluster 0	PLAREA = low
	PLAREA = high
Cluster 1	PLAREA = medium

Πίνακας 4.3.3 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των ασυμπτωματικών ασθενών με K-Modes.

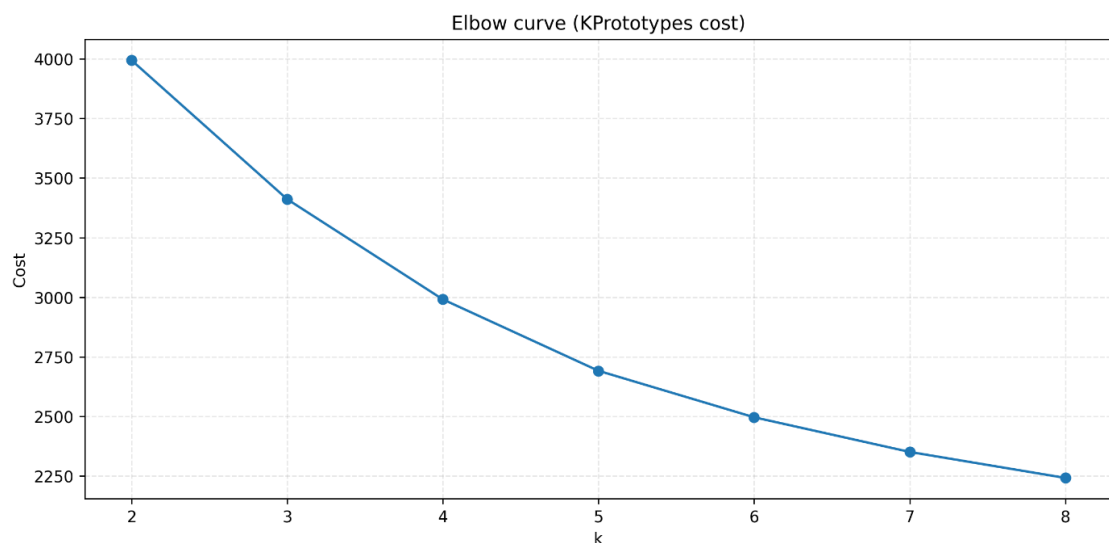
Οι κανόνες δείχνουν ότι το Cluster 0 χαρακτηρίζεται κυρίως από χαμηλές ή υψηλές τιμές της μεταβλητής PLAREA, υποδεικνύοντας μεγαλύτερη ετερογένεια ως προς την επιφάνεια της αθηρωματικής πλάκας. Αντίθετα, το Cluster 1 περιγράφεται κυρίως από μέτριες τιμές PLAREA, γεγονός που διαφοροποιεί τη συγκεκριμένη συστάδα από την υπόλοιπη ομάδα ασθενών.

Συνολικά, οι κανόνες TE2Rules παρέχουν μια πιο ερμηνεύσιμη περιγραφή των χαρακτηριστικών που διαφοροποιούν τις δύο συστάδες των ασυμπτωματικών ασθενών και συμβάλλουν στην καλύτερη κατανόηση των προτύπων που εντοπίστηκαν μέσω της ομαδοποίησης.

4.3.2 Αποτελέσματα ομαδοποίησης στο σύνολο των ασυμπτωματικών ασθενών με τον αλγόριθμο K-Prototypes

Η ανάλυση συστάδων με χρήση του αλγορίθμου K-Prototypes εφαρμόστηκε στο σύνολο των ασυμπτωματικών ασθενών, με στόχο τη συνδυαστική αξιοποίηση κατηγορικών και αριθμητικών χαρακτηριστικών της αθηρωματικής πλάκας. Για την εκτέλεση της ανάλυσης χρησιμοποιήθηκαν οι μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA, τόσο στην κατηγορική, όσο και στην αριθμητική τους μορφή, μετά από κατάλληλη κανονικοποίηση των αριθμητικών χαρακτηριστικών.

Η επιλογή του αριθμού των συστάδων πραγματοποιήθηκε με βάση τη μέθοδο elbow και τις εσωτερικές μετρικές αξιολόγησης clustering. Όπως παρουσιάζεται στο Σχήμα 4.3.12, παρατηρείται σημαντική μείωση του κόστους από $k=2$ προς μεγαλύτερες τιμές του k , ενώ στη συνέχεια η μείωση γίνεται πιο ομαλή, υποδεικνύοντας ως καταλληλότερη επιλογή τις δύο συστάδες.



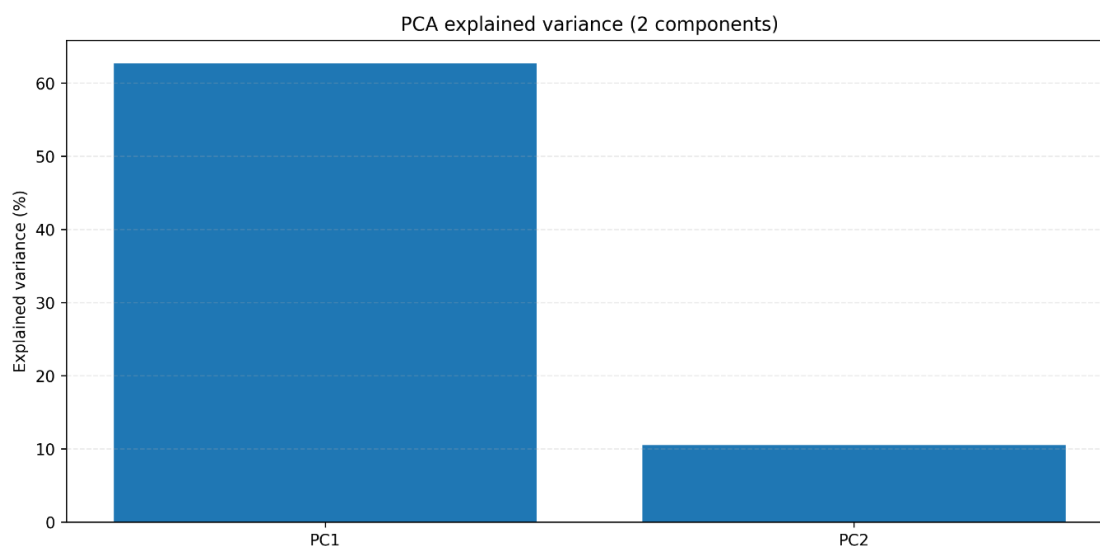
Σχήμα 4.3.12 Καμπύλη elbow για τον αλγόριθμο K-Prototypes στο σύνολο των ασυμπτωματικών ασθενών.

Η ανάλυση των μετρικών ποιότητας clustering παρουσιάζεται στο Σχήμα 4.3.13. Ο δείκτης Silhouette εμφανίζει τη μεγαλύτερη τιμή για $k=2$, ενώ παράλληλα ο δείκτης Davies–Bouldin παρουσιάζει σχετικά χαμηλές τιμές και ο δείκτης Calinski–Harabasz τη μεγαλύτερη τιμή του για μικρό αριθμό clusters. Τα αποτελέσματα αυτά υποστηρίζουν την επιλογή δύο συστάδων για την περιγραφή της δομής των δεδομένων.



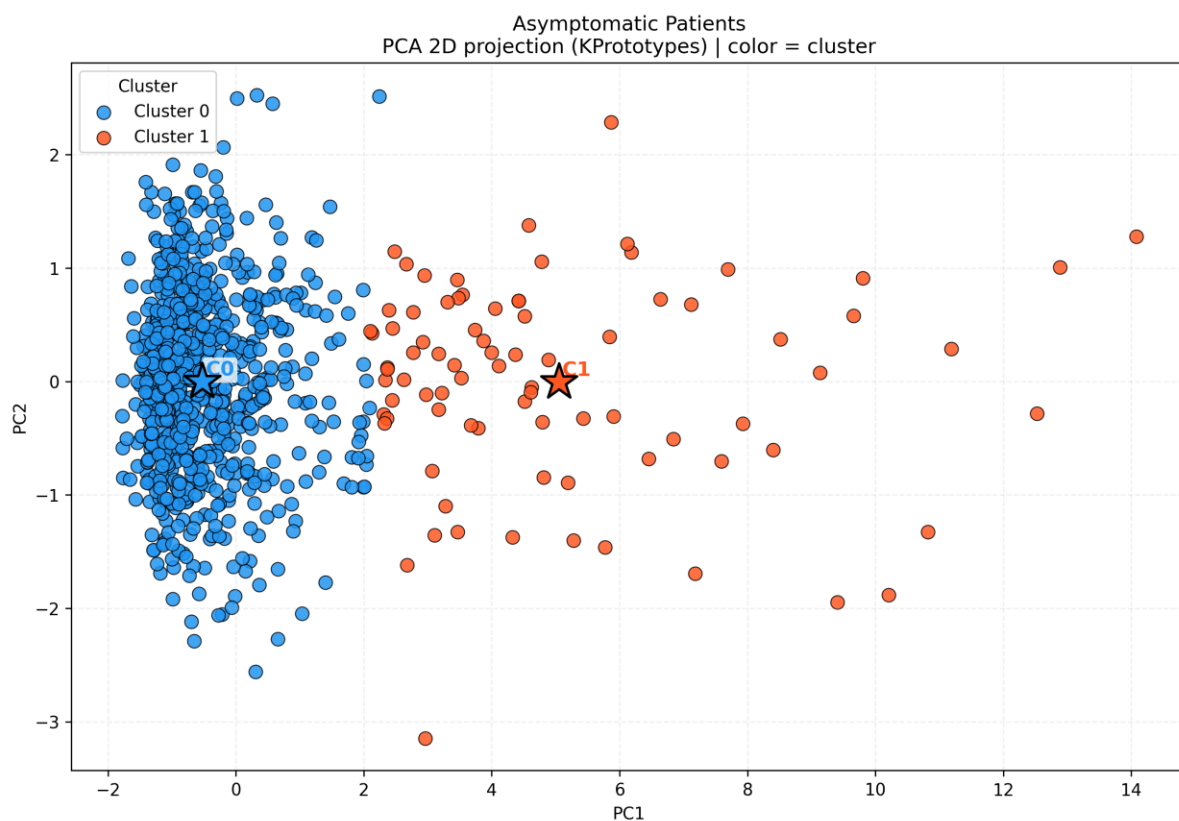
Σχήμα 4.3.13 Μετρικές αξιολόγησης ομαδοποίησης ως συνάρτηση του αριθμού συστάδων k στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.

Η πρώτη κύρια συνιστώσα εξηγεί περίπου το 63% της συνολικής διακύμανσης των δεδομένων και η δεύτερη περίπου το 10%, όπως παρουσιάζεται στο Σχήμα 4.3.14. Συνεπώς, η δισδιάστατη αναπαράσταση PCA αποτυπώνει ικανοποιητικό μέρος της πληροφορίας των δεδομένων.



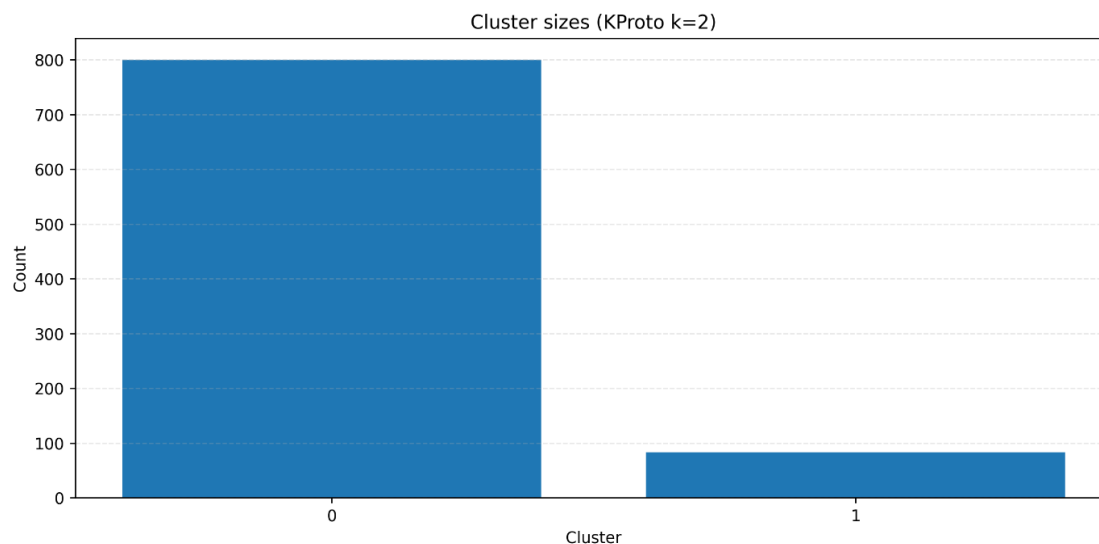
Σχήμα 4.3.14 Ποσοστό εξηγούμενης διακύμανσης από τις κύριες συνιστώσες στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.

Η δισδιάστατη απεικόνιση PCA των ασθενών παρουσιάζεται στο Σχήμα 4.3.15. Η προβολή δείχνει σχετικά σαφή διαχωρισμό της Cluster 1 από την κύρια ομάδα ασθενών κατά μήκος της πρώτης κύριας συνιστώσας (PC1), γεγονός που υποδηλώνει ότι συγκεκριμένα χαρακτηριστικά της αθηρωματικής πλάκας διαφοροποιούν τη συγκεκριμένη υποομάδα ασθενών.



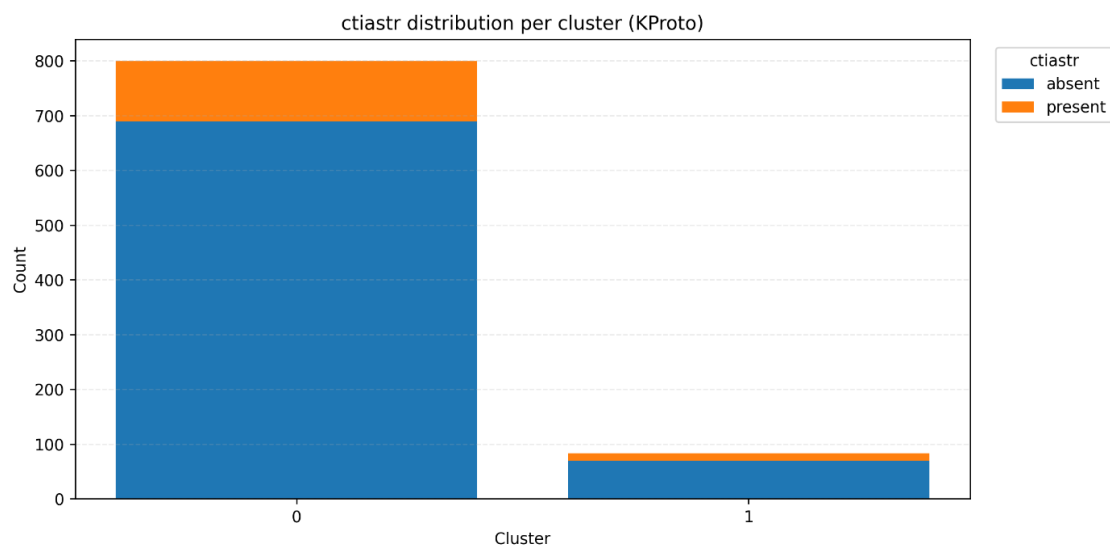
Σχήμα 4.3.15 Προβολή των συστάδων στον χώρο των δύο πρώτων κύριων συνιστωσών στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.

Η κατανομή των ασθενών στις δύο συστάδες παρουσιάζεται στο Σχήμα 4.3.16. Παρατηρείται ότι από τη μία το, Cluster 0 περιλαμβάνει τη μεγάλη πλειονότητα των ασθενών, και από την άλλη το Cluster 1 αποτελεί μικρότερη, αλλά περισσότερο διακριτή ομάδα ασθενών.

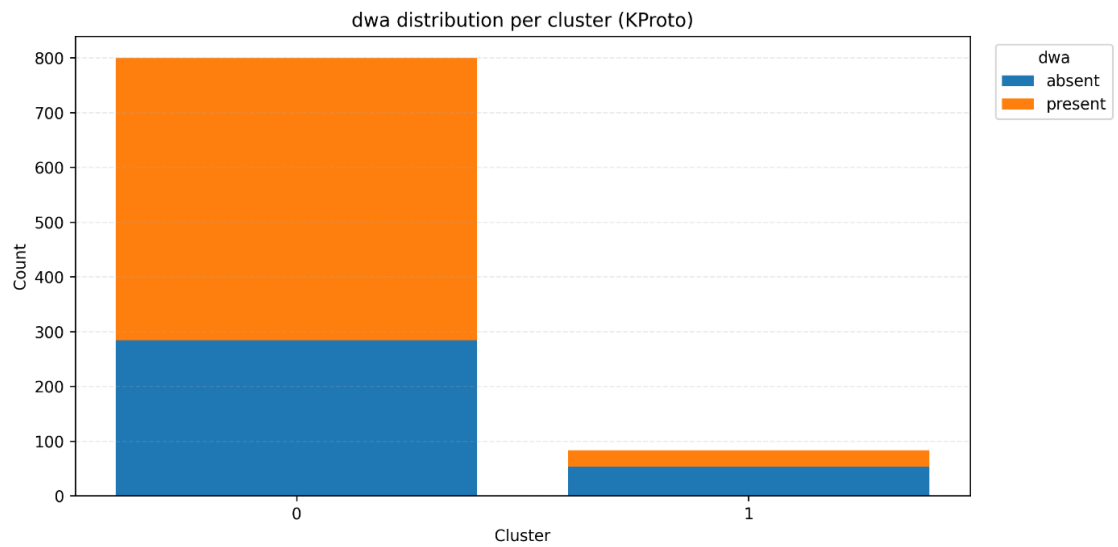


Σχήμα 4.3.16 Κατανομή των ασθενών ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.

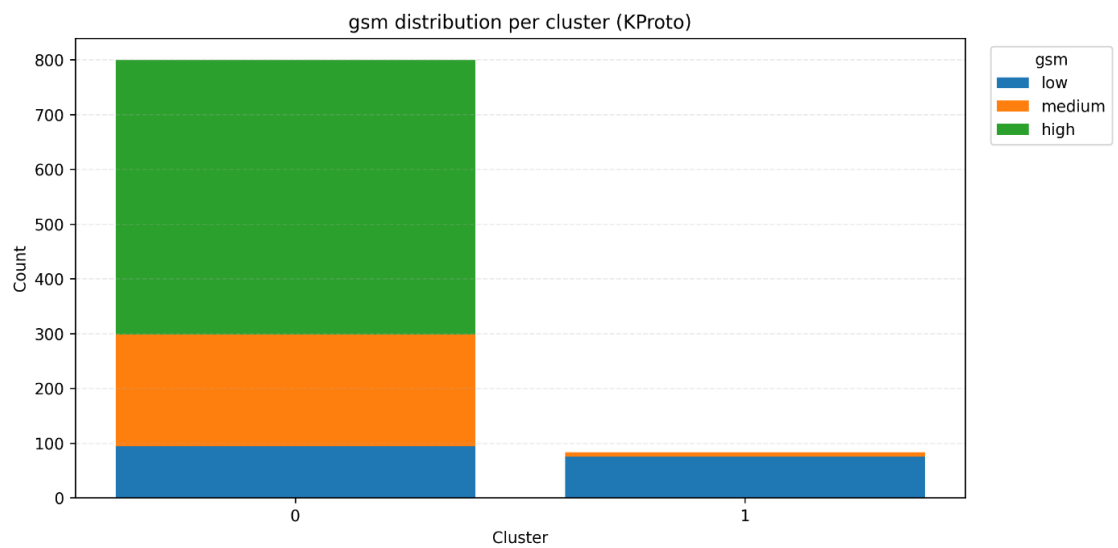
Η ανάλυση της κατανομής των χαρακτηριστικών ανά cluster στο σύνολο των ασυμπτωματικών ασθενών παρουσιάζεται στα Σχήματα 4.3.17 - 4.3.22. Τα διαγράμματα απεικονίζουν την κατανομή των μεταβλητών CTIASTR, DWA, GSM, JBA, PLAREA και STENOSIS για κάθε cluster που προέκυψε από τον αλγόριθμο K-Prototypes.



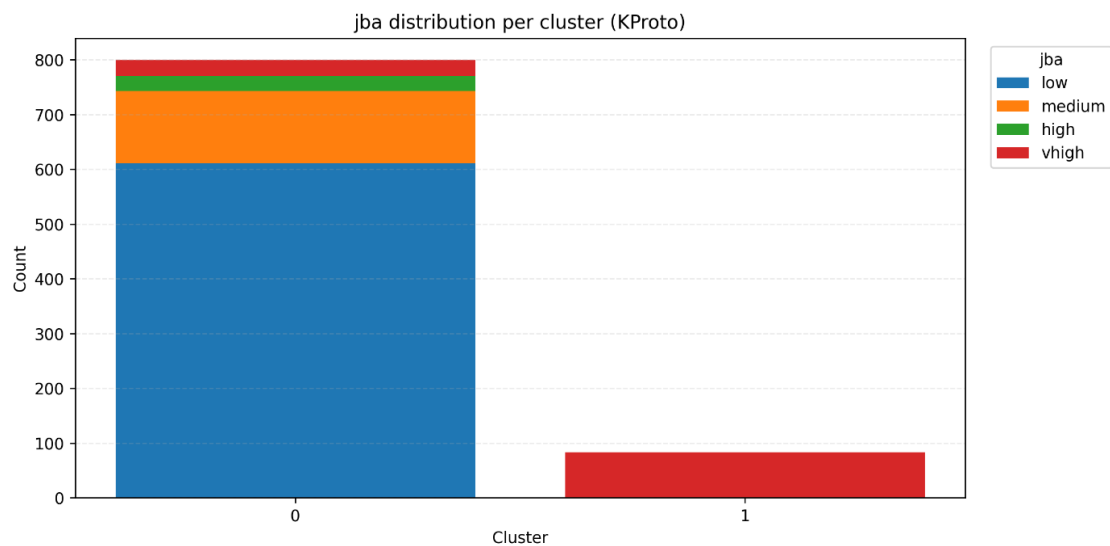
Σχήμα 4.3.17 Κατανομή CTIASTR ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.



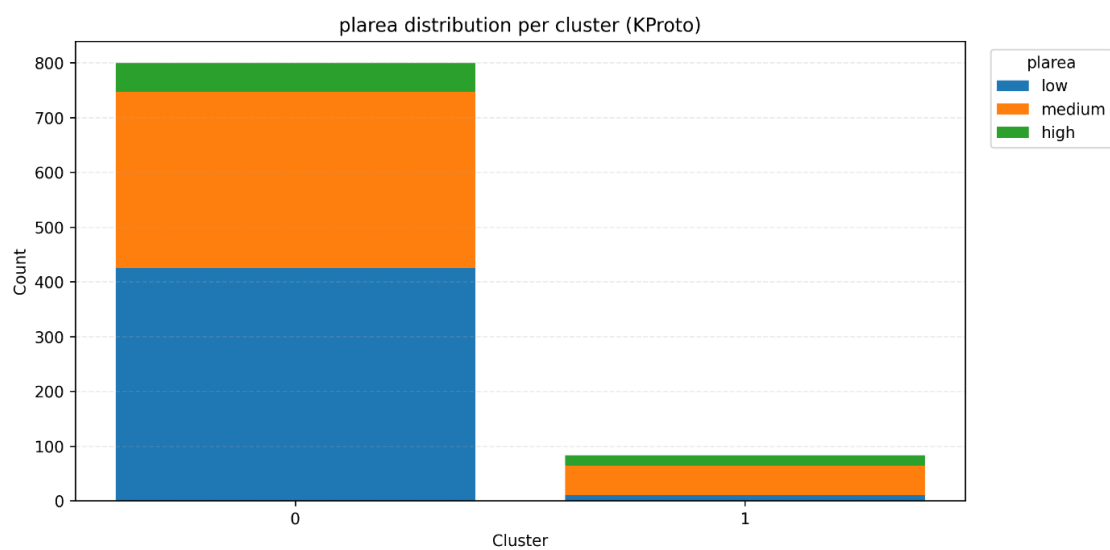
Σχήμα 4.3.18 Κατανομή DWA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.



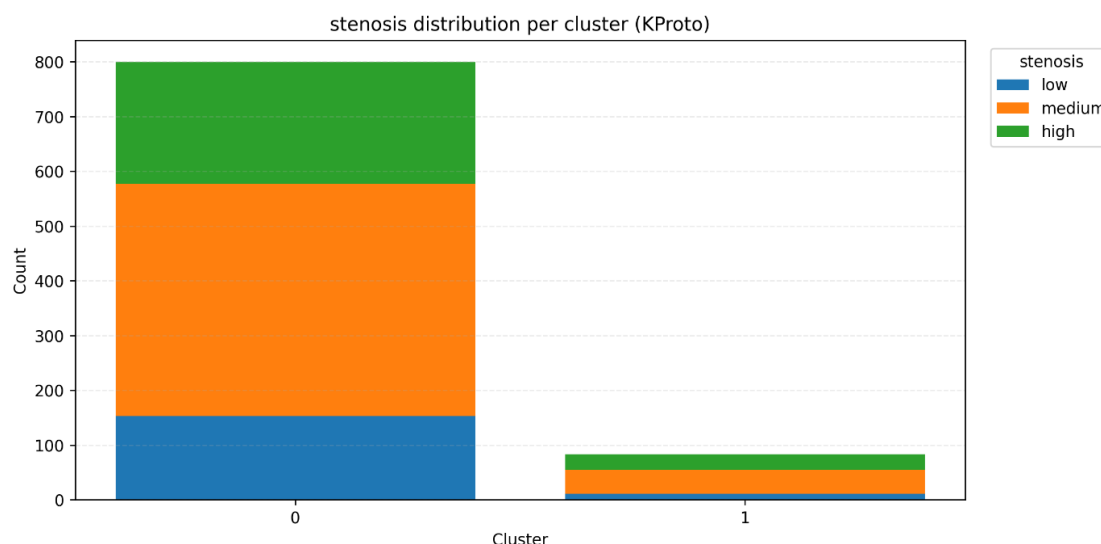
Σχήμα 4.3.19 Κατανομή GSM ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.



Σχήμα 4.3.20 Κατανομή JBA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.



Σχήμα 4.3.21 Κατανομή PLAREA ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.



Σχήμα 4.3.22 Κατανομή STENOSIS ανά συστάδα στο σύνολο των ασυμπτωματικών ασθενών K-Prototypes.

Η αξιολόγηση της ποιότητας του clustering στο σύνολο των ασυμπτωματικών ασθενών, πραγματοποιήθηκε με χρήση των δεικτών Silhouette, Davies–Bouldin και Calinski–Harabasz. Για $k=2$, ο δείκτης Silhouette έλαβε τιμή 0.59, ο δείκτης Davies–Bouldin τιμή 0.77 και ο δείκτης Calinski–Harabasz τιμή 661.90. Οι τιμές των μετρικών υποδηλώνουν ικανοποιητική ποιότητα ομαδοποίησης, με σχετικά καλό διαχωρισμό μεταξύ των συστάδων και αυξημένη συνοχή των ασθενών εντός κάθε cluster, γεγονός που υποστηρίζει την καταλληλότητα της επιλογής $k=2$ για το συγκεκριμένο σύνολο δεδομένων.

Παράλληλα, αξιολογήθηκε η σταθερότητα του μοντέλου μέσω πολλαπλών επαναλήψεων της διαδικασίας clustering ($n=20$), με χρήση του δείκτη Adjusted Rand Index (ARI). Η μέση τιμή του Pairwise ARI ήταν 1.00, με ελάχιστη και μέγιστη τιμή επίσης 1.00, γεγονός που δείχνει απόλυτη σταθερότητα των αποτελεσμάτων του αλγορίθμου K-Prototypes στο συγκεκριμένο σύνολο δεδομένων.

Τα αριθμητικά αποτελέσματα των μετρικών παρουσιάζονται στους Πίνακες 4.3.4 και 4.3.5.

Group	n	k	Silhouette	Davies -Bouldin	Calinski - Harabasz
Asymptomatic dataset	882	2	0.59	0.77	661.90

Πίνακας 4.3.4 Εσωτερικές μετρικές ομαδοποίησης K -Prototypes στο σύνολο των ασυμπτωματικών ασθενών

Group	Runs	ARI (Mean)	ARI (Min)	ARI (Max)
Asymptomatic dataset	20	1.00	1.00	1.00

Πίνακας 4.3.5 Σταθερότητα της ομαδοποίησης K-Prototypes στο σύνολο των ασυμπτωματικών ασθενών

4.3.3 Αποτελέσματα κανόνων TE2Rules στο σύνολο των ασυμπτωματικών ασθενών με τον αλγόριθμο K-Prototypes

Η μέθοδος TE2Rules εφαρμόστηκε στις συστάδες που προέκυψαν από την ανάλυση K-Prototypes στο σύνολο των ασυμπτωματικών ασθενών, με $k=2$. Στόχος ήταν η εξαγωγή ερμηνεύσιμων κανόνων που περιγράφουν τα βασικά χαρακτηριστικά κάθε συστάδας.

Η μεταβλητή-στόχος παραμένει η ίδια όπως και στα προηγούμενα μοντέλα που είναι ,η μεταβλητή Cluster. Ως μεταβλητές εισόδου χρησιμοποιήθηκαν οι μεταβλητές STENOSIS, GSM, PLAREA, DWA, CTIASTR και JBA. Για κάθε συστάδα εφαρμόστηκε προσέγγιση one-vs-rest, ώστε να εντοπιστούν οι κανόνες που διαφοροποιούν τη συγκεκριμένη συστάδα από την υπόλοιπη ομάδα ασθενών.

Οι κανόνες που προέκυψαν παρουσιάζονται στον Πίνακα 4.3.6.

Cluster	Κανόνες TE2Rules
Cluster 0	$JBA \leq \text{high}$
	$DWA = \text{present AND PLAREA} \leq \text{medium}$
	$PLAREA = \text{low}$
	$GSM \geq \text{medium}$
	$DWA = \text{present AND STENOSIS} \leq \text{medium}$
Cluster 1	$GSM = \text{low AND JBA} = \text{very high AND STENOSIS} = \text{high}$
	$GSM = \text{low AND JBA} = \text{very high AND PLAREA} \geq \text{medium AND STENOSIS} = \text{low}$
	$DWA = \text{absent AND JBA} = \text{very high AND PLAREA} = \text{high}$
	$CTIASTR = \text{present AND DWA} = \text{absent AND GSM} = \text{low AND PLAREA} \geq \text{medium}$
	$CTIASTR = \text{absent AND GSM} \geq \text{medium AND JBA} = \text{very high AND PLAREA} = \text{high AND STENOSIS} \leq \text{medium}$
	$CTIASTR = \text{present AND JBA} = \text{very high AND PLAREA} = \text{high AND STENOSIS} \geq \text{medium}$
	$CTIASTR = \text{absent AND DWA} = \text{absent AND GSM} \geq \text{medium AND JBA} = \text{very high AND STENOSIS} \leq \text{medium}$

Πίνακας 4.3.6 Αντιπροσωπευτικοί κανόνες TE2Rules για το σύνολο των ασυμπτωματικών ασθενών με K-Prototypes.

Κεφάλαιο 5: Ανάλυση και Συζήτηση Αποτελεσμάτων

5.1	Συνολική ερμηνεία της ανάλυσης συστάδων	117
5.2	Ανάλυση μορφολογικών προτύπων στο σύνολο των ασθενών	119
5.2.1	Αποτελέσματα K-Modes στο σύνολο των ασθενών	119
5.2.2	Αποτελέσματα K-Prototypes στο σύνολο των ασθενών	121
5.2.3	Σύγκριση K-Modes & K-Prototypes στο σύνολο ασθενών	124
5.3	Ανάλυση μορφολογικών προτύπων στο σύνολο των συμπτωματικών ασθενών	126
5.3.1	Αποτελέσματα K-Modes στους συμπτωματικούς ασθενείς	126
5.3.2	Αποτελέσματα K-Prototypes στους συμπτωματικούς ασθενείς	129
5.3.3	Σύγκριση K-Modes & K-Prototypes στους συμπτωματικούς ασθενείς	132
5.4	Ανάλυση μορφολογικών προτύπων στο σύνολο των ασυμπτωματικών ασθενών	134
5.4.1	Αποτελέσματα K-Modes στους ασυμπτωματικούς ασθενείς	134
5.4.2	Αποτελέσματα K-Prototypes στους ασυμπτωματικούς ασθενείς	137
5.4.3	Σύγκριση K-Modes & K-Prototypes στους ασυμπτωματικούς ασθενείς	140
5.5	Σύγκριση μεταξύ συμπτωματικών και ασυμπτωματικών ασθενών	142
5.6	Ερμηνεία της επικάλυψης των συστάδων	145
5.7	Σύνδεση με τη βιβλιογραφία	147

5.1 Συνολική ερμηνεία της ανάλυσης συστάδων

Η εφαρμογή των αλγορίθμων K-Modes και K-Prototypes στο σύνολο των ασθενών, καθώς και στις επιμέρους ομάδες συμπτωματικών και ασυμπτωματικών ασθενών, ανέδειξε την παρουσία διαφορετικών μορφολογικών προτύπων της αθηρωματικής πλάκας. Παρόλο που στις προβολές PCA παρατηρήθηκε σε αρκετές περιπτώσεις επικάλυψη μεταξύ των συστάδων, τα αποτελέσματα έδειξαν ότι ορισμένα χαρακτηριστικά της πλάκας συμβάλλουν σταθερά στον διαχωρισμό των ασθενών σε διαφορετικές ομάδες.

Τα χαρακτηριστικά που φάνηκαν να επηρεάζουν περισσότερο τη διαδικασία ομαδοποίησης ήταν ο δείκτης Juxtaluminal Black Area (JBA), η ηχογένεια της πλάκας (GSM), η επιφάνεια της πλάκας (PLAREA) και ο βαθμός στένωσης (STENOSIS). Οι κανόνες που εξήχθησαν μέσω της μεθόδου TE2Rules έδειξαν ότι οι συστάδες που εμφάνιζαν χαρακτηριστικά συμβατά με αυξημένο μορφολογικό κίνδυνο συσχετίζονταν συστηματικά με πολύ υψηλές τιμές JBA, χαμηλές τιμές GSM και αυξημένες τιμές PLAREA. Ωστόσο, οι συστάδες χαμηλότερου κινδύνου εμφάνιζαν συχνότερα χαμηλότερες τιμές JBA και υψηλότερες τιμές GSM.

Η σύγκριση των δύο αλγορίθμων έδειξε ότι ο K-Prototypes παρουσίασε συνολικά καλύτερη απόδοση σε σχέση με τον K-Modes, τόσο ως προς τις μετρικές ποιότητας clustering, όσο και ως προς τη σταθερότητα των αποτελεσμάτων. Η δυνατότητα ταυτόχρονης αξιοποίησης αριθμητικών και κατηγορικών χαρακτηριστικών επέτρεψε την ανάδειξη περισσότερο διαφοροποιημένων και συνεκτικών μορφολογικών προτύπων, οδηγώντας σε πιο σταθερές συστάδες και σε κανόνες TE2Rules με μεγαλύτερη πολυπλοκότητα και κλινική ερμηνευσιμότητα. Σε αντίθεση με τον K-Modes, όπου οι κανόνες βασίζονταν πιο συχνά σε μεμονωμένα χαρακτηριστικά, οι κανόνες του K-Prototypes περιλάμβαναν πιο σύνθετους συνδυασμούς μεταβλητών, προσφέροντας πιο ολοκληρωμένη περιγραφή των μορφολογικών προτύπων των ασθενών. Το εύρημα αυτό ήταν ιδιαίτερα εμφανές στο σύνολο των ασυμπτωματικών ασθενών, όπου η χρήση αποκλειστικά κατηγορικών δεδομένων οδήγησε σε χαμηλότερο βαθμό διαχωρισμού και σταθερότητας των συστάδων.

Παράλληλα, η εφαρμογή της μεθόδου TE2Rules συνέβαλε ουσιαστικά στην ερμηνεία των αποτελεσμάτων της ομαδοποίησης. Οι κανόνες που προέκυψαν ήταν σε μεγάλο βαθμό συνεπείς μεταξύ των διαφορετικών αναλύσεων και ανέδειξαν επαναλαμβανόμενους συνδυασμούς χαρακτηριστικών που σχετίζονται με κάθε συστάδα. Παρότι οι αναλύσεις ανέδειξαν σαφείς τάσεις διαφοροποίησης μεταξύ των συστάδων, σε αρκετές περιπτώσεις παρατηρήθηκε επικάλυψη μεταξύ των clusters στις προβολές PCA. Το εύρημα αυτό υποδηλώνει ότι τα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας δεν διαχωρίζονται σε απόλυτα διακριτές κατηγορίες, αλλά παρουσιάζουν περισσότερο συνεχές φάσμα μεταβολών. Επομένως, ασθενείς με ενδιάμεσα μορφολογικά χαρακτηριστικά ενδέχεται να εμφανίζουν κοινά στοιχεία μεταξύ διαφορετικών συστάδων, γεγονός που αντανακλά την πολυπαραγοντική και ετερογενή φύση της καρωτιδικής νόσου.

Συμπερασματικά, τα αποτελέσματα της ανάλυσης έδειξαν ότι και οι δύο αλγόριθμοι ήταν σε θέση να αναγνωρίσουν διακριτά μορφολογικά πρότυπα της αθηρωματικής πλάκας. Ωστόσο, ο αλγόριθμος K-Prototypes παρουσίασε συνολικά καλύτερη απόδοση ως προς την ποιότητα και τη σταθερότητα της ομαδοποίησης. Συγχρόνως, οδήγησε σε περισσότερο εκφραστικούς και κλινικά συνεκτικούς κανόνες μέσω της μεθόδου TE2Rules. Η αξιοποίηση τόσο αριθμητικών όσο και κατηγορικών χαρακτηριστικών φαίνεται να προσφέρει πιο ολοκληρωμένη αναπαράσταση της μορφολογικής ετερογένειας των ασθενών. Ως εκ τούτου, ενισχύεται η δυνατότητα ερμηνείας των αποτελεσμάτων και η ανάδειξη προτύπων που σχετίζονται με αυξημένο μορφολογικό κίνδυνο.

5.2 Ανάλυση μορφολογικών προτύπων στο σύνολο των ασθενών

5.2.1 Αποτελέσματα K-Modes στο σύνολο των ασθενών

Η εφαρμογή του αλγορίθμου K-Modes στο σύνολο των ασθενών οδήγησε στον σχηματισμό δύο συστάδων ($k=2$), οι οποίες παρουσίασαν διαφορετικά μορφολογικά χαρακτηριστικά ως προς τη δομή της αθηρωματικής πλάκας. Η επιλογή των δύο clusters βασίστηκε τόσο στην καμπύλη elbow (Σχήμα 4.1.1), όσο και στις μετρικές αξιολόγησης του clustering (Πίνακας 4.1.1), όπου παρατηρήθηκε ικανοποιητική ισορροπία μεταξύ ποιότητας ομαδοποίησης και απλότητας του μοντέλου.

Η ανάλυση PCA έδειξε ότι οι δύο πρώτες κύριες συνιστώσες εξηγούσαν συνολικά περίπου το 56% της συνολικής διακύμανσης των δεδομένων (Σχήμα 4.1.3), με την πρώτη συνιστώσα να εξηγεί περίπου το 31% και τη δεύτερη περίπου το 25%. Η δισδιάστατη προβολή των clusters (Σχήμα 4.1.4) έδειξε ότι, παρόλο που υπήρχε μερική επικάλυψη μεταξύ των ομάδων, διακρίνονταν διαφορετικές περιοχές συγκέντρωσης των ασθενών, γεγονός που υποδηλώνει την ύπαρξη διαφορετικών μορφολογικών προτύπων της πλάκας.

Η αξιολόγηση της ποιότητας της ομαδοποίησης έδωσε τιμή Silhouette ίση με 0.30, Davies–Bouldin ίση με 1.97 και Calinski–Harabasz ίση με 187.7 (Πίνακας 4.1.1). Οι τιμές αυτές υποδηλώνουν μέτριο, αλλά αποδεκτό βαθμό διαχωρισμού μεταξύ των συστάδων, κάτι που θεωρείται αναμενόμενο λόγω της ετερογένειας των ιατρικών δεδομένων και της χρήσης αποκλειστικά κατηγορικών χαρακτηριστικών. Επιπρόσθετα, η ανάλυση σταθερότητας έδειξε μέση τιμή Adjusted Rand Index (ARI) ίση με 0.69 (Πίνακας 4.1.2), γεγονός που δείχνει σχετικά σταθερή συμπεριφορά του αλγορίθμου σε επαναλαμβανόμενες εκτελέσεις.

Η κατανομή των ασθενών μεταξύ των δύο clusters δεν ήταν ισομερής (Σχήμα 4.1.5). Το Cluster 0 περιλάμβανε περίπου το 75% των ασθενών του συνόλου, ενώ το Cluster 1 αντιστοιχούσε περίπου στο 25% των ασθενών. Αφενός, το εύρημα αυτό υποδηλώνει ότι η πλειονότητα των ασθενών εμφάνιζε κοινά μορφολογικά χαρακτηριστικά, αφετέρου ένα μικρότερο υποσύνολο παρουσίαζε περισσότερο επιβαρυνμένο μορφολογικό προφίλ.

Αναλυτικότερα, το Cluster 0 χαρακτηριζόταν κυρίως από χαμηλές τιμές JBA, υψηλότερες τιμές GSM και χαμηλές ή μέτριες τιμές PLAREA (Σχήματα 4.1.8 - 4.1.10). Επιπλέον, παρατηρήθηκε μεγαλύτερη παρουσία ασθενών χωρίς CTIASTR (Σχήμα 4.1.6), γεγονός που παραπέμπει σε σχετικά σταθερότερη μορφολογία αθηρωματικής πλάκας. Οι κανόνες που εξήχθησαν μέσω της μεθόδου TE2Rules επιβεβαίωσαν το συγκεκριμένο πρότυπο (Πίνακας 4.1.3), καθώς οι πιο αντιπροσωπευτικοί κανόνες του Cluster 0 ήταν οι «JBA = low», «GSM = high» και «CTIASTR = absent AND PLAREA = low».

Όσον αφορά στο Cluster 1 παρουσιάζεται σημαντικά μεγαλύτερη παρουσία πολύ υψηλών τιμών JBA, χαμηλών τιμών GSM και αυξημένων τιμών PLAREA (Σχήματα 4.1.8-4.1.10). Τα χαρακτηριστικά αυτά σχετίζονται με περισσότερο ευάλωτη μορφολογία πλάκας και αυξημένο μορφολογικό κίνδυνο. Οι κανόνες TE2Rules που περιγράφουν το συγκεκριμένο cluster (Πίνακας 4.1.3) περιλαμβάνει κυρίως τους κανόνες «JBA = very high», «GSM $\leq$ medium AND JBA = medium ή high AND PLAREA = medium» και «GSM = low AND PLAREA $\geq$ medium». Τα αποτελέσματα αυτά δείχνουν ότι, ο συνδυασμός χαμηλής ηχογένειας και αυξημένου JBA, αποτέλεσε βασικό στοιχείο διαφοροποίησης των δύο ομάδων.

Ιδιαίτερα έντονη ήταν η διαφοροποίηση των clusters ως προς το χαρακτηριστικό JBA (Σχήμα 4.1.9). Το Cluster 1 συγκέντρωσε τη μεγάλη πλειονότητα των ασθενών με πολύ υψηλές τιμές JBA, ενώ στο Cluster 0 κυριάρχησαν χαμηλές τιμές JBA. Αντίστοιχη εικόνα παρατηρήθηκε και στο GSM (Σχήμα 4.1.8), όπου το Cluster 0 εμφάνιζε κυρίως υψηλές τιμές ηχογένειας, ενώ το Cluster 1 χαρακτηριζόταν συχνότερα από χαμηλές τιμές GSM.

Συνολικά, τα αποτελέσματα του K-Modes στο σύνολο των ασθενών έδειξαν ότι, ο αλγόριθμος ήταν σε θέση να διαχωρίσει δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας. Το πρώτο αντιστοιχούσε σε ασθενείς με πιο σταθερά μορφολογικά χαρακτηριστικά. Το δεύτερο περιλάμβανε ασθενείς με περισσότερο επιβαρυσμένο μορφολογικό προφίλ, το οποίο χαρακτηριζόταν κυρίως από πολύ υψηλό JBA, χαμηλό GSM και αυξημένη επιφάνεια πλάκας.

5.2.2 Αποτελέσματα K-Prototypes στο σύνολο των ασθενών

Η εφαρμογή του αλγορίθμου K-Prototypes στο σύνολο των ασθενών οδήγησε επίσης στον σχηματισμό δύο συστάδων ($k=2$), οι οποίες παρουσίασαν σαφή διαφοροποίηση, ως προς τα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας. Η επιλογή των δύο clusters βασίστηκε τόσο στην καμπύλη elbow (Σχήμα 4.1.12), όσο και στην αξιολόγηση των μετρικών ποιότητας της ομαδοποίησης (Πίνακας 4.1.4), όπου παρατηρήθηκε καλύτερος διαχωρισμός των ομάδων σε σύγκριση με την αντίστοιχη ανάλυση K-Modes.

Η καμπύλη elbow (Σχήμα 4.1.12) κατέδειξε σημαντική μείωση του κόστους από $k=2$ έως $k=4$. Μετά το $k=2$ η μείωση του κόστους γινόταν περισσότερο σταδιακή. Το αποτέλεσμα αυτό, σε συνδυασμό με τις μετρικές αξιολόγησης, υποστήριξε την επιλογή δύο clusters, ως κατάλληλου συμβιβασμού μεταξύ απόδοσης και πολυπλοκότητας του μοντέλου.

Η ανάλυση PCA έδειξε ότι η πρώτη κύρια συνιστώσα εξηγούσε περίπου το 54% της συνολικής διακύμανσης, ενώ η δεύτερη περίπου το 13%, με συνολικό ποσοστό εξηγούμενης διακύμανσης περίπου 67% (Σχήμα 4.1.14). Η προβολή των ασθενών στον διδιάστατο χώρο PCA (Σχήμα 4.1.15) παρουσίασε πιο καθαρό διαχωρισμό των clusters σε σχέση με το K-Modes, ιδιαίτερα κατά μήκος της πρώτης κύριας συνιστώσας. Παρόλο που, υπήρχε περιορισμένη επικάλυψη μεταξύ των ομάδων, οι περισσότερες παρατηρήσεις του Cluster 1 εντοπίζονταν σε διακριτή περιοχή του χώρου χαρακτηριστικών, γεγονός που υποδηλώνει εντονότερη διαφοροποίηση του συγκεκριμένου υποσυνόλου ασθενών.

Η αξιολόγηση της ποιότητας της ομαδοποίησης έδωσε τιμή Silhouette ίση με 0.46, Davies-Bouldin ίση με 1.00 και Calinski-Harabasz ίση με 535.87 (Πίνακας 4.1.4). Οι τιμές αυτές υποδηλώνουν σημαντικά καλύτερο βαθμό διαχωρισμού και συνοχής των clusters σε σχέση με το K-Modes. Η υψηλότερη τιμή Silhouette και η χαμηλότερη τιμή Davies-Bouldin δείχνουν ότι οι ασθενείς κάθε cluster εμφάνιζαν μεγαλύτερη ομοιότητα μεταξύ τους και σαφέστερο διαχωρισμό από τις υπόλοιπες ομάδες.

Ταυτόχρονα, η ανάλυση σταθερότητας έδειξε εξαιρετικά υψηλή επανάληψη του μοντέλου, με μέση τιμή Adjusted Rand Index (ARI) ίση με 0.99 και εύρος τιμών από 0.99 έως 1.00 μεταξύ των επαναλαμβανόμενων εκτελέσεων (Πίνακας 4.1.5). Το αποτέλεσμα αυτό υποδηλώνει ιδιαίτερα σταθερή συμπεριφορά του K-Prototypes, ενισχύοντας έτσι την αξιοπιστία της ομαδοποίησης.

Η κατανομή των ασθενών μεταξύ των δύο clusters δεν ήταν ισομερής (Σχήμα 4.1.16) αφού όπως παρατηρήθηκε το Cluster 0 περιλαμβάνει τη μεγάλη πλειονότητα των ασθενών (περίπου 85%), ενώ το Cluster 1 αποτελεί μικρότερο υποσύνολο ασθενών (περίπου 15%) με περισσότερο επιβαρυσμένα μορφολογικά χαρακτηριστικά.

Αναλυτικότερα, το Cluster 0 χαρακτηριζόταν κυρίως από χαμηλές ή μέτριες τιμές JBA, υψηλές τιμές GSM και χαμηλές έως μέτριες τιμές PLAREA (Σχήματα 4.1.19-4.1.21). Επιπλέον, οι περισσότεροι ασθενείς της συστάδας δεν εμφάνιζαν CTIASTR, ενώ παρατηρήθηκε αυξημένη παρουσία DWA (Σχήματα 4.1.17 και 4.1.18). Οι κανόνες TE2Rules επιβεβαίωσαν το συγκεκριμένο πρότυπο (Πίνακας 4.1.6), καθώς οι πιο αντιπροσωπευτικοί κανόνες του Cluster 0 ήταν οι «JBA $\leq$ high», «CTIASTR = absent AND DWA = present AND PLAREA $\leq$ medium», «PLAREA = low» και «GSM $\geq$ medium». Συνολικά, το Cluster 0 φαίνεται να αντιστοιχεί σε ασθενείς με σχετικά σταθερότερη μορφολογία αθηρωματικής πλάκας.

Ωστόσο, το Cluster 1 παρουσίαζε ιδιαίτερα υψηλή συγκέντρωση ασθενών με πολύ υψηλές τιμές JBA, χαμηλές τιμές GSM και αυξημένες τιμές PLAREA (Σχήματα 4.1.19–4.1.21). Επιπλέον, στο συγκεκριμένο cluster παρατηρήθηκε μικρότερη παρουσία DWA και μεγαλύτερη συγκέντρωση ασθενών με μορφολογικά χαρακτηριστικά που σχετίζονται με αυξημένη ευαλωτότητα της πλάκας. Οι κανόνες TE2Rules που περιέγραφαν τη συστάδα (Πίνακας 4.1.6) περιλάμβαναν κυρίως τους «DWA = absent AND JBA = very high», «GSM $\leq$ medium AND JBA = very high AND PLAREA = high», «GSM = low AND JBA = very high» και «JBA = very high AND PLAREA = high».

Ιδιαίτερα έντονη διαφοροποίηση παρατηρήθηκε στο χαρακτηριστικό JBA (Σχήμα 4.1.20). Σχεδόν όλοι οι ασθενείς με πολύ υψηλό JBA συγκεντρώνονταν στο Cluster 1, ενώ το Cluster 0 κυμαίνονταν από χαμηλές και μέτριες τιμές JBA. Αντίστοιχα, το GSM διαφοροποιούσε σημαντικά τις δύο ομάδες (Σχήμα 4.1.19), καθώς το Cluster 0 εμφάνιζε κυρίως υψηλές τιμές ηχογένειας, ενώ το Cluster 1 χαρακτηριζόταν κυρίως από χαμηλό GSM. Παρόμοια εικόνα παρατηρήθηκε και στο PLAREA (Σχήμα 4.1.21), όπου το Cluster 1 εμφάνιζε αυξημένη παρουσία μεσαίων και υψηλών τιμών επιφάνειας πλάκας.

Συνοψίζοντας, τα αποτελέσματα του K-Prototypes έδειξαν ότι ο αλγόριθμος ήταν σε θέση να διαχωρίσει αποτελεσματικά δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας στο σύνολο των ασθενών. Το πρώτο πρότυπο αντιστοιχούσε σε ασθενείς με πιο σταθερή μορφολογία και υψηλότερη ηχογένεια της πλάκας. Το δεύτερο αντιπροσώπευε μικρότερο αλλά περισσότερο επιβαρυνμένο υποσύνολο ασθενών, το οποίο χαρακτηριζόταν κυρίως από πολύ υψηλό JBA, χαμηλό GSM και αυξημένη επιφάνεια πλάκας.

5.2.3 Σύγκριση K-Modes & K-Prototypes στο σύνολο ασθενών

Η σύγκριση των αλγορίθμων K-Modes και K-Prototypes στο σύνολο των ασθενών έδειξε ότι και οι δύο μέθοδοι ήταν ικανές να αναγνωρίσουν δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας. Παρόλα αυτά, ο K-Prototypes εμφάνισε συνολικά καλύτερη εικόνα ως προς τον διαχωρισμό των συστάδων και τη σταθερότητα των αποτελεσμάτων.

Υπενθυμίζεται ότι και στις δύο περιπτώσεις επιλέχθηκαν δύο συστάδες ($k=2$), γεγονός που δείχνει ότι τα δεδομένα οργανώνονταν κυρίως γύρω από δύο βασικά μορφολογικά προφίλ ασθενών. Επιπλέον, τόσο στον K-Modes, όσο και στον K-Prototypes παρατηρήθηκε μία μεγάλη συστάδα που περιλάμβανε την πλειονότητα των ασθενών και μία μικρότερη συστάδα με περισσότερο επιβαρυσμένα χαρακτηριστικά αθηρωματικής πλάκας.

Ωστόσο, ο διαχωρισμός των clusters ήταν σαφέστερος στην περίπτωση του K-Prototypes. Η τιμή Silhouette αυξήθηκε από 0.30 στον K-Modes σε 0.46 στον K-Prototypes, εντούτοις η τιμή Davies-Bouldin μειώθηκε αισθητά από 1.97 σε 1.00. Την ίδια στιγμή, ο δείκτης Calinski-Harabasz παρουσίασε σημαντική αύξηση, από 187.7 σε 535.87. Οι διαφορές αυτές δείχνουν ότι ο K-Prototypes δημιούργησε πιο συνεκτικές και καλύτερα διαχωρισμένες ομάδες ασθενών.

Αντίστοιχη εικόνα παρατηρήθηκε και ως προς τη σταθερότητα των αποτελεσμάτων. Από τη μία ο K-Modes εμφάνισε μέση τιμή ARI ίση με 0.69, γεγονός που υποδηλώνει σχετικά καλή αλλά όχι απόλυτα σταθερή συμπεριφορά μεταξύ επαναλαμβανόμενων εκτελέσεων. Από την άλλη, ο K-Prototypes παρουσίασε εξαιρετικά υψηλή σταθερότητα, με μέση τιμή ARI ίση με 0.99 και ελάχιστες διαφοροποιήσεις μεταξύ των runs. Το αποτέλεσμα αυτό δείχνει ότι η ταυτόχρονη αξιοποίηση αριθμητικών και κατηγορικών χαρακτηριστικών συνέβαλε σημαντικά στη βελτίωση της συνέπειας της ομαδοποίησης.

Η διαφορά αυτή φαίνεται να σχετίζεται με τον τρόπο λειτουργίας των δύο αλγορίθμων. Ο K-Modes χρησιμοποιεί αποκλειστικά κατηγορικές μεταβλητές, γεγονός που περιορίζει την ποσότητα πληροφορίας που αξιοποιείται για τον διαχωρισμό των ασθενών. Όμως η δυνατότητα του K-Prototypes να συνδυάζει αριθμητικά και κατηγορικά δεδομένα, επιτρέπει πιο λεπτομερή αποτύπωση των διαφορών μεταξύ των μορφολογικών χαρακτηριστικών της πλάκας.

Παρά τις διαφορές στην απόδοση, και οι δύο μέθοδοι ανέδειξαν παρόμοια μορφολογικά πρότυπα. Τα χαρακτηριστικά που φάνηκαν να συμβάλλουν περισσότερο στον διαχωρισμό των clusters ήταν κυρίως το JBA, το GSM και το PLAREA. Συγκεκριμένα, οι συστάδες υψηλότερου μορφολογικού κινδύνου χαρακτηρίζονταν σταθερά από πολύ υψηλό JBA, χαμηλό GSM και αυξημένη επιφάνεια πλάκας. Επίσης, οι συστάδες χαμηλότερου κινδύνου εμφάνιζαν συχνότερα χαμηλότερες τιμές JBA και υψηλότερη ηχογένεια.

Παρόμοια εικόνα προέκυψε και από τους κανόνες TE2Rules, καθώς και στις δύο προσεγγίσεις εμφανίστηκαν επαναλαμβανόμενοι συνδυασμοί χαρακτηριστικών, που περιέγραφαν τις ίδιες μορφολογικές τάσεις. Έτσι, ενισχύεται η αξιοπιστία των αποτελεσμάτων. Ωστόσο, στον αλγόριθμο K-Prototypes οι κανόνες ήταν συνήθως πιο σύνθετοι και πιο σαφείς ως προς τη διαφοροποίηση των clusters, αντανakλώντας τον καλύτερο διαχωρισμό που παρατηρήθηκε και στις μετρικές αξιολόγησης.

Συνολικά, τα αποτελέσματα δείχνουν ότι ο K-Prototypes υπερείχε του K-Modes στο σύνολο των ασθενών, παρουσιάζοντας καλύτερο διαχωρισμό, υψηλότερη σταθερότητα και πιο συνεκτικά μορφολογικά πρότυπα. Παρόλα αυτά, και οι δύο αλγόριθμοι κατέληξαν σε παρόμοια κλινικά συμπεράσματα ως προς τα βασικά χαρακτηριστικά που σχετίζονται με αυξημένο μορφολογικό κίνδυνο της αθηρωματικής πλάκας.

5.3 Ανάλυση μορφολογικών προτύπων στο σύνολο των συμπτωματικών ασθενών

5.3.1 Αποτελέσματα K-Modes στους συμπτωματικούς ασθενείς

Η εφαρμογή του αλγορίθμου K-Modes στο σύνολο των συμπτωματικών ασθενών οδήγησε στον σχηματισμό δύο συστάδων ($k=2$), οι οποίες παρουσίασαν διακριτές διαφοροποιήσεις ως προς τα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας. Η επιλογή του αριθμού των clusters βασίστηκε τόσο στην καμπύλη elbow (Σχήμα 4.2.1), όσο και στις μετρικές αξιολόγησης clustering (Σχήμα 4.2.2). Παρατηρήθηκε ότι, η επιλογή των δύο clusters παρείχε τον καλύτερο συμβιβασμό μεταξύ απλότητας του μοντέλου και ποιότητας ομαδοποίησης.

Η καμπύλη elbow έδειξε σημαντική μείωση του κόστους από $k=2$ έως $k=4$, ενώ για μεγαλύτερες τιμές του k η μείωση γινόταν περισσότερο σταδιακή. Παράλληλα, οι δείκτες αξιολόγησης clustering παρουσίασαν σχετικά σταθερή συμπεριφορά για μικρές τιμές του k , υποστηρίζοντας την επιλογή των δύο clusters ως κατάλληλη για την περιγραφή της δομής των δεδομένων.

Η ανάλυση PCA έδειξε ότι οι δύο πρώτες κύριες συνιστώσες εξηγούσαν συνολικά περίπου το 55% της συνολικής διακύμανσης των δεδομένων, με την πρώτη κύρια συνιστώσα να εξηγεί περίπου το 31% και τη δεύτερη περίπου το 24% (Σχήμα 4.2.3). Η προβολή PCA των ασθενών (Σχήμα 4.2.4) έδειξε ότι, παρόλο που υπήρχε μερική επικάλυψη μεταξύ των clusters, διακρίνονταν διαφορετικές περιοχές συγκέντρωσης των ασθενών, γεγονός που υποδηλώνει την ύπαρξη διαφορετικών μορφολογικών προτύπων αθηρωματικής πλάκας στους συμπτωματικούς ασθενείς.

Η αξιολόγηση της ποιότητας της ομαδοποίησης έδωσε τιμή Silhouette ίση με 0.21, Davies-Bouldin ίση με 3.29 και Calinski-Harabasz ίση με 8.28 (Πίνακας 4.2.1). Οι τιμές αυτές υποδηλώνουν σχετικά μέτριο βαθμό διαχωρισμού μεταξύ των clusters, κάτι που θεωρείται αναμενόμενο λόγω της υψηλής ετερογένειας που παρουσιάζουν οι συμπτωματικοί ασθενείς και της χρήσης αποκλειστικά κατηγορικών χαρακτηριστικών στην ομαδοποίηση.

Παράλληλα, η ανάλυση σταθερότητας έδειξε μέση τιμή Adjusted Rand Index (ARI) ίση με 0.79, με εύρος τιμών από 0.37 έως 1.00 μεταξύ των επαναλαμβανόμενων εκτελέσεων του αλγορίθμου (Πίνακας 4.2.2). Το αποτέλεσμα αυτό δείχνει σχετικά καλή επανάληψη της ομαδοποίησης, αν και παρατηρήθηκαν διαφοροποιήσεις μεταξύ ορισμένων runs.

Η κατανομή των ασθενών μεταξύ των δύο clusters δεν ήταν ισομερής (Σχήμα 4.2.5). Το Cluster 0 περιλάμβανε περίπου το 68% των συμπτωματικών ασθενών, ενώ το Cluster 1 αντιστοιχούσε περίπου στο 32% των ασθενών. Το εύρημα αυτό υποδηλώνει ότι η πλειονότητα των συμπτωματικών ασθενών παρουσίαζε κοινά μορφολογικά χαρακτηριστικά, αλλά ένα μικρότερο υποσύνολο εμφάνιζε διαφορετικό και περισσότερο επιβαρυνόμενο μορφολογικό προφίλ.

Αναλυτικότερα, το Cluster 0 χαρακτηριζόταν από μεγαλύτερη παρουσία χαμηλών τιμών GSM, μέτριων τιμών στένωσης και πολύ υψηλών τιμών JBA (Σχήματα 4.2.8, 4.2.9 και 4.2.11). Ακόμα, στο συγκεκριμένο cluster παρατηρήθηκε αυξημένη παρουσία DWA (Σχήμα 4.2.7) και οι τιμές PLAREA κατανέμονταν κυρίως στις κατηγορίες medium και high (Σχήμα 4.2.10). Οι κανόνες TE2Rules επιβεβαίωσαν το συγκεκριμένο πρότυπο, καθώς οι πιο αντιπροσωπευτικοί κανόνες της συστάδας περιλάμβαναν τους «GSM = low», «JBA $\geq$ high AND STENOSIS $\leq$ medium» και «GSM = high AND JBA = medium ή high AND STENOSIS = medium». Συνολικά, το Cluster 0 φαίνεται να αντιστοιχεί σε ασθενείς με περισσότερο επιβαρυνμένη μορφολογία πλάκας και αυξημένα χαρακτηριστικά ευαλωτότητας.

Εντούτοις, το Cluster 1 παρουσίαζε μεγαλύτερη συγκέντρωση ασθενών με medium ή high τιμές GSM, υψηλές τιμές STENOSIS και κυρίως μέτριες ή υψηλές τιμές JBA (Σχήματα 4.2.8, 4.2.9 και 4.2.11). Επιπρόσθετα, παρατηρήθηκε μικρότερη παρουσία πολύ υψηλών τιμών JBA σε σχέση με το Cluster 0, ενώ οι ασθενείς του συγκεκριμένου cluster εμφάνιζαν επίσης αυξημένες τιμές PLAREA. Οι κανόνες TE2Rules που περιέγραφαν τη συστάδα περιλάμβαναν κυρίως τους «GSM = medium ή high AND JBA = medium ή high», «GSM = medium AND STENOSIS = high» και «JBA $\leq$ high AND STENOSIS = high». Το Cluster 1 φαίνεται να αντιπροσωπεύει διαφορετικό μορφολογικό πρότυπο συμπτωματικών ασθενών, στο οποίο κυριαρχεί η υψηλή στένωση σε συνδυασμό με περισσότερο ενδιάμεσες τιμές JBA και υψηλότερη ηχογένεια συγκριτικά με το Cluster 0.

Ιδιαίτερα έντονη διαφοροποίηση μεταξύ των δύο clusters παρατηρήθηκε στα χαρακτηριστικά GSM και JBA. Το Cluster 0 συγκέντρωνε σημαντικά περισσότερους ασθενείς με χαμηλό GSM και πολύ υψηλό JBA, χαρακτηριστικά που σχετίζονται με περισσότερο εύαλωτη μορφολογία πλάκας. Αντίθετα, το Cluster 1 εμφάνιζε συχνότερα medium ή high τιμές GSM και μικρότερη συγκέντρωση very high JBA, γεγονός που υποδηλώνει διαφορετικό μορφολογικό προφίλ.

Συνοπτικά, τα αποτελέσματα του K-Modes στους συμπτωματικούς ασθενείς έδειξαν ότι ο αλγόριθμος ήταν σε θέση να αναγνωρίσει δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας. Το πρώτο πρότυπο αντιστοιχούσε σε ασθενείς με περισσότερο επιβαρυσμένα χαρακτηριστικά ευαλωτότητας, κυρίως λόγω χαμηλού GSM και πολύ υψηλού JBA, ενώ το δεύτερο περιλάμβανε ασθενείς με υψηλότερη στένωση και περισσότερο ενδιάμεσες μορφολογικές τιμές.

5.3.2 Αποτελέσματα K-Prototypes στους συμπτωματικούς ασθενείς

Η εφαρμογή του αλγορίθμου K-Prototypes στο σύνολο των συμπτωματικών ασθενών οδήγησε στον σχηματισμό δύο συστάδων ($k=2$), οι οποίες παρουσίασαν διακριτές διαφοροποιήσεις ως προς τα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας. Η επιλογή του αριθμού των clusters βασίστηκε τόσο στην καμπύλη elbow (Σχήμα 4.2.12) όσο και στις μετρικές αξιολόγησης clustering (Σχήμα 4.2.13), όπου διαπιστώθηκε ότι, η επιλογή των δύο clusters παρείχε την καλύτερη ισορροπία μεταξύ ποιότητας ομαδοποίησης και ερμηνευσιμότητας των αποτελεσμάτων.

Η καμπύλη elbow έδειξε σημαντική μείωση του κόστους από $k=2$ έως $k=5$, ενώ για μεγαλύτερες τιμές του k η μείωση γινόταν πιο σταδιακή. Παράλληλα, οι δείκτες αξιολόγησης ομαδοποίησης παρουσίασαν σχετικά σταθερή συμπεριφορά για μικρές τιμές του k , υποστηρίζοντας την επιλογή των δύο clusters ως κατάλληλη για την περιγραφή της δομής των δεδομένων.

Επίσης, η ανάλυση PCA έδειξε ότι οι δύο πρώτες κύριες συνιστώσες εξηγούσαν συνολικά περίπου το 55% της συνολικής διακύμανσης των δεδομένων, με την πρώτη κύρια συνιστώσα να εξηγεί περίπου το 35% και τη δεύτερη περίπου το 20% (Σχήμα 4.2.14). Η προβολή PCA των ασθενών (Σχήμα 4.2.15) έδειξε ότι, παρά τη μερική επικάλυψη μεταξύ των clusters, παρατηρούνταν διαφορετικές περιοχές συγκέντρωσης των ασθενών στον χώρο χαρακτηριστικών, στοιχείο που υποδηλώνει την ύπαρξη διαφορετικών μορφολογικών προτύπων αθηρωματικής πλάκας στους συμπτωματικούς ασθενείς.

Η αξιολόγηση της ποιότητας της ομαδοποίησης έδωσε τιμή Silhouette ίση με 0.15, Davies-Bouldin ίση με 2.03 και Calinski-Harabasz ίση με 22.23 (Πίνακας 4.2.4). Οι τιμές αυτές υποδηλώνουν μέτριο βαθμό διαχωρισμού μεταξύ των clusters. Ωστόσο, η ποιότητα της ομαδοποίησης θεωρείται ικανοποιητική δεδομένης της υψηλής ετερογένειας των συμπτωματικών ασθενών και της ταυτόχρονης χρήσης αριθμητικών και κατηγορικών χαρακτηριστικών.

Συγχρόνως, η ανάλυση σταθερότητας έδειξε πολύ υψηλή μέση τιμή Adjusted Rand Index (ARI) ίση με 0.99, με εύρος τιμών από 0.93 έως 1.00 μεταξύ των επαναλαμβανόμενων εκτελέσεων του αλγορίθμου (Πίνακας 4.2.5). Το αποτέλεσμα αυτό δείχνει εξαιρετικά υψηλή σταθερότητα και επαναληψιμότητα της ομαδοποίησης.

Η κατανομή των ασθενών μεταξύ των δύο clusters ήταν σχεδόν ισομερής (Σχήμα 4.2.16). Το Cluster 0 περιλάμβανε 58 ασθενείς, ενώ το Cluster 1 περιλάμβανε 56 ασθενείς, γεγονός που δείχνει ότι η ομαδοποίηση δεν επηρεάστηκε από σημαντική ανισορροπία στο μέγεθος των συστάδων.

Η κατανομή της μεταβλητής CTIASTR παρουσίασε σχετικά παρόμοια εικόνα και στα δύο clusters (Σχήμα 4.2.17), χωρίς να παρατηρείται έντονη διαφοροποίηση μεταξύ των συστάδων. Αντίθετα, στη μεταβλητή DWA (Σχήμα 4.2.18) παρατηρήθηκε μεγαλύτερη παρουσία της κατηγορίας “present” στο Cluster 0, ενώ στο Cluster 1 εμφανίστηκε αυξημένη συχνότητα της κατηγορίας “absent”.

Ιδιαίτερα έντονη διαφοροποίηση παρατηρήθηκε στη μεταβλητή GSM (Σχήμα 4.2.19). Το Cluster 0 χαρακτηριζόταν κυρίως από medium και high τιμές GSM, ενώ το Cluster 1 παρουσίαζε σημαντικά μεγαλύτερη συγκέντρωση ασθενών με low τιμές GSM. Το εύρημα αυτό δείχνει διαφοροποίηση ως προς την ηχογένεια των αθηρωματικών πλακών μεταξύ των δύο ομάδων.

Ακόμη πιο έντονη διαφοροποίηση παρατηρήθηκε στη μεταβλητή JBA (Σχήμα 4.2.20). Το Cluster 1 παρουσίασε έντονη επικράτηση very high τιμών JBA, ενώ το Cluster 0 εμφάνιζε κυρίως low, medium και high τιμές. Το αποτέλεσμα αυτό καταδεικνύει αυξημένη ανομοιογένεια της πλάκας στο Cluster 1 και πιθανώς περισσότερο επιβαρυνμένο μορφολογικό προφίλ.

Η μεταβλητή PLAREA (Σχήμα 4.2.21) παρουσίασε πιο ισορροπημένη κατανομή μεταξύ των δύο clusters, με επικράτηση medium και high τιμών και στις δύο συστάδες. Αντίστοιχα, η μεταβλητή STENOSIS (Σχήμα 4.2.22) εμφάνισε σχετικά παρόμοια κατανομή μεταξύ των clusters, με επικράτηση medium και high τιμών. Το αποτέλεσμα αυτό δείχνει ότι, η στένωση δεν αποτέλεσε τον κυριότερο παράγοντα διαφοροποίησης της ομαδοποίησης.

Οι κανόνες TE2Rules επιβεβαίωσαν τα παραπάνω πρότυπα μορφολογικής διαφοροποίησης. Οι κανόνες του Cluster 0 σχετίζονταν κυρίως με medium ή high τιμές GSM και χαμηλότερες κατηγορίες JBA, ενώ οι κανόνες του Cluster 1 χαρακτηρίζονταν από low GSM και very high JBA.

Συνοψίζοντας, τα αποτελέσματα του K-Prototypes έδειξαν ότι ο αλγόριθμος ήταν σε θέση να αναγνωρίσει δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας στους συμπτωματικούς ασθενείς. Το πρώτο πρότυπο αντιστοιχούσε σε ασθενείς με μεγαλύτερη ηχογένεια και χαμηλότερη ανομοιογένεια πλάκας, ενώ το δεύτερο συνδεόταν με χαμηλό GSM και πολύ υψηλό JBA, χαρακτηριστικά που συνδέονταν με περισσότερο ευάλωτη μορφολογία πλάκας. Παράλληλα, η υψηλή σταθερότητα της ομαδοποίησης και η συμφωνία των κανόνων TE2Rules με τα μορφολογικά χαρακτηριστικά των συστάδων ενισχύουν την αξιοπιστία και την ερμηνευσιμότητα των αποτελεσμάτων της ανάλυσης.

5.3.3 Σύγκριση K-Modes & K-Prototypes στους συμπτωματικούς ασθενείς

Η σύγκριση των αλγορίθμων K-Modes και K-Prototypes στο σύνολο των συμπτωματικών ασθενών έδειξε ότι και οι δύο προσεγγίσεις ήταν σε θέση να αναγνωρίσουν δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας. Παρόλα αυτά, παρατηρήθηκαν σημαντικές διαφοροποιήσεις τόσο στην ποιότητα και σταθερότητα της ομαδοποίησης, όσο και στον τρόπο με τον οποίο περιεγράφηκαν τα μορφολογικά χαρακτηριστικά των ασθενών.

Αρχικά, και οι δύο αλγόριθμοι κατέληξαν στην επιλογή δύο συστάδων ($k=2$), στοιχείο που υποδηλώνει ότι τα δεδομένα των συμπτωματικών ασθενών οργανώνονταν κυρίως γύρω από δύο βασικά μορφολογικά προφίλ. Ωστόσο, η κατανομή των ασθενών στα clusters διαφοροποιήθηκε μεταξύ των δύο μεθόδων. Συγκεκριμένα στον K-Modes παρατηρήθηκε περισσότερο ανισομερές κατανομή, ενώ στον K-Prototypes οι δύο συστάδες ήταν σχεδόν ισομεγέθεις, εύρημα που υποδηλώνει πιο ισορροπημένο διαχωρισμό των δεδομένων.

Οι μετρικές αξιολόγησης έδειξαν, επίσης, διαφορετική συμπεριφορά μεταξύ των δύο αλγορίθμων. Ο K-Modes παρουσίασε ελαφρώς υψηλότερη τιμή Silhouette (0.21 έναντι 0.15), υποδηλώνοντας κάπως καλύτερο τοπικό διαχωρισμό μεταξύ των clusters. Αντίθετα, ο K-Prototypes εμφάνισε χαμηλότερη τιμή Davies-Bouldin (2.03 έναντι 3.29) και υψηλότερη τιμή Calinski-Harabasz (22.23 έναντι 8.28), καταδεικνύοντας καλύτερη συνολική συνοχή και σαφέστερο διαχωρισμό των συστάδων στον πολυδιάστατο χώρο χαρακτηριστικών.

Ακόμη πιο έντονη ήταν η διαφοροποίηση ως προς τη σταθερότητα των αποτελεσμάτων. Αναλυτικότερα, ο K-Modes παρουσίασε μέση τιμή Adjusted Rand Index (ARI) ίση με 0.79, με αισθητές διαφοροποιήσεις μεταξύ των επαναλαμβανόμενων εκτελέσεων του αλγορίθμου. Από την άλλη, ο K-Prototypes εμφάνισε εξαιρετικά υψηλή σταθερότητα, με μέση τιμή ARI ίση με 0.99 και ελάχιστες αποκλίσεις μεταξύ των runs. Το αποτέλεσμα αυτό ενισχύει ότι, η ταυτόχρονη αξιοποίηση αριθμητικών και κατηγορικών χαρακτηριστικών οδήγησε σε πιο συνεπή και επαναλαμβανόμενη ομαδοποίηση.

Παρά τις διαφορές στις μετρικές και στη σταθερότητα, και οι δύο αλγόριθμοι ανέδειξαν παρόμοια μορφολογικά πρότυπα στους συμπτωματικούς ασθενείς. Συγκεκριμένα, τα clusters αυξημένου μορφολογικού κινδύνου χαρακτηρίζονταν σταθερά από χαμηλές τιμές GSM και πολύ υψηλές τιμές JBA, χαρακτηριστικά που σχετίζονται με περισσότερο ευάλωτη μορφολογία αθηρωματικής πλάκας. Το εύρημα αυτό εμφανίστηκε με συνέπεια τόσο στις κατανομές των χαρακτηριστικών, όσο και στους κανόνες TE2Rules.

Ωστόσο, ο K-Prototypes ανέδειξε πιο σύνθετα και περισσότερο διακριτά μορφολογικά πρότυπα. Οι κανόνες TE2Rules του K-Prototypes περιλάμβαναν συχνότερα συνδυασμούς πολλαπλών χαρακτηριστικών, όπως ταυτόχρονη παρουσία χαμηλού GSM, very high JBA, αυξημένου PLAREA και απουσίας DWA. Αντιθέτως, οι κανόνες του K-Modes βασίζονταν κυρίως σε μεμονωμένα χαρακτηριστικά ή σε πιο απλούς συνδυασμούς μεταβλητών.

Συγχρόνως, οι προβολές PCA έδειξαν ότι ο K-Prototypes παρουσίαζε πιο καθαρή οργάνωση των clusters στον χώρο χαρακτηριστικών, παρόλο που και στις δύο περιπτώσεις παρατηρήθηκε μερική επικάλυψη μεταξύ των ομάδων. Το αποτέλεσμα αυτό αναδεικνύει ότι η αξιοποίηση αριθμητικών δεδομένων βοήθησε στην καλύτερη αποτύπωση της μορφολογικής ετερογένειας των συμπτωματικών ασθενών.

Με βάση τα πιο πάνω, τα αποτελέσματα δείχνουν ότι και οι δύο αλγόριθμοι ήταν ικανοί να αναγνωρίσουν βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας στους συμπτωματικούς ασθενείς. Παρόλα αυτά, ο K-Prototypes παρουσίασε συνολικά μεγαλύτερη σταθερότητα, καλύτερο συνολικό διαχωρισμό των clusters και περισσότερο εκφραστικούς κανόνες TE2Rules, γεγονός που ενισχύει την άποψη ότι η συνδυαστική αξιοποίηση αριθμητικών και κατηγορικών χαρακτηριστικών οδηγεί σε πιο ολοκληρωμένη περιγραφή της μορφολογικής ετερογένειας των συμπτωματικών ασθενών.

5.4 Ανάλυση μορφολογικών προτύπων στο σύνολο των ασυμπτωματικών ασθενών

5.4.1 Αποτελέσματα K-Modes στους ασυμπτωματικούς ασθενείς

Η εφαρμογή του αλγορίθμου K-Modes στο σύνολο των ασυμπτωματικών ασθενών οδήγησε στον σχηματισμό δύο συστάδων ($k=2$), οι οποίες παρουσίασαν διακριτές διαφοροποιήσεις ως προς τα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας. Η επιλογή του αριθμού των clusters βασίστηκε τόσο στην καμπύλη elbow (Σχήμα 4.3.1), όσο και στις μετρικές αξιολόγησης clustering (Σχήμα 4.3.2), όπου παρατηρήθηκε ότι η επιλογή δύο συστάδων παρείχε τον καλύτερο συνδυασμό απλότητας και ποιότητας ομαδοποίησης.

Η καμπύλη elbow έδειξε σημαντική μείωση του κόστους για μικρές τιμές του k και ειδικότερα από $k=2$ έως $k=5$, ενώ στη συνέχεια παρατηρήθηκε ότι η μείωση γινόταν πιο ομαλή. Παράλληλα, οι μετρικές αξιολόγησης της ομαδοποίησης παρουσίασαν σχετικά σταθερή συμπεριφορά στις μικρές τιμές του k , στοιχείο που ενίσχυσε την επιλογή των δύο συστάδων ως κατάλληλη για την περιγραφή της δομής των δεδομένων στους ασυμπτωματικούς ασθενείς.

Από την ανάλυση PCA διαφάνηκε ότι, οι δύο πρώτες κύριες συνιστώσες εξηγούσαν συνολικά περίπου το 57% της συνολικής διακύμανσης των δεδομένων, με την πρώτη κύρια συνιστώσα να εξηγεί περίπου το 31% και τη δεύτερη περίπου το 26% (Σχήμα 4.3.3). Η προβολή PCA των ασθενών (Σχήμα 4.3.4) ανέδειξε μερική επικάλυψη μεταξύ των ομάδων, ωστόσο, ήταν εμφανείς διαφορετικές περιοχές συγκέντρωσης των ασθενών στον χώρο χαρακτηριστικών. Επιπλέον, ασθενείς με χρόνο εμφάνισης συμβάντος μικρότερο και μεγαλύτερο των τριών ετών εμφανίζονταν και στις δύο συστάδες, γεγονός που υποδηλώνει ότι ο χρόνος παρακολούθησης δεν αποτέλεσε τον βασικό παράγοντα διαφοροποίησης της ομαδοποίησης.

Η αξιολόγηση της ποιότητας της ομαδοποίησης έδωσε τιμή Silhouette ίση με 0.27, Davies-Bouldin ίση με 2.56 και Calinski-Harabasz ίση με 124.04 (Πίνακας 4.3.1). Οι τιμές αυτές μαρτυρούν ικανοποιητικό βαθμό συνοχής και διαχωρισμού μεταξύ των clusters, δείχνοντας ότι ο αλγόριθμος ήταν σε θέση να αναγνωρίσει βασικά μορφολογικά πρότυπα στο σύνολο των ασυμπτωματικών ασθενών.

Συγχρόνως, η ανάλυση σταθερότητας έδειξε μέση τιμή Adjusted Rand Index (ARI) ίση με 0.46, με εύρος τιμών από 0.05 έως 1.00 μεταξύ των επαναλαμβανόμενων εκτελέσεων του αλγορίθμου (Πίνακας 4.3.2). Το αποτέλεσμα αυτό υποδηλώνει μέτρια σταθερότητα ομαδοποίησης, αφού παρατηρήθηκε σχετικά αυξημένη μεταβλητότητα μεταξύ διαφορετικών εκτελέσεων.

Η κατανομή των ασθενών μεταξύ των δύο clusters δεν ήταν πλήρως ισομερής (Σχήμα 4.3.5). Το Cluster 0 περιλάμβανε περίπου το 57% των ασυμπτωματικών ασθενών, ενώ το Cluster 1 αντιστοιχούσε περίπου στο 43% των ασθενών. Το εύρημα αυτό δείχνει αφενός ότι, η πλειονότητα των ασυμπτωματικών ασθενών συγκεντρωνόταν σε ένα κυρίαρχο μορφολογικό πρότυπο, και αφετέρου ένα μικρότερο υποσύνολο εμφάνιζε διαφορετικά χαρακτηριστικά πλάκας.

Συγκεκριμένα, το Cluster 0 χαρακτηριζόταν κυρίως από υψηλές τιμές GSM, χαμηλές τιμές JBA και παρουσία χαμηλών ή υψηλών τιμών PLAREA (Σχήματα 4.3.8, 4.3.9 και 4.3.10). Επιπρόσθετα, στο συγκεκριμένο cluster εντοπίστηκε μεγαλύτερη παρουσία DWA (Σχήμα 4.3.7), με τις τιμές STENOSIS να κατανέμονται κυρίως στις κατηγορίες medium και high (Σχήμα 4.3.11). Οι κανόνες TE2Rules επιβεβαίωσαν το συγκεκριμένο πρότυπο, μιας και οι πιο αντιπροσωπευτικοί κανόνες της συστάδας περιλάμβαναν τους «PLAREA = low» και «PLAREA = high» (Πίνακας 4.3.3). Άρα, το Cluster 0 φαίνεται να αντιστοιχεί σε ασθενείς με περισσότερο σταθερό μορφολογικό προφίλ πλάκας και υψηλότερη ηχογένεια.

Στον αντίποδα, το Cluster 1 παρουσίαζε κυρίως μέτριες τιμές PLAREA, αυξημένη συγκέντρωση very high τιμών JBA και σχετικά χαμηλότερες τιμές GSM, συγκριτικά με το Cluster 0 (Σχήματα 4.3.8, 4.3.9 και 4.3.10). Επιπλέον, οι ασθενείς της συγκεκριμένης συστάδας εμφάνιζαν μικρότερη παρουσία DWA (Σχήμα 4.3.7) και οι τιμές STENOSIS παρέμεναν κυρίως στις κατηγορίες medium και high (Σχήμα 4.3.11). Οι κανόνες TE2Rules που περιέγραφαν τη συστάδα περιλάμβαναν κυρίως τον κανόνα «PLAREA = medium» (Πίνακας 4.3.3). Έτσι φαίνεται ότι, το Cluster 1 αντιπροσωπεύει διαφορετικό μορφολογικό πρότυπο ασυμπτωματικών ασθενών, με αυξημένη ανομοιογένεια της πλάκας και περισσότερο ενδιάμεσα χαρακτηριστικά.

Ιδιαίτερα έντονη διαφοροποίηση μεταξύ των δύο clusters παρατηρήθηκε σε μεγάλο βαθμό στα χαρακτηριστικά PLAREA και JBA. Αναλυτικότερα, το Cluster 1 συγκέντρωνε μεγαλύτερη παρουσία πολύ υψηλών τιμών JBA και μέτριων τιμών PLAREA, ενώ από την άλλη το Cluster 0 εμφάνιζε συχνότερα χαμηλές τιμές JBA και υψηλότερες τιμές GSM. Από τα ευρήματα αυτά φαίνεται ότι, ακόμη και στους ασυμπτωματικούς ασθενείς υπάρχουν διακριτά μορφολογικά πρότυπα πλάκας με διαφορετικό βαθμό μορφολογικής επιβάρυνσης.

Εν κατακλείδι, τα αποτελέσματα του K-Modes στους ασυμπτωματικούς ασθενείς έδειξαν ότι ο αλγόριθμος ήταν σε θέση να αναγνωρίσει δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας. Το πρώτο πρότυπο αντιστοιχούσε σε ασθενείς με περισσότερο σταθερά χαρακτηριστικά και υψηλότερη ηχογένεια, και το δεύτερο περιλάμβανε ασθενείς με αυξημένη ανομοιογένεια της πλάκας με μεγαλύτερη παρουσία very high τιμών JBA.

5.4.2 Αποτελέσματα K-Prototypes στους ασυμπτωματικούς ασθενείς

Η εφαρμογή του αλγορίθμου K-Prototypes στο σύνολο των ασυμπτωματικών ασθενών οδήγησε στον σχηματισμό δύο συστάδων ($k=2$), αξιοποιώντας ταυτόχρονα αριθμητικά και κατηγορικά χαρακτηριστικά των δεδομένων. Η επιλογή των δύο συστάδων βασίστηκε τόσο στην καμπύλη elbow (Σχήμα 4.3.12), όσο και στις μετρικές αξιολόγησης clustering (Σχήμα 4.3.13), όπου φάνηκε ότι η τιμή $k=2$ προσέφερε τον καλύτερο συνδυασμό ποιότητας ομαδοποίησης και απλότητας του μοντέλου.

Η καμπύλη elbow έδειξε σημαντική μείωση του κόστους από $k=2$ έως $k=4$, ενώ στη συνέχεια η μείωση γινόταν πιο ομαλή. Παράλληλα, οι δείκτες αξιολόγησης clustering παρουσίασαν τις καλύτερες επιδόσεις για δύο συστάδες, γεγονός που ενίσχυσε την επιλογή του $k=2$ ως κατάλληλη για την περιγραφή της δομής των δεδομένων στους ασυμπτωματικούς ασθενείς.

Η ανάλυση PCA έδειξε ότι οι δύο πρώτες κύριες συνιστώσες εξηγούσαν συνολικά περίπου το 73% της συνολικής διακύμανσης των δεδομένων, με την πρώτη κύρια συνιστώσα να εξηγεί περίπου το 63% και τη δεύτερη περίπου το 10% (Σχήμα 4.3.14). Η προβολή PCA των ασθενών (Σχήμα 4.3.15) ανέδειξε αρκετά καθαρό διαχωρισμό μεταξύ των δύο clusters, σε μεγαλύτερο βαθμό σε σχέση με τον K-Modes. Παρόλο που, ασθενείς με χρόνο εμφάνισης συμβάντος μικρότερο και μεγαλύτερο των τριών ετών εμφανίζονταν και στις δύο συστάδες, ο χρόνος παρακολούθησης δεν φάνηκε να αποτελεί τον βασικό παράγοντα διαφοροποίησης της ομαδοποίησης.

Η αξιολόγηση της ποιότητας της ομαδοποίησης έδωσε τιμή Silhouette ίση με 0.59, Davies-Bouldin ίση με 0.77 και Calinski-Harabasz ίση με 661.90 (Πίνακας 4.3.4). Οι τιμές αυτές υποδηλώνουν πολύ καλή συνοχή και σαφή διαχωρισμό μεταξύ των clusters, δείχνοντας ότι ο K-Prototypes κατάφερε να αποτυπώσει αποτελεσματικά διαφορετικά μορφολογικά πρότυπα στους ασυμπτωματικούς ασθενείς.

Συγχρόνως, η ανάλυση σταθερότητας έδειξε εξαιρετικά σταθερή συμπεριφορά του αλγορίθμου, καθώς η τιμή του Adjusted Rand Index (ARI) ήταν ίση με 1.00 σε όλες τις επαναλαμβανόμενες εκτελέσεις (Πίνακας 4.3.5). Το εύρημα αυτό υποδηλώνει ότι η ομαδοποίηση παρέμεινε απολύτως συνεπής σε κάθε εκτέλεση του αλγορίθμου.

Η κατανομή των ασθενών μεταξύ των δύο clusters δεν ήταν ισομερής (Σχήμα 4.3.16). Το Cluster 0 περιλάμβανε τη μεγάλη πλειονότητα των ασθενών, ενώ το Cluster 1 αντιστοιχούσε σε σαφώς μικρότερο αριθμό ασθενών. Το αποτέλεσμα αυτό δείχνει από τη μία, ότι οι περισσότεροι ασυμπτωματικοί ασθενείς συγκεντρώνονταν σε ένα κυρίαρχο μορφολογικό πρότυπο και από την άλλη ένα μικρότερο υποσύνολο παρουσίαζε πιο ιδιαίτερα χαρακτηριστικά πλάκας.

Πιο συγκεκριμένα, το Cluster 0 χαρακτηριζόταν κατά κύριο λόγο από υψηλές τιμές GSM, χαμηλές ή μέτριες τιμές JBA και αυξημένη παρουσία DWA (Σχήματα 4.3.18, 4.3.19 και 4.3.20). Επιπλέον, οι τιμές PLAREA συγκεντρώνονταν κυρίως στις κατηγορίες low και medium (Σχήμα 4.3.21), ενώ οι τιμές STENOSIS κατανέμονταν κυρίως στις κατηγορίες medium και high (Σχήμα 4.3.22). Οι κανόνες TE2Rules επιβεβαίωσαν το συγκεκριμένο προφίλ, καθώς οι περισσότεροι χαρακτηριστικοί κανόνες της συστάδας περιλάμβαναν τους « $JBA \leq high$ », « $PLAREA = low$ » και « $GSM \geq medium$ » (Πίνακας 4.3.6). Συνολικά, το Cluster 0 φαίνεται να αντιστοιχεί σε ασθενείς με πιο σταθερό μορφολογικό προφίλ πλάκας και υψηλότερη ηχογένεια.

Αντιθέτως, το Cluster 1 παρουσίαζε έντονη συγκέντρωση very high τιμών JBA, χαμηλές τιμές GSM και συχνότερη απουσία DWA (Σχήματα 4.3.18, 4.3.19 και 4.3.20). Παράλληλα, οι τιμές PLAREA εντοπίζονταν κατά κύριο λόγο στις κατηγορίες medium και high (Σχήμα 4.3.21), ενώ οι τιμές STENOSIS παρέμεναν κυρίως στις κατηγορίες medium και high (Σχήμα 4.3.22). Οι κανόνες TE2Rules της συγκεκριμένης συστάδας περιλάμβαναν περισσότερο σύνθετους συνδυασμούς χαρακτηριστικών, όπως « $GSM = low \text{ AND } JBA = very high \text{ AND } STENOSIS = high$ » και « $DWA = absent \text{ AND } JBA = very high \text{ AND } PLAREA = high$ » (Πίνακας 4.3.6). Το Cluster 1 φαίνεται να αντιπροσωπεύει ένα πιο επιβαρυσμένο μορφολογικό πρότυπο πλάκας, με μεγαλύτερη ανομοιογένεια και χαρακτηριστικά που συνδέονται με αυξημένη μορφολογική αστάθεια.

Η πιο έντονη διαφοροποίηση μεταξύ των δύο clusters παρατηρήθηκε εντονότερα στα χαρακτηριστικά JBA, GSM και DWA. Το Cluster 1 συγκέντρωνε σχεδόν αποκλειστικά very high τιμές JBA και χαμηλές τιμές GSM, ενώ το Cluster 0 εμφάνιζε κυρίως υψηλές τιμές GSM και χαμηλότερες τιμές JBA. Επιπλέον, η παρουσία DWA ήταν αισθητά αυξημένη στο Cluster 0, σε σχέση με το Cluster 1. Τα ευρήματα αυτά δείχνουν ότι ακόμη και στους ασυμπτωματικούς ασθενείς μπορούν να αναγνωριστούν διαφορετικά μορφολογικά πρότυπα πλάκας με διαφορετικό βαθμό μορφολογικής επιβάρυνσης.

Συνολικά, τα αποτελέσματα του K-Prototypes στους ασυμπτωματικούς ασθενείς έδειξαν ότι ο αλγόριθμος μπόρεσε να αναγνωρίσει με σαφήνεια δύο βασικά μορφολογικά πρότυπα αθηρωματικής πλάκας. Το πρώτο πρότυπο αντιστοιχούσε στην πλειονότητα των ασθενών και χαρακτηριζόταν από πιο σταθερά μορφολογικά χαρακτηριστικά και υψηλότερη ηχογένεια. Το δεύτερο πρότυπο, περιλάμβανε μικρότερο αριθμό ασθενών με αυξημένη ανομοιογένεια της πλάκας, χαμηλότερο GSM και πολύ υψηλές τιμές JBA.

5.4.3 Σύγκριση K-Modes & K-Prototypes στους ασυμπτωματικούς ασθενείς

Η σύγκριση των αλγορίθμων K-Modes και K-Prototypes στο σύνολο των ασυμπτωματικών ασθενών έδειξε ότι και οι δύο μέθοδοι μπόρεσαν να αναγνωρίσουν διακριτά μορφολογικά πρότυπα αθηρωματικής πλάκας. Παρόλα αυτά, ο K-Prototypes παρουσίασε συνολικά καλύτερη εικόνα τόσο ως προς την ποιότητα της ομαδοποίησης, όσο και ως προς τη σταθερότητα των αποτελεσμάτων.

Σημειώνεται ότι και οι δύο αλγόριθμοι κατέληξαν στη δημιουργία δύο clusters ($k=2$), γεγονός που υποδηλώνει ότι στους ασυμπτωματικούς ασθενείς φαίνεται να υπάρχουν δύο κύρια μορφολογικά πρότυπα πλάκας. Ωστόσο, ο βαθμός διαχωρισμού μεταξύ των συστάδων ήταν αισθητά διαφορετικός ανάμεσα στις δύο προσεγγίσεις.

Στην περίπτωση του K-Modes, οι μετρικές αξιολόγησης έδειξαν μέτρια ποιότητα clustering, με Silhouette ίση με 0.27, Davies-Bouldin ίση με 2.56 και Calinski-Harabasz ίση με 124.04. Επιπλέον, η ανάλυση σταθερότητας παρουσίασε μέση τιμή ARI ίση με 0.46, με αρκετά μεγάλη μεταβλητότητα μεταξύ διαφορετικών εκτελέσεων του αλγορίθμου. Τα αποτελέσματα αυτά καταδεικνύουν ότι, ο K-Modes κατάφερε να εντοπίσει βασικές διαφοροποιήσεις μεταξύ των ασθενών, χωρίς όμως να επιτυγχάνει ιδιαίτερα σαφή ή σταθερό διαχωρισμό των clusters.

Από την άλλη, ο K-Prototypes παρουσίασε σημαντικά καλύτερες επιδόσεις. Η τιμή Silhouette αυξήθηκε σε 0.59, η τιμή Davies-Bouldin μειώθηκε σε 0.77 και η τιμή Calinski-Harabasz αυξήθηκε σε 661.90, στοιχείο που υποδηλώνει περισσότερο καθαρό διαχωρισμό και μεγαλύτερη συνοχή των συστάδων. Παράλληλα, η ανάλυση σταθερότητας έδειξε απόλυτα σταθερή συμπεριφορά, με ARI ίσο με 1.00, σε όλες τις επαναλαμβανόμενες εκτελέσεις του αλγορίθμου.

Η διαφορά αυτή, πιθανότατα, οφείλεται στο ότι ο K-Prototypes συνδυάζει αριθμητικά και κατηγορικά χαρακτηριστικά, επιτρέποντας πιο ολοκληρωμένη αναπαράσταση της μορφολογίας της αθηρωματικής πλάκας. Αντίθετα, ο K-Modes βασίζεται αποκλειστικά σε κατηγορικές μεταβλητές, γεγονός που ενδέχεται να περιορίζει την ικανότητά του να αποτυπώνει πιο λεπτομερές διαφοροποιήσεις μεταξύ των ασθενών.

Παρά τις διαφορές στις επιδόσεις, και οι δύο αλγόριθμοι ανέδειξαν παρόμοια μορφολογικά πρότυπα. Η μία συστάδα περιλάμβανε κυρίως ασθενείς με υψηλότερες τιμές GSM, χαμηλότερες τιμές JBA και γενικά σταθερότερο μορφολογικό προφίλ πλάκας. Στον αντίποδα, η δεύτερη συστάδα συγκέντρωνε ασθενείς με very high τιμές JBA, χαμηλότερες τιμές GSM και μεγαλύτερη ανομοιογένεια της πλάκας. Οι κανόνες TE2Rules επιβεβαίωσαν τα συγκεκριμένα πρότυπα και στις δύο μεθόδους, ενισχύοντας την ερμηνευσιμότητα των αποτελεσμάτων.

Επιπλέον, η προβολή PCA έδειξε ότι ο K-Prototypes πέτυχε σαφέστερο οπτικό διαχωρισμό των clusters σε σχέση με τον K-Modes, εύρημα το οποίο συμφωνεί και με τις καλύτερες τιμές των δεικτών αξιολόγησης. Παράλληλα, και στις δύο προσεγγίσεις παρατηρήθηκε παρουσία ασθενών με χρόνο εμφάνισης συμβάντος τόσο μικρότερο, όσο και μεγαλύτερο των τριών ετών μέσα στις ίδιες συστάδες, γεγονός που υποδηλώνει ότι ο χρόνος παρακολούθησης δεν αποτέλεσε τον κύριο παράγοντα διαφοροποίησης των clusters.

Συνοψίζοντας, τα αποτελέσματα δείχνουν ότι ο K-Prototypes παρείχε πιο σταθερή και ποιοτική ομαδοποίηση στους ασυμπτωματικούς ασθενείς σε σύγκριση με τον K-Modes. Η δυνατότητα ταυτόχρονης αξιοποίησης αριθμητικών και κατηγορικών χαρακτηριστικών φαίνεται να βελτίωσε σημαντικά την ικανότητα του αλγορίθμου να αναγνωρίζει διακριτά μορφολογικά πρότυπα αθηρωματικής πλάκας.

5.5 Σύγκριση μεταξύ συμπτωματικών και ασυμπτωματικών ασθενών

Η σύγκριση μεταξύ των συμπτωματικών και των ασυμπτωματικών ασθενών ανέδειξε ουσιαστικές διαφορές τόσο στα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας, όσο και στη συμπεριφορά των αλγορίθμων ομαδοποίησης. Παρόλο που και στις δύο ομάδες εντοπίστηκαν διακριτά μορφολογικά πρότυπα, οι συμπτωματικοί ασθενείς παρουσίαζαν γενικά περισσότερο επιβαρυνμένη και ετερογενή μορφολογία πλάκας, συγκριτικά με τους ασυμπτωματικούς.

Ένα από τα πιο σταθερά ευρήματα σε όλες τις αναλύσεις ήταν ο σημαντικός ρόλος των χαρακτηριστικών JBA, GSM και PLAREA στη διαφοροποίηση των clusters. Τόσο στους συμπτωματικούς, όσο και στους ασυμπτωματικούς ασθενείς, οι συστάδες που αντιστοιχούσαν σε αυξημένο μορφολογικό κίνδυνο χαρακτηρίζονταν συστηματικά από very high τιμές JBA, χαμηλές τιμές GSM και αυξημένες τιμές PLAREA. Αντίθετα, οι συστάδες χαμηλότερου μορφολογικού κινδύνου εμφάνιζαν συνήθως υψηλότερες τιμές GSM, χαμηλότερες τιμές JBA και συνολικά πιο σταθερό μορφολογικό προφίλ πλάκας.

Στους συμπτωματικούς ασθενείς, τα χαρακτηριστικά που σχετίζονται με αυξημένη ευαλωτότητα της πλάκας εμφανίζονταν συχνότερα και με μεγαλύτερη ένταση. Οι αντίστοιχες συστάδες συγκέντρωναν περισσότερους ασθενείς με very high JBA και low GSM. Ταυτόχρονα, παρουσίαζαν μεγαλύτερη διασπορά στον χώρο χαρακτηριστικών. Η εικόνα αυτή πιθανόν αντανακλά τη μεγαλύτερη πολυπλοκότητα και ετερογένεια της μορφολογίας της πλάκας σε ασθενείς που έχουν ήδη εμφανίσει νευρολογικά συμπτώματα.

Αντίθετα, στους ασυμπτωματικούς ασθενείς παρατηρήθηκαν περισσότερο συνεκτικά και ομοιογενή μορφολογικά πρότυπα. Η πλειονότητα των ασθενών συγκεντρωνόταν σε clusters με υψηλότερες τιμές GSM και χαμηλότερες τιμές JBA, στοιχείο που παραπέμπει σε πιο σταθερή μορφολογία πλάκας. Παρόλο που και στους ασυμπτωματικούς ασθενείς εντοπίστηκαν συστάδες με πιο επιβαρυνμένα χαρακτηριστικά, οι διαφοροποιήσεις μεταξύ των ομάδων εμφανίζονταν γενικά περισσότερο καθαρές και σταθερές.

Η διαφορά αυτή αποτυπώθηκε και στις μετρικές αξιολόγησης clustering. Στους ασυμπτωματικούς ασθενείς, ιδιαίτερα με τον K-Prototypes, παρατηρήθηκαν υψηλότερες τιμές Silhouette και Calinski-Harabasz, καθώς και χαμηλότερες τιμές Davies-Bouldin. Τα πιο πάνω καταδεικνύουν καλύτερο διαχωρισμό και μεγαλύτερη συνοχή των clusters. Σε αντίθεση με αυτό, στους συμπτωματικούς ασθενείς οι μετρικές ήταν γενικά χαμηλότερες, δείχνοντας μεγαλύτερη επικάλυψη και δυσκολότερο διαχωρισμό των συστάδων.

Παρόμοια εικόνα παρατηρήθηκε και στην ανάλυση σταθερότητας. Οι ασυμπτωματικοί ασθενείς παρουσίασαν σημαντικά μεγαλύτερη σταθερότητα clustering, ιδιαίτερα με τον K-Prototypes, όπου ο δείκτης ARI παρέμεινε ίσος με 1.00 σε όλες τις εκτελέσεις. Αντίθετα, στους συμπτωματικούς ασθενείς παρατηρήθηκε μεγαλύτερη μεταβλητότητα μεταξύ διαφορετικών runs, κυρίως στον K-Modes. Το εύρημα αυτό ενισχύει την άποψη ότι η μορφολογία της πλάκας στους συμπτωματικούς ασθενείς είναι περισσότερο ετερογενής και δύσκολα διαχωρίσιμη.

Επιπροσθέτως, η εφαρμογή της μεθόδου TE2Rules βοήθησε σημαντικά στην ερμηνεία των αποτελεσμάτων και στις δύο ομάδες ασθενών. Οι κανόνες που προέκυψαν ανέδειξαν επαναλαμβανόμενους συνδυασμούς χαρακτηριστικών που διαφοροποιούσαν σταθερά τις συστάδες. Στους συμπτωματικούς ασθενείς οι κανόνες ήταν συνήθως περισσότερο σύνθετοι και περιλάμβαναν συνδυασμούς very high JBA, low GSM, αυξημένου PLAREA και απουσίας DWA. Αντίθετα, στους ασυμπτωματικούς ασθενείς οι κανόνες περιέγραφαν συχνότερα πιο σταθερά μορφολογικά χαρακτηριστικά.

Συγχρόνως, η σύγκριση των δύο αλγορίθμων έδειξε ότι ο K-Prototypes υπερείχε σταθερά του K-Modes τόσο στους συμπτωματικούς, όσο και στους ασυμπτωματικούς ασθενείς. Η δυνατότητα συνδυαστικής αξιοποίησης αριθμητικών και κατηγορικών χαρακτηριστικών φαίνεται ότι επέτρεψε πιο ολοκληρωμένη αποτύπωση της μορφολογίας της αθηρωματικής πλάκας και καλύτερο διαχωρισμό των ασθενών. Επιπλέον, ο K-Prototypes παρήγαγε πιο σταθερά clusters και περισσότερο εκφραστικούς κανόνες TE2Rules, ιδιαίτερα στις αναλύσεις των ασυμπτωματικών ασθενών.

Συνολικά, τα αποτελέσματα δείχνουν ότι υπάρχουν σαφείς μορφολογικές διαφοροποιήσεις μεταξύ συμπτωματικών και ασυμπτωματικών ασθενών. Οι συμπτωματικοί ασθενείς εμφάνιζαν συχνότερα χαρακτηριστικά που σχετίζονται με αυξημένη μορφολογική αστάθεια της πλάκας, όπως very high JBA και low GSM. Από την άλλη, οι ασυμπτωματικοί ασθενείς παρουσίαζαν γενικά πιο σταθερό μορφολογικό προφίλ. Η ανάλυση clustering, και ιδιαίτερα η εφαρμογή του K-Prototypes, ανέδειξε αποτελεσματικά αυτές τις διαφοροποιήσεις και συνέβαλε στην καλύτερη κατανόηση της μορφολογικής ετερογένειας της καρωτιδικής νόσου.

5.6 Ερμηνεία της επικάλυψης των συστάδων

Παρόλο που οι αλγόριθμοι K-Modes και K-Prototypes κατάφεραν να αναγνωρίσουν διακριτά μορφολογικά πρότυπα αθηρωματικής πλάκας, στις περισσότερες αναλύσεις παρατηρήθηκε κάποιος βαθμός επικάλυψης μεταξύ των clusters στις προβολές PCA. Το φαινόμενο αυτό εμφανίστηκε τόσο στο σύνολο των ασθενών, όσο και στις επιμέρους ομάδες συμπτωματικών και ασυμπτωματικών ασθενών, αν και με διαφορετική ένταση ανάλογα με τον αλγόριθμο και την ομάδα ασθενών.

Η παρουσία επικάλυψης μεταξύ των συστάδων δεν ερμηνεύεται απαραίτητα ότι η ομαδοποίηση απέτυχε. Αντίθετα, σε ιατρικά δεδομένα, όπως αυτά τα δεδομένα της καρωτιδικής νόσου, θεωρείται εύλογο διαφορετικοί ασθενείς να μοιράζονται κοινά μορφολογικά χαρακτηριστικά, χωρίς όμως να ανήκουν ακριβώς στην ίδια κατηγορία. Η μορφολογία της αθηρωματικής πλάκας δεν φαίνεται να χωρίζεται σε απόλυτα διακριτές ομάδες, αλλά μάλλον να ακολουθεί ένα συνεχές φάσμα μεταβολών και ενδιάμεσων μορφολογικών καταστάσεων.

Η επικάλυψη αυτή ήταν περισσότερο εμφανής στους συμπτωματικούς ασθενείς, όπου οι μετρικές clustering παρουσίασαν γενικά χαμηλότερες τιμές και οι προβολές PCA έδειχναν μεγαλύτερη διασπορά των παρατηρήσεων. Η εικόνα αυτή, πιθανόν, αντανακλά τη μεγαλύτερη πολυπλοκότητα και ετερογένεια της νόσου στους συμπτωματικούς ασθενείς. Διαφορετικοί συνδυασμοί μορφολογικών χαρακτηριστικών μπορεί να οδηγούν σε παρόμοια κλινική εικόνα, γεγονός που καθιστά δυσκολότερο τον πλήρη διαχωρισμό των clusters.

Αντίθετα, στους ασυμπτωματικούς ασθενείς η επικάλυψη ήταν γενικά μικρότερη, ιδιαίτερα στις αναλύσεις με τον K-Prototypes. Οι συστάδες εμφανίζονταν περισσότερο συνεκτικές και καλύτερα διαχωρισμένες, εύρημα που συμφωνεί και με τις υψηλότερες τιμές των δεικτών αξιολόγησης clustering. Η εικόνα αυτή υποδηλώνει ότι στους ασυμπτωματικούς ασθενείς τα μορφολογικά πρότυπα της πλάκας παρουσιάζουν μεγαλύτερη ομοιογένεια.

Θεωρείται σημαντικό να ληφθεί υπόψη ότι, η PCA χρησιμοποιήθηκε αποκλειστικά για σκοπούς οπτικοποίησης των δεδομένων σε δύο διαστάσεις. Κατά τη διαδικασία αυτή χάνεται μέρος της πληροφορίας που υπάρχει στον αρχικό πολυδιάστατο χώρο χαρακτηριστικών. Επομένως, η οπτική επικάλυψη που παρατηρείται στις δισδιάστατες προβολές, δεν ερμηνεύεται απαραίτητα ότι οι συστάδες δεν διαχωρίζονται ικανοποιητικά στον πλήρη χώρο των δεδομένων.

Η μικρότερη επικάλυψη που παρατηρήθηκε στον K-Prototypes, ίσως, να σχετίζεται με τη δυνατότητα του αλγορίθμου να αξιοποιεί ταυτόχρονα αριθμητικά και κατηγορικά χαρακτηριστικά. Η συνδυαστική χρήση περισσότερων δεδομένων φαίνεται ότι επέτρεψε καλύτερη αποτύπωση των διαφορών μεταξύ των ασθενών και οδήγησε σε περισσότερο διακριτά μορφολογικά πρότυπα.

Επιπλέον, οι κανόνες TE2Rules έδειξαν ότι ακόμη και στις περιπτώσεις όπου υπήρχε οπτική επικάλυψη των clusters, υπήρχαν συγκεκριμένοι συνδυασμοί χαρακτηριστικών που διαφοροποιούσαν σταθερά τις συστάδες. Χαρακτηριστικά όπως very high JBA, low GSM και αυξημένο PLAREA εμφανίζονταν επαναλαμβανόμενα στα clusters αυξημένου μορφολογικού κινδύνου, γεγονός που ενισχύει την ερμηνευσιμότητα και τη συνέπεια των αποτελεσμάτων.

Συνολικά, η επικάλυψη των συστάδων φαίνεται να αντανακλά κυρίως τη φυσική ετερογένεια της καρωτιδικής νόσου και τη συνεχή μεταβολή των μορφολογικών χαρακτηριστικών της αθηρωματικής πλάκας. Παρά την παρουσία μερικής επικάλυψης, οι αλγόριθμοι clustering και ιδιαίτερα ο K-Prototypes, κατάφεραν να αναδείξουν σαφή μορφολογικά πρότυπα και να διαχωρίσουν ομάδες ασθενών με διαφορετικό βαθμό μορφολογικής επιβάρυνσης.

5.7 Σύνδεση με τη βιβλιογραφία

Τα αποτελέσματα της παρούσας μελέτης βρίσκονται σε ουσιαστική συμφωνία με τη σύγχρονη διεθνή βιβλιογραφία, επιβεβαιώνοντας ότι συγκεκριμένα μορφολογικά χαρακτηριστικά της αθηρωματικής πλάκας έχουν συσχετιστεί με την αστάθεια της καρωτιδικής νόσου και τον αυξημένο κίνδυνο εμφάνισης αγγειακών εγκεφαλικών επεισοδίων [1,2]. Ανάμεσα στα χαρακτηριστικά που αναδείχθηκαν ως σημαντικότεροι παράγοντες διαφοροποίησης των clusters συγκαταλέγονταν το JBA, το GSM και το PLAREA, τα οποία παρουσίασαν σταθερή συμβολή σχεδόν σε όλες τις αναλύσεις, ανεξάρτητα από τον αλγόριθμο ή την ομάδα ασθενών.

Οι μελέτες της ερευνητικής ομάδας ACSRS (Asymptomatic Carotid Stenosis Study) έχουν δείξει ότι η παρουσία αυξημένου Juxtaluminal Black Area (JBA), ιδιαίτερα όταν συνοδεύεται από απουσία ή ρήξη του ηχογενούς καλύμματος της πλάκας, σχετίζεται με σημαντικά αυξημένο κίνδυνο εμφάνισης εγκεφαλικού επεισοδίου [1]. Παράλληλα, χαμηλές τιμές του δείκτη GSM έχουν συσχετιστεί με περισσότερο ευάλωτες και λιπώδεις πλάκες, ενώ αυξημένη επιφάνεια πλάκας (PLAREA) αντανακλά μεγαλύτερο αθηρωματικό φορτίο και πιο επιβαρυνμένη μορφολογία [1]. Τα ευρήματα αυτά επιβεβαιώθηκαν και στην παρούσα μελέτη, όπου τα clusters αυξημένου μορφολογικού κινδύνου χαρακτηρίζονταν συστηματικά από πολύ υψηλές τιμές JBA, χαμηλές τιμές GSM και αυξημένες τιμές PLAREA.

Παρόμοια συμπεράσματα παρουσιάζονται και στη μελέτη των Prentzas et al. (2025), η οποία εστίασε στη χρήση Explainable Artificial Intelligence (XAI) για την πρόβλεψη υψηλού κινδύνου ασυμπτωματικών καρωτιδικών πλακών [2]. Η συγκεκριμένη εργασία ανέδειξε ότι χαρακτηριστικά όπως το JBA, το GSM, η παρουσία DWA και το PLAREA αποτελούν σημαντικούς δείκτες μορφολογικής αστάθειας της πλάκας και μπορούν να συμβάλουν ουσιαστικά στη διαστρωμάτωση κινδύνου των ασθενών. Παρότι, η παρούσα μελέτη βασίστηκε σε τεχνικές μη επιβλεπόμενης μάθησης και όχι σε προγνωστικά μοντέλα, κατέληξε σε αντίστοιχα μορφολογικά πρότυπα, γεγονός που ενισχύει την αξιοπιστία των συγκεκριμένων υπερηχογραφικών δεικτών.

Τα αποτελέσματα της μελέτης συμφωνούν, επίσης, με τη βιβλιογραφία που αφορά τη χρήση τεχνικών μηχανικής μάθησης στην ανάλυση της καρωτιδικής νόσου [3–6]. Προηγούμενες εργασίες έχουν δείξει ότι η αξιοποίηση υπερηχογραφικών χαρακτηριστικών και δεδομένων υφής εικόνας, σε συνδυασμό με μοντέλα μηχανικής μάθησης, μπορεί να βελτιώσει σημαντικά την αναγνώριση ασθενών υψηλού κινδύνου [6]. Η παρούσα εργασία ενισχύει περαιτέρω αυτή την προσέγγιση, δείχνοντας ότι ακόμη και μέσω *unsupervised learning* τεχνικών είναι δυνατό να αναδειχθούν διακριτά μορφολογικά πρότυπα με ουσιαστική κλινική σημασία.

Παράλληλα, η μελέτη αυτή ενισχύει τη σύγχρονη βιβλιογραφική τάση που υπογραμμίζει τη σημασία της ερμηνευσιμότητας στην ιατρική τεχνητή νοημοσύνη [2,10]. Οι Prentzas et al. (2025) επεσήμαναν ότι πολλά παραδοσιακά μοντέλα μηχανικής μάθησης λειτουργούν ως συστήματα «μαύρου κουτιού» (*black-box*), γεγονός που περιορίζει την εμπιστοσύνη και την κλινική αξιοποίησή τους [2]. Στην παρούσα εργασία, η εφαρμογή των TE2Rules [10] συνέβαλε ουσιαστικά στην ερμηνεία των αποτελεσμάτων *clustering*, επιτρέποντας την εξαγωγή σαφών και κατανοητών κανόνων που περιέγραφαν τα μορφολογικά χαρακτηριστικά κάθε συστάδας. Με τον τρόπο αυτό, η ανάλυση δεν περιορίστηκε μόνο στον υπολογιστικό διαχωρισμό των ασθενών, αλλά παρείχε και ερμηνεύσιμη κλινική πληροφορία με πιθανή πρακτική αξία.

Επιπλέον, η μεγαλύτερη ετερογένεια που παρατηρήθηκε στους συμπτωματικούς ασθενείς συμφωνεί με προηγούμενες ιστοπαθολογικές και υπερηχογραφικές μελέτες, σύμφωνα με τις οποίες οι συμπτωματικές πλάκες εμφανίζουν πιο σύνθετη και ασταθή μορφολογία [1,6]. Η αυξημένη επικάλυψη των *clusters*, η μεγαλύτερη διασπορά στις προβολές PCA και οι χαμηλότερες τιμές των δεικτών *clustering* στους συμπτωματικούς ασθενείς πιθανόν αντανακλούν αυτή τη μεγαλύτερη βιολογική και μορφολογική πολυπλοκότητα της νόσου.

Η υπεροχή του αλγορίθμου K-Prototypes, έναντι του K-Modes συμφωνεί επίσης με τη γενικότερη βιβλιογραφία clustering μικτού τύπου δεδομένων [7,9,15]. Η δυνατότητα ταυτόχρονης αξιοποίησης αριθμητικών και κατηγορικών χαρακτηριστικών επιτρέπει πληρέστερη αναπαράσταση των δεδομένων και περιορίζει την απώλεια πληροφορίας που προκαλεί η αποκλειστική κατηγοριοποίηση των μεταβλητών. Στην παρούσα εργασία, αυτό αποτυπώθηκε τόσο στις υψηλότερες τιμές των δεικτών Silhouette και Calinski–Harabasz [11,13], όσο και στις χαμηλότερες τιμές Davies–Bouldin [12], αλλά και στη σημαντικά αυξημένη σταθερότητα των clusters.

Συνολικά, τα αποτελέσματα της μελέτης επιβεβαιώνουν σε μεγάλο βαθμό τα ευρήματα της σύγχρονης βιβλιογραφίας σχετικά με τη μορφολογία της καρωτιδικής πλάκας, τη διαστρωμάτωση κινδύνου εγκεφαλικού επεισοδίου και τη χρήση Explainable Machine Learning τεχνικών στην ιατρική ανάλυση δεδομένων. Παράλληλα, η εργασία συμβάλλει περαιτέρω στην κατανόηση της μορφολογικής ετερογένειας της καρωτιδικής νόσου και αναδεικνύει τη χρησιμότητα των τεχνικών clustering και Explainable AI στην έγκαιρη αναγνώριση ασθενών με αυξημένο μορφολογικό κίνδυνο.

Κεφάλαιο 6: Συμπεράσματα και Μελλοντικές Κατευθύνσεις

6.1	Συνολικά συμπεράσματα της εργασίας	151
6.2	Αξιολόγηση της μεθοδολογίας και περιορισμοί	154
6.3	Μελλοντικές κατευθύνσεις	157

6.1 Συνολικά συμπεράσματα της εργασίας

Σκοπός της παρούσας μελέτης ήταν η διερεύνηση της μορφολογικής ετερογένειας της καρωτιδικής αθηρωματικής πλάκας μέσω τεχνικών μη επιβλεπόμενης μάθησης (unsupervised learning), με στόχο την αναγνώριση διακριτών μορφολογικών προτύπων που σχετίζονται με διαφορετικό βαθμό μορφολογικής επιβάρυνσης και ευαλωτότητας της πλάκας. Για τον σκοπό αυτό εφαρμόστηκαν οι αλγόριθμοι K-Modes και K-Prototypes σε σύνολο δεδομένων ασθενών με καρωτιδική νόσο, ενώ παράλληλα αξιοποιήθηκε η μέθοδος ερμηνευσιμότητας TE2Rules για την ερμηνεία των αποτελεσμάτων της ομαδοποίησης.

Τα αποτελέσματα της ανάλυσης έδειξαν ότι και οι δύο αλγόριθμοι ήταν σε θέση να αναγνωρίσουν διακριτά μορφολογικά πρότυπα αθηρωματικής πλάκας, τόσο στο σύνολο των ασθενών όσο και στις επιμέρους ομάδες συμπτωματικών και ασυμπτωματικών ασθενών. Οι συστάδες που προέκυψαν διαφοροποιούνταν κυρίως ως προς χαρακτηριστικά όπως το Juxtaluminal Black Area (JBA), η ηχογένεια της πλάκας (GSM), η επιφάνεια της πλάκας (PLAREA) και, σε ορισμένες περιπτώσεις, ο βαθμός στένωσης (STENOSIS). Τα χαρακτηριστικά αυτά εμφανίστηκαν επανειλημμένα ως οι σημαντικότεροι παράγοντες διαφοροποίησης των ομάδων ασθενών.

Ένα από τα σημαντικότερα ευρήματα της μελέτης ήταν ότι οι συστάδες αυξημένου μορφολογικού κινδύνου χαρακτηρίζονταν σταθερά από πολύ υψηλές τιμές JBA, χαμηλές τιμές GSM και αυξημένες τιμές PLAREA. Αντίθετα, οι συστάδες χαμηλότερου μορφολογικού κινδύνου εμφάνιζαν συχνότερα υψηλότερες τιμές GSM, χαμηλότερες τιμές JBA και περισσότερο σταθερό μορφολογικό προφίλ πλάκας. Το εύρημα αυτό εμφανίστηκε με συνέπεια σε όλες σχεδόν τις αναλύσεις, ενισχύοντας τη σημασία των συγκεκριμένων χαρακτηριστικών ως βασικών δεικτών μορφολογικής αστάθειας της αθηρωματικής πλάκας.

Η συγκριτική αξιολόγηση των δύο αλγορίθμων έδειξε ότι ο K-Prototypes παρουσίασε συνολικά καλύτερη απόδοση σε σχέση με τον K-Modes, τόσο ως προς την ποιότητα του clustering, όσο και ως προς τη σταθερότητα των αποτελεσμάτων. Η δυνατότητα ταυτόχρονης αξιοποίησης αριθμητικών και κατηγορικών χαρακτηριστικών επέτρεψε πληρέστερη αναπαράσταση των μορφολογικών χαρακτηριστικών των ασθενών και οδήγησε σε περισσότερο συνεκτικές και καλύτερα διαχωρισμένες συστάδες. Η υπεροχή αυτή ήταν ιδιαίτερα εμφανής στο σύνολο των ασυμπτωματικών ασθενών, όπου ο K-Prototypes εμφάνισε πολύ υψηλή σταθερότητα και σαφέστερο διαχωρισμό των clusters.

Παράλληλα, η εφαρμογή της μεθόδου Explainable Artificial Intelligence μέσω του TE2Rules συνέβαλε ουσιαστικά στην ερμηνεία των αποτελεσμάτων της ομαδοποίησης. Οι κανόνες που εξήχθησαν περιέγραψαν με σαφή και κατανοητό τρόπο τα μορφολογικά χαρακτηριστικά κάθε συστάδας, επιτρέποντας τη σύνδεση των παραγόμενων συστάδων με συγκεκριμένους και κλινικά ερμηνεύσιμους συνδυασμούς χαρακτηριστικών. Με τον τρόπο αυτό, η ανάλυση δεν περιορίστηκε μόνο στον διαχωρισμό των ασθενών σε ομάδες, αλλά παρείχε και ερμηνεύσιμη πληροφορία με πιθανή κλινική αξία.

Επιπλέον, η μελέτη ανέδειξε σημαντικές διαφοροποιήσεις μεταξύ συμπτωματικών και ασυμπτωματικών ασθενών. Συγκεκριμένα, οι συμπτωματικοί ασθενείς παρουσίασαν μεγαλύτερη μορφολογική ετερογένεια, αυξημένη επικάλυψη μεταξύ των clusters στις προβολές PCA και χαμηλότερες τιμές στις μετρικές αξιολόγησης clustering, κάτι που πιθανόν αντανάκλα τη μεγαλύτερη πολυπλοκότητα της νόσου στους ασθενείς που έχουν ήδη εμφανίσει νευρολογικά συμπτώματα. Από την άλλη, οι ασυμπτωματικοί ασθενείς εμφάνισαν περισσότερο ομοιογενή και καλύτερα διαχωρισμένα μορφολογικά πρότυπα.

Συνολικά, η παρούσα εργασία έδειξε ότι οι τεχνικές μη επιβλεπόμενης μάθησης μπορούν να αποτελέσουν χρήσιμο εργαλείο για την ανάλυση της μορφολογικής ετερογένειας της καρωτιδικής νόσου και την αναγνώριση ομάδων ασθενών με διαφορετικό μορφολογικό προφίλ. Συγχρόνως, η συνδυαστική χρήση τεχνικών clustering και Explainable Artificial Intelligence ενισχύει σημαντικά την ερμηνευσιμότητα των αποτελεσμάτων και δημιουργεί προοπτικές για μελλοντική αξιοποίηση αντίστοιχων προσεγγίσεων σε συστήματα υποστήριξης κλινικών αποφάσεων και εξατομικευμένης εκτίμησης κινδύνου εγκεφαλικού επεισοδίου.

6.2 Αξιολόγηση της μεθοδολογίας και περιορισμοί

Η παρούσα εργασία αξιοποίησε προηγμένες τεχνικές μη επιβλεπόμενης μάθησης (unsupervised learning) για τη συστηματική μελέτη της μορφολογικής ετερογένειας της καρωτιδικής αθηρωματικής πλάκας, συνδυάζοντας αλγορίθμους ομαδοποίησης (clustering) και μεθοδολογίες Εξηγήσιμης Τεχνητής Νοημοσύνης (Explainable Artificial Intelligence - XAI). Η συγκεκριμένη προσέγγιση παρουσίασε σημαντικά πλεονεκτήματα ως προς την ανάλυση και ερμηνεία των δεδομένων. Ωστόσο, συνοδεύεται και από ορισμένους μεθοδολογικούς και κλινικούς περιορισμούς, οι οποίοι πρέπει να λαμβάνονται υπόψη κατά την ερμηνεία και γενίκευση των αποτελεσμάτων.

Ένα από τα σημαντικότερα πλεονεκτήματα της προτεινόμενης μεθοδολογίας ήταν η δυνατότητα ανάλυσης πολύπλοκων μορφολογικών χαρακτηριστικών, χωρίς την ανάγκη ύπαρξης προκαθορισμένων κλινικών κατηγοριών ή ετικετών κινδύνου. Η χρήση τεχνικών μη επιβλεπόμενης μάθησης επέτρεψε την αντικειμενική και αυτόματη αναγνώριση υποκείμενων μορφολογικών προτύπων της αθηρωματικής πλάκας, αναδεικνύοντας ομάδες ασθενών με διαφορετικό βαθμό μορφολογικής επιβάρυνσης. Η προσέγγιση αυτή είναι ιδιαίτερα σημαντική σε ιατρικά δεδομένα υψηλής ετερογένειας, όπου τα όρια μεταξύ διαφορετικών μορφολογικών καταστάσεων δεν είναι πάντοτε σαφώς διακριτά.

Παράλληλα, η συνδυαστική εφαρμογή των αλγορίθμων K-Modes και K-Prototypes επέτρεψε τη σύγκριση διαφορετικών στρατηγικών clustering για δεδομένα κατηγορικού και μικτού τύπου. Ιδιαίτερα σημαντική ήταν η αξιοποίηση του αλγορίθμου K-Prototypes, καθώς επέτρεψε την ταυτόχρονη επεξεργασία αριθμητικών και κατηγορικών χαρακτηριστικών. Η δυνατότητα αυτή οδήγησε σε πληρέστερη αναπαράσταση της μορφολογίας της αθηρωματικής πλάκας και συνέβαλε στη δημιουργία πιο σταθερών, συνεκτικών και καλύτερα διαχωρισμένων συστάδων.

Επιπλέον, σημαντικό πλεονέκτημα της μελέτης αποτέλεσε η ενσωμάτωση της Εξηγήσιμης Τεχνητής Νοημοσύνης μέσω της μεθόδου TE2Rules. Η εξαγωγή σαφών και εύκολα κατανοητών λογικών κανόνων ενίσχυσε σημαντικά την κλινική ερμηνευσιμότητα της ομαδοποίησης, περιορίζοντας τον χαρακτήρα «μαύρου κουτιού» (black-box), ο οποίος συχνά χαρακτηρίζει τα μοντέλα μηχανικής μάθησης. Με τον τρόπο αυτό, οι παραχθείσες συστάδες συνδέθηκαν άμεσα με συγκεκριμένους συνδυασμούς μορφολογικών χαρακτηριστικών της πλάκας, γεφυρώνοντας το χάσμα μεταξύ υπολογιστικής ανάλυσης και καθημερινής ιατρικής πράξης.

Παρά τα σημαντικά αυτά πλεονεκτήματα, η παρούσα εργασία παρουσιάζει και ορισμένους περιορισμούς. Αρχικά, το χρησιμοποιούμενο κλινικό σύνολο δεδομένων εμφάνιζε ανισορροπία μεταξύ συμπτωματικών και ασυμπτωματικών ασθενών, με τους ασυμπτωματικούς να αποτελούν τη μεγάλη πλειονότητα του πληθυσμού. Η ανισορροπία αυτή ενδέχεται να επηρέασε τη διαδικασία ομαδοποίησης και τον τρόπο με τον οποίο αναγνωρίστηκαν τα μορφολογικά πρότυπα, ιδιαίτερα στις μικρότερες υποομάδες ασθενών.

Επιπλέον, το μέγεθος του δείγματος των συμπτωματικών ασθενών ήταν σχετικά περιορισμένο σε σύγκριση με το συνολικό σύνολο δεδομένων. Ο παράγοντας αυτός, σε συνδυασμό με τη μεγαλύτερη μορφολογική πολυπλοκότητα που εμφανίζουν οι ασταθείς πλάκες μετά από κλινικό επεισόδιο, πιθανόν συνέβαλε στη μεγαλύτερη μεταβλητότητα και στις χαμηλότερες τιμές των δεικτών ποιότητας clustering που παρατηρήθηκαν στις συγκεκριμένες αναλύσεις.

Ένας ακόμη σημαντικός περιορισμός σχετίζεται με τη χρήση της Μεθόδου Κύριων Συνιστωσών (Principal Component Analysis - PCA) αποκλειστικά για σκοπούς οπτικοποίησης. Η προβολή ενός πολυδιάστατου χώρου χαρακτηριστικών σε δύο διαστάσεις συνεπάγεται αναπόφευκτη απώλεια μέρους της πληροφορίας.

Ως αποτέλεσμα, η οπτική επικάλυψη των συστάδων στις γραφικές παραστάσεις μπορεί σε ορισμένες περιπτώσεις να είναι μεγαλύτερη από εκείνη που πραγματικά υπάρχει στον αρχικό χώρο των δεδομένων. Για τον λόγο αυτό, οι προβολές PCA χρειάζεται να ερμηνεύονται υποστηρικτικά και όχι ως μοναδικό κριτήριο αξιολόγησης της ποιότητας της ομαδοποίησης.

Επιπρόσθετα, οι αλγόριθμοι clustering εμφανίζουν ευαισθησία τόσο στην αρχικοποίηση των συστάδων, όσο και στην επιλογή του βέλτιστου αριθμού clusters. Παρότι στην παρούσα εργασία χρησιμοποιήθηκαν μετρικές αξιολόγησης και επαναλαμβανόμενοι έλεγχοι σταθερότητας για τη βελτίωση της αξιοπιστίας των αποτελεσμάτων, δεν μπορεί να αποκλειστεί το ενδεχόμενο διαφορετικές υπερπαραμέτροι ή διαφορετικά σύνολα δεδομένων να οδηγούσαν σε ελαφρώς διαφοροποιημένα μορφολογικά πρότυπα.

Επιπλέον, μέρος της πληροφορίας των δεδομένων απλοποιήθηκε μέσω της κατηγοριοποίησης ορισμένων συνεχών μεταβλητών, κυρίως στις αναλύσεις K-Modes. Παρότι η διαδικασία αυτή διευκόλυνε την εφαρμογή κατηγορικών αλγορίθμων, ενδέχεται να οδήγησε σε απώλεια μικρότερων, αλλά πιθανώς κλινικά σημαντικών διαφοροποιήσεων μεταξύ των ασθενών.

Τέλος, η παρούσα μελέτη βασίστηκε σε ένα συγκεκριμένο κλινικό σύνολο δεδομένων και δεν πραγματοποιήθηκε εξωτερική επικύρωση των αποτελεσμάτων σε ανεξάρτητο πληθυσμό ασθενών. Επομένως, απαιτούνται μελλοντικές μελέτες σε μεγαλύτερα, πολυκεντρικά και περισσότερο ισορροπημένα σύνολα δεδομένων, ώστε να αξιολογηθεί περαιτέρω η γενίκευση και η σταθερότητα των μορφολογικών προτύπων που αναδείχθηκαν.

Λαμβάνοντας υπόψη τους παραπάνω περιορισμούς, τα ευρήματα της παρούσας εργασίας καταδεικνύουν με σαφήνεια ότι ο συνδυασμός τεχνικών clustering και Explainable Artificial Intelligence μπορεί να προσφέρει ουσιαστική και ερμηνεύσιμη πληροφορία για τη μορφολογική ανάλυση της καρωτιδικής αθηρωματικής πλάκας, συμβάλλοντας σημαντικά στην καλύτερη κατανόηση της ετερογένειας της καρωτιδικής νόσου.

6.3 Μελλοντικές κατευθύνσεις

Τα αποτελέσματα της παρούσας εργασίας ανέδειξαν τη σημαντική δυναμική των τεχνικών μη επιβλεπόμενης μάθησης και της Εξηγήσιμης Τεχνητής Νοημοσύνης στην ανάλυση της μορφολογικής ετερογένειας της καρωτιδικής αθηρωματικής πλάκας. Παρόλα αυτά, υπάρχουν αρκετές ερευνητικές κατευθύνσεις που θα μπορούσαν να επεκτείνουν και να ενισχύσουν περαιτέρω τη συγκεκριμένη προσέγγιση.

Μία από τις σημαντικότερες μελλοντικές κατευθύνσεις αφορά την αξιοποίηση μεγαλύτερων και πολυκεντρικών συνόλων δεδομένων. Η εφαρμογή των αλγορίθμων σε δεδομένα που προέρχονται από διαφορετικά νοσοκομεία ή ερευνητικά κέντρα θα μπορούσε να βελτιώσει σημαντικά τη γενικευσιμότητα των αποτελεσμάτων και να επιτρέψει την ανεξάρτητη επιβεβαίωση των μορφολογικών προτύπων που αναδείχθηκαν στην παρούσα μελέτη. Στο πλαίσιο αυτό, ιδιαίτερη σημασία έχει η εξωτερική επικύρωση των αποτελεσμάτων σε ανεξάρτητα σύνολα δεδομένων, ώστε να αξιολογηθεί η σταθερότητα και η γενικευσιμότητα των μορφολογικών προτύπων που αναδείχθηκαν.

Παράλληλα, ιδιαίτερο ενδιαφέρον παρουσιάζει η ενσωμάτωση επιπλέον κλινικών, βιοχημικών και απεικονιστικών χαρακτηριστικών. Μελλοντικές αναλύσεις θα μπορούσαν να συμπεριλάβουν δημογραφικά δεδομένα, παράγοντες καρδιαγγειακού κινδύνου, στοιχεία από το ιατρικό ιστορικό των ασθενών, καθώς και πιο εξειδικευμένους δείκτες ιατρικής εικόνας (radiomics), με στόχο τη δημιουργία πιο ολοκληρωμένων μορφολογικών προφίλ ασθενών.

Επιπλέον, θα μπορούσαν να διερευνηθούν και άλλες τεχνικές clustering, όπως ιεραρχικές μέθοδοι ομαδοποίησης, αλγόριθμοι βασισμένοι στην πυκνότητα, όπως ο DBSCAN, ή ακόμη και σύγχρονες προσεγγίσεις βαθιάς μάθησης (deep learning) για την εξαγωγή χαρακτηριστικών και τη μείωση της διάστασης των δεδομένων. Η σύγκριση διαφορετικών τεχνικών ενδέχεται να οδηγήσει στην αναγνώριση ακόμη πιο σύνθετων μορφολογικών προτύπων και σε καλύτερο διαχωρισμό των συστάδων.

Σημαντική προοπτική αποτελεί, επίσης, η αξιοποίηση διαχρονικών δεδομένων (longitudinal data) και η μελέτη της χρονικής εξέλιξης της μορφολογίας της αθηρωματικής πλάκας. Η παρακολούθηση των ίδιων ασθενών σε διαδοχικά χρονικά σημεία θα μπορούσε να συμβάλει στην καλύτερη κατανόηση του τρόπου με τον οποίο μεταβάλλονται τα μορφολογικά χαρακτηριστικά της πλάκας με την πάροδο του χρόνου, καθώς και της σχέσης των μεταβολών αυτών με τον κίνδυνο εμφάνισης αγγειακού εγκεφαλικού επεισοδίου.

Παράλληλα, ιδιαίτερα ενδιαφέρουσα κατεύθυνση αποτελεί η ανάπτυξη υβριδικών προγνωστικών μοντέλων που θα συνδυάζουν τεχνικές clustering, αλγορίθμους επιβλεπόμενης μάθησης και μεθόδους Explainable AI. Μια τέτοια προσέγγιση θα μπορούσε να οδηγήσει όχι μόνο στην αναγνώριση μορφολογικών προτύπων, αλλά και στην πρόβλεψη μελλοντικών κλινικών συμβάντων με μεγαλύτερη ακρίβεια και καλύτερη ερμηνευσιμότητα.

Επιπλέον, η περαιτέρω ανάπτυξη Explainable AI μεθόδων στην ιατρική απεικόνιση αποτελεί ιδιαίτερα σημαντικό πεδίο έρευνας. Η εξαγωγή περισσότερο λεπτομερών και κλινικά ερμηνεύσιμων κανόνων μπορεί να ενισχύσει την εμπιστοσύνη των ιατρών στα συστήματα τεχνητής νοημοσύνης και να διευκολύνει τη μελλοντική ενσωμάτωσή τους στην καθημερινή κλινική πρακτική.

Τέλος, μία ιδιαίτερα σημαντική μελλοντική εφαρμογή αφορά την ανάπτυξη Συστημάτων Υποστήριξης Κλινικών Αποφάσεων (Clinical Decision Support Systems – CDSS), τα οποία θα αξιοποιούν τεχνικές clustering και Explainable AI για την αυτόματη εκτίμηση του μορφολογικού κινδύνου της καρωτιδικής πλάκας. Τέτοιου είδους συστήματα θα μπορούσαν να συμβάλουν στην έγκαιρη αναγνώριση ασθενών υψηλού κινδύνου και να υποστηρίξουν περισσότερο εξατομικευμένες στρατηγικές πρόληψης και θεραπευτικής παρέμβασης.

Συνολικά, η παρούσα εργασία ανέδειξε τη δυνατότητα των τεχνικών clustering και Explainable AI στην ανάλυση της καρωτιδικής νόσου. Η περαιτέρω ανάπτυξη και συνδυαστική αξιοποίηση των τεχνολογιών αυτών αναμένεται να συμβάλει ουσιαστικά σε πληρέστερη κατανόηση της μορφολογικής ετερογένειας της αθηρωματικής πλάκας, όπως και στη βελτίωση της κλινικής διαστρωμάτωσης κινδύνου των ασθενών.

Βιβλιογραφία

- [1] A. N. Nicolaides, et al., "Asymptomatic internal carotid artery stenosis and risk of stroke," *The Lancet*, vol. 376, no. 9746, pp. 1074-1084, 2010.
- [2] N. Prentzas, C. S. Skouteli, E. Kyriacou, M. Griffin, A. Nicolaides, C. P. Loizou, A. Kakas, and C. S. Pattichis, "Explainable artificial intelligence in the prediction of high-risk asymptomatic carotid plaques based on ultrasonic image features," *International Angiology*, vol. 44, no. 4, pp. 291-300, August 2025.
- [3] T. M. Mitchell, *Machine Learning*, New York, NY: McGraw-Hill, 1997.
- [4] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, Cambridge, MA: MIT Press, 2012.
- [5] EIT Health, "Machine learning in healthcare: uses, benefits and pioneers in the field," 2024. [Online]. Available: <https://eithealth.eu/news-article/machine-learning-in-healthcare-uses-benefits-and-pioneers-in-the-field/>. [Accessed 2026].
- [6] Rajkomar, S., J. Dean, and I. Kohane, "Machine learning in medicine: a review," 2019. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8822225/>. [Accessed 2026].
- [7] PMC, "Cluster analysis in biomedical research," 2016. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4776444/>. [Accessed 2026].
- [8] Columbia University Mailman School of Public Health, "K-means cluster analysis," 2022. [Online]. Available: <https://www.publichealth.columbia.edu/research/population-health-methods/k-means-cluster-analysis>. [Accessed 2026].
- [9] Huang, Z., "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283-304, 1998.
- [10] Agarwal, R., N. Tagasovska, and M. Fiterau, "TE2Rules: Explaining Tree Ensembles with Rules," *arXiv preprint arXiv:2206.14359*, 2022.
- [11] Rousseeuw, P. J., "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.

- [12] Davies, D. L., and D. W. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vols. PAMI-1, no. 2, pp. 224-227, 1979.
- [13] Calinski, T., and J. Harabasz, "A dendrite method for cluster analysis," Communications in Statistics, vol. 3, no. 1, pp. 1-27, 1974.
- [14] MathWorks, "Machine Learning," 2024. [Online]. Available: <https://www.mathworks.com/discovery/machine-learning.html>. [Accessed 2026].
- [15] Huang, Z., "Clustering large data sets with mixed numeric and categorical values," in Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 1997.
- [16] Saba, L., et al., "Artificial Intelligence in Symptomatic Carotid Plaque Detection: A Narrative Review," Applied Sciences, vol. 13, no. 7, p. 4321, 2023.

Παράρτημα Α

Στο Παράρτημα Α παρουσιάζονται τα *cluster modes* που προέκυψαν από την εφαρμογή των αλγορίθμων K-Modes και K-Prototypes στα διαφορετικά υποσύνολα δεδομένων. Τα αποτελέσματα αυτά αποτυπώνουν τις επικρατούσες τιμές των μορφολογικών χαρακτηριστικών σε κάθε συστάδα και συμβάλλουν στην κατανόηση των μορφολογικών προτύπων που αναδείχθηκαν μέσω της ανάλυσης συστάδων.

Συγκεκριμένα στα σχήματα Α-1 – Α-6, παρουσιάζονται τα *cluster modes* για:

- το συνολικό σύνολο δεδομένων,
- το υποσύνολο των ασυμπτωματικών ασθενών και
- το υποσύνολο των συμπτωματικών ασθενών,
- τόσο για τον αλγόριθμο K-Modes όσο και για τον K-Prototypes.

	A	B	C	D	E	F	G
1	Cluster ▾	stenosis ▾	gsm ▾	plarea ▾	dwa ▾	ctiastr ▾	jba ▾
2	0	medium	high	low	present	absent	low
3	1	medium	low	medium	present	absent	vhigh

Σχήμα Α- 1 cluster modes – full dataset K-Modes

	A	B	C	D	E	F	G
1	Cluster ▾	stenosis ▾	gsm ▾	plarea ▾	dwa ▾	ctiastr ▾	jba ▾
2	0	medium	high	low	present	absent	low
3	1	medium	high	medium	present	absent	low

Σχήμα Α- 2 cluster modes – asymptomatic dataset K-Modes

	A	B	C	D	E	F	G
1	Cluster ▾	stenosis ▾	gsm ▾	plarea ▾	dwa ▾	ctiastr ▾	jba ▾
2	0	medium	low	medium	present	absent	vhigh
3	1	high	medium	medium	present	absent	medium

Σχήμα Α- 3 cluster modes – symptomatic dataset K-Modes

	A	B	C	D	E	F	G
1	Cluster ▾	stenosis ▾	gsm ▾	plarea ▾	dwa ▾	ctiastr ▾	jba ▾
2		0 medium	high	low	present	absent	low
3		1 medium	low	medium	absent	absent	vhigh

Σχήμα A- 4 cluster modes – full dataset K-Prototypes

	A	B	C	D	E	F	G
1	Cluster ▾	stenosis ▾	gsm ▾	plarea ▾	dwa ▾	ctiastr ▾	jba ▾
2		0 medium	high	low	present	absent	low
3		1 medium	low	medium	absent	absent	vhigh

Σχήμα A- 5 cluster modes – asymptomatic dataset K-prototypes

	A	B	C	D	E	F	G
1	Cluster ▾	stenosis ▾	gsm ▾	plarea ▾	dwa ▾	ctiastr ▾	jba ▾
2		0 medium	medium	medium	present	absent	medium
3		1 medium	low	medium	present	absent	vhigh

Σχήμα A- 6 cluster modes – symptomatic dataset K-Prototypes

Παράρτημα Β

Ενδεικτική μορφή του αρχείου ACSRS_JBA_Categorical

Το αρχείο ACSRS_JBA_Categorical περιλαμβάνει τις βασικές κατηγορικές μεταβλητές που χρησιμοποιήθηκαν στην ανάλυση συστάδων με τον αλγόριθμο K-Modes. Οι μεταβλητές αυτές περιγράφουν μορφολογικά και υπερηχογραφικά χαρακτηριστικά της αθηρωματικής πλάκας, καθώς και την κλινική κατάσταση των ασθενών.

Στο Σχήμα Β.1 παρουσιάζεται ενδεικτικά η δομή του αρχείου και η μορφή των κατηγορικών δεδομένων που χρησιμοποιήθηκαν στην ανάλυση.

	A	B	C	D	E	F	G	H
1	caseID	stenosis	gsm	plarea	dwa	ctiastr	jba	patstat
2	1	medium	low	medium	present	present	high	asympt
3	2	medium	low	medium	present	absent	low	asympt
4	3	medium	medium	low	present	absent	low	asympt
5	4	medium	low	medium	present	absent	high	sympt
6	5	medium	medium	high	present	absent	medium	asympt
7	6	high	medium	medium	present	present	low	asympt
8	7	medium	medium	medium	present	absent	low	asympt
9	8	low	medium	low	present	absent	medium	sympt
10	9	high	low	medium	absent	absent	medium	asympt
11	10	medium	medium	high	present	absent	medium	asympt
12	11	medium	high	medium	present	absent	low	asympt
13	12	medium	high	low	present	absent	low	asympt
14	13	medium	low	medium	absent	absent	vhhigh	asympt
15	14	medium	medium	medium	present	absent	low	asympt
16	15	low	medium	high	present	present	medium	sympt
17	16	medium	high	low	absent	absent	low	asympt
18	17	high	medium	high	present	absent	vhhigh	sympt
19	18	high	high	high	present	absent	vhhigh	sympt
20	19	high	medium	medium	present	absent	vhhigh	sympt
21	20	high	high	low	present	absent	low	sympt
22	21	high	medium	high	present	absent	medium	sympt

Σχήμα Β- 1 Ενδεικτική μορφή του αρχείου ACSRS_JBA_Categorical.

Ενδεικτική μορφή του αρχείου acsrsJba_EventsLastFuAge

Το αρχείο acsrsJba_EventsLastFuAge περιλαμβάνει τις βασικές κατηγορικές και αριθμητικές μεταβλητές που χρησιμοποιήθηκαν στην ανάλυση συστάδων με τον αλγόριθμο K-Prototypes. Οι μεταβλητές αυτές περιγράφουν μορφολογικά και υπερηχογραφικά χαρακτηριστικά της αθηρωματικής πλάκας, καθώς και την κλινική κατάσταση των ασθενών.

Στο Σχήμα Β.2 παρουσιάζεται ενδεικτικά η δομή του αρχείου και η μορφή των κατηγορικών και αριθμητικών δεδομένων που χρησιμοποιήθηκαν στην ανάλυση.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	caseID	stenosis	STN	gsm	GSMN	plarea	PLAREAN	dwa	DWAN	ctiastr	CTIASTRN	jba	JBAN	patstat	events	lastfu	cut72	agecl	age
3	1	medium	85	low	11	medium	59.12	present	1	present	1	high	9.15	asympt	0	14	1	3	76
4	2	medium	75	low	11	medium	53.31	present	1	absent	0	low	3.03	asympt	0	14	0	2	66
7	5	medium	85	medium	20	high	105.21	present	1	absent	0	medium	5.27	asympt	0	13	1	3	76
8	6	high	95	medium	30	medium	58.68	present	1	present	1	low	2.44	asympt	0	5	0	3	71
12	10	medium	75	medium	24	high	80.61	present	1	absent	0	medium	6.08	asympt	0	14	0	2	63
13	11	medium	80	high	40	medium	69.7	present	1	absent	0	low	0.89	asympt	0	14	1	3	73
15	13	medium	80	low	0	medium	78.22	absent	0	absent	0	vhigh	41.01	asympt	0	9	0	2	64
17	15	low	50	medium	25	high	106.27	present	1	present	1	medium	7.63	sympt	2	11	0	2	68
22	21	high	97	medium	26	high	133.56	present	1	absent	0	medium	6.47	sympt	3	6	1	3	76
24	23	high	90	low	4	high	95.45	present	1	absent	0	vhigh	16.94	sympt	2	2	0	2	68
25	25	medium	83	high	49	low	35.04	present	1	absent	0	low	0.38	asympt	0	6	1	3	77
27	29	low	55	low	7	low	35.42	present	1	absent	0	high	8.89	asympt	0	2	0	2	61
28	30	high	90	high	67	medium	68.17	absent	0	absent	0	low	0.23	asympt	0	1	1	3	76
31	34	medium	80	low	9	medium	76.52	present	1	absent	0	high	9.03	asympt	0	3	0	2	67
34	40	low	50	low	12	low	7.1	absent	0	absent	0	medium	5.76	asympt	0	8	1	4	81
35	42	medium	70	high	52	low	12.56	present	1	absent	0	low	0.16	asympt	0	16	0	1	57
36	43	medium	75	medium	22	low	19.2	present	1	absent	0	low	1.17	asympt	0	3	0	1	57
37	44	low	50	high	52	low	24.79	present	1	absent	0	low	1.24	asympt	0	4	1	3	75
39	47	medium	85	high	42	medium	72.12	present	1	present	1	medium	4.32	sympt	3	4	0	2	66
41	55	medium	80	low	13	high	106.16	present	1	absent	0	vhigh	12.09	asympt	0	16	0	1	53
43	58	medium	80	low	14	medium	79.41	absent	0	present	1	vhigh	14.91	asympt	0	5	1	3	73
45	61	high	90	low	7	medium	60.27	present	1	absent	0	vhigh	14.86	asympt	0	5	0	2	68
46	62	medium	80	high	119	low	35.91	absent	0	absent	0	low	0.2	asympt	0	11	1	3	76
47	63	high	90	high	39	medium	47.23	present	1	absent	0	low	2.51	asympt	0	15	1	4	87
50	70	high	95	medium	19	medium	68.64	present	1	present	1	high	9.76	asympt	0	7	0	3	72

Σχήμα Β- 2 Ενδεικτική μορφή του αρχείου acsrsJba_EventsLastFuAge

Παράρτημα Γ

Γ-1: Ορισμός παραμέτρων και χαρακτηριστικών ανάλυσης

Στο παρακάτω απόσπασμα παρουσιάζονται οι βασικές παράμετροι της υλοποίησης του αλγορίθμου K-Modes, καθώς και τα μορφολογικά χαρακτηριστικά που χρησιμοποιήθηκαν στη διαδικασία ομαδοποίησης των ασθενών.

```
51 # =====
52 # CONFIG
53 # =====
54 @dataclass(frozen=True)
55 class Config:
56     input_file: str = "ACSRs_JBA_Categorical.xlsx"
57     base_dir: str = "outputs_k2_full"
58
59     # features used for clustering
60     features: Tuple[str, ...] = ("stenosis", "gsm", "plarea", "dwa", "ctiastr", "jba")
61
62     # optional: keep only rows where patstat is valid
63     filter_patstat: bool = True
64     patstat_col: str = "patstat"
65     patstat_values: Tuple[str, ...] = ("symp", "asympt")
66
67     # kmodes
68     k: int = 2
69     diag_n_init: int = 5 #diagnostic plots (elbow/metrics)
```

Σχήμα Γ- 1 Αρχείο patients_analysis_full_kmodes.py. Ορισμός βασικών παραμέτρων και χαρακτηριστικών που χρησιμοποιήθηκαν στην ανάλυση clustering με τον αλγόριθμο K-Modes.

Γ-2: Παράμετροι αξιολόγησης και σταθερότητας

```
70 km_init: str = "huang"
71 km_n_init: int = 20
72
73 # diagnostics (k search)
74 k_min: int = 2
75 k_max: int = 10
76
77 # evaluation
78 silhouette_sample_size: int = 2000
79 stability_runs: int = 20
80 random_state: int = 42
81
```

Σχήμα Γ- 2 Αρχείο patients_analysis_full_kmodes.py. Παράμετροι αξιολόγησης και σταθερότητας της ανάλυσης clustering.

Γ.3 Παράμετροι διαχωρισμού υποσυνόλων και clustering

```
43 # =====
44 # CONFIG
45 # =====
46 @dataclass(frozen=True)
47 class Config:
48     input_file: str = "acsrsJba_EventsLastFuAge.xlsx"
49     base_dir: str = "outputs_split_kmodes"
50
51     target_col: str = "patstat"
52     lastfu_col: str = "lastfu"
53     lastfu_cutoff: int = 6 # <=6 εξάμηνα = <=3 χρόνια
54
55     features: Tuple[str, ...] = (
56         "stenosis", "gsm", "plarea", "dwa", "ctiastr", "jba"
57     )
58
59     k: int = 2
60     km_init: str = "Huang"
61     km_n_init: int = 20
```

Σχήμα Γ- 3 Αρχείο split_analysis_kmodes.py. Παράμετροι διαχωρισμού symptomatic/asymptomatic ασθενών και ρυθμίσεις clustering.

Γ-4: Διαχωρισμός δεδομένων σε υποομάδες ασθενών

Στο παρακάτω απόσπασμα παρουσιάζεται η διαδικασία διαχωρισμού του συνόλου δεδομένων σε συμπτωματικούς και ασυμπτωματικούς ασθενείς, καθώς και η ξεχωριστή εφαρμογή της ανάλυσης clustering σε κάθε υποομάδα.

```
521
522 logging.info("Reading: %s", cfg.input_file)
523 df = pd.read_excel(cfg.input_file)
524 df = df[df[cfg.target_col].isin(["sympt", "asympt"])]
525
526 df_sympt = df[df[cfg.target_col] == "sympt"].copy()
527 df_asympt = df[df[cfg.target_col] == "asympt"].copy()
528
529 logging.info("Total: %d | Symptomatic: %d | Asymptomatic: %d",
530             len(df), len(df_sympt), len(df_asympt))
531
532 run_group_analysis(df_sympt, "symptomatic", "Symptomatic Patients", cfg)
533 run_group_analysis(df_asympt, "asymptomatic", "Asymptomatic Patients", cfg)
534
535 logging.info("Done. Results in: %s", cfg.base_dir)
```

Σχήμα Γ- 4 αρχείο split_analysis_kmodes.py. Απόσπασμα κώδικα για τον διαχωρισμό του συνόλου δεδομένων σε συμπτωματικούς και ασυμπτωματικούς ασθενείς και την εφαρμογή της ανάλυσης clustering σε κάθε υποομάδα.

Γ-5: Εφαρμογή του αλγορίθμου K-Modes

Στο παρακάτω απόσπασμα παρουσιάζεται η βασική διαδικασία εφαρμογής του αλγορίθμου K-Modes για την ομαδοποίηση των κατηγορικών δεδομένων και την παραγωγή των αναθέσεων συστάδων.

```
334
335 # =====
336 # FIT + STABILIZE
337 # =====
338 def fit_kmodes(X_cat: pd.DataFrame, cfg: Config) -> Tuple[KModes, np.ndarray]:
339     km = KModes(n_clusters=cfg.k, init=cfg.km_init, n_init=cfg.km_n_init, verbose=0)
340     return km, km.fit_predict(X_cat)
341
```

Σχήμα Γ- 5 Αρχείο split_analysis_kmodes.py. Απόσπασμα κώδικα για την εφαρμογή του αλγορίθμου K-Modes και την παραγωγή των αναθέσεων συστάδων.

Γ-6: Υπολογισμός μετρικών αξιολόγησης clustering

Στο παρακάτω απόσπασμα παρουσιάζεται η διαδικασία υπολογισμού των βασικών μετρικών αξιολόγησης που χρησιμοποιήθηκαν για την εκτίμηση της ποιότητας των συστάδων.

```
357 # EVALUATION
358 # =====
359 def evaluate_internal(
360     df_out: pd.DataFrame, features: Sequence[str],
361     cluster_col: str, silhouette_sample_size: int, random_state: int,
362 ) -> Dict[str, Any]:
363     labels = df_out[cluster_col].to_numpy()
364     n, k = len(labels), int(len(np.unique(labels)))
365     X_cat = df_out.loc[:, features].astype(str)
366
367     enc = OrdinalEncoder(handle_unknown="use_encoded_value", unknown_value=-1)
368     X_ord = enc.fit_transform(X_cat).astype(float)
369     X_int = _encode_to_int_matrix(X_cat)
370
371     dbi = float(davies_bouldin_score(X_ord, labels)) if k > 1 else np.nan
372     chi = float(calinski_harabasz_score(X_ord, labels)) if k > 1 else np.nan
```

Σχήμα Γ- 6 αρχείο split_analysys_kmodes.py. Απόσπασμα κώδικα για τον υπολογισμό των μετρικών Davies–Bouldin και Calinski–Harabasz μετά από κωδικοποίηση των κατηγορικών μεταβλητών.

Γ-7: Αξιολόγηση σταθερότητας clustering με χρήση ARI

Στο παρακάτω απόσπασμα παρουσιάζεται η διαδικασία αξιολόγησης της σταθερότητας των αποτελεσμάτων clustering μέσω πολλαπλών εκτελέσεων του αλγορίθμου και υπολογισμού του δείκτη Adjusted Rand Index (ARI).

```
405 def evaluate_stability(  
416     km = KModes(n_clusters=k, init=km_init, n_init=km_n_init,  
417                 verbose=0, random_state=run_i)  
418     yp = km.fit_predict(Xp)  
419     y = np.empty_like(yp)  
420     y[perm] = yp  
421     all_labels.append(y)  
422  
423     aris = [  
424         adjusted_rand_score(all_labels[i], all_labels[j])  
425         for i in range(n_runs) for j in range(i + 1, n_runs)  
426     ]  
427     aris = np.asarray(aris, dtype=float)  
428     return {  
429         "n_runs": n_runs,  
430         "pairwise_ari_mean": float(np.mean(aris)),  
431         "pairwise_ari_min": float(np.min(aris)),  
432         "pairwise_ari_max": float(np.max(aris)),  
433     }
```

Σχήμα Γ- 7 αρχείο split_analysis_kmodes.py. Απόσπασμα κώδικα για την αξιολόγηση της σταθερότητας των συστάδων μέσω πολλαπλών εκτελέσεων και υπολογισμού του δείκτη Adjusted Rand Index (ARI).

Γ-8: Ορισμός κατηγορικών και αριθμητικών χαρακτηριστικών στον K-Prototypes

Στο παρακάτω απόσπασμα παρουσιάζεται ο ορισμός των κατηγορικών και αριθμητικών χαρακτηριστικών που χρησιμοποιήθηκαν στην εφαρμογή του αλγορίθμου K-Prototypes.

```
43 # =====  
44 # CONFIG  
45 # =====  
46 @dataclass(frozen=True)  
47 class Config:  
48     input_file: str = "acsrsJba_EventsLastFuAge.xlsx"  
49     base_dir: str = "outputs_split_kproto"  
50  
51     target_col: str = "patstat"  
52     lastfu_col: str = "lastfu"  
53     lastfu_cutoff: int = 6 # <=6 εξάμηνα = <=3 χρόνια  
54  
55     cat_features: Tuple[str, ...] = (  
56         "stenosis", "gsm", "plarea", "dwa", "ctiastr", "jba"  
57     )  
58     num_features: Tuple[str, ...] = (  
59         "STN", "GSMN", "PLAREAN", "DWAN", "CTIASTRN", "JBAN", "age"  
60     )
```

Σχήμα Γ- 8 αρχείο split_analysis_kproto.py. Απόσπασμα κώδικα με τον ορισμό των κατηγορικών και αριθμητικών μεταβλητών που χρησιμοποιήθηκαν στον αλγόριθμο K-Prototypes.

Γ-9: Κανονικοποίηση αριθμητικών μεταβλητών με RobustScaler

Στο παρακάτω απόσπασμα παρουσιάζεται η διαδικασία προετοιμασίας των δεδομένων για την εφαρμογή του αλγορίθμου K-Prototypes, συμπεριλαμβανομένης της κανονικοποίησης των αριθμητικών μεταβλητών με χρήση της μεθόδου RobustScaler και του διαχωρισμού αριθμητικών και κατηγορικών χαρακτηριστικών.

```
104 def prepare_data(  
105     num_features, cat_features,  
106     df, num_indices, cat_indices  
107 ):  
108     """  
109     KPrototypes expects a single array with numericals first,  
110     then categoricals. Returns array + cat_indices.  
111     Returns only rows with no missing values in selected features.  
112     """  
113     all_features = list(num_features) + list(cat_features)  
114     df_sub = df[all_features].dropna()  
115     # scale numericals  
116     scaler = RobustScaler()  
117     X_num = scaler.fit_transform(df_sub[list(num_features)].values.astype(float))  
118     # categoricals as object  
119     X_cat = df_sub[list(cat_features)].astype(str).values  
120     X = np.hstack([X_num, X_cat])  
121     cat_indices = list(range(len(num_features), len(all_features)))
```

Σχήμα Γ- 9 αρχείο analysis_kprototypes.py. Απόσπασμα κώδικα για την προεπεξεργασία, κανονικοποίηση και προετοιμασία των δεδομένων πριν από την εφαρμογή του αλγορίθμου K-Prototypes.

Γ-10: Εφαρμογή του αλγορίθμου K-Prototypes

Στο παρακάτω απόσπασμα παρουσιάζεται η διαδικασία εφαρμογής του αλγορίθμου K-Prototypes, καθώς και ο υπολογισμός των βασικών μετρικών αξιολόγησης για διαφορετικές τιμές του αριθμού συστάδων.

```
322  
323     for kk in ks:  
324         kp = KPrototypes(n_clusters=kk, init=km_init, n_init=km_n_init, verbose=0)  
325         labels = kp.fit_predict(X, categorical=cat_indices)  
326         costs.append(float(kp.cost_))  
327  
328         if len(np.unique(labels)) <= 1:  
329             silhouettes.append(np.nan); dbis.append(np.nan); chis.append(np.nan)  
330             continue  
331  
332         silhouettes.append(float(silhouette_score(X_num, labels)))  
333         dbis.append(float(davies_bouldin_score(X_num, labels)))  
334         chis.append(float(calinski_harabasz_score(X_num, labels)))
```

Σχήμα Γ- 10 αρχείο analysis_kprototypes.py. Απόσπασμα κώδικα για την εφαρμογή του αλγορίθμου K-Prototypes και τον υπολογισμό των μετρικών αξιολόγησης clustering.

Γ-11: Κωδικοποίηση μεταβλητών και παραμετροποίηση της TE2Rules

Στο παρακάτω απόσπασμα παρουσιάζεται η διαδικασία χειροκίνητης κωδικοποίησης των κατηγορικών μεταβλητών σε αριθμητική μορφή, καθώς και οι βασικές παράμετροι που χρησιμοποιήθηκαν για την εφαρμογή της μεθόδου TE2Rules και του μοντέλου Random Forest.

```
11 from te2rules.explainer import ModelExplainer
12
13 MAPPINGS = {
14     "stenosis": {"low": 0, "medium": 1, "high": 2},
15     "gsm":      {"low": 0, "medium": 1, "high": 2},
16     "plarea":   {"low": 0, "medium": 1, "high": 2},
17     "dwa":      {"absent": 0, "present": 1},
18     "ctiastr":  {"absent": 0, "present": 1},
19     "jba":      {"low": 0, "medium": 1, "high": 2, "vhigh": 3},
20 }
21
22 FEATURES      = ["stenosis", "gsm", "plarea", "dwa", "ctiastr", "jba"]
23 K             = 2
24 RF_N_ESTIMATORS = 50
25 RF_MAX_DEPTH   = 5
26
```

Σχήμα Γ- 11 αρχείο te2rules_full.py. Απόσπασμα κώδικα για την κωδικοποίηση των κατηγορικών μεταβλητών και την παραμετροποίηση της TE2Rules.

Γ-12: Εφαρμογή της μεθόδου TE2Rules και εξαγωγή κανόνων

Στο παρακάτω απόσπασμα παρουσιάζεται η διαδικασία εκπαίδευσης του μοντέλου Random Forest και η εφαρμογή της μεθόδου TE2Rules για την εξαγωγή ερμηνεύσιμων κανόνων τύπου «if-then» για κάθε συστάδα, ακολουθώντας την προσέγγιση one-vs-rest.

```
71 def run_te2rules_for_cluster(X_values, y_binary, cluster_id, out_dir, label):
72     model = RandomForestClassifier(
73         n_estimators=RF_N_ESTIMATORS,
74         max_depth=RF_MAX_DEPTH,
75         random_state=42,
76     )
77     model.fit(X_values, y_binary)
78     explainer = ModelExplainer(model, feature_names=FEATURES)
79     rules = explainer.explain(X_values, y_binary)
80
81     txt_path = os.path.join(out_dir, f"te2rules_cluster_{cluster_id}_vs_rest.txt")
82     with open(txt_path, "w", encoding="utf-8") as f:
83         f.write(f"{' '*60}\n{label}\nRULES FOR CLUSTER {cluster_id} vs REST\n{' '*60}\n\n")
84         f.write(f"Number of rules found: {len(rules)}\n\n")
85         for i, r in enumerate(rules, 1):
86             f.write(f"Rule {i}:\n{r}\n\n")
87
```

Σχήμα Γ- 12 αρχείο te2rules_full.py. Απόσπασμα κώδικα για την εφαρμογή της TE2Rules και την εξαγωγή κανόνων ερμηνευσιμότητας ανά συστάδα.

Παράρτημα Δ

Πιο κάτω ακολουθεί το link για το GitHub repository της εργασίας, με όλους τους κώδικες και τα αποτελέσματα από την ανάλυση δεδομένων.

<https://github.com/FlorentiaPattidou/medical-data-clustering-analysis/tree/main>