ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

ΚΑΙ

ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

"ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΕ ΒΙΟΪΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΩΑΝΝΑ ΑΓΓΕΛΙΔΟΥ

Επιβλέπουσα Καθηγήτρια

Φιλία Βόντα

Αναπληρώτρια Καθηγήτρια Ε.Μ.Π.

Αθήνα, Ιούνιος 2018

ΕΥΧΑΡΙΣΤΙΕΣ

Μέσα από την παρούσα διπλωματική εργασία θα ήθελα να ευχαριστήσω ιδιαιτέρως την επιβλέπουσα καθηγήτρια, κ. Φιλία Βόντα, τόσο για το ότι μου ανέθεσε ένα τόσο επιστημονικά ενδιαφέρον θέμα, όσο και για την πολύτιμη βοήθεια που μου προσέφερε, σε όλα τα στάδια της εργασίας. Επίσης, θα ήταν παράλειψη αν δεν ευχαριστούσα την οικογένειά μου για την στήριξη που μου προσέφερε στα χρόνια των σπουδών μου στο Πολυτεχνείο και φυσικά στο στάδιο της διπλωματικής.

ΠΕΡΙΛΗΨΗ

Στην παρούσα διπλωματική εργασία, παρουσιάζεται η μέθοδος «Ανάλυσης κατά Συστάδες». Στόχος της μεθόδου αυτής είναι να ομαδοποιήσει τις παρατηρήσεις σε συστάδες (clusters) με τέτοιο τρόπο, ώστε κάθε συστάδα να παρουσιάζει μεγάλη ομοιογένεια.

Στο πρώτο μέρος της εργασίας, γίνεται μια θεωρητική παρουσίαση της μεθόδου. Αρχικά, περιγράφονται οι διάφορες τεχνικές της Συσταδοποίησης και αναλύονται οι πολλαπλές εφαρμογές της Ανάλυσης κατά Συστάδες στις διάφορες επιστήμες. Στη συνέχεια, ακολουθεί μια εκτενής επεξήγηση των μαθηματικών εργαλείων που χρησιμοποιούνται στις τεχνικές αυτής της μεθόδου. Έχοντας αναφερθεί, λοιπόν, σε όλες τις απαραίτητες γνώσεις που πρέπει να έχει κάποιος προς ανάγνωση και κατανόηση της μεθόδου, αναλύονται διεξοδικά οι πιο γνωστές μέθοδοι Ανάλυσης σε Συστάδες, δηλαδή η Ιεραρχική Ανάλυση και η Μη Ιεραρχική Ανάλυση.

Στο δεύτερο μέρος της εργασίας, εφαρμόζουμε την Ταξινόμηση κατά Συστάδες, χρησιμοποιώντας το στατιστικό πρόγραμμα R-Studio, στο «Σετ δεδομένων έκφρασης πρωτεΐνης ποντικών», το οποίο δημιουργήθηκε από πειράματα από τους Higuera et al και Ahmed et al. Στο πείραμα συμμετείχαν ποντίκια που έπασχαν από σύνδρομο Down (Down Syndrome- DS) (το DS προκαλείται από την παρουσία ενός επιπρόσθετου χρωμοσώματος, την τρισωμία) και υγιή ποντίκια (control). Ορισμένα ποντίκια διεγέρθηκαν στη μάθηση και άλλα όχι, ενώ σε κάποια από αυτά χορηγήθηκε μεμαντίνη για να διαπιστωθεί αν μπορούν να ανακτήσουν την ικανότητα της μάθησης. Έτσι, τα ποντίκια χωρίστηκαν σε οκτώ κλάσεις. Σκοπός της εργασίας αυτής είναι να διαπιστώσει εάν η γνώση των πρωτεϊνών οδηγεί από μόνη της στο διαχωρισμό των ποντικιών σε κλάσεις, εάν διαχωρίζονται τα υγιή από τα DS ποντίκια ή εάν διαχωρίζονται με κάποιο άλλο τρόπο. Έτσι, μέσα από όλη αυτή τη διαδικασία προκύπτουν συμπεράσματα για το συγκεκριμένο σύνολο δεδομένων, αλλά και γενικά για τη μέθοδο της Ανάλυσης σε Συστάδες και τη χρήση της.

The present thesis will proceed to examine the method of “Cluster Analysis”. The goal of this method is to group the observations in clusters in such a way that each cluster would display homogeneity to a relatively great extent.

In the first part of this thesis, the above mentioned method is examined from a theoretical point of view. Firstly, the different techniques of clustering are utterly described and the multiple applications of Cluster Analysis in different disciplines are mentioned. Furthermore, a detailed demonstration of the mathematical tools that are used in the techniques of this method follows. Having, thus, already referred to all the necessary background one shall have in order to read and understand this method, the most well-known methods of Cluster Analysis will be extensively examined, namely the Hierarchical Analysis and the Non-Hierarchical Analysis.

In the second part of this thesis, we apply the Cluster Classification using the R-Studio statistical program to the "Mouse Protein Expression Data Set", which was created by experiments by Higuera et al. and Ahmed et al. The experiment involved mice suffering from Down Syndrome (DS) (the DS is caused by the presence of an additional chromosome, trisomy) and healthy mice (control). Some mice were stimulated to learn and others did not, while some were given memantine to see their ability to adapt to learning. Thus, the mice were divided into eight classes. The purpose of this work is to find out whether protein-only knowledge leads to the separation of mice in these classes, if in general the healthy mice are separated from the DS mice or if they are separated differently. Thus, through this procedure, conclusions for this specific data set but also in general for the method of Cluster Analysis and its use, are drawn.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η επιτακτική ανάγκη των σύγχρονων επιστημονικών μελετών της από κοινού διερεύνησης της επίδρασης μεγάλου όγκου μεταβλητών, οι οποίες μετρώνται σε ένα δείγμα οδήγησε στην ανακάλυψη ειδικών τεχνικών για την επεξεργασία των στοιχείων, τις επονομαζόμενες *πολυμεταβλητές τεχνικές (Multivariate techniques)*. Οι μελέτες αυτές μπορεί να αφορούν στην έρευνα αγοράς στις προτιμήσεις των καταναλωτών, στα οργανοληπτικά χαρακτηριστικά τροφίμων, στα βιολογικά χαρακτηριστικά ενός πληθυσμού (πχ μετρήσεις του ύψους και του βάρους κάποιων ατόμων) κα. Με άλλα λόγια, αντικείμενο της Πολυμεταβλητής Ανάλυσης είναι οι στατιστικές μέθοδοι συλλογής, περιγραφής και ανάλυσης δεδομένων πολλών μεταβλητών σε ένα πλήθος ατόμων ή γενικότερα πειραματικών μονάδων (Πετρίδης, 2015).

Αυτή η ανάγκη της ανάπτυξης των πολυμεταβλητών τεχνικών μπορεί φυσικά να ακολουθεί κάποιες θεωρητικές προσεγγίσεις, αλλά μπορεί καμιά φορά να αφορά απλά στην προσπάθεια κατανόησης και στην ταξινόμηση μεγάλου αριθμού δεδομένων. Όταν, για παράδειγμα, έχουμε σαν δείγμα στη διάθεσή μας πολύ μεγάλο όγκο πληροφοριών, τότε συχνά έχουμε πλεονασμό πληροφορίας. Αυτό συμβαίνει γιατί το ίδιο χαρακτηριστικό των παρατηρήσεων περιγράφεται και μετράται από πολλές μεταβλητές. Γι' αυτό και το πρώτο βήμα στην ανάλυσή μας είναι να προσπαθήσουμε να βρούμε και να κατανοήσουμε μια δομή στα δεδομένα, το οποίο επιτυγχάνεται με την ταξινόμηση. Οι πολυμεταβλητές μέθοδοι, οι οποίες ασχολούνται με την παραπάνω διαδικασία και πετυχαίνουν τη συνοπτική παρουσίαση των δεδομένων ενός δείγματος, διευκρινίζοντας όμως παράλληλα και βασικές συσχετίσεις και διαστάσεις μεταξύ τους, ονομάζονται στη βιβλιογραφία *τεχνικές μείωσης των δεδομένων (data reduction techniques)* ή και πιο απλά *μέθοδοι ταξινόμησης (classification methods)* (Ηλιοπούλου, 2015).

Πρέπει να τονιστεί εδώ, όμως, ότι η ταξινόμηση αυτή, δεν συνεπάγεται τη διαίρεση ενός ομοιογενούς συνόλου δεδομένων απλά σε υποομάδες, αλλά τον καθορισμό ομάδων, οι οποίες ανταποκρίνονται στην πραγματικότητα. Επίσης, είναι σημαντικό να καταλάβει ο αναγνώστης ότι δεν οδηγούμαστε πάντα στην ίδια ταξινόμηση και πάντα θα υπάρχει μια ποικιλία από εναλλακτικές ταξινομήσεις για το ίδιο σύνολο δεδομένων. Τα ανθρώπινα όντα, για

παράδειγμα, μπορούν να ταξινομηθούν σε αρσενικά και θηλυκά, ή με βάση το μορφωτικό επίπεδο, την κοινωνική τάξη, την ηλικία, ακόμα και το χρώμα δέρματος. Ο κατάλογος είναι ατελείωτος και είναι προφανές ότι το είδος της ταξινόμησης που προκύπτει από μια ανάλυση εξαρτάται σε μεγάλο βαθμό από τις μεταβλητές που χρησιμοποιήθηκαν για να αναπαραστήσουν το αντικείμενο. Αυτό είναι ένα κρίσιμο σημείο, δεδομένου ότι μια κακή επιλογή μεταβλητών μπορεί να οδηγήσει σε μια κακή ομαδοποίηση για το συγκεκριμένο σκοπό. Επίσης, διαφορετικές ταξινομήσεις δεν μπορούν να συλλέξουν ίδιο σύνολο ατόμων σε ομάδες. Στις πρακτικότερες εφαρμογές όμως της ανάλυσης σε ομάδες, ο ερευνητής γνωρίζει αρκετά για το πρόβλημα που έχει να αντιμετωπίσει και έχει πλέον αρκετά εργαλεία στα χέρια του, ώστε να διακρίνει τις καλές από τις κακές ομαδοποιήσεις (Καράγεωργα, 2012).

Με την πρόοδο της τεχνολογίας, την επινόηση και τη χρήση στατιστικών λογισμικών προγραμμάτων Η/Υ, οι πολυμεταβλητές τεχνικές έγιναν πολύ γρήγορα δημοφιλείς. Η συνεισφορά των τεχνικών αυτών προάγει ιδιαιτέρως τα αποτελέσματα της έρευνας, προσδίδοντας επιστημονική καταξίωση και υψηλή ικανοποίηση στον εκάστοτε ερευνητή.

Οι πολυμεταβλητές μέθοδοι, λοιπόν, περιλαμβάνουν πολλές τεχνικές, αναλόγως του προβλήματος που προσπαθούν να αντιμετωπίσουν και ο κάθε οργανωτής μιας έρευνας καλείται να αποφασίσει για τα συγκεκριμένα δεδομένα ποιά είναι η βέλτιστη από αυτές. Η *Παραγοντική Ανάλυση (factor analysis), η Ανάλυση Κύριων Συνιστωσών, η Διαχωριστική Ανάλυση (discriminant analysis) και η Ανάλυση κατά Συστάδες (cluster analysis)* είναι οι πιο γνωστές από αυτές. Εμείς στην παρούσα διπλωματική εργασία θα ασχοληθούμε με την Ανάλυση κατά Συστάδες.

1 ΚΕΦΑΛΑΙΟ: ΕΙΣΑΓΩΓΗ

1.1 ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ

Η Ανάλυση Συστάδων (Cluster Analysis) είναι μια από τις πιο βασικές μεθόδους Ανάλυσης Δεδομένων (Norušis, 2011). Στόχος της είναι ο διαμερισμός ενός συνόλου δεδομένων σε υποσύνολα, τις συστάδες. Κάθε συστάδα πρέπει να είναι ομοιογενής ώστε οι παρατηρήσεις που ανήκουν σε αυτήν να είναι όμοιες μεταξύ τους και κάθε συστάδα πρέπει να διαφέρει από μια άλλη με βάση κάποια χαρακτηριστικά (Sharma, 1996). Δηλαδή, αντικείμενα ή παρατηρήσεις που ανήκουν στην ίδια συστάδα έχουν παρόμοια χαρακτηριστικά (Μαύρου, 2012). Μιας και οι μεταβλητές στα δεδομένα είναι οι στήλες και οι παρατηρήσεις οι γραμμές, η Ανάλυση Συστάδων ουσιαστικά ομαδοποιεί παρόμοιες γραμμές.

Τα δεδομένα στη μέθοδο αυτή μπορούν να είναι είτε ποσοτικά, είτε ποιοτικά, όμως για τα τελευταία απαιτούνται εξειδικευμένες τεχνικές. Αφού η Ανάλυση Συστάδων είναι μια διαδικασία ομαδοποίησης ενός συνόλου δεδομένων με βάση ένα μέτρο ομοιότητας, γίνεται αντιληπτό ότι η ταξινόμηση αυτή εξαρτάται κατά πολύ από το είδος των δεδομένων και την εφαρμογή της μεθόδου. Δεδομένα διαφορετικών εφαρμογών είναι πιθανό να ταξινομηθούν διαφορετικά. Γι' αυτό, η Ανάλυση Συστάδων μπορεί να γίνει αρκετά περίπλοκη για τον ερευνητή, μιας και θα πρέπει να αποφασίσει ποιός είναι ο πιο κατάλληλος αλγόριθμος για την κάθε εφαρμογή. Για την απόφαση αυτή είναι αναγκαία η γνώση των δεδομένων και η γνώση αυτή χρησιμοποιείται ξανά και ξανά στα διάφορα βήματα της διαδικασίας.

Το πρώτο και πιο βασικό βήμα στη μέθοδο αυτή είναι αρχικά η περιγραφή και η επιλογή των κατάλληλων χαρακτηριστικών στα δεδομένα και έπειτα ο καθορισμός του μέτρου ομοιότητας με το οποίο θα γίνουν οι διάφορες συγκρίσεις ανάμεσα στις παρατηρήσεις. Σε δεύτερη φάση, πρέπει να εξετάσουμε και να καθορίσουμε το είδος της τεχνικής ταξινόμησης που θα χρησιμοποιήσουμε (Ιεραρχική ή Μη). Έπειτα, θα πρέπει να αποφασίσουμε τη

μέθοδο ένωσης των συστάδων και τέλος, θα πρέπει να ερμηνευτεί η λύση του αλγορίθμου για την τελική παραγωγή των ομάδων (Ηλιοπούλου, 2015).

Κάποια σημαντικά θέματα πρέπει να διευκρινιστούν σε αυτό το σημείο. Αρχικά, αναλόγως το μέτρο ομοιότητας και τη μέθοδο ομαδοποίησης που πρόκειται να επιλεχθούν, οι συστάδες που προκύπτουν είναι μάλλον διαφορετικές και γι' αυτό ο ερευνητής πρέπει να λάβει υπόψη του τη φύση των δεδομένων, καθώς και το πρόβλημα το οποίο εξετάζει. Πολλές φορές απαιτούνται αρκετές δοκιμές της Ανάλυσης Συστάδων προσθαφαιρώντας κάποιες μεταβλητές ή χρησιμοποιώντας άλλα μέτρα ομοιότητας κάθε φορά, ώστε να εξακριβωθεί η σταθερότητα της ταξινόμησης.

Επίσης, όταν χρησιμοποιούμε σαν μέθοδο ομαδοποίησης μια Ιεραρχική μέθοδο, σημαντικό ρόλο παίζει ο πίνακας αποστάσεων (πίνακας ομοιότητας) $O = (dij)$, ο οποίος είναι συμμετρικός και έχει ως στοιχεία του τις αποστάσεις ανάμεσα στις παρατηρήσεις $X_1, \ldots, X_n$, αντί τις ίδιες τις παρατηρήσεις.

Τέλος, όπως είπαμε, το τελικό αποτέλεσμα πρέπει να ερμηνευτεί. Στη μέθοδο της Ανάλυσης Συστάδων δίνονται ονομασίες στις συστάδες που προκύπτουν. Για το λόγο αυτό, μελετώνται οι τιμές των μεταβλητών στην κάθε συστάδα και βάσει της εμπειρίας του κάθε ερευνητή, διαπιστώνεται εάν οι συστάδες αποτελούν ένα απλό αποτέλεσμα του αλγορίθμου ή εάν όντως υπάρχουν στην πραγματικότητα.

Συνοψίζοντας, τα βήματα που ακολουθεί η Ανάλυση Συστάδων για τη δημιουργία ομάδων είναι τα εξής (Everitt, Landau, Leese, & Stahl, 2011)**:**

1. επιλογή των αντικειμένων που αποτελούν αντιπροσωπευτικό δείγμα προς ομαδοποίηση
2. επιλογή των μεταβλητών ανάλογα με το σκοπό της ανάλυσης
3. αντιμετώπιση τυχόν ελλειπουσών τιμών
4. τυποποίηση των μεταβλητών για καλύτερη σύγκριση
5. επιλογή του μέτρου ομοιότητας
6. επιλογή της μεθόδου ένωσης των συστάδων
7. προσδιορισμός του αριθμού των συστάδων (στις Μη Ιεραρχικές μεθόδους)
8. συμπεράσματα για τις σχηματισμένες συστάδες

Τα βήματα της μεθόδου θα γίνουν πιο σαφή στα παρακάτω κεφάλαια.

1.2 ΚΑΤΗΓΟΡΙΕΣ ΜΕΘΟΔΩΝ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΣΤΑΔΩΝ

Στην επιστημονική βιβλιογραφία υπάρχουν πολλές διαφορετικές μέθοδοι Ανάλυσης Συστάδων. Οι διαφορές των μεθόδων αυτών έγκεινται στις επαγωγικές τους αρχές αλλά και στον τρόπο σχηματισμού των ομάδων. Ένας βασικός λόγος ύπαρξης τόσο μεγάλου αριθμού μεθόδων για την τεχνική αυτή είναι το ότι δεν υπάρχει κάποιος αυστηρός ορισμός της έννοιας της συστάδας (Estivill-Castro & Yang, 2000).

Οι Han, Kamber και Pei (2011) μιλούν για πέντε κατηγορίες μεθόδων Ανάλυσης Συστάδων:

Ιεραρχικές Μέθοδοι (hierarchical methods) Οι μέθοδοι αυτές εκτελούν μια διαδικασία συνεχών διασπάσεων ή συγχωνεύσεων συστάδων. Προχωρούν δηλαδή ή με διαδοχικές διαιρέσεις ή με διαδοχικές ενώσεις των παρατηρήσεων για να επιτύχουν την επιθυμητή συσταδοποίηση.

Διαχωριστικές μέθοδοι (partitioning methods) Οι μέθοδοι αυτές διαχωρίζουν τα αντικείμενα σε k συστάδες. Σε αυτή την περίπτωση, το πλήθος των ομάδων καθορίζεται από τον ερευνητή. Εφαρμόζεται μια συνεχής διαδικασία κατά την οποία τα αντικείμενα του δείγματος μετακινούνται από τη μια ομάδα στην άλλη μέχρι να μην υπάρχει καμία αλλαγή. Η ποιότητα της μεθόδου για την κάθε λύση πιθανών συστάδων ελέγχεται με τη βοήθεια ενός κριτηρίου, η τιμή του οποίου σε κάθε επανάληψη μειώνεται. Ο πιο διάσημος αλγόριθμος διαχωριστικής μεθόδου είναι ο *k- Means*.

Μέθοδοι βασισμένες στην πυκνότητα (density based methods) Στις μεθόδους αυτές οι αλγόριθμοι ελέγχουν την πυκνότητα των παρατηρήσεων στο χώρο και αναλόγως δημιουργούνται οι ομάδες, οι οποίες καλύπτουν τα πυκνά σημεία του χώρου. Η γειτονιά κάθε παρατήρησης που ανήκει σε μια ομάδα είναι συγκεκριμένης διαμέτρου και πρέπει να περιλαμβάνει έναν συγκεκριμένο αριθμό παρατηρήσεων. Η κάθε ομάδα συνεχίζει να μεγαλώνει όσο η γειτονιά των παρακείμενων παρατηρήσεων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι βασισμένες στην πυκνότητα είναι γνωστές για τη δημιουργία μη

κυρτών, περίπλοκων σχημάτων και για τη δυνατότητά τους να απομονώνουν τις ακραίες τιμές.

Μέθοδοι πλέγματος (grid based methods) Οι μέθοδοι αυτές χωρίζουν το χώρο του δείγματος σε κελιά, τα οποία αποτελούν ένα πλέγμα. Οι παρατηρήσεις πλέον αντιπροσωπεύονται από τα κελιά στα οποία ανήκουν και η αναζήτηση για δημιουργία συστάδων γίνεται στα κελιά και όχι στις παρατηρήσεις. Στις μεθόδους πλέγματος, το πλήθος των κελιών είναι αυτό που καθορίζει το χρόνο επεξεργασίας και όχι το πλήθος των παρατηρήσεων. Επειδή τα κελιά είναι πολύ λιγότερα από τις παρατηρήσεις κατά κανόνα, οι μέθοδοι αυτές είναι πολύ γρήγορες. Ένα σημαντικό θέμα προς συζήτηση όμως στις μεθόδους αυτές είναι και ο καθορισμός κατάλληλου μεγέθους στα κελιά.

Μέθοδοι βασισμένες σε μοντέλα (model based methods) Όπως υπονοεί και το όνομά των μεθόδων αυτών, σε αυτή την περίπτωση γίνεται χρήση μοντέλων. Στόχος των μοντέλων αυτών είναι η βελτιστοποίηση της προσαρμογής μεταξύ δεδομένων και μοντέλων. Το μοντέλο εκπαιδεύεται για τη μη συμμετοχή των αντικειμένων σε ομάδες με μη επιβλεπόμενη μάθηση. Η πιο γνωστή μέθοδος από αυτές τις τεχνικές είναι οι επονομαζόμενοι Αυτοοργανούμενοι Χάρτες (Self Organizing Maps), ένας ειδικός δηλαδή τύπος νευρωνικών δικτύων.

✓ Ο k-means και οι Ιεραρχικοί αλγόριθμοι είναι οι πιο διαδεδομένοι αλγόριθμοι.

1.3 ΕΦΑΡΜΟΓΕΣ

Η Ανάλυση Συστάδων είναι μία οικογένεια μεθόδων ταξινόμησης, οι εφαρμογές της οποίας απλώνονται σε πολλές επιστήμες. Οι πρώτες εφαρμογές προέρχονται από τη Ζωολογία, όπου η μέθοδος λέγεται συνήθως και αριθμητική ταξινομία (numerical taxonomy), και τη Βιολογία (ταξιδόμηση των ειδών αναλόγως με τα διάφορα περιβαλλοντικά γεγονότα ή ομαδοποίηση πρωτεϊνών και γονιδίων που έχουν ίδια λειτουργία). Άλλες σημαντικές εφαρμογές βρίσκουμε στην Ιατρική (ομαδοποίηση ασθενών, συμπτωμάτων ή θεραπειών, ανάλυση εικόνας), στην Ψυχιατρική (ορθή διάγνωση των

συμπτωμάτων της σχιζοφρένειας και της παράνοιας για πιο επιτυχημένη και ουσιαστική θεραπεία), στη Βιοπληροφορική και στην τεχνητή νοημοσύνη για την αναγνώριση προτύπων (pattern recognition) (Everitt et al., 2011), στα Νευρωνικά Δίκτυα (neural networks), στη μάθηση μηχανών (machine learning), στην άντληση και διερεύνηση δεδομένων (data mining), το Ίντερνετ (ομαδοποίηση σχετιζόμενων αρχείων για browsing, ταξινόμηση weblog για εύρεση παρόμοιων προτύπων προσπέλασης), στην Αστρονομία, στην Κλιματολογία (η κατανόηση του κλίματος στον πλανήτη μας απαιτεί την εύρεση μοτίβων στην ατμόσφαιρα και τους ωκεανούς- η Ανάλυση Συστάδων βρίσκει μοτίβα στην ατμοσφαιρική πίεση των πολικών περιοχών και σε σημεία των ωκεανών που επηρεάζουν σημαντικά το εδαφικό κλίμα) (Caccam & Refran, 2012), ακόμα και στην Αρχαιολογία (πέτρινα αγγεία και εργαλεία διαφορετικών περιόδων). (Ηλιοπούλου, 2015; Πετρίδης, 2015)

Επίσης, ένας κλάδος στον οποίο χρησιμοποείται ιδιαιτέρως η Ανάλυση Συστάδων είναι η διαφήμιση και ιδιαίτερα η έρευνα- τμηματοποίηση της αγοράς. Με τον όρο τμηματοποίηση της αγοράς εννοούμε την ταξινόμηση- διαχωρισμό των καταναλωτών σε ομάδες με βάση την καταναλωτική τους συμπεριφορά.

Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η μέθοδος αυτή για τους γεωγράφους, καθώς εφαρμόζεται σε προβλήματα οριοθέτησης περιφερειών (Johnston, 1980). Πιο συγκεκριμένα, η Ανάλυση Συστάδων εφαρμόζεται σε θέματα που έχουν να κάνουν με τη χωρική πληροφορία και την ανάλυσή της. Τα αντικείμενα σε αυτή την περίπτωση είναι γεωγραφικές περιοχές με μεγάλο όγκο χαρακτηριστικών και η ομαδοποίηση καταλήγει στον καθορισμό ομοιογενών περιφερειών, με τέτοιο τρόπο ώστε οι παρατηρήσεις με πανομοιότυπα χαρακτηριστικά να βρίσκονται στην ίδια περιφέρεια. Κάτι τέτοιο είναι εξαιρετικά χρήσιμο αφού:

- Οι κατοικίες μπορούν να ταξινομηθούν ανάλογα με τη γεωγραφική θέση, την αξία και τον τύπο τους.
- Θεματικοί χάρτες, σχεδιασμένοι μέσω αυτής της μεθόδου, αναγνωρίζουν τμήματα γης με όμοια χρήση, όπως για παράδειγμα αστικές, βιομηχανικές ή αγροτικές περιοχές. Τέτοια στοιχεία μπορεί να φανούν ιδιαιτέρως χρήσιμα στους αρμόδιους κρατικούς φορείς για τη δημιουργία υποδομών και την άσκηση πολιτικής.

- Η μέθοδος αυτή βοηθάει στην ομαδοποίηση χωρικών δεδομένων (δεδομένα τηλεσκόπησης) (Μηλιαρέσης & Ηλιοπούλου, 2004).
- Τέλος, μια εξαιρετικά χρήσιμη εφαρμογή της μεθόδου Ανάλυσης Συστάδων είναι και ο εντοπισμός ομάδων παρόμοιων επιχειρήσεων. Μια τέτοια ομάδα αποτελούν επιχειρήσεις στον ίδιο γεωγραφικό χώρο, οι οποίες ενώνονται με κάποια κοινά χαρακτηριστικά και λειτουργούν μεταξύ τους συμπληρωματικά ή/και ανταγωνιστικά. Η πιο διάσημη περίπτωση μιας τέτοιας επιχειρηματικής συστάδας είναι αυτή της Silicon Valley. Μια τέτοια συγκέντρωση επιχειρήσεων μπορεί να αποφέρει πολλά οφέλη, όπως η ανάπτυξη ειδικών υποδομών, χρήσιμων για τον συγκεκριμένο κλάδο, η διάδοση νέων τεχνολογιών και βέλτιστων πρακτικών, η μείωση των εξόδων μεταφοράς, ο μεγάλος ανταγωνισμός, ο οποίος συμβάλλει στη βελτίωση της ποιότητας, η επίτευξη μιας οικονομίας σε κλίμακα, καθώς και η κοινή αξιοποίηση του τοπικού εργατικού δυναμικού, που αποκτά σταδιακά εξειδικευμένες δεξιότητες και γνώσεις (Κύρκος, 2015).

1.4 ΧΡΗΣΙΜΟΤΗΤΑ

Ως αυτόνομη αναλυτική εργασία, η μέθοδος της Ανάλυσης Συστάδων επιτρέπει στον ερευνητή να ταξινομήσει τα δεδομένα σε ομάδες όμοιων παρατηρήσεων. Έπειτα, ο αναλυτής μπορεί να επικεντρωθεί στην κάθε συστάδα, να αναγνωρίσει τα κοινά χαρακτηριστικά που τη δημιούργησαν και να εξάγει χρήσιμη γνώση και συμπεράσματα για τη λήψη αποφάσεων που αφορούν το εκάστοτε πρόβλημα.

Ανεξάρτητα όμως από την αξία της ως αυτόνομο εργαλείο ανάλυσης δεδομένων, η μέθοδος αυτή είναι ικανή σε συνδυασμό με άλλες Αναλύσεις Δεδομένων να αποτελέσει ενδιάμεσο στάδιο προεπεξεργασίας, μιας και οι αλγόριθμοί της μπορούν να ταξινομούν τις παρατηρήσεις σύμφωνα με την ομοιότητά τους. Έτσι, μπορεί να χρησιμοποιηθεί στον εντοπισμό ακραίων τιμών (outliers) (Ng & Han, 1994; Shekhar & Chawla, 2003). Τα outliers απομακρύνονται από το σύνολο των δεδομένων και έτσι προκύπτει ένα καλύτερο σύνολο, βελτιωμένο.

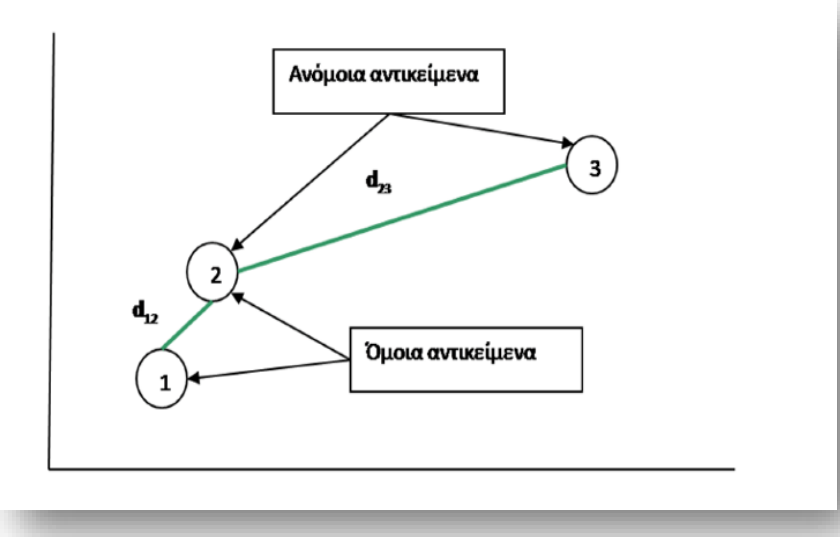
2 ΚΕΦΑΛΑΙΟ: ΜΑΘΗΜΑΤΙΚΑ ΕΡΓΑΛΕΙΑ

2.1 ΑΠΟΣΤΑΣΗ

Στην Ανάλυση Συστάδων, στην οποία οι παρατηρήσεις ομαδοποιούνται με βάση την ομοιότητά τους, είναι προφανές ότι ένα από τα βασικότερα ζητήματα που πρέπει να αποφασιστούν είναι ο καθορισμός των μέτρων ομοιότητας που θα χρησιμοποιηθούν. Ένας τέτοιος τρόπος καθορισμού του βαθμού ομοιότητας είναι η απόσταση.

Η απόσταση αποτελεί για την ανάλυση δεδομένων θεμελιώδη έννοια της πολυμεταβλητής και όχι μόνο ανάλυσης. Κύριος στόχος της είναι να υπολογίσει πόσο απέχουν δύο παρατηρήσεις, να διαπιστώσει δηλαδή το εάν μοιάζουν ή όχι οι παρατηρήσεις αυτές.

Έστω η απλή περίπτωση, όπου η κάθε παρατήρηση έχει δύο μόνο (αριθμητικές) μεταβλητές X και Y. Έτσι, κάθε παρατήρηση αναπαριστάται στο χώρο XY ως ένα σημείο. Σύμφωνα με τον ορισμό της απόστασης, δυο σημεία τα οποία βρίσκονται «κοντά» στο επίπεδο είναι όμοια, ενώ αυτά που είναι μακριά είναι ανόμοια. Στο παρακάτω σχήμα απεικονίζονται τρία σημεία στον χώρο XY. Όπως φαίνεται, τα σημεία 1 και 2 είναι όμοια, ενώ τα 2 και 3 ανόμοια (Κύρκος, 2015).



Εικόνα 1 Απόσταση σημείων στο χώρο XY

Όταν οι παρατηρήσεις έχουν n μεταβλητές, τότε αυτές είναι σημεία στον χώρο των n διαστάσεων.

Στην πραγματικότητα, επειδή τα δεδομένα που εξετάζουμε κάθε φορα δεν έχουν μόνο αριθμητικά γνωρίσματα, έχουν αναπτυχθεί κατάλληλα μέτρα υπολογισμού της απόστασης ανάλογα με το εάν τα γνωρίσματα αυτά είναι αριθμητικά, δυαδικά ή ονομαστικά.

2.1.1 Απόσταση αριθμητικών μεταβλητών

Έστω παρατηρήσεις με n μεταβλητές- γνωρίσματα η κάθε μια. Η απόσταση των x_a και x_b συμβολίζεται ως $d(x_a, x_b)$.

ΕΥΚΛΕΙΔΕΙΑ ΑΠΟΣΤΑΣΗ

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} (x_{aj} - x_{bj})^2}$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a, x_b και x_{aj} η τιμή της j μεταβλητής της παρατήρησης x_a.

ΕΥΚΛΕΙΔΕΙΑ ΑΠΟΣΤΑΣΗ ΜΕ ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ

Όταν η κλίμακα των παρατηρήσεων είναι διαφορετική, δεν μπορεί να γίνει σωστά η μέτρηση της απόστασης. Για το λόγο αυτό φέρνουμε όλες τις μεταβλητές σε συγκρίσιμη κλίμακα. Αυτό το επιτυγχάνουμε διαιρώντας την κάθε μεταβλητή με την τυπική της απόκλιση s_j.

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} \left(\frac{x_{aj} - x_{bj}}{s_j}\right)^2}$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a, x_b και s_j η τυπική απόκλιση της μεταβλητής j.

Η απόσταση αυτή επιτρέπει πιο καλές συγκρίσεις μεταξύ των μεταβλητών και άρα είναι πιο ενδιαφέρουσα. Το μόνο της μειονέκτημα είναι ότι δε λαμβάνει υπόψη της στον τύπο τις πιθανές συνδιακυμάνσεις μεταξύ των μεταβλητών. Άρα, θα ήταν οφέλιμη μια απόσταση, η οποία θα μέτραγε τις συνδιακυμάνσεις.

ΕΥΚΛΕΙΔΕΙΑ ΑΠΟΣΤΑΣΗ ΜΕ ΣΥΝΤΕΛΕΣΤΗ ΒΑΡΥΤΗΤΑΣ

Όταν θέλουμε να προσδώσουμε ιδιαίτερη βαρύτητα σε ορισμένα απο τα γνωρίσματα, χρησιμοποιούμε τον παρακάτω τύπο, ο οποίος λαμβάνει υπόψην του το συντελεστή βαρύτητας της κάθε μεταβλητής

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} w_j * \left(x_{aj} - x_{bj}\right)^2}$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a , x_b και w_j ο συντελεστής βαρύτητας του γνωρίσματος j.

ΑΠΟΣΤΑΣΗ MAHALANOBIS

Μια απόσταση λοιπόν που μετράει τις συνδιακυμάνσεις είναι η *απόσταση Mahalanobis* και δίνεται από τον παρακάτω τύπο:

$$d(x_a, x_b) = (x_a - x_b)^T S^{-1} (x_a - x_b)$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a , x_b και S είναι ο δειγματικός πίνακας διακυμάνσεων.

ΑΠΟΣΤΑΣΗ MANHATTAN

Μια παραλλαγή της Ευκλείδειας απόστασης είναι η *απόσταση Manhattan*, η οποία δίνεται από τον παρακάτω τύπο:

$$d(x_a, x_b) = \sum_{j=1}^{n} |x_{aj} - x_{bj}|$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a, x_b και x_{aj} η τιμή της j μεταβλητής της παρατήρησης x_a.

ΑΠΟΣΤΑΣΗ MINKOWSKI:

Γενίκευση των παραπάνω αποστάσεων αποτελεί η *απόσταση Minkowski* (αν q=1 προκύπτει η απόσταση Manhattan ενώ αν q=2 η ευκλείδεια απόσταση), η οποία ορίζεται από τον τύπο:

$$d(x_a, x_b) = \sqrt[q]{\sum_{j=1}^{n} |x_{aj} - x_{bj}|^q}$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a, x_b και x_{aj} η τιμή της j μεταβλητής της παρατήρησης x_a.

Εκτός από την ευκλείδεια απόσταση και τις παραλλαγές της, υπάρχουν και άλλα μέτρα ομοιότητας όπως οι αποστάσεις, οι οποίες βασίζονται σε συσχετισμό (συντελεστές συσχέτισης). Οι συντελεστές αυτοί χρησιμοποιούνται ευρέως σε βιοϊατρικά δεδομένα γονιδιακής έκφρασης. Οι πιο γνωστοί συντελεστές γραμμικής συσχέτισης είναι οι εξής:

ΣΥΝΤΕΛΕΣΤΗΣ ΓΡΑΜΜΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ r ΤΟΥ PEARSON

Η *συσχέτιση Pearson* μετρά το βαθμό γραμμικής σχέσης μεταξύ δύο μεταβλητών και υπολογίζεται από τον τύπο

$$r = \frac{\sum_{j=1}^{n}(x_{aj} - \overline{x_a})(x_{bj} - \overline{x_b})}{\sqrt{\sum_{j=1}^{n}(x_{aj} - \overline{x_a})^2 \sum_{j=1}^{n}(x_{bj} - \overline{x_b})^2}}$$

όπου x_{aj} η τιμή της j μεταβλητής της παρατήρησης x_a και $\overline{x_a}$ η μέση τιμή της.

ΣΥΝΤΕΛΕΣΤΗΣ ΓΡΑΜΜΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ sp ΤΟΥ SPEARMAN

Ο *συντελεστής συσχέτισης sp του Spearman* ορίζεται από τον ίδιο τύπο με αυτόν του συντελεστή συσχέτισης Pearson, αλλά στη θέση των μεταβλητών βάζουμε την τάξη τους, δηλαδή τη σειρά κατάταξής τους και ο sp υπολογίζεται από τον τύπο

$$sp = \frac{\sum_{j=1}^{n}({x_{aj}}' - \overline{x_a}')({x_{bj}}' - \overline{x_b}')}{\sqrt{\sum_{j=1}^{n}({x_{aj}}' - \overline{x_a}')^2 \sum_{j=1}^{n}({x_{bj}}' - \overline{x_b}')^2}}$$

όπου ${x_{aj}}'$ η σειρά κατάταξης της j μεταβλητής της παρατήρησης x_a και $\overline{x_a}'$ η μέση τιμή των σειρών κατάταξης της μεταβλητής j.

Στις ίσες τιμές μεταβλητών απονέμεται κατάταξη ίση με το μέσο όρο των θέσεών τους στην αύξουσα σειρά της κατάταξης.

ΣΥΝΤΕΛΕΣΤΗΣ ΓΡΑΜΜΙΚΗΣ ΣΥΣΧΕΤΙΣΗΣ τ ΤΟΥ KENDALL

Ο *συντελεστής συσχέτισης Kendall* μετρά την αντιστοιχία μεταξύ της κατάταξης των μεταβλητών των παρατηρήσεων x_a και x_b. Ο συνολικός αριθμός πιθανών ζευγών του x_a με του x_b είναι $n(n-1)/2$, όπου n είναι το μέγεθος των x_a και x_b. Αρχικά διατάσσουμε τα ζεύγη στις τιμές x_a. Εάν τα x_a και x_b συσχετιστούν, τότε θα έχουν τις ίδιες σχετικές κατατάξεις. Έπειτα, για κάθε x_{bj}, μετράμε το πλήθος των συμπαγών ζευγών n_c (αυτών δηλαδή που σχετίζονται) και περιγράφονται από την ανισότητα: $x_{bj} > x_{bi}$ και το πλήθος

των μη συμπαγών ζευγών n_d (αυτών δηλαδή που δεν σχετίζονται) και περιγράφονται από την ανισότητα: $x_{bj} < x_{bi}$. Έτσι, ο συντελεστής συσχέτισης Kendall ορίζεται ως εξής:

$$\tau = \frac{(n_c - c_d)}{\frac{1}{2}n(n-1)}$$

ΠΑΡΑΤΗΡΗΣΕΙΣ

- Η Ευκλείδεια απόσταση δεν επηρεάζεται από την προσθήκη νέων παρατηρήσεων και αποδίδει καλά όταν στα δεδομένα υπάρχουν συμπαγείς ή απομονωμένες συστάδες (Jianchang Mao & Jain, 1996).
- Η ύπαρξη γραμμικής συσχέτισης μεταξύ των μεταβλητών μπορεί να επιφέρει στρεβλώσεις στον υπολογισμό της απόστασης. Το πρόβλημα αυτό λύνεται ως ένα βαθμό με την απόσταση Mahalanobis.
- Οι συντελεστές συσχέτισης είναι ποσότητες που δείχνουν το βαθμό συνδιακύμανσης δυο μεγεθών. Είναι θετικοί όταν τα δυο μεγέθη διακυμαίνονται με τον ίδιο τρόπο και αρνητικοί όταν διακυμαίνονται αντίθετα.

2.1.2 Απόσταση δυαδικών μεταβλητών

Σε πραγματικά δεδομένα, είναι σύνηθες να υπάρχουν εκτός από τα αριθμητικά και άλλες μεταβλητές, άλλων τύπων, όπως οι *ονομαστικές* και οι *δυαδικές*.

Οι δυαδικές (binary) μπορούν να πάρουν δυο δυνατές τιμές, την τιμή 1 και την τιμή 0. Οι τιμές μιας δυαδικής μεταβλητής αντιπροσωπεύουν μια πληροφορία, όπου η κάθε μια δυνατή τιμή αναπαριστά μια κατάσταση ίσης σημασίας ή αξίας. Παράδειγμα τέτοιας μεταβλητής είναι το «φύλο», όπου η τιμή 1 συμβολίζει το αρσενικό και η τιμή 0 το θηλυκό. Μια τέτοια μεταβλητή είναι και συμμετρική. Υπάρχουν όμως και δυαδικές μεταβλητές, στις οποίες οι δυο αυτές καταστάσεις δεν έχουν ίση αξία, δεν είναι ισότιμες. Παράδειγμα τέτοιων

μεταβλητών είναι εκείνες που περιγράφουν την ύπαρξη ή μη ενός γεγονότος (πχ η χρεοκοπία ή μη μιας επιχείρησης). Συνήθως, η ύπαρξη ενός συμβάντος είναι πιο σπάνια και γι' αυτό η κατάσταση αυτή συμβολίζεται με την τιμή 1. Τέτοιου τύπου δυαδικές μεταβλητές, όπου οι τιμές τους δεν είναι ίσης αξίας, ονομάζονται μη συμμετρικές.

Έστω ένα σύνολο δεδομένων, το οποίο περιέχει μόνο δυαδικές μεταβλητές. Έστω επίσης τα x_a και x_b αντικείμενα αυτού του συνόλου. Τα x_a και x_b μπορούν να παίρνουν ίδιες τιμές (0 ή 1) ή και διαφορετικές.
Οι πιθανοί συνδυασμοί των τιμών των δύο μεταβλητών είναι οι εξής:

- k είναι το πλήθος των δυάδων όπου και το x_a και το x_b έχουν την τιμή 1
- n είναι το πλήθος των δυάδων όπου και το x_a και το x_b έχουν την τιμή 0
- l είναι το πλήθος των δυάδων όπου το x_a έχει την τιμή 1 και το x_b έχει την τιμή 0
- m είναι το πλήθος των δυάδων όπου το x_a έχει την τιμή 0 και το x_b έχει την τιμή 1

Ο παρακάτω πίνακας συνάφειας συνοψίζει όσα ειπώθηκαν παραπάνω:

		Αντικείμενο x_b		
		1	***0***	***Άθροισμα***
Αντικείμενο x_a	***1***	k	l	$k+l$
	0	m	n	$m+n$
	Άθροισμα	$k+m$	$l+n$	$k+l+m+n$

Πίνακας 1 Σύνοψη συνδυασμού των τιμών που μπορούν να πάρουν οι δυαδικές μεταβλητές

Εάν οι δυαδικές μεταβλητές είναι συμμετρικές, τότε η απόσταση των x_a και x_b δίνεται από τον *συντελεστή simple matching (simple matching coefficient)*, ο οποίος δίνεται από τον εξής τύπο:

$$d(x_a, x_b) = \frac{l + m}{k + l + n + m}$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a, x_b και k, l, m, n όπως ορίζονται στον πίνακα 1.

Για τις μεταβλητές που δεν είναι συμμετρικές, έχουν προταθεί διάφορα μέτρα, με πιο γνωστό αυτό του *συντελεστή Jaccard* (παρακάτω τύπος). Στον τύπο αυτό παραλείπεται ο συνδυασμός των τιμών όταν είναι ίσες με 0, διότι είναι μικρότερης σημασίας.

$$d(x_a, x_b) = \frac{l + m}{k + l + m}$$

2.1.3 Απόσταση ονομαστικών μεταβλητών

Ονομαστικές (nominal) λέγονται οι μεταβλητές που δέχονται ονομαστικές τιμές, δηλαδή λέξεις.

Έστω δυο αντικείμενα x_a και x_b με n ονομαστικά γνωρίσματα το καθένα. Κατ' αντιστοιχία με τις δυαδικές μεταβλητές και το συντελεστή simple matching, μπορούμε να υπολογίσουμε την απόσταση των δυο αυτών αντικειμένων ως εξής:

$$d(x_a, x_b) = \frac{n - m}{n}$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a, x_b και m, n όπως ορίζονται στον πίνακα 1.

Ένας άλλος τρόπος να αντιμετωπίσουμε το θέμα του υπολογισμού της απόστασης τέτοιων μεταβλητών είναι με την εισαγωγή ψευδομεταβλητών. Για μια ονομαστική μεταβλητή, η οποία μπορεί να πάρει k πιθανές τιμές, δημιουργούμε k ψευδομεταβλητές. Εάν κάποιο αντικείμενο αυτής της

μεταβλητής πάρει μια συγκεκριμένη τιμή στην ονομαστική μεταβλητή, τότε η αντίστοιχη ψευδομεταβλητή παίρνει την τιμή 1 και οι υπόλοιπες ψευδομεταβλητές παίρνουν την τιμή 0. Αφού γίνουν οι απαραίτητες μετατροπές για όλα τα αντικείμενα, ο υπολογισμός γίνεται με τον τρόπο που ειπώθηκε παραπάνω.

2.1.4 Απόσταση διατακτικών μεταβλητών

Οι *διατακτικές (ordinal)* μεταβλητές είναι μεταβλητές, οι οποίες δέχονται τιμές που δηλώνουν κάποια σειρά ή κατάταξη. Ένα παράδειγμα διατακτικών μεταβλητών είναι οι βαθμοί που παίρνουν οι μαθητές στο δημοτικό. Ένας μαθητής μπορεί να βαθμολογηθεί με Α, Β, Γ. Οι τιμές αυτές μπορεί να είναι λεκτικές, όμως υποδηλώνουν και κάποια διάταξη.

Μπορούμε να εκμεταλλευτούμε το γεγονός ότι οι διατακτικές μεταβλητές υποδηλώνουν μια σειρά και να τις αντιμετωπίσουμε σαν αριθμητικές και να αντικαταστήσουμε (αν η μεταβλητή έχει n τιμές) τη χαμηλότερη θέση με 1, την επόμενη με 2, μέχρι και την υψηλότερη, η οποία λαμβάνει την τιμή n. Η διαδικασία αυτή όμως έχει το μειονέκτημα ότι μεταβλητές με πολλές διατακτικές τιμές μπορούν να οδηγήσουν σε μεγάλες διαφορές μεταξύ των αντικειμένων και άρα να επηρεαστεί δυσανάλογα η απόσταση. Για να αποφύγουμε αυτό το πρόβλημα, κανονικοποιούμε τις αριθμητικές τιμές στο διάστημα [0,...,1]. Ο μετασχηματισμός των τιμών γίνεται σύμφωνα με τον τύπο:

$$m_{new} = \frac{m-1}{n-1}$$

όπου m_{new} είναι η νέα τιμή, m είναι η τιμή πριν την κανονικοποίηση και n είναι το πλήθος δυνατών τιμών της διατακτικής μεταβλητής.

Αφού γίνει ο μετασχηματισμός των διατακτικών τιμών και η αντιστοίχισή τους σε αριθμητικές τιμές της περιοχής [0,...,1], μπορούμε να υπολογίσουμε την απόσταση δυο αντικειμένων με την Ευκλείδεια απόσταση ή όποια άλλη παραλλαγή της επιλέξουμε.

2.1.5 Απόσταση μεταβλητών μεικτού τύπου

Οι μέχρι τώρα υπολογισμοί αποστάσεων που αναφέρθηκαν θεωρούν ότι όλες οι μεταβλητές είναι ίδιου τύπου. Σε πραγματικά δεδομένα όμως, οι μεταβλητές είναι διαφόρων (μεικτών) τύπων και άρα μπορεί να είναι αριθμητικές, δυαδικές, ονομαστικές κα. Για να μπορέσουμε να υπολογίσουμε λοιπόν την απόσταση αντικειμένων που έχουν διαφορετικούς τύπους μεταβλητών θα πρέπει να γίνει ένας συνδυασμός των προηγούμενων τύπων.

Η απόσταση δύο αντικειμένων x_a και x_b με n μεταβλητές διαφορετικών τύπων μπορεί να υπολογιστεί από την εξίσωση:

$$d(x_a, x_b) = \frac{\sum_{j=1}^{n} \delta_{abj} \Delta_{abj}}{\sum_{j=1}^{n} \delta_{abj}}$$

όπου $d(x_a, x_b)$ είναι η απόσταση των x_a, x_b.

Το δ_{abj} παίρνει τιμές ως ακολούθως:

- Τιμή = 0 εάν η τιμή του x_a (x_{aj}) ή του x_b (x_{bj}) στη μεταβλητή j λείπει.
- Τιμή = 0 εάν η μεταβλητή j είναι μη συμμετρική και η τιμή των x_a και x_b στη μεταβλητή j είναι ίση με 0 (x_{aj}= x_{bj}= 0).
- Τιμή = 1 σε οποιαδήποτε άλλη περίπτωση

Ο υπολογισμός της τιμής του Δ_{abj} εξαρτάται από τον τύπο της μεταβλητής j:

- Εάν η μεταβλητή j είναι δυαδική ή ονομαστική το Δ_{abj} παίρνει την τιμή 0 εάν $x_{\alpha j}$ = x_{bj}. Διαφορετικά παίρνει την τιμή 1.
- Εάν η μεταβλητή j είναι αριθμητική, τότε το Δ_{abj} υπολογίζεται σύμφωνα με την παρακάτω εξίσωση, όπου max_j είναι η μέγιστη τιμή της μεταβλητής j και min_j είναι η ελάχιστη τιμή της μεταβλητής j.

$$\Delta_{abj} = \frac{|x_{aj} - x_{bj}|}{\max_j - \min_j}$$

- Εάν η μεταβλητή j είναι διατακτική, τότε οι τιμές της μετασχηματίζονται και ανάγονται στην περιοχή $[0,\ldots,1]$ και ακολούθως το Δ_{abj} υπολογίζεται με τρόπο αντίστοιχο των αριθμητικών μεταβλητών.

2.2 ΜΕΘΟΔΟΙ ΕΝΩΣΗΣ ΣΥΣΤΑΔΩΝ

2.2.1 Απλή Σύνδεση

Η μέθοδος της *Απλής Σύνδεσης (Simple Linkage)* ή αλλιώς η μέθοδος του κοντινότερου γείτονα είναι αυτή που μετράει σαν απόσταση δυο συστάδων τη μικρότερη απόσταση από οποιοδήποτε μέλος της πρώτης συστάδας προς οποιοδήποτε μέλος της δεύτερης συστάδας (Sneath & Sokal, 1973; Κύρκος, 2015).

$$d(C_1, C_2) = \min_{x_a \epsilon C_1, x_b \epsilon C_2} d(x_a, x_b)$$

όπου C_1, C_2 είναι οι δυο συστάδες, x_a και x_b είναι τα κοντινότερα σημεία των συστάδων αυτών και d(x_a, x_b) είναι η απόσταση μεταξύ των x_a και x_b.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Ένα βασικό πλεονέκτημα της μεθόδου είναι ότι μπορεί να εντοπίσει και να χειριστεί μη ελλειπτικά σημεία (non elliptical) .

ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Ένα σύνηθες πρόβλημα της απλής σύνδεσης είναι ότι συνενώνει

συστάδες, οι οποίες έχουν δυο κοντινά σημεία και πολλά άλλα που βρίσκονται σε μεγάλες αποστάσεις. Ένα άλλο πρόβλημα είναι ότι μπορεί να προκληθεί η δημιουργία μιας επιμήκους συστάδας και να προστίθενται συνεχώς νέα σημεία στην «ουρά» της. Επίσης, εάν μεταξύ δυο πραγματικών συστάδων υπάρχουν μεμονωμένα σημεία που δημιουργούν μια «γέφυρα», τότε οι συστάδες αυτές θα ενωθούν. Το αποτέλεσμα αυτής της διαδικασίας είναι ότι τα σημεία που βρίσκονται στα δύο άκρα της συστάδας θα απέχουν πολύ μεταξύ τους. Το πρόβλημα αυτό είναι γνωστό ως *φαινόμενο της αλυσίδας (chaining phenomenon)*. Τέλος, είναι πολύ ευαίσθητη η μέθοδος αυτή σε θόρυβο και σε απομακρυσμένες τιμές (outliers) (Tan, Steinbach, & Kumar, 2006).

2.2.2 Πλήρης Σύνδεση

Η μέθοδος της *Πλήρους Σύνδεσης (Complete Linkage)* ή αλλιώς και μέθοδος του μακρινότερου γείτονα είναι αυτή που ως απόσταση δύο συστάδων αντιλαμβάνεται τη μεγαλύτερη απόσταση μεταξύ οποιουδήποτε μέλους της πρώτης συστάδας και οποιουδήποτε μέλους της δεύτερης συστάδας. Με πιο απλά λόγια, η απόσταση μεταξύ δύο συστάδων είναι η απόσταση ανάμεσα στα δύο πιο απομακρυσμένα σημεία τους. Ο μαθηματικός ορισμός της μεθόδου αυτής δίνεται από τον παρακάτω τύπο:

$$d(C_1, C_2) = \max_{x_a \epsilon C_1, x_b \epsilon C_2} d(x_a, x_b)$$

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Με τη μέθοδο της πλήρους σύνδεσης μπορούμε να αποφύγουμε τα προβλήματα που παρουσιάζονται με την απλή σύνδεση, όπως η δημιουργία επιμηκών συστάδων. Εδώ, η μέθοδος της πλήρους σύνδεσης τείνει να δημιουργήσει συμπαγείς και σφαιρικές συστάδες με συγκρίσιμη διάμετρο. Αυτό συμβαίνει, γιατί από όλες τις υποψήφιες προς συνένωση συστάδες, επιλέγει εκείνες τις δύο, που δημιουργούν τη συστάδα με τη μικρότερη διάμετρο. Είναι

καλό να χρησιμοποιούμε αυτή τη μέθοδο όταν γνωρίζουμε ότι αντικείμενα της ίδιας συστάδας μπορεί να βρίσκονται σε μεγάλες αποστάσεις μεταξύ τους.

ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Ένα μειονέκτημα της μεθόδου είναι η ευαισθησία της στην ύπαρξη αντικειμένων με ακραίες τιμές. Εάν υπάρχει ένα αντικείμενο με ακραίες τιμές σε μια συστάδα, τότε δύσκολα αυτή η συστάδα θα συγχωνευθεί με κάποιαν άλλη. Επίσης, τείνει να διασπά μεγάλες συστάδες και είναι πολύ πιθανό να οδηγήσει σε κυκλικά σχήματα.

2.2.3 Σύνδεση Μέσου Όρου

Η μέθοδος της *Σύνδεσης Μέσου Όρου (Average Linkage)* ορίζει σαν απόσταση δύο συστάδων τη μέση απόσταση όλων των συνδυασμών των ζευγών αντικειμένων, όπου το πρώτο αντικείμενο ανήκει στην πρώτη συστάδα και το δεύτερο αντικείμενο ανήκει στη δεύτερη συστάδα (Murtagh, 1984). Ο μαθηματικός ορισμός της απόστασης Μέσου Όρου δίνεται από τον τύπο

$$d(C_1, C_2) = \frac{\sum_{x_a \epsilon C_1} \sum_{x_b \epsilon C_2} d(x_a, x_b)}{N_{c_1} N_{C_2}}$$

όπου $d(C_{1,} C_2)$ είναι η απόσταση των δυο συστάδων C_1, C_2, $d(x_a, x_b)$ είναι η απόσταση μεταξύ των αντικειμένων x_a, x_b και N_{C_1}, N_{C_2} είναι το πλήθος των στοιχείων των συστάδων C_1και C_2 αντίστοιχα.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Η απόσταση μέσου όρου αποτελεί ενδιάμεση λύση ανάμεσα στην ευαισθησία στα αντικείμενα με ακραίες τιμές της μεθόδου της πλήρους σύνδεσης και στην τάση δημιουργίας επιμήκων συστάδων της απλής σύνδεσης. Χάρη στον υπολογισμό της μέσης απόστασης μεταξύ των ζευγών,

δε δημιουργείται το φαινόμενο της αλυσίδας. Επίσης, εξομαλύνεται η επιρροή των αντικειμένων με ακραίες τιμές.

ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Από άποψη υπολογιστικού κόστους, η μέθοδος είναι ακριβή καθώς υπολογίζει τις αποστάσεις όλων των δυνατών ζευγών. Ένα άλλο μειονέκτημα είναι ότι μπορεί να διασπάσει υπαρκτές επιμήκεις συστάδες ενώ παράλληλα ευνοεί τις κυκλικές.

2.2.4 Απόσταση Κεντρικών Σημείων

Σύμφωνα με την προσέγγιση αυτή, η απόσταση μεταξύ δυο συστάδων είναι η απόσταση των μέσων σημείων των δύο συστάδων. Ο τύπος που ορίζει την απόσταση αυτών των σημείων είναι ο εξής:

$$d(C_1, C_2) = d(m_1, m_2)$$

όπου $d(C_1, C_2)$ είναι η απόσταση των δυο συστάδων C_1, C_2 και m_1, m_2 είναι τα κεντρικά σημεία των συστάδων C_1 και C_2.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Η μέθοδος της απόστασης των κεντρικών σημείων έχει το πλεονέκτημα ότι δεν επηρρεάζεται σημαντικά από την ύπαρξη αντικειμένων με ακραίες τιμές.

2.2.5 Μέθοδος Ward

Η *μέθοδος του Ward* (1963) διαφέρει σημαντικά από τις προηγούμενες μεθόδους, καθώς δεν υπολογίζει κάποια «απόσταση» μεταξύ των συστάδων αλλά σχηματίζει ομάδες μεγιστοποιώντας την ομοιογένεια μέσα στις ομάδες. Το

μέτρο που εφαρμόζεται είναι το άθροισμα του τετραγωνικού σφάλματος, και επιδίωξη της μεθόδου είναι η ελαχιστοποίηση του. Με πιο απλά λόγια, ελαχιστοποιεί τη μεταβλητότητα μεταξύ των τιμών σε κάθε ομάδα με μία υπολογιστική διαδικασία γνωστή και ως τετραγωνικό σφάλμα.

Το ίδιο κριτήριο χρησιμοποιείται και από τον αλγόριθμο k-Means, οπότε η μέθοδος Ward μπορεί να θεωρηθεί το ιεραρχικό ανάλογο του k-Means.

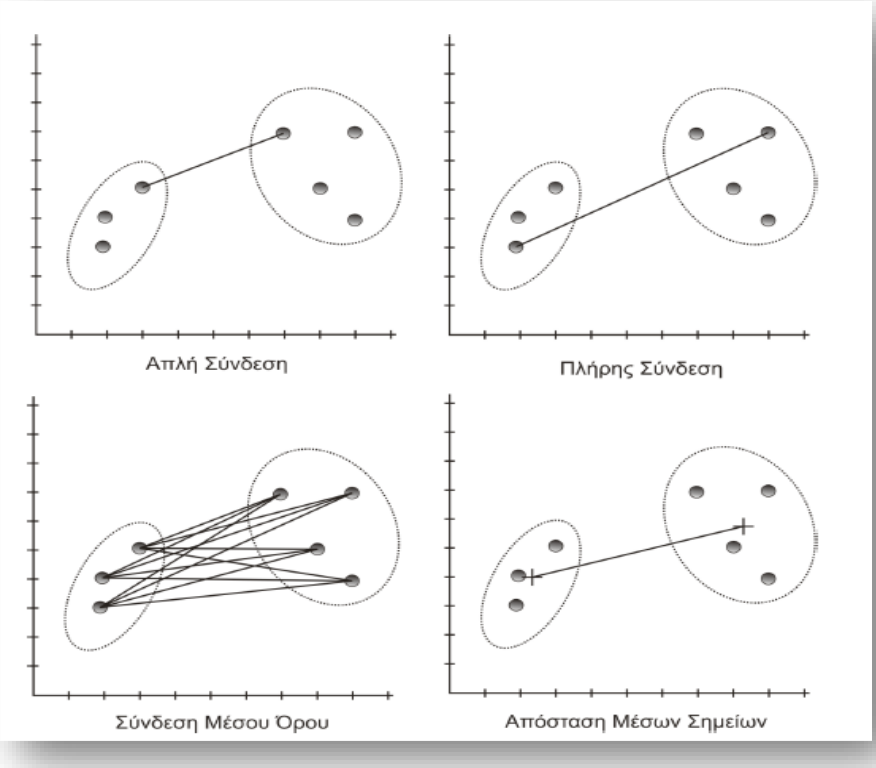
Το τετραγωνικό σφάλμα δίνεται από τον τύπο:

$$E = \sum_{x \epsilon C_i} (x - m_i)^2$$

όπου C_i είναι μια συστάδα και m_i είναι το κεντρικό σημείο της.

Η μέθοδος, για να συνενώσει δύο συστάδες από συνολικό πλήθος k συστάδων, ελέγχει τα δυνατά $k(k-1)/2$ ζεύγη συστάδων τα οποία μπορούν να δημιουργηθούν, και επιλέγει το ζεύγος, το οποίο όταν ενωθεί θα μας δώσει τη συστάδα με το ελάχιστο τετραγωνικό σφάλμα.

✓ Η μέθοδος του Ward έχει την τάση να παράγει ισοπληθείς ομάδες.



Εικόνα 2 Μέθοδοι ένωσης των συστάδων

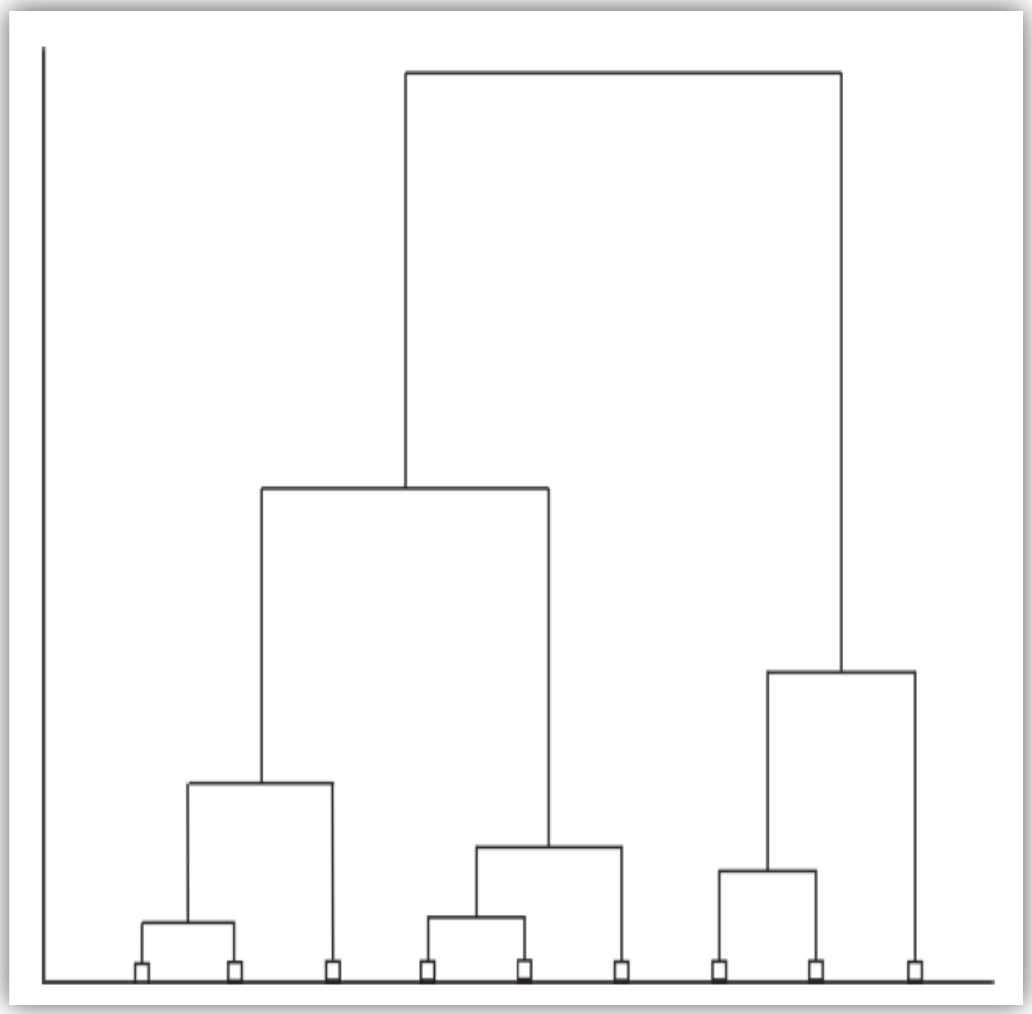
2.3 ΔΕΝΔΡΟΓΡΑΜΜΑΤΑ

Τα δενδρογράμματα είναι ένας γραφικός τρόπος αναπαράστασης των δεδομένων. Ειδικά στην Ανάλυση Συστάδων, είναι ένας γραφικός τρόπος αναπαράστασης των δεδομένων και καταγράφει ακολουθίες από διαδοχικές συγχωνεύσεις ή διασπάσεις.

Ένα δενδρόγραμμα μοιάζει με δένδρο, στα φύλλα του οποίου (δηλαδή στο κατώτερό του επίπεδο), βρίσκονται τα αντικείμενα. Κάθε κόμβος του δέντρου αναπαριστά μια ομάδα (συστάδα) και επιπλέον, κάθε κόμβος του διαχωρίζει δύο κλάδους. Στη συσσωρευτική ομαδοποίηση (η οποία θα αναλυθεί εκτενώς σε επόμενο κεφάλαιο), κάθε κόμβος με τα κλαδιά και τα παιδιά του συμβολίζει τη συγχώνευση των συστάδων-παιδιών και τη δημιουργία της συστάδας γονέα, ενώ στη διαιρετική κάθε κόμβος συμβολίζει τη διάσπαση του κόμβου-γονέα και τη δημιουργία των συστάδων-παιδιών.

Πρέπει να ειπωθεί εδώ ότι ο βαθμός της (αν)ομοιότητας με το επίπεδο αυξάνεται μονότονα και το δέντρο σχεδιάζεται με τέτοιο τρόπο ώστε να αποτυπώνεται στη διαφορά του ύψους του κάθε επιπέδου του επακριβώς η αύξηση αυτής της (αν)ομοιότητας.

Τέλος, είναι σημαντικό να σημειωθεί ότι ο κάθε χρήστης μπορεί να επιλέξει ένα επίπεδο μόνο του δενδρογράμματος, ένα κομμάτι του δηλαδή μόνο και να αποφασίσει αυτός για το διαμοιρασμό των αντικειμένων σε συστάδες. Όμως, πρέπει να γνωρίζει επίσης ότι διαφορετικές επιλογές μεθόδων συσταδοποίησης ή ακόμα και μικροαλλαγές στα δεδομένα μπορούν να δημιουργήσουν διαφορετικά δενδρογράμματα (Κύρκος, 2015).



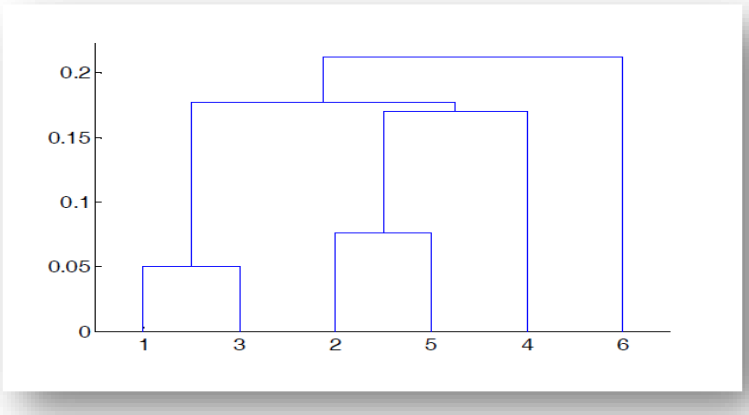
Εικόνα 3 Παράδειγμα μορφής Δενδρογράμματος

3 ΚΕΦΑΛΑΙΟ: ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

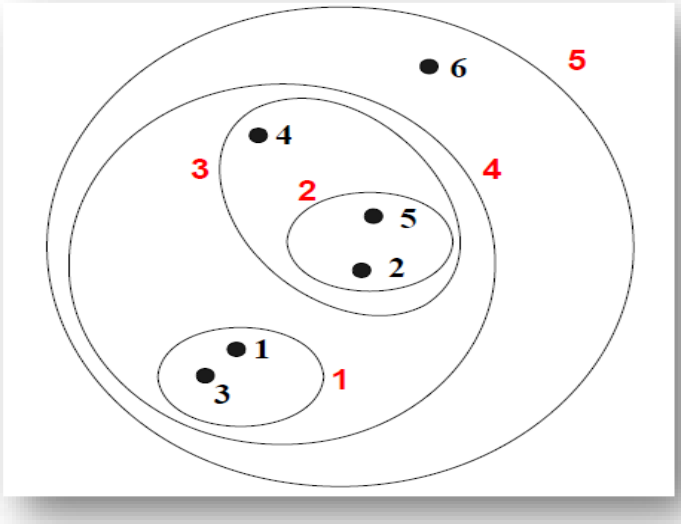
3.1 ΙΕΡΑΡΧΙΚΗ ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ

Οι *Ιεραρχικές Μέθοδοι* είναι μέθοδοι οι οποίες δημιουργούν ένα σύνολο από εμφωλευμένες ομάδες (Tan et al., 2006). Το τελικό αποτέλεσμα αναπαρίσταται με ένα ιεραρχικό δένδρο, ένα δενδρόγραμμα στο οποίο ο x-άξονας είναι τα αντικείμενα (οι παρατηρήσεις) και ο y-άξονας είναι οι αποστάσεις μεταξύ των αντικειμένων. Εάν στα δεδομένα μας υπάρχουν n παρατηρήσεις, τότε το δένδρο έχει $n-1$ επίπεδα. Η Ιεραρχική Ανάλυση Συστάδων, λοιπόν, δημιουργεί μια ιεραρχία από επίπεδα που περιγράφουν ένα συγκεκριμένο τρόπο διαχωρισμού των παρατηρήσεων σε συστάδες.

Το πιο βασικό χαρακτηριστικό αυτών των μεθόδων είναι ότι μόλις ένα αντικείμενο ανατεθεί σε μια ομάδα, δεν μπορεί να μετακινηθεί από αυτήν, αλλά ούτε και να ενωθεί με αντικείμενα άλλης ομάδας. Επίσης, στις Ιεραρχικές Μεθόδους, ο ερευνητής δε γνωρίζει εξ αρχής τον αριθμό των συστάδων που θα παραχθούν. Τέλος, επειδή σε κάθε βήμα οι Ιεραρχικές Μέθοδοι, προκειμένου να υπολογίσουν τις αποστάσεις των παρατηρήσεων, δημιουργούν έναν πίνακα ομοιοτήτων, χρειάζονται πολύ χώρο και χρόνο στον Η/Υ (μεγάλη πολυπλοκότητα) και γι'αυτό δε χρησιμοποιούνται για μεγάλο όγκο δεδομένων. Οι πιο γνωστές Ιεραρχικές Μέθοδοι είναι οι διαιρετικές και οι συσσωρευτικές.



Εικόνα 4 Παράδειγμα Δενδρογράμματος (β)



Εικόνα 5 Ταξινόμηση των παρατηρήσεων με χρήση Ιεραρχικής Ανάλυσης

3.2 ΔΙΑΙΡΕΤΙΚΕΣ ΜΕΘΟΔΟΙ (DIVISIVE METHODS)

Οι *διαιρετικές μέθοδοι (divisive methods)* ξεκινούν με μια ενιαία ομάδα, η οποία περιέχει όλες τις παρατηρήσεις. Η πρώτη αυτή συστάδα χωρίζεται σε δυο υποσυστάδες με τέτοιο τρόπο ώστε η παρατήρηση, η οποία βρίσκεται στη μεγαλύτερη απόσταση από όλες τις άλλες, να αποχωρεί από τη συστάδα αυτή και να δημιουργεί μια καινούρια συστάδα μόνη της. Η επαναληπτική αυτή διαδικασία εκτελείται μέχρι κάθε παρατήρηση να σχηματίζει μια ξεχωριστή ομάδα από μόνη της.

Αφού οι μέθοδοι αυτές ξεκινούν από το πιο ψηλό ιεραρχικό επίπεδο και «κατεβαίνουν», λέμε ότι ακολουθούν μια «από πάνω προς τα κάτω» λογική (top down).

3.3 ΣΥΣΣΩΡΕΥΤΙΚΕΣ ΜΕΘΟΔΟΙ (AGGLOMERATIVE METHODS)

Οι *συσσωρευτικές μέθοδοι (agglomerative methods)* ακολουθούν την

αντίστροφη διαδικασία των διαιρετικών μεθόδων. Ξεκινούν θεωρώντας κάθε παρατήρηση μια ξεχωριστή συστάδα. Σε κάθε βήμα-επανάληψη οι παρατηρήσεις που βρίσκονται πιο κοντά ενώνονται, δημιουργώντας μία συστάδα εκ νέου. Έτσι, συγκρίνοντας τις συστάδες που προκύπτουν κάθε φορά, οι πιο όμοιες συγχωνεύονται. Η επαναληπτική αυτή διαδικασία εκτελείται μέχρι να ενωθούν όλες οι παρατηρήσεις σε μία ενιαία ομάδα (Κύρκος, 2015).

Αφού οι μέθοδοι αυτές ξεκινούν από το πιο χαμηλό ιεραρχικό επίπεδο των συγχωνεύσεων και ανά επίπεδο ανέρχονται, λέμε ότι ακολουθούν μια «από κάτω προς τα πάνω» λογική (bottom up). Οι συσσωρευτικές μέθοδοι είναι οι πιο διαδεδομένες μέθοδοι Ιεραρχικής Συσταδοποίησης.

3.4 ΣΗΜΑΝΤΙΚΑ ΣΗΜΕΙΑ ΠΡΙΝ ΤΗΝ ΑΝΑΛΥΣΗ

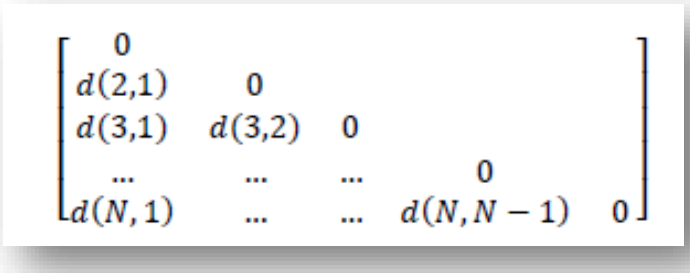
Ο ερευνητής πριν ξεκινήσει τη διαδικασία εκτέλεσης μιας Ιεραρχικής Μεθόδου πρέπει να αποφασίσει ορισμένα σημαντικά θέματα.

ΕΠΙΛΟΓΗ ΜΕΤΡΟΥ ΑΠΟΣΤΑΣΗΣ

Για τον υπολογισμό της ομοιότητας των συστάδων είναι αναγκαίο ένα μέτρο, οπότε θα πρέπει να αποφασίσει ποια απόσταση θα χρησιμοποιήσει για τα συγκεκριμένα δεδομένα. Μιλήσαμε σε προηγούμενο κεφάλαιο για την επιλογή της κατάλληλης απόστασης αναλόγως το είδος των μεταβλητών, οπότε θεωρούμε πως ο εκάστοτε ερευνητής διαλέγει μια απόσταση με βάση τα κριτήρια αυτά.

ΠΙΝΑΚΑΣ ΟΜΟΙΟΤΗΤΑΣ

Για να δημιουργηθούν οι συστάδες δημιουργείται ένας πίνακας (αν)ομοιότητας $n \times n$, από n παρατηρήσεις. Κάθε στοιχείο του πίνακα αντιπροσωπεύει την απόσταση δύο παρατηρήσεων. Ο πίνακας ομοιότητας έχει την ακόλουθη μορφή:



Εικόνα 6 Παράδειγμα ενός πίνακα ομοιότητας

όπου $d(x_1,x_2)$ είναι η απόσταση των παρατηρήσεων x_1 και x_2. Τα στοιχεία της διαγωνίου είναι μηδενικά διότι η απόσταση της κάθε παρατήρησης από τον εαυτό της είναι μηδενική $(d(x_i,x_i)=0)$. Επίσης, ο πίνακας είναι συμμετρικός (η απόσταση δυο παρατηρήσεων είναι συμμετρική $\left(d(x_i,x_j)=\ d(x_j,x_i)\right)$), γι' αυτό και φαίνονται μόνο τα στοιχείο του πίνακα κάτω από τη διαγώνιο.

ΕΠΙΛΟΓΗ ΜΕΘΟΔΟΥ ΕΝΩΣΗΣ ΤΩΝ ΟΜΑΔΩΝ

Ένα άλλο σημείο, το οποίο πρέπει ο ερευνητής να αποφασίσει είναι αυτό του τρόπου με τον οποίο θα ενώνονται οι ομάδες (είτε μέσω συγχώνευσης παρατηρήσεων είτε μέσω συγχώνευσης ομάδων). Αυτό είναι επίσης ένα κομμάτι για το οποίο μιλήσαμε εκτενώς στο δεύτερο κεφάλαιο, οπότε ο αναγνώστης μπορεί να ανατρέξει εκεί για πληροφορίες.

ΠΑΡΑΤΗΡΗΣΕΙΣ ΣΤΙΣ ΜΕΘΟΔΟΥΣ

- Όπως έχουμε πει, για όλες τις μεθόδους ένωσης ομάδων έχουμε ανάγκη από έναν πίνακα αποστάσεων, μέσω του οποίου υπολογίζουμε επαναληπτικά τις νέες αποστάσεις. Εξαίρεση όμως αποτελεί η centroid μέθοδος, η οποία χρησιμοποιεί το κέντρο μιας συστάδας. Όταν έχουμε συνεχή δεδομένα, κέντρο θεωρείται ο μέσος των μεταβλητών, ενώ όταν δεν είναι συνεχή χρησιμοποιούμε τη διάμεσο ή την κορυφή. Υπάρχουν όμως και περιπτώσεις που δεν είναι ούτε αυτό δυνατό και τότε καλό είναι να επιλέγουμε κάποια άλλη μέθοδο αντ' αυτής.

- Μετά από πειράματα προσομοίωσης για σύγκριση μεθόδων, οι μέθοδοι Ward και η average linkage έχουν την καλύτερη απόδοση, ενώ η μέθοδος του κοντινότερου γείτονα έχει τη χειρότερη. Όμως, τα συμπεράσματα αυτά δεν είναι απόλυτα. Ανάλογα με τη μορφή του δείγματος και το είδος των μεταβλητών, μπορεί κάποια άλλη μέθοδος από τις επικρατέστερες να αποδίδει καλύτερα. Πχ η μέθοδος του Ward και ο αλγόριθμος K-means τείνουν να δημιουργούν συστάδες με παρόμοια διακύμανση. Επίσης, συνήθως οι πιο πολλές μέθοδοι δε μπορούν να δημιουργήσουν συστάδες με περίεργα σχήματα. Όταν συμβαίνει αυτό, η μέθοδος του κοντινότερου γείτονα είναι αυτή που αποδίδει καλύτερα (Βερύκιος, Καγκλής, & Σταυρόπουλος, 2016).

ΑΛΓΟΡΙΘΜΟΣ ΣΥΣΣΩΡΕΥΤΙΚΗΣ ΜΕΘΟΔΟΥ

Ο γενικός αλγόριθμος της συσσωρευτικής μεθόδου Ανάλυσης Συστάδων έχει ως εξής:

1. Αρχικά, κάθε παρατήρηση από τα n σημεία αποτελεί μια ξεχωριστή συστάδα.
2. Υπολογισμός του πίνακα ομοιότητας (στον πίνακα αυτό καταγράφονται οι αποστάσεις μεταξύ των παρατηρήσεων- συστάδων).

Επανάλαβε:

3. Εύρεση της min απόστασης και ένωση των δυο αυτών παρατηρήσεων που έχουν τη min απόσταση. Η τιμή αυτή συμβολίζεται με $\big(d(U,V)\big)$. Και είναι η απόσταση των δύο συστάδων U και V.
4. Έτσι, παράγεται μια ενιαία συστάδα UV από τις συστάδες U και V. Εάν η min απόσταση περιλαμβάνει μια παρατήρηση και μια ήδη δημιουργηθείσα ομάδα, τότε απλά προσθέτουμε την παρατήρηση αυτή στην ομάδα. Εάν περιλαμβάνει δυο ομάδες τότε ενώνουμε τις ομάδες αυτές. Στον πίνακα ομοιοτήτων, διαγράφονται οι στήλες και οι γραμμές που αντιστοιχούν στις ομάδες U και V, και στη θέση τους προστίθεται μια στήλη και μια γραμμή για την καινούρια συστάδα UV.

5. Ενημερώνεται ο πίνακας γειτνίασης, επαναϋπολογίζοντας τις αποστάσεις των συστάδων.
6. Η διαδικασία επαναλαμβάνεται μέχρι όλες οι παρατηρήσεις να σχηματίζουν μία μόνο συστάδα.

✓ Είναι σημαντικό να τονιστεί στο σημείο αυτό, ότι οι μέθοδοι συσσωρευτικής Ανάλυσης Συστάδων διαφέρουν μεταξύ τους αναλόγως με το μέτρο απόστασης το οποίο χρησιμοποιούν.

ΤΑ ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΩΝ ΙΕΡΑΡΧΙΚΩΝ ΜΕΘΟΔΩΝ

1. Δε χρειάζεται να γνωρίζουμε εκ των προτέρων τον αριθμό των συστάδων.
2. Δημιουργώντας πολλαπλά επίπεδα συστάδων, μπορούμε να έχουμε όποιο αριθμό συστάδων εμείς επιθυμούμε, κόβοντας απλώς το δενδρόγραμμά μας στο αντίστοιχο επίπεδο.
3. Οι μέθοδοι Ιεραρχικής Ταξινόμησης χαρακτηρίζονται από καλή προσαρμοστικότητα και αυτό γιατί είναι σε θέση να εντοπίσουν επιμήκεις, καλά διαχωρισμένες και ομόκεντρες ομάδες

ΤΑ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΩΝ ΙΕΡΑΡΧΙΚΩΝ ΜΕΘΟΔΩΝ

1. Σε κάθε επανάληψη του αλγορίθμου τους, οι Ιεραρχικές Μέθοδοι πρέπει να ελέγχουν πολλές αποστάσεις ώστε να ενημερώνουν τον πίνακα ομοιοτήτων. Κάτι τέτοιο λοιπόν, όπως έχει ειπωθεί και σε προηγούμενη ενότητα, έχει τεράστιο υπολογιστικό κόστος $O(n^2)$, γιαυτό και δε χρησιμοποιούνται οι μέθοδοι αυτές για μεγάλα δεδομένα (Tan et al., 2006)
2. Η ενέργεια που πραγματοποιείται στην κάθε επανάληψη δεν είναι αντιστρέψιμη. Μόλις ενταχθούν δυο αντικείμενα σε μια συστάδα, παραμένουν σε αυτήν και δε γίνεται να διαχωριστούν ή να ενταχθούν αργότερα σε άλλες.

4 ΚΕΦΑΛΑΙΟ: ΜΗ ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

4.1 ΜΕΘΟΔΟΣ K-MEANS

Σκοπός των αλγορίθμων Μη Ιεραρχικής Ταξινόμησης είναι να χωρίζουν το πολυεπίπεδο των δεδομένων σε μικρότερες περιοχές, αντιστοιχίζοντας σε κάθε μια περιοχή και μια συστάδα. Στη μεγάλη αυτή κατηγορία αλγορίθμων ταξινόμησης ή αλλιώς *αλγορίθμων διαμέρισης (partitioning algorithms)* ανήκει και ο αλγόριθμος *K-means*.

Η μέθοδος *K-means* προτάθηκε από τον MacQueen (1967). Στόχος της είναι να μοιράσει ένα σύνολο παρατηρήσεων σε έναν προκαθορισμένο αριθμό ομάδων k, με τέτοιο τρόπο ώστε να αυξάνεται η ομοιότητα εντός των ομάδων αυτών. Έτσι, ο αλγόριθμος υπολογίζει επαναληπτικά το κέντρο της κάθε συστάδας (centroid) και οι παρατηρήσεις εντάσσονται κάθε φορά στη συστάδα που έχει το πιο κοντινό κέντρο (Κύρκος, 2015). Γι' αυτό και οι συστάδες που δημιουργούνται διαφέρουν πριν και μετά την κάθε επανάληψη (Tan et al., 2006). Υπενθυμίζουμε στο σημείο αυτό ότι κέντρο μιας συστάδας καλούμε το διάνυσμα των μέσων, δηλαδή τη μέση τιμή κάθε μεταβλητής όλων των παρατηρήσεων των δεδομένων. Έπειτα, υπολογίζεται η απόσταση της κάθε παρατήρησης από τα κέντρα των συστάδων (συνήθως η απόσταση αυτή είναι η Ευκλείδεια) και έτσι οι παρατηρήσεις κατατάσσονται στη συστάδα εκείνη το κέντρο της οποίας είναι πιο κοντά. Μόλις καταταχθούν όλες οι παρατηρήσεις στις συστάδες που ανήκουν, υπολογίζουμε ξανά τα κέντρα. Η διαδικασία αυτή συνεχίζεται επαναληπτικά και ολοκληρώνεται όταν δεν υπάρχουν διαφορές μεταξύ δυο διαδοχικών επαναλήψεων.

Είναι σημαντικό να τονιστεί εδώ ότι το γεγονός ότι πρέπει να γνωρίζουμε εξ αρχής τον αριθμό των συστάδων, θέτει σημαντικούς περιορισμούς αφού θα πρέπει ή με κάποιον τρόπο πριν την εκτέλεση του αλγορίθμου να αποφασίσουμε τον αριθμό των συστάδων ή να τρέξουμε πολλές φορές τον αλγόριθμο, με διαφορετικές επιλογές κάθε φορά ώστε να αποφασίσουμε τη βέλτιστη επιλογή. Όλα αυτά θα συζητηθούν πιο αναλυτικά παρακάτω.

ΑΛΓΟΡΙΘΜΟΣ ΜΕΘΟΔΟΥ K-MEANS

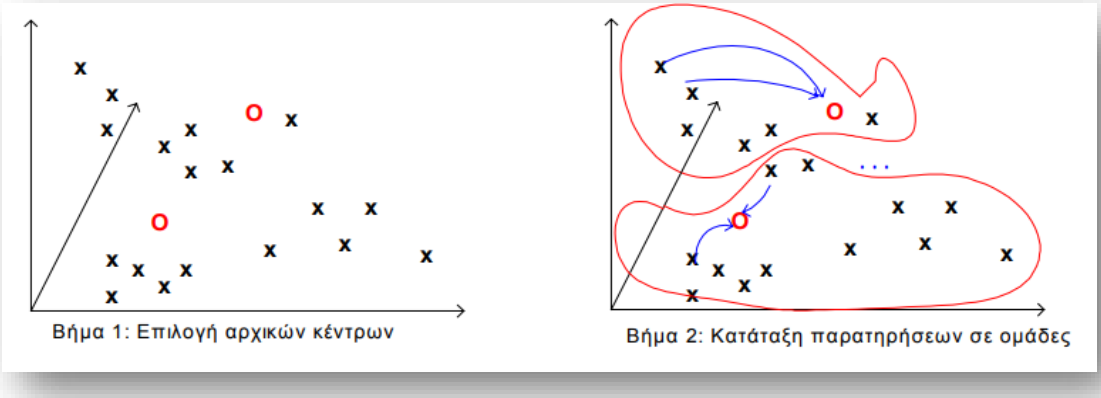
Ο γενικός αλγόριθμος της Μεθόδου K-Means έχει ως εξής (Κύρκος, 2015):

1. Αρχικά, επιλέγουμε τον αριθμό k των συστάδων προς δημιουργία και ορίζουμε τα αρχικά κέντρα.

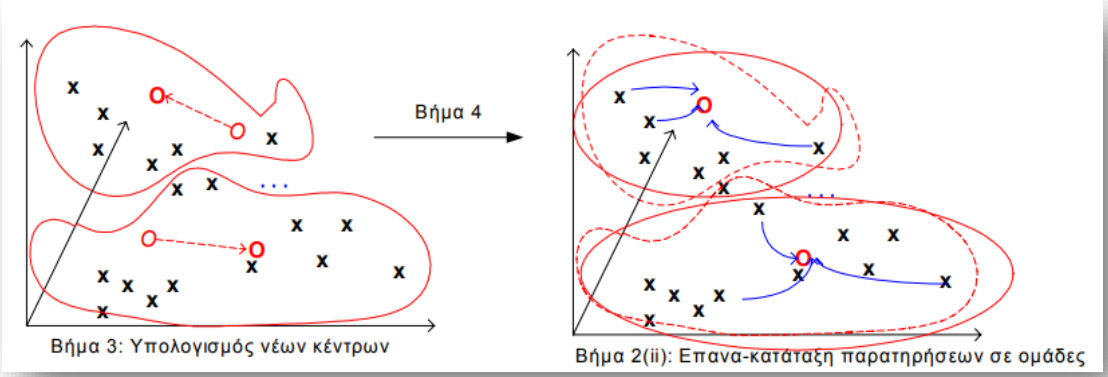
Επανάλαβε:

2. Κατατάσσουμε κάθε παρατήρηση στη συστάδα, το κέντρο της οποίας είναι πιο κοντά στην παρατήρηση αυτή. Συνήθως, χρησιμοποιείται ο τύπος της Ευκλείδειας απόστασης.
3. Εναπαϋπολογίζουμε τα κέντρα της κάθε συστάδας.
4. Ο αλγόριθμος τερματίζεται όταν τα νέα κέντρα δε διαφέρουν απ' τα κέντρα του προηγούμενου βήματος.

Τα παρακάτω γραφήματα απεικονίζουν τα 4 βήματα του αλγορίθμου:



Εικόνα 7 Διαδικασία εκτέλεσης του αλγορίθμου K-means (α)



Εικόνα 8 Διαδικασία εκτέλεσης του αλγορίθμου K-means (β)

ΠΑΡΑΤΗΡΗΣΕΙΣ

- Στη μέθοδο k-Means αποθηκεύουμε μόνα τα κέντρα και δε χρησιμοποιούμε κάποιο πίνακα ομοιότητας γι' αυτό και χρειάζεται πολύ λιγότερο χώρο στον Η/Υ.
- Η πολυπλοκότητα της μεθόδου είναι $O(I * n * k * d)$ όπου I είναι ο αριθμός των επαναλήψεων, n ο αριθμός των παρατηρήσεων, k ο αριθμός των συστάδων και d ο αριθμός των μεταβλητών κάθε παρατήρησης.
- Επίσης, το σημαντικό με αυτό τον αλγόριθμο είναι ότι από τις πρώτες επαναλήψεις έρχεται πολύ κοντά στην τελική λύση και στις επαναλήψεις που απομένουν οι όποιες διαφορές έχουν να κάνουν με τη μετακίνηση κάποιων ελάχιστων παρατηρήσεων, οι οποίες ταλαντεύονται μεταξύ δυο συστάδων. Αυτό έχει σαν αποτέλεσμα να μη χρειάζεται μεγάλος αριθμός επαναλήψεων, μιας και η βασική δομή σχηματίζεται άμεσα.
- Τέλος, όπως ειπώθηκε και πιο πάνω, μπορεί η μέθοδος να χρησιμοποιεί συνήθως την Ευκλείδεια απόσταση, όμως μπορεί να χρησιμοποιηθούν αναλόγως τα δεδομένα και άλλες αποστάσεις. Για παράδειγμα, στα μη συνεχή δεδομένα, στα οποία δεν μπορούμε να υπολογίσουμε το μέσο της συστάδας, χρησιμοποιούμε άλλα μέτρα, όπως τη διάμεσο (medoid) σε κατηγορικά δεδομένα διάταξης (ordinal data) ή την κορυφή (mode) (δηλαδή την τιμή που εμφανίζεται πιο συχνά) σε ονομαστικά δεδομένα. Σε δεδομένα μεικτού τύπου χρησιμοποιούμε ως κέντρο μιας συστάδας τους μέσους των συνεχών μεταβλητών και τις κορυφές των κατηγορικών (Πετρίδης, 2015).

4.2 ΚΡΙΤΗΡΙΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΩΝ ΣΥΣΤΑΔΩΝ

Προκειμένου ο εκάστοτε ερευνητής να μπορεί να ελέγξει την εγκυρότητα της επιλογής του αριθμού των συστάδων που έχει, αναπτύχθηκε ένας μεγάλος αριθμός στατιστικών συναρτήσεων, οι οποίες αντιστοιχούν στα βήματα της επιλογής των τελικών συστάδων. Με τη χρήση της τεχνολογίας όμως και των

εξελιγμένων στατιστικών προγραμμάτων που έχουν δημιουργηθεί, είναι δυνατό να χρησιμοποιηθούν οι συναρτήσεις αυτές εξ αρχής, πριν δηλαδή την εκτέλεση του αλγορίθμου, ώστε ο ερευνητής να αποφασίσει για τη βέλτιστη επιλογή εγκαίρως, γλιτώνοντας έτσι χρόνο και κόπο στην έρευνά του. Μέσα από τις γραφικές παραστάσεις που δημιουργούν τα προγράμματα, αναζητάμε ένα μεγάλο άλμα στην τιμή του στατιστικού στοιχείου που μελετάμε, ουσιαστικά μια «γωνία» ή τη μέγιστη τιμή του γραφήματος, ανάλογα με το κριτήριο. Οι πιο γνωστές από αυτές τις στατιστικές συναρτήσεις είναι οι εξής (Καράγεωργα, 2012; Πετρίδης, 2015):

ΜΕΣΗ ΤΕΤΡΑΓΩΝΙΚΗ ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ ΝΕΟΥ CLUSTER (RMSSTD):

Το *RMSSTD (Root-mean-square standard deviation)* είναι η συγκεντρωτική (pooled) τυπική απόκλιση όλων των παρατηρήσεων που σχηματίζουν κάθε συστάδα και ο τύπος της είναι:

$$Pooled\ Variance = \frac{Pooled\ SS(Άθροισμα\ τετραγώνων)\ για\ όλες\ τις\ μεταβλητές}{Pooled\ βαθμοί\ ελευθερίας\ για\ όλες\ τις\ μεταβλητές}$$

και

$$RMSSTD = \sqrt{Pooled\ Variance}$$

Αφού ο αλγόριθμος Ανάλυσης Συστάδων στοχεύει στη δημιουργία ομοιογενών συστάδων, το RMSSTD κριτήριο κάθε συστάδας θα πρέπει να είναι όσο πιο μικρό γίνεται. Έτσι, περιπτώσεις μεγαλύτερων τιμών υποδηλώνουν πιθανή μη ομοιογένεια στη συστάδα.

ΗΜΙΜΕΤΡΙΚΗ R–ΤΕΤΡΑΓΩΝΟ (SPR)

Όπως έχει ειπωθεί, σε κάθε βήμα της επαναληπτικής διαδικασίας, κάθε νέα συστάδα δημιουργείται από τη συγχώνευση δυο άλλων συστάδων. Η διαφορά μεταξύ του συγκεντρωτικού SS_W (within sum of squares) της νέας συστάδας και του αθροίσματος των συγκεντρωτικών SS_W των δυο συστάδων που συγχωνεύονται, ονομάζεται *απώλεια της ομοιογένειας SPR (Semipartial R-*

square). Όταν η απώλεια ισούται με 0, η νέα συστάδα έχει προκύψει από τη συγχώνευση δυο τέλεια ομοιογενών συστάδων, ενώ όταν η απώλεια αυτή είναι μεγάλη, η νέα συστάδα έχει προκύψει από τη συγχώνευση δυο μη ομοιογενών συστάδων.

R- ΤΕΤΡΑΓΩΝΟ (R-SQUARE)

$$RS = \frac{SS_b}{SS_t}$$

όπου SS_b (sum of squares between clusters) είναι ο βαθμός στον οποίο η κάθε συστάδα διαφέρει από την άλλη και $SS_t = SS_b + SS_W$ (total sum of squares).

Από τον τύπο του SS_t καταλαβαίνουμε ότι όσο πιο μεγάλο είναι το SS_b τόσο πιο μικρό είναι το SS_W. Άρα, σε επίπεδο συστάδων, όσο πιο μεγάλες είναι οι διαφορές μεταξύ τους, τόσο πιο ομοιογενής είναι η κάθε συστάδα και αντιστρόφως. Έτσι, η τιμή του RS, η οποία μετρά το βαθμό στον οποίο οι συστάδες διαφέρουν μεταξύ τους, κυμαίνεται από 1 εως 0, με το τελευταίο να σημαίνει ότι δεν υπάρχουν διαφορές μεταξύ των συστάδων και με το 1 να αντιπροσωπεύει τις μέγιστες διαφορές.

ΑΠΟΣΤΑΣΗ ΔΥΟ ΣΥΣΤΑΔΩΝ

Η *απόσταση centroid (CD)* είναι ουσιαστικά η Ευκλείδεια απόσταση ανάμεσα στα centroid των δυο συστάδων οι οποίες πρόκειται να ενωποιηθούν. Στην περίπτωση της απλής σύνδεσης, η CD είναι η min Ευκλείδεια απόσταση ($MIND$) ανάμεσα σε όλα τα πιθανά ζευγάρια παρατηρήσεων. Αντίστοιχα στην περίπτωση της πλήρους σύνδεσης, η CD είναι η max Ευκλείδεια απόσταση ($MAXD$), ενώ για τη μέθοδο Ward, η CD είναι το άθροισμα των τετραγώνων μεταξύ των δύο ομάδων (between groups). Προκειμένου να ενωθούν δυο συστάδες λοιπόν, η τιμή της CD θα πρέπει να είναι μικρή.

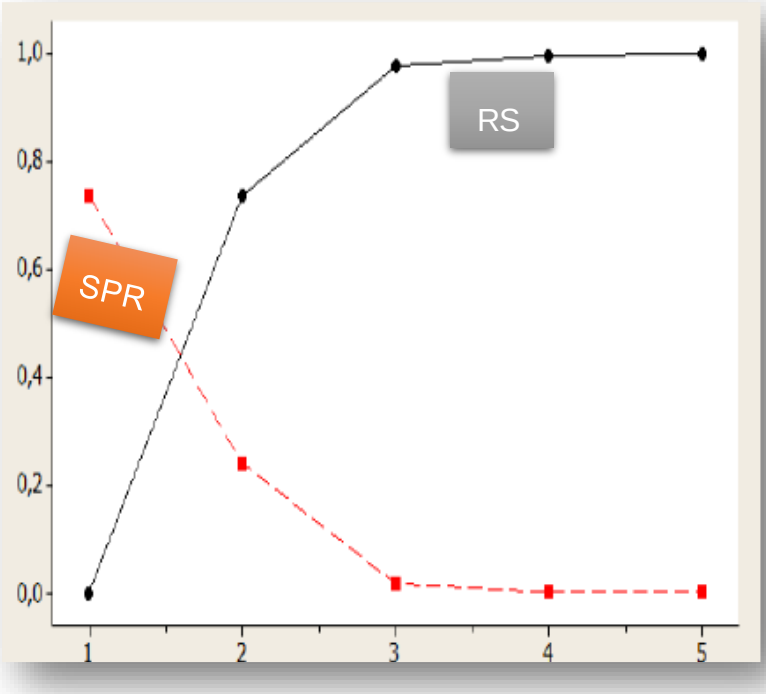
Ο παρακάτω πίνακας περιέχει μια σύνοψη όλων των στατιστικών που αναφέρθηκαν παραπάνω.

Στατιστικό	Έννοια που μετράται	Σχόλια
RMSSTD	Ομοιογένεια του νέου cluster	Η τιμή θα πρέπει να είναι μικρή
SPR	Ομοιογένεια των συγχωνευμένων cluster	Η τιμή θα πρέπει να είναι μικρή
RS	Ετερογένεια των cluster.	Η τιμή θα πρέπει να είναι υψηλή
CD	Ετερογένεια των συγχωνευμένων cluster	Η τιμή θα πρέπει να είναι μικρή

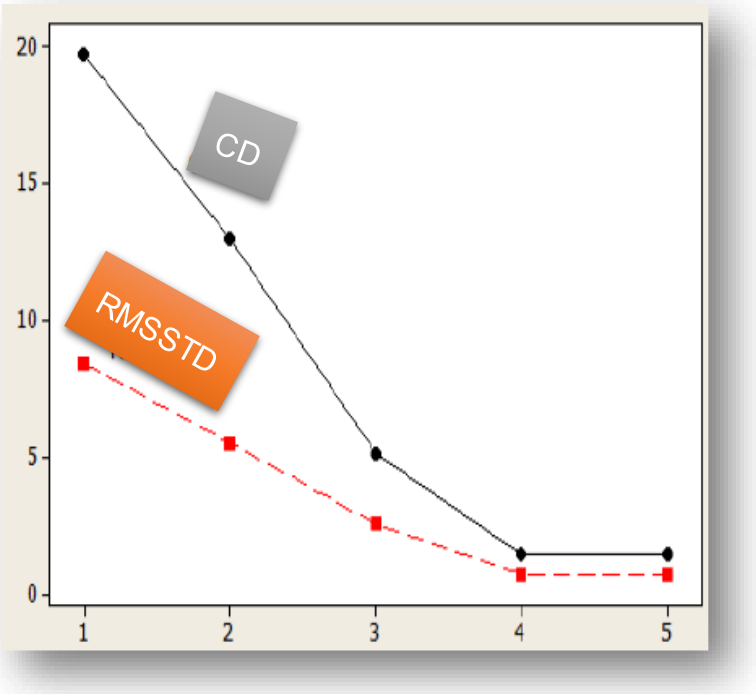
Πίνακας 2 Σύνοψη στατιστικών για την ομοιογένεια ή μη μιας συστάδας (cluster)

Το παρακάτω σχήμα δίνει παραδείγματα των διαγραμμάτων RS, SPR, RMSSTD και CD. Από αυτά, είναι σαφές ότι υπάρχει μια έντονη μεταβολή μεταξύ της λύσης των δύο συστάδων έναντι των τριών. Άρα, συμπεραίνουμε ότι το σύνολο των δεδομένων πάνω στα οποία εφαρμόστηκαν οι συναρτήσεις, χωρίζεται καλά σε 3 συστάδες, όπως αυτό φαίνεται από το διάγραμμα του RS και τις συστάδες.

Τέλος, από τις χαμηλές τιμές των SPR, RMSSTD και CD καταλαβαίνουμε ότι οι συστάδες μας είναι και ομοιογενείς.



Εικόνα 9 Διαγράμματα SPR και RS



Εικόνα 10 Διαγράμματα CD και RMSSTD

ΑΘΡΟΙΣΜΑ WSS

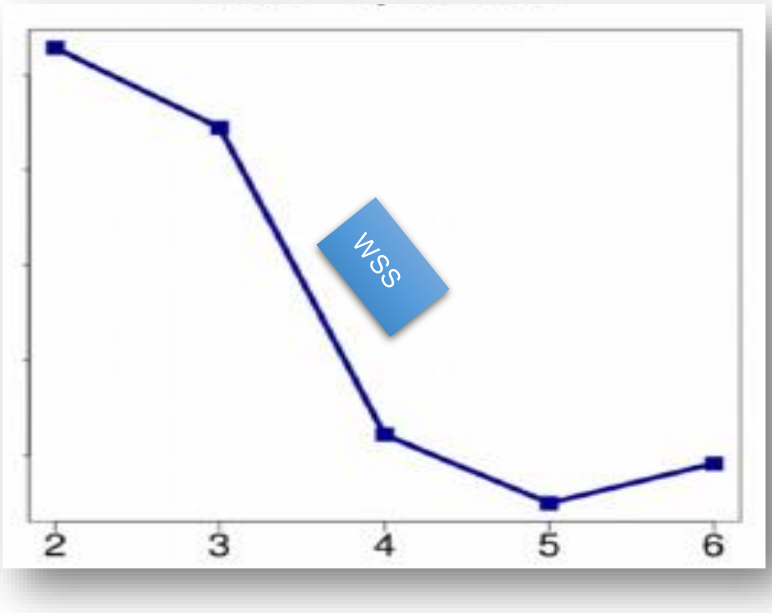
Τέλος, ένα πολύ σημαντικό κριτήριο, το οποίο χρησιμοποιείται αρκετά συχνά, είναι αυτό του αθροίσματος *WSS (Within-Sum-of-Squares).* Πρόκειται για ένα αντικειμενικό κριτήριο, το οποίο με τη βοήθεια μιας ποσότητας μετρά το βαθμό συνεκτικότητας, δηλαδή πόσο συμπαγείς είναι οι δημιουργούμενες ομάδες (Thorndike, 1953). Η ποσότητα αυτή ορίζεται ως εξής:

$$WSS = \sum_{i=1}^{k} \sum_{x \epsilon C_i} (x - m_i)^2$$

όπου i η κάθε συστάδα, x η κάθε παρατήρηση και m_i το κέντρο της συστάδας στην οποία ανήκει η κάθε παρατήρηση x. Ισούται, δηλαδή, με το άθροισμα των τετραγώνων των αποστάσεων των παρατηρήσεων από τα k κέντρα των συστάδων στις οποίες ανήκουν.

Καθώς μεγαλώνει ο αριθμός των συστάδων, το άθροισμα WSS μειώνεται. Κάτι τέτοιο είναι αναμενόμενο διότι όσο πιο μεγάλος είναι ο αριθμός των συστάδων, τόσο μικρότερες είναι οι αποστάσεις των παρατηρήσεων που ανήκουν στην εκάστοτε συστάδα. Επειδή, λοιπόν, η μείωση αυτή δεν είναι σταθερή, ο τρόπος για να αποφασίσει ο ερευνητής ποιος είναι ο βέλτιστος αριθμός συστάδων είναι, παρατηρώντας στην γραφική παράσταση του WSS συναρτήσει του k το σημείο στο οποίο δημιουργείται μια απότομη μείωση. Το σημείο αυτό θεωρείται το βέλτιστο k (*Μέθοδος του αγκώνα- Elbow Method*) (Νικολάου, 2015).

Στο παρακάτω σχήμα, απεικονίζεται ένα παράδειγμα μιας γραφικής παράστασης του WSS συναρτήσει του k για κάποιο σύνολο δεδομένων. Από ότι φαίνεται, ο βέλτιστος αριθμός των ομάδων είναι ο 5.

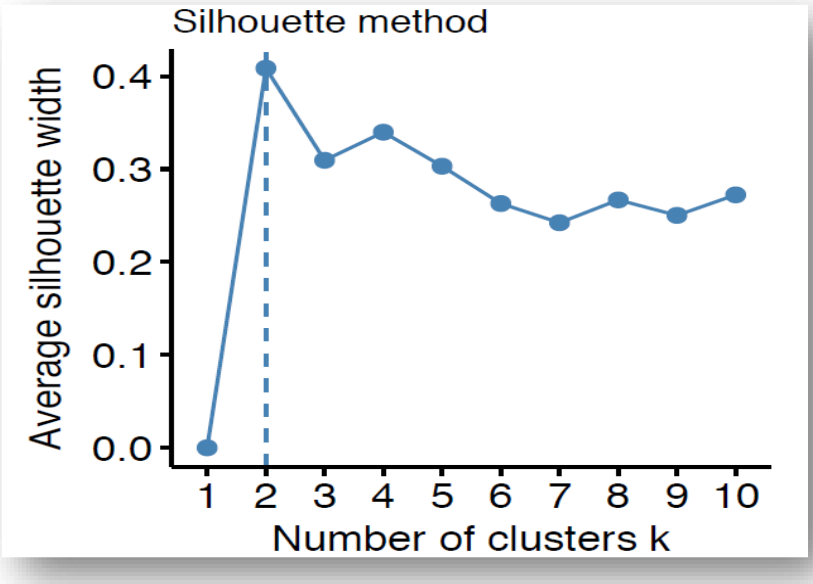


Εικόνα 11 Διάγραμμα WSS

ΜΕΘΟΔΟΣ ΣΙΛΟΥΕΤΑΣ (SILHOUETTE METHOD)

Η *μέθοδος Silhouette* μετρά την ποιότητα μιας ομαδοποίησης. Δηλαδή, καθορίζει το πόσο καλά κάθε αντικείμενο βρίσκεται μέσα στο σύμπλεγμα του.

Ένα υψηλό μέσο πλάτος σιλουέτας υποδηλώνει καλή ομαδοποίηση. Η μέθοδος μέσου Silhouette πλάτους (Average Silhouette Width) υπολογίζει και απεικονίζει στο αντίστοιχο διάγραμμα τη μέση σιλουέτα των παρατηρήσεων για τις αντίστοιχες τιμές του k. Ο βέλτιστος αριθμός συμπλεγμάτων k είναι αυτός που μεγιστοποιεί τη μέση σιλουέτα σε ένα εύρος πιθανών τιμών για το k (Kaufman & Rousseeuw, 1990). Παρακάτω δίνεται η μορφή ενός τέτοιου διαγράμματος.



Εικόνα 12 Διάγραμμα Silhouette Method

ΔΕΙΚΤΗΣ DUNN (Dunn index - D)

Ο *δείκτης Dunn* υπολογίζεται από τον τύπο

$$D = \frac{\min separation}{\max diameter}$$

όπου

- **min** $separation$ η min απόσταση μεταξύ της κάθε παρατήρησης της κάθε κλάσης με την κάθε παρατήρηση των υπολοίπων κλάσεων

- $\max diameter$ η μέγιστη απόσταση μεταξύ των συστάδων, δηλαδή η μέγιστη διάμετρος (η οποία υπολογίζεται από την απόσταση μεταξύ των παρατηρήσεων που ανήκουν στην ίδια κλάση).

✓ Ο δείκτης Dunn πρέπει να μεγιστοποιηθεί.

4.3 ΕΠΙΛΟΓΗ ΑΡΧΙΚΩΝ ΚΕΝΤΡΩΝ

Τα αρχικά κέντρα των συστάδων μπορούν είτε να οριστούν τυχαία από τον ερευνητή είτε να υπολογιστούν μέσα από κάποιο αλγόριθμο. Ένας τέτοιος αλγόριθμος είναι και ο παρακάτω και εκτελείται πριν από την έναρξη του αλγορίθμου της *K-Means* (Norušis, 2011).

ΑΛΓΟΡΙΘΜΟΣ ΑΡΧΙΚΩΝ ΚΕΝΤΡΩΝ

1. Επίλεξε k παρατηρήσεις ως αρχικά κέντρα.

Για κάθε παρατήρηση έλεγξε τις παρακάτω συνθήκες:

2. Αν η min από τις αποστάσεις της παρατήρησης αυτής από τα ήδη υπάρχοντα κέντρα είναι πιο μεγάλη από την απόσταση των δυο πιο κοντινών ήδη υπαρχόντων κέντρων, τότε η παρατήρηση αυτή αντικαθιστά τα ήδη υπάρχοντα κέντρα. Έστω δηλαδή $d(x, y)$ η απόσταση των παρατηρήσεων x και y και $c_j,\ j = 1,\dots,k,\ j$ = τα υπάρχοντα κέντρα. Υπολογίζουμε για την κάθε παρατήρηση i τις αποστάσεις της $d_j = d(x_i, c_j),\ j = 1,\dots,k$. Επίσης, υπολογίζουμε τις αποστάσεις ανάμεσα στα υπάρχοντα κέντρα, δηλαδή $d_{ij} = d(c_i, c_j),\ i, j = 1,\dots,k,\ i \neq j$. Τέλος, έλεγξε εάν $\min_j(d_j) > \min_{i,j}(d_{ij})$ και εάν η ανισότητα αυτή για κάποιο κέντρο ισχύει, τότε αντικατάστησε την παρατήρηση αυτή με το κέντρο που είναι πιο κοντά της.

- ✓ Με τον αλγόριθμο αυτό, ελέγχοντας όλες τις παρατηρήσεις μία μία, τα κέντρα που καθορίζονται, είναι και τα αρχικά κέντρα που θα χρησιμοποιήσει ο αλγόριθμος *K-Means* κατά την έναρξή του.

4.4 ΕΚΤΙΜΗΣΗ ΤΗΣ ΠΟΙΟΤΗΤΑΣ ΤΗΣ ΛΥΣΗΣ

Για να ελέγξουμε την ποιότητα της λύσης του αλγορίθμου, το πιο σύνηθες είναι να μετρήσουμε το άθροισμα των τετραγώνων του σφάλματος (Sum of Squared Error (SSE)) (Tan et al., 2006). Για κάθε παρατήρηση, ως σφάλμα ορίζεται η απόστασή της από την πιο κοντινή συστάδα. Για να υπολογίσουμε, λοιπόν, το SSE υπολογίζουμε το άθροισμα των τετραγώνων των σφαλμάτων της κάθε παρατήρησης, όπως φαίνεται και στον παρακάτω τύπο:

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} (x - m_i)^2$$

όπου C_i , $i = 1, \dots, k$ είναι οι συστάδες, x οι παρατηρήσεις και m_i το κέντρο της συστάδας C_i.

Στην πραγματικότητα, ο αλγόριθμος προσπαθεί να ελαπώσει επαναληπτικά την απόσταση όλων των παρατηρήσεων ως προς ένα σημείο της συστάδας. Επιπλέον, είναι εύκολο να αποδειχτεί ότι το σημείο που ελαχιστοποιείται το SSE για κάθε συστάδα είναι ο μέσος όρος

$$c_i = \frac{\sum_{x \epsilon C_i} x}{M_i}$$

όπου M_i το πλήθος των στοιχείων της συστάδας i.

Έτσι, με δεδομένες δύο συστάδες, μπορούμε να επιλέξουμε εκείνη που έχει το μικρότερο σφάλμα.

Τέλος, ο τρόπος για βελτίωση της ταξινόμησης, δηλαδή για μείωση του SSE είναι να αυξήσουμε τον αριθμό k. Σε γενικές γραμμές όμως, μια καλή ταξινόμηση με μικρό k είναι πιθανό να έχει μικρότερο SSE σε σχέση με μια κακή ταξινόμηση μεγαλύτερου k.

ΤΑ ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ K-MEANS

- Είναι εύκολα κατανοητός και απλός.
- Τα αντικείμενα χωρίζονται σε συστάδες αυτόματα.
- Δημιουργεί συστάδες πιο συμπαγείς (ομοιογενείς) σε σχέση με αυτές των Ιεραρχικών Μεθόδων, και ιδιαίτερα σε περιπτώσεις σφαιρικής μορφής.
- Όπως ειπώθηκε και πιο πάνω, είναι πολύ πιο γρήγορος συγκριτικά με τις Ιεραρχικές Μεθόδους αφού έχει πολυπλοκότητα $O(I * n * k * d)$ και γι' αυτό το λόγο προτιμάται σε περιπτώσεις μεγάλου όγκου δεδομένων.

ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ K-MEANS

- Το βασικότερο μειονέκτημα της μεθόδου είναι ότι ο ερευνητής πρέπει να προκαθορίσει τον αριθμό των συστάδων.
- Η τελική λύση εξαρτάται σε μεγάλο βαθμό από την επιλογή των αρχικών κέντρων. Διαφορετικά αρχικά κέντρα μπορούν να οδηγήσουν σε διαφορετικές τελικές λύσεις του αλγορίθμου.
- Επηρεάζεται πολύ από την ύπαρξη ακραίων τιμών (outliers).
- Η μέθοδος αυτή τείνει να δημιουργεί ίδιου μεγέθους και σφαιρικές συστάδες. Έτσι, δεν πρέπει να χρησιμοποιείται για συστάδες με διαφορετικά μεγέθη ή περίπλοκα σχήματα.

✓ Για την αντιμετώπιση των μειονεκτημάτων της μεθόδου *K-Means*, και ιδιαίτερα για το πρόβλημα του προκαθορισμού των συστάδων, έχουν συζητηθεί διάφορες λύσεις. Μια από αυτές θα μπορούσε να είναι η εφαρμογή σε πρώτη φάση του αλγορίθμου της Ιεραρχικής Μεθόδου για

να δούμε πόσες συστάδες θα σχηματιστούν και στη συνέχεια η εφαρμογή του αλγορίθμου της μεθοδου *K-Means,* χρησιμοποιώντας τον αριθμό των συστάδων που βρήκαμε αρχικά. Φυσικά, όπως έχουμε ήδη αναφέρει, με τα τεχνολογικά εργαλεία που έχουμε στα χέρια μας πλέον, δεν χρειάζεται να εκτελέσουμε κάποια τέτοια διαδικασία, αφού μέσα από τις γραφικές παραστάσεις που μπορούν να σχεδιαστούν, μπορούμε να επιλέξουμε αυτόματα και γρήγορα τον βέλτιστο αριθμό συστάδων αλλά και τα αρχικά μας κέντρα.

4.5 ΜΕΘΟΔΟΣ CLARA

Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος k-Means είναι ευαίσθητος στην ύπαρξη ακραίων τιμών. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι ο αλγόριθμος *k-Medoids*, ο οποίος χρησιμοποιεί ως κέντρο ένα υπαρκτό σημείο του σετ δεδομένων και όχι ένα υπολογιζόμενο μέσο σημείο. Μια από τις πρώτες εκδοχές του αλγόριθμου k-Medoids ήταν η μέθοδος *Partitioning Around Medoids – PAM* (Kaufman & Rousseeuw, 1990).

Αναλυτικότερα, στον αλγόριθμο k-Medoids επιλέγονται αρχικά k σημεία ως κέντρα (medoids). Τα υπόλοιπα σημεία κατατάσσονται στη συστάδα του πλησιέστερου κέντρου. Μια συνάρτηση κόστους μετρά το άθροισμα των αποστάσεων όλων των σημείων από το κέντρο της συστάδας τους. Σε μια επαναληπτική διαδικασία, σημεία τα οποία δεν είναι κέντρα δοκιμάζονται ως πιθανά κέντρα. Εάν για ένα σημείο το κόστος γίνεται μικρότερο, τότε το σημείο αυτό γίνεται το νέο κέντρο στη θέση του προηγούμενου. Ο αλγόριθμος k-Medoids λειτουργεί πιο αποτελεσματικά από τον k-Means, όταν στα δεδομένα υπάρχουν αντικείμενα με ακραίες τιμές. Ωστόσο, το κόστος υπολογισμού των medoids είναι σημαντικά μεγαλύτερο από το κόστος υπολογισμού των μέσων τιμών. Για τον λόγο αυτό, ο αλγόριθμος k-Medoids δεν αποδίδει καλά με μεγάλα σύνολα δεδομένων.

Μια βελτίωση του αλγορίθμου, η οποία αντιμετωπίζει αυτό το πρόβλημα, είναι η μέθοδος *CLARA (Clustering LARge Applcations)* (Kaufman & Rousseeuw, 1990). Η μέθοδος CLARA δε χρησιμοποιεί ολόκληρο το σύνολο

δεδομένων. Αντιθέτως, εκτελεί τυχαία δειγματοληψία και επιλέγει ένα υποσύνολο του. Το υποσύνολο δεδομένων υπόκειται σε Ανάλυση Συστάδων, σύμφωνα με τη μέθοδο PAM. Λόγω της τυχαίας δειγματοληψίας, είναι αρκετά πιθανό ότι τα medoids που θα υπολογιστούν, θα είναι όμοια με αυτά που θα προέκυπταν από την επεξεργασία ολόκληρου του συνόλου δεδομένων. Ο αλγόριθμος επιλέγει πολλά υποσύνολα δεδομένων και επιστρέφει το καλύτερο αποτέλεσμα.

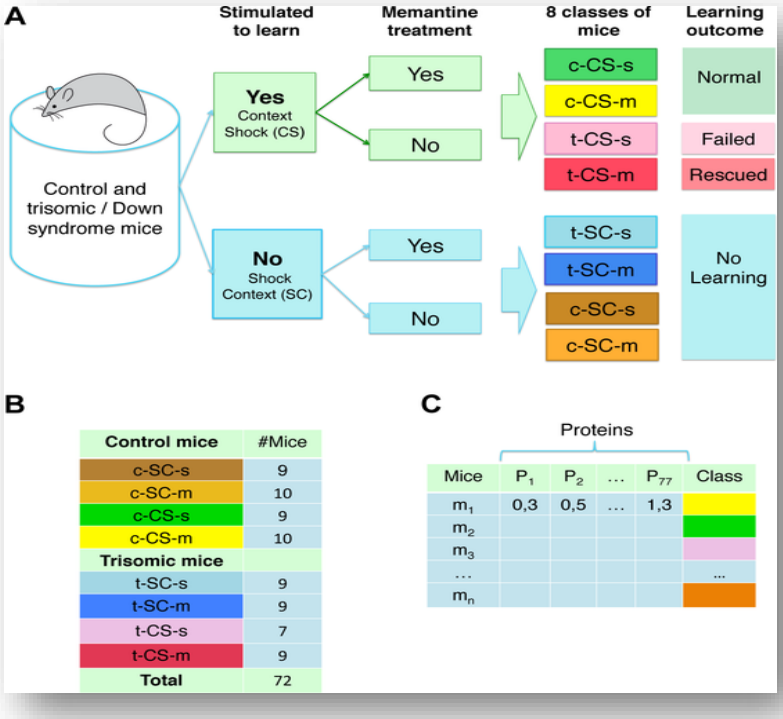
5 ΚΕΦΑΛΑΙΟ: ΕΦΑΡΜΟΓΗ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΣΤΑΔΩΝ ΣΕ ΒΙΟΪΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

Το επιλεγμένο σύνολο δεδομένων *«Σετ δεδομένων έκφρασης πρωτεΐνης ποντικών»* δημιουργήθηκε από πειράματα από τους Higuera, Gardiner, & Cios (2015). Αυτές οι έρευνες αποσκοπούσαν στην κατανόηση των επιπτώσεων του συνδρόμου Down (DS) στη μάθηση μέσω της ανάλυσης της έκφρασης πρωτεΐνης σε ποντίκια. Το DS στον άνθρωπο προκαλείται από την παρουσία ενός επιπρόσθετου χρωμοσώματος, το οποίο ονομάζεται τρισωμία. Η έκφραση της πρωτεΐνης διακόπτεται από την ανθρώπινη τρισωμία, οδηγώντας σε φυσικές και πνευματικές εκδηλώσεις που συνδέονται με το DS. Λόγω της μεγάλης διάδοσης του συνδρόμου (1 στις 1000 γεννήσεις παγκοσμίως) και των επιπτώσεων στην υγεία, υπάρχει ισχυρή ανάγκη για περαιτέρω κατανόηση και αντιμετώπιση της πάθησης αυτής.

Για τη δημιουργία του σετ δεδομένων, οι Higuera et al. χρησιμοποίησαν ποντίκια τρισωμίας (Ts65Dn) αλλά και ελέγχου, δηλαδή υγιή ποντίκια (control mice), εκθέτοντάς τα σε συνθήκες περιβάλλοντος φόβου. Κάποια από αυτά εκχήθηκαν με μεμαντίνη (memantine) και κάποια με αλατούχο ορό (saline). Τέλος, κάποια έλαβαν ερεθίσματα μάθησης (CS) ενώ άλλα όχι (SC). Τα ποντίκια στη συνέχεια υποβλήθηκαν σε ευθανασία και μετρήθηκαν τα επίπεδα 77 πρωτεϊνών ανιχνεύσιμων στον εγκεφαλικό φλοιό τους. Καταγράφηκαν 15 μετρήσεις για κάθε πρωτεΐνη ανά ποντίκι. Άρα, το σύνολο δεδομένων έχει διαστάσεις 1080×77 . Από τις τρεις παραπάνω δυαδικές μεταβλητές δημιουργήθηκαν οι οκτώ ομάδες ποντικιών (πραγματικές κλάσεις). Οι ομάδες αυτές είναι:

- c-CS-s: ποντίκια ελέγχου, διεγερμένα για μάθηση, στα οποία εγχύθηκε αλατούχος ορός (9 ποντίκια)
- c-CS-m: ποντίκια ελέγχου, διεγερμένα για μάθηση, στα οποία εγχύθηκε μεμαντίνη (10 ποντίκια)
- c-SC-s: ποντίκια ελέγχου, μη διεγερμένα για μάθηση, στα οποία εγχύθηκε αλατούχος ορός (9 ποντίκια)

- c-SC-m: ποντίκια ελέγχου, μη διεγερμένα για μάθηση, στα οποία εγχύθηκε μεμαντίνη (10 ποντίκια)
- t-CS-s: ποντίκια τρισωμίας, διεγερμένα για μάθηση, στα οποία εγχύθηκε αλατούχος ορός (7 ποντίκια)
- t-CS-m: ποντίκια τρισωμίας, διεγερμένα για μάθηση, στα οποία εγχύθηκε μεμαντίνη (9 ποντίκια)
- t-SC-s: ποντίκια τρισωμίας, μη διεγερμένα για μάθηση, στα οποία εγχύθηκε αλατούχος ορός (9 ποντίκια)
- t-SC-m: ποντίκια τρισωμίας, μη διεγερμένα για μάθηση, στα οποία εγχύθηκε μεμαντίνη (9 ποντίκια)



Εικόνα 13 Απεικόνιση των πραγματικών κλάσεων των δεδομένων

Σκοπός της εργασίας αυτής είναι να διαπιστώσει εάν η γνώση των πρωτεϊνών οδηγεί από μόνη της στο διαχωρισμό των ποντικιών στις κλάσεις αυτές, εάν διαχωρίζονται τα υγιή από τα DS ποντίκια, εάν διαχωρίζονται διαφορετικά ή αν μπορούμε να πάρουμε κάποια πληροφορία για τη μάθηση και το φάρμακο που δόθηκε. Για το σκοπό αυτό θα χρησιμοποιηθεί η μέθοδος

Ανάλυσης Συστάδων, η εφαρμογή της οποίας θα γίνει μέσω του στατιστικού προγράμματος R-Studio.

Βιβλιοθήκες R-studio

Κατά τη διάρκεια της ανάλυσης χρησιμοποιήθηκαν οι βιβλιοθήκες gdata, VIM, mice, ggplot2, colorspace, ElemStatLearn, cluster, factoextra,magrittr, clustertend, NbClust, clValid, dendextend, gsheet, rmarkdown, car, dplyr, gridExtra, htmltools, reshape2, granova, psychometric, Hmisc, outliers, GGally, survminer και pvclust.

Δεδομένα

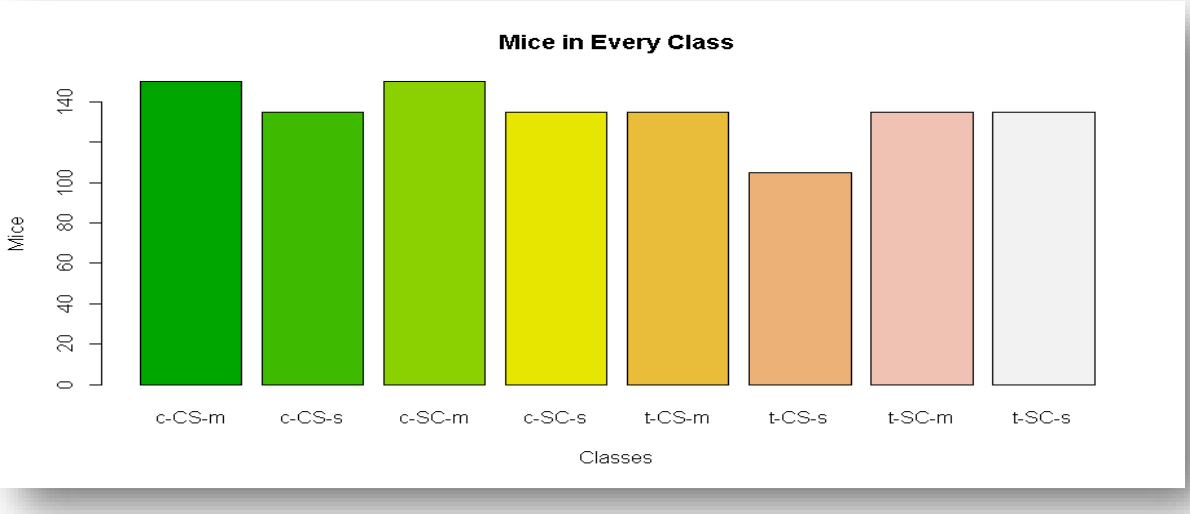
Το αρχείο Excel 'dataSet.xls' ελήφθη από τον διαδικτυακό αποθηκευτικό χώρο ("https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression") της UCI για εκμάθηση μηχανών (machine learning) και εισήχθη στο διαδραστικό περιβάλλον R Studio ως "ds".

```
mydata = read.xls("C:/Users/dataSet.xls",  perl="C:/Perl/bin/perl.exe",
                  header=T,row.names=1,   method="tab", na.strings=c("NA",""))
```

5.1 ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

```
## Περιγραφική για τις αριθμητικές μεταβλητές
>summary(mydata)


## Plot για την κατανομή των ποντικιών στις πραγματικές κλάσεις πριν την επεξεργασία του
σετ
   (Εικόνα 14)
>classes1 <- mydata$class
>plot(classes1,col=terrain.colors(8),legend=levels(classes1),width=0.25)
>title(main="Mice in Every Class",xlab="Classes",ylab="Mice")

## Περιγραφική για τις κατηγορικές μεταβλητές (levels και ποσοστά κάθε level)
>levels(factor(mydata$Genotype))
"Control" "Ts65Dn"
> levels(factor(mydata$Treatment))
 "Memantine"  "Saline"
> levels(factor(mydata$Behavior))
 "C/S" "S/C"
> levels(factor(mydata$class))
"c-CS-m" "c-CS-s" "c-SC-m" "c-SC-s" "t-CS-m" "t-CS-s" "t-SC-m" "t-SC-s"
```

```
> prop.table(table(mydata$Genotype))
Control    Ts65Dn
0.5277778 0.4722222
> prop.table(table(mydata$Behavior))
     C/S       S/C
0.4861111 0.5138889
> prop.table(table(mydata$class))
    c-CS-m       c-CS-s       c-SC-m       c-SC-s       t-CS-m       t-CS-s       t-SC-m
0.13888889 0.12500000 0.13888889 0.12500000 0.12500000 0.09722222 0.12500000
    t-SC-s
0.12500000
> prop.table(table(mydata$Treatment))
Memantine    Saline
0.5277778 0.4722222
```



Εικόνα 14 Plot για την κατανομή των ποντικιών στις πραγματικές κλάσεις (πριν την επεξεργασία των δεδομένων)

Ελλειπείς Τιμές (NA)

```
## Πόσο % των γραμμών έχει NA?
>sum(apply(mydata,  1, function(x)sum(is.na(x)))!=0)/nrow(mydata)
0.4888889
```

Όπως βλέπουμε το 48% των γραμμών έχει ελλειπείς τιμές, γι' αυτό και δεν θα τις αφαιρέσουμε από το σύνολο δεδομένων, μιας και θα χάσουμε πολλές πληροφορίες.

```
## Πώς κατανέμονται  τα NA στις γραμμές
>table(apply(mydata,  1, function(x)sum(is.na(x))))

  0   1   2   3   4   5 43
552 165 120 117 104  19  3
```

Από όπι βλέπουμε, από 3 γραμμές λείπουν 43 από τις 77 πρωτεΐνες, οπότε αυτά τα 3 ποντίκια μόνο τα αφαιρούμε από το δείγμα και στη συνέχεια αντικαθιστούμε όλα τα υπόλοιπα NA με το μέσο όρο, ο οποίος υπολογίζεται με βάση τις διαθέσιμες τιμές της κλάσης στην οποία ανήκει το εκάστοτε ποντίκι.

```
## Αφαιρώ τις 3 γραμμές με 43 NAs.
>mydata2<-mydata[!(apply(mydata,  1, function(x)sum(is.na(x)))==43),]

## «Γεμίζω» τα NAs
>mydata3<- mydata2[,-((ncol(mydata2)-3):ncol(mydata2))]
>for(i in 1:nrow(mydata3)){
  for(j in 1:ncol(mydata3)){
    if(is.na(mydata3[i,j])){
      mydata3[i,j] <- tapply(mydata3[,j],classes,  mean, na.rm=T)[classes[j]]
}}}
```

Τέλος, επειδή η κλίμακα των μεταβλητών είναι διαφορετική, πρέπει να τις μετατρέψουμε ώστε να τις φέρουμε σε συγκρίσιμη μορφή, για να πετύχουμε καλύτερη απόδοση. Έτσι, κανονικοποιώ τα δεδομένα μου.

```
## Κανονικοποιώ το σετ δεδομένων
>scale(mydata3)
>df=mydata3
```

5.2 ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ

ΤΑΣΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Το παρακάτω διάγραμμα δείχνει το πόσο καλά μπορεί να συσταδοποιηθεί το συγκεκριμένο σετ δεδομένων με την μέθοδο Ανάλυσης Συστάδων. Βασίζεται στον αλγόριθμο της οπτικής εκτίμησης της τάσης της συσταδοποίησης (visual assessment of cluster tendency- VAT) (Bezdek & Hathaway, 2002). Σύμφωνα με αυτόν, το επίπεδο χρώματος είναι ανάλογο με την τιμή της ανομοιότητας μεταξύ των παρατηρήσεων: καθαρό κόκκινο εάν $dist(x_i, x_j) = 0$ και καθαρό μπλε εάν $dist(x_i, x_j) = 1$. Τα αντικείμενα που ανήκουν στο ίδιο σύμπλεγμα εμφανίζονται με διαδοχική σειρά (Kassambara, 2017). Πιο αναλυτικά, στο διάγραμμα αυτό, όσο πιο έντονες είναι οι χρωματικές διαφορές των δύο χρωματικών ομάδων που σχηματίζονται, τόσο πιο καλά είναι τα δεδομένα μας για συσταδοποίηση. Εάν για παράδειγμα, σε όλο το διάγραμμα ήταν παντού

έντονο το μπλε και κόκκινο χρώμα και δεν σχημάτιζαν- όπως παρακάτω- κάποιο μοτίβο, δε θα ήταν καλά τα δεδομένα για ταξινόμηση με τη μέθοδο Ανάλυσης Συστάδων.

```
## Διάγραμμα Τάσης Συσταδοποίησης
>fviz_dist(dist(df), show_labels = FALSE)+
  labs(title = "MICE PROTEINS DATA")
```

Εικόνα 15 Τάση συσταδοποίησης (Tendency of Clustering)

Η παραπάνω εικόνα ανομοιότητας επιβεβαιώνει ότι υπάρχει δομή συμπλέγματος στο σετ δεδομένων που επεξεργαζόμαστε εδώ, άρα ταξινομείται καλά με τη μέθοδο που χρησιμοποιούμε.

ΕΠΙΛΟΓΗ ΚΑΛΥΤΕΡΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΑΞΙΝΟΜΗΣΗΣ

Από το πακέτο clValid της R (Brock, Pihur & Datta, 2008), ο παρακάτω κώδικας υπολογίζει τον καλύτερο αλγόριθμο ταξινόμησης και τον καλύτερο αριθμό συστάδων για το συγκεκριμένο σετ δεδομένων λαμβάνοντας υπόψη τη συνεκτικότητα (η οποία αντιστοιχεί στο βαθμό που τα αντικείμενα τοποθετούνται στην ίδια ομάδα σε σχέση με τους πλησιέστερους γείτονές τους, έχει τιμή μεταξύ 0 και άπειρο και πρέπει να ελαχιστοποιηθεί), τη μέθοδο της Σιλουέτας και το δείκτη Dunn.

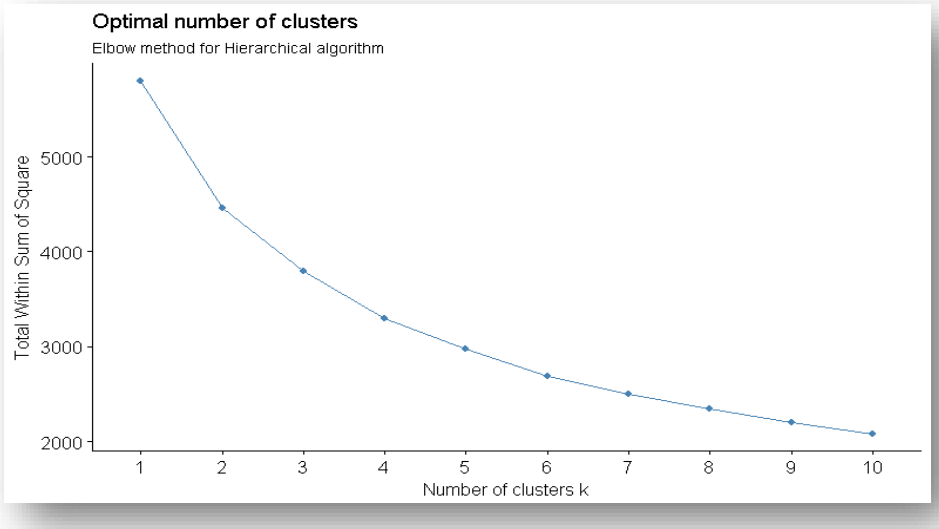
```
>clmethods <- c("hierarchical","kmeans","pam","clara")
>intern <- clValid(df, nClust = 2:8,maxitems = 1077,
                   clMethods = clmethods, validation = "internal")
>summary(intern)
Optimal Scores:

             Score   Method       Clusters
Connectivity 3.0290  hierarchical  2
Dunn         0.2534  hierarchical  3
Silhouette   0.5406  hierarchical  2
```

Όπως βλέπουμε, ο αλγόριθμος προτείνει την Ιεραρχική μέθοδο Ταξινόμησης, με 2 ή 3 συστάδες.

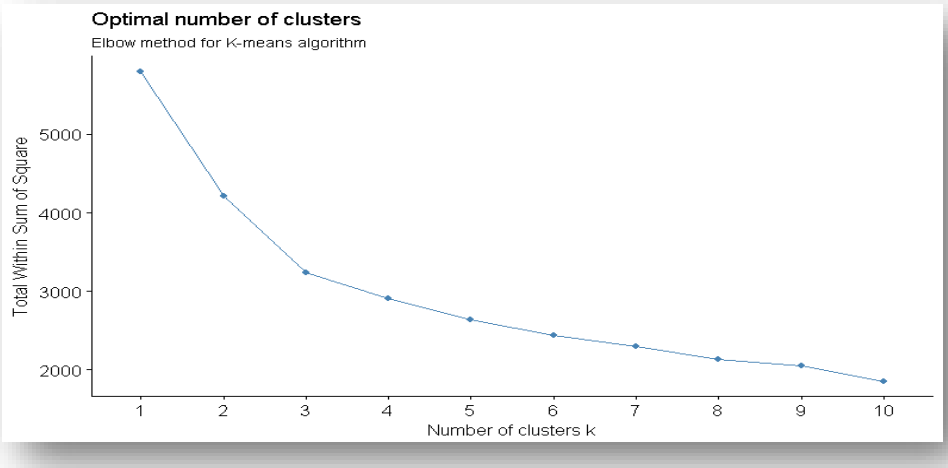
ΒΕΛΤΙΣΤΟΣ ΑΡΙΘΜΟΣ ΣΥΣΤΑΔΩΝ- Κ

- **ΜΕΘΟΔΟΣ ΤΟΥ ΑΓΚΩΝΑ (ELBOW METHOD)**

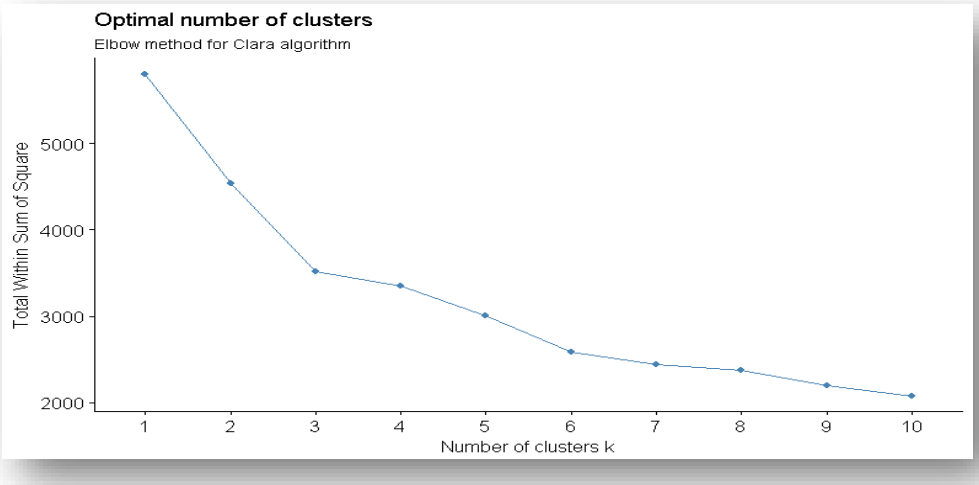
```
fviz_nbclust(df, hcut,                    ## Αντίστοιχα kmeans ή clara
   method = "wss") +
  labs(subtitle = "Elbow method for Hierarchical algorithm")
```



Εικόνα 16 Διάγραμμα WSS (Elbow Method) για την Ιεραρχική Ταξινόμηση



Εικόνα 17 Διάγραμμα WSS (Elbow Method) για τον αλγόριθμο K-Means

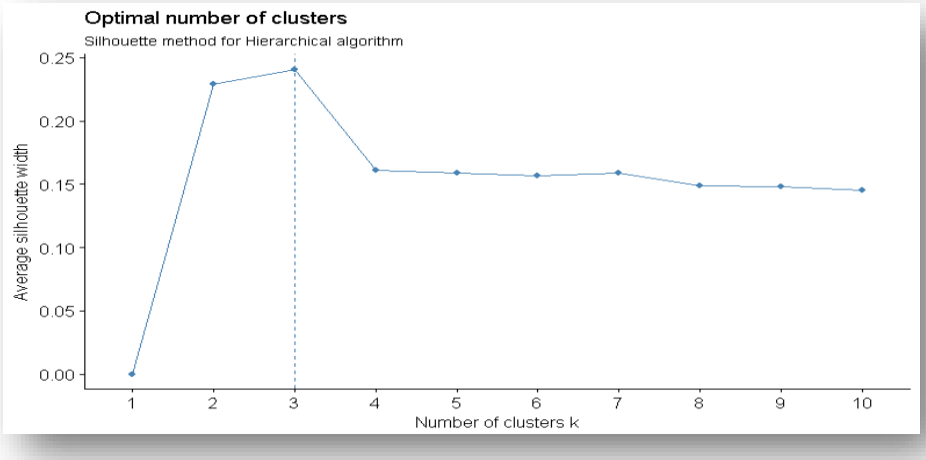


Εικόνα 18 Διάγραμμα WSS (Elbow Method) για τον αλγόριθμο Clara

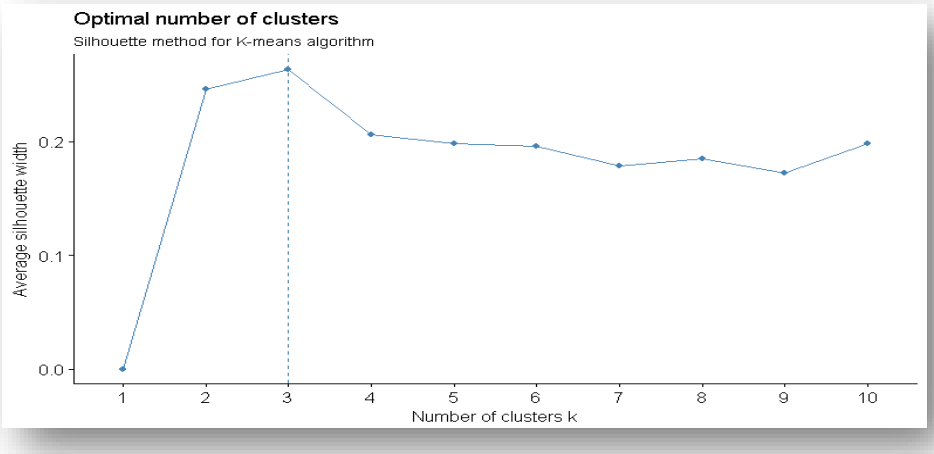
Από τα παραπάνω διαγράμματα βλέπουμε ότι για την Ιεραρχική μέθοδο, ο βέλτιστος αριθμός συστάδων είναι το 2, για την K-Means το 3 πιο έντονα και αμέσως μετά και το 2 και τέλος για την Clara το 3.

- **ΜΕΘΟΔΟΣ ΣΙΛΟΥΕΤΑΣ (SILHOUETTE METHOD)**

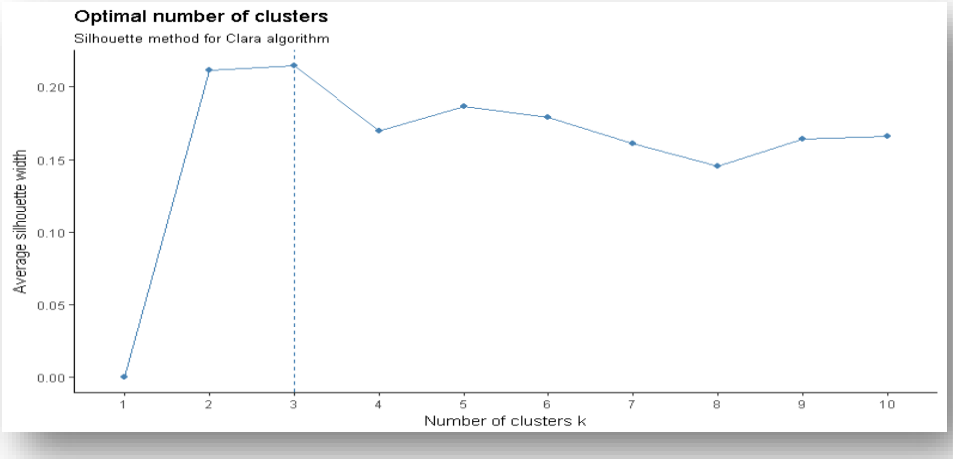
```
fviz_nbclust(df, hcut,                    ## Αντίστοιχα kmeans ή clara
 method = "silhouette")+
 labs(subtitle = "Silhouette method for Hierarchical algorithm")
```



Εικόνα 19 Διάγραμμα Silhouette για την Ιεραρχική Ταξινόμηση



Εικόνα 20 Διάγραμμα Silhouette για τον αλγόριθμο K-Means



Εικόνα 21 Διάγραμμα Silhouette για τον αλγόριθμο Clara

Στα παραπάνω διαγράμματα αναζητούμε το σημείο εκείνο που μεγιστοποιείται η γραμμή. Από ότι βλέπουμε, ο βέλτιστος αριθμός συστάδων για την Ιεραρχική μέθοδο και για την K-Means είναι το 3 και για την Clara το 3, με το 2 όμως να είναι σχεδόν στην ίδια ευθεία με το 3.

Σύμφωνα με όλα τα παραπάνω συμπεράσματα, θα αναζητήσουμε λύσεις με τις τρεις αυτές μεθόδους Ανάλυσης Συστάδων για k=2,3,8. Ο λόγος που θα αναζητήσουμε και για k=8, παρόπι δε μας το υπέδειξαν τα μέχρι τώρα μέτρα και διαγράμματα, είναι ότι αυτός είναι ο αριθμός στον οποίο έχουν χωριστεί στην πραγματικότητα οι παρατηρήσεις μας και θέλουμε να δούμε τι συμπεράσματα θα μπορούσαμε να βγάλουμε από μια τέτοια ομαδοποίηση.

DISTANCE

```
>res.dist <- get_dist(df, method = "spearman")        ## Αντίστοιχα manhattan, pearson,
                                                          kendall, euklidean κτλ
>res.hc <- hclust(d = res.dist, method = "average ")  ## Αντίστοιχα ward.D2, single,
                                                          complete, median κτλ

>cor(res.dist, cophenetic(res.hc))
0.7075158
```

Με τον παραπάνω κώδικα ελέγχουμε τον καλύτερο συνδυασμό απόστασης και μεθόδου ένωσης των ομάδων. Τιμές άνω του 0,75 θεωρούνται αρκετά καλές. Από όλους τους συνδυασμούς έχουμε τους 4 καλύτερους:

1. Spearman – average: *0.7075158*
2. *Pearson – average: 0.6547807*
3. Kendall – average: 0.6527228
4. Ευκλείδεια- average: 0.6440159

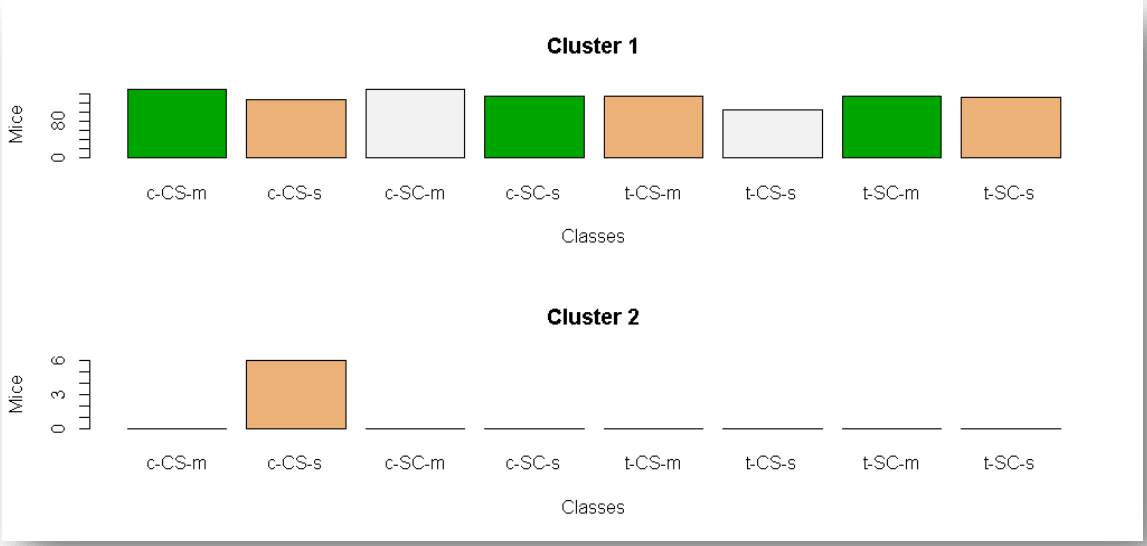
✓ *Θα ξεκινήσουμε λοιπόν την ανάλυσή μας με την Ιεραρχική Ταξινόμηση, η οποία βρέθηκε ως η καλύτερη προσαρμογή στα δεδομένα μας, χρησιμοποιώντας αυτές τις 4 αποστάσεις και στη συνέχεια θα εφαρμόσουμε και τους αλγορίθμους K-Means και Clara.*

5.2.1 Ιεραρχική Ταξινόμηση με απόσταση SPEARMAN

- **k=2**

```
## Ο παρακάτω κώδικας είναι αντίστοιχος και για τις υπόλοιπες περιπτώσεις αλγορίθμων ταξινόμησης method και αριθμού συστάδων k, οπότε θα αναφερθεί μία μόνο φορά.

>res.dist_spearman=get_dist(df, method = "spearman")        ## Αντίστοιχα pearson, kendall,
                                                                                    euclidean
>res.hc_spearman=hclust(d = res.dist_sperman, method = "average")

>table(grp2_spearman)                                    ## Αριθμός ποντικιών σε κάθε cluster
   1   2
1071   6

## Κόβω το δενδρόγραμμα σε 2 clusters
>grp2_spearman=cutree(res.hc_spearman,k = 2        )## Αντίστοιχα grp_3 και grp_8 για
                                                                                    k=3 και k=8

## Δενδρόγραμμα χωρισμένο σε 2 clusters, με διαφορετικό χρώμα το καθένα
>fviz_dend(res.hc_spearman,  k = 2,                        ## Αντίστοιχα για k=3,8
        cex = 0.5, k_colors ="jco", color_labels_by_k = TRUE, rect = TRUE,
        rect_border = c("#2E9FDF", "#00AFBB"), rect_fill = TRUE)

## Τα plots που ακολουθούν αναπαριστούν πόσες παρατηρήσεις έχουν μπει στο κάθε cluster και ποιές πραγματικές κλάσεις εμφανίζονται στο καθένα από αυτά.
> par(mfrow=c(2,1))
>plot(classes[grp2_spearman==1],col=terrain.colors(3))
>title(main="Cluster 1",xlab="Classes",ylab="Mice")
>plot(classes[grp2_spearman==2],col=terrain.colors(3))
>title(main="Cluster 2",xlab="Classes",ylab="Mice")
```

Όπως βλέπουμε και από την εντολή *table(grp2_spearman)* αλλά και από το παρακάτω διάγραμμα, σχεδόν όλα τα ποντίκια πάνε στο 1ο cluster και μόνο 6 πάνε στο 2ο . Κάτι τέτοιο διαφαίνεται και στο αντίστοιχο δενδόγραμμα. Άρα, ο διαχωρισμός που πρότεινε ο αλγόριθμος βέλτιστης μεθόδου δεν είναι στην πράξη καλός.



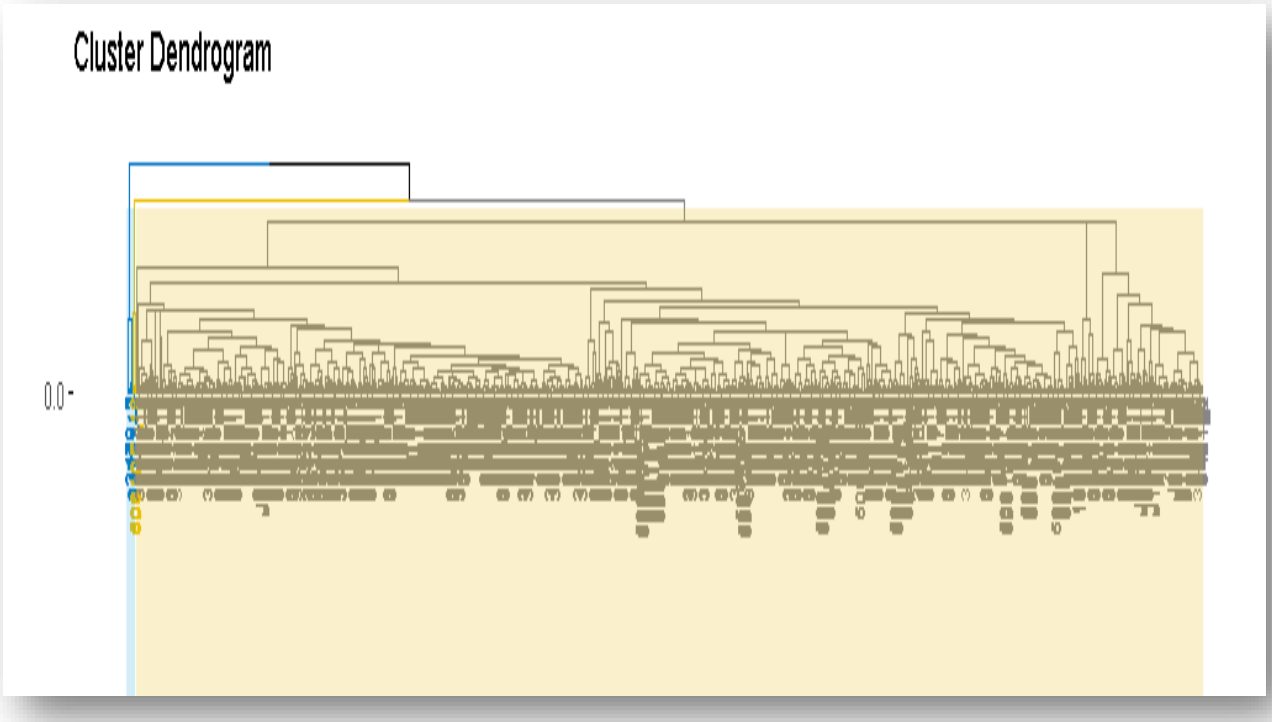
Εικόνα 22 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster

ΣΧΟΛΙΟ

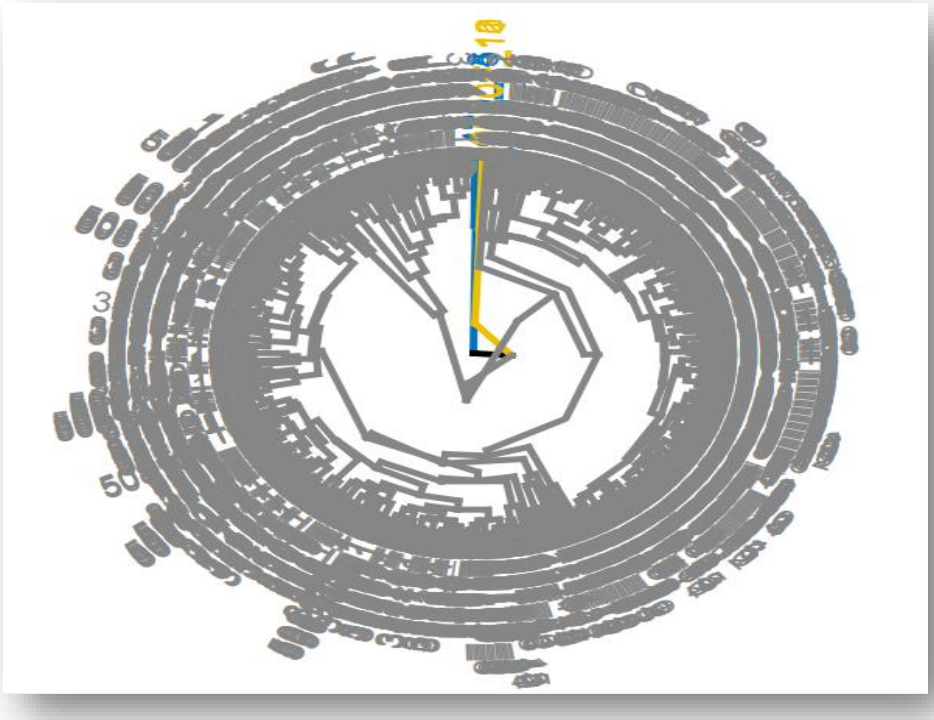
Η απόσταση Spearman, όπως ειπώθηκε σε προηγούμενη ενότητα, θεωρείται η καλύτερη σε συνδυασμό με την Ιεραρχική Ταξινόμηση για την προσαρμογή των δεδομένων σε ομάδες και αφού ούτε εδώ δεν γίνεται καλός διαχωρισμός για k=2, παραλείπεται το αντίστοιχο κομμάτι και στις υπόλοιπες περιπτώσεις Ιεραρχικής Ταξινόμησης με τις άλλες αποστάσεις.

- **k=3**

```
> table(grp3_spearman)
   1   2   3
1068   6   3

## Τέλος, τα barplots δείχνουν με τη σειρά τους ποιος τύπος γενότυπου, φαρμάκου και συμπεριφοράς στη μάθηση υπάρχει στην κάθε συστάδα αντίστοιχα.
>par(mfrow=c(1,2))
>plot(genotype[grp3_spearman==1],col=terrain.colors(2)    ## Αντίστοιχα για treatment και behavior
>title(main="Cluster 1",xlab="Genotype",ylab="Mice")
>plot(genotype[grp3_spearman==2],col=terrain.colors(2))   ## Αντίστοιχα για treatment και behavior
>title(main="Cluster 2",xlab="Genotype",ylab="Mice")

## Κυκλικό Δενδρόγραμμα χωρισμένο σε 3 clusters, με διαφορετικό χρώμα το καθένα
>fviz_dend(res.hc_spearman,  k =3,                    ## Αντίστοιχα για k=8
        cex = 0.5, k_colors ="jco", color_labels_by_k = TRUE, rect = TRUE,
        rect_border = c("#2E9FDF", "#00AFBB"), rect_fill = TRUE, type = "circular")
```

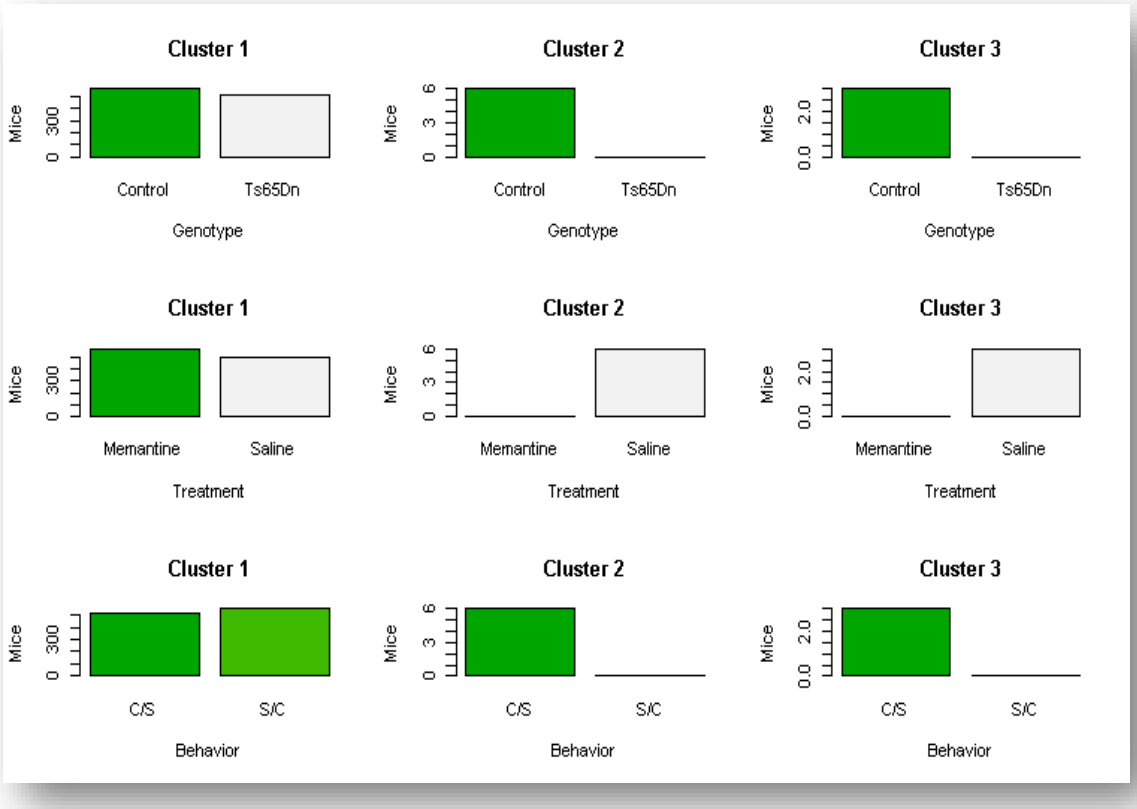


Εικόνα 23 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Spearman για k=3



Εικόνα 24 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Spearman για k=3

Εικόνα 25 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster

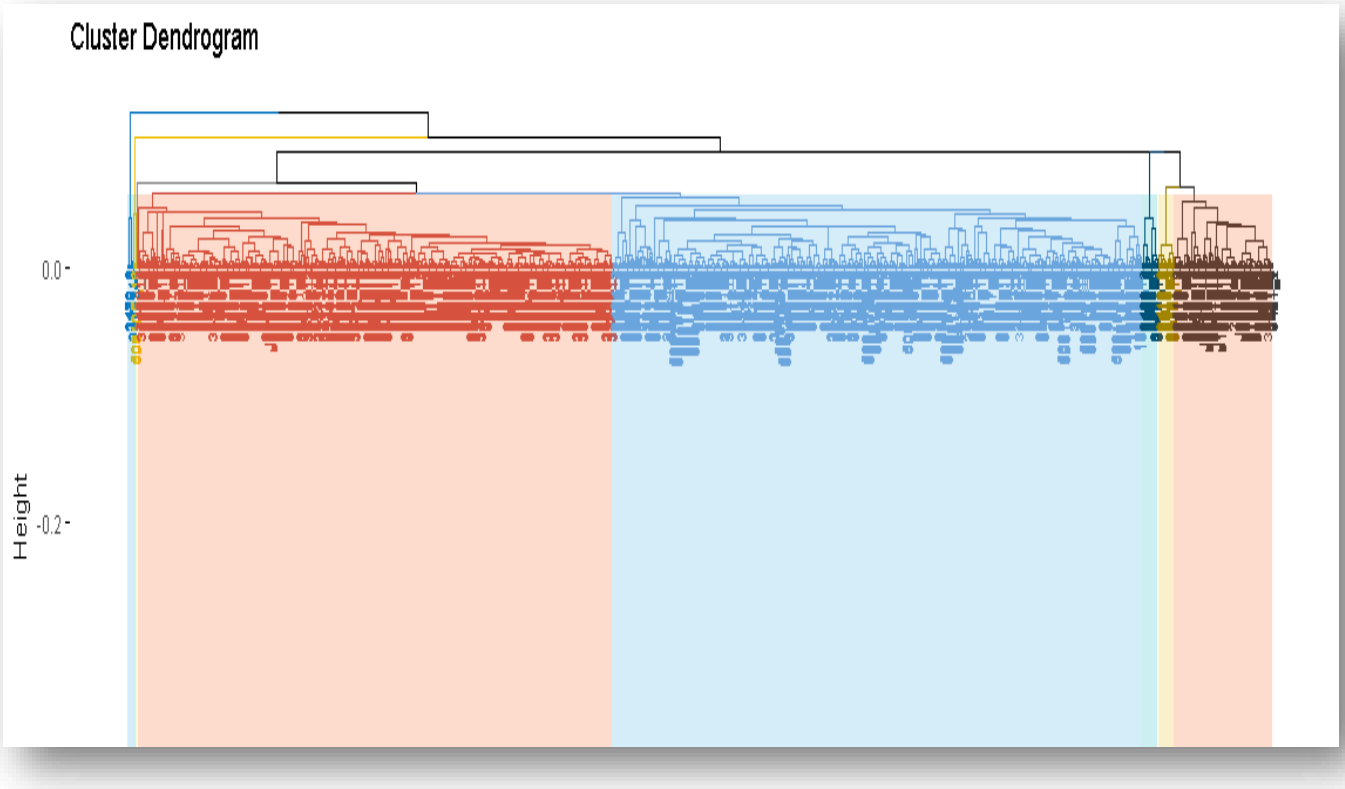


Εικόνα 26 Ποντίκια του κάθε γενότυπου (α), θεραπείας (β) και συμπεριφοράς (γ) στο κάθε cluster

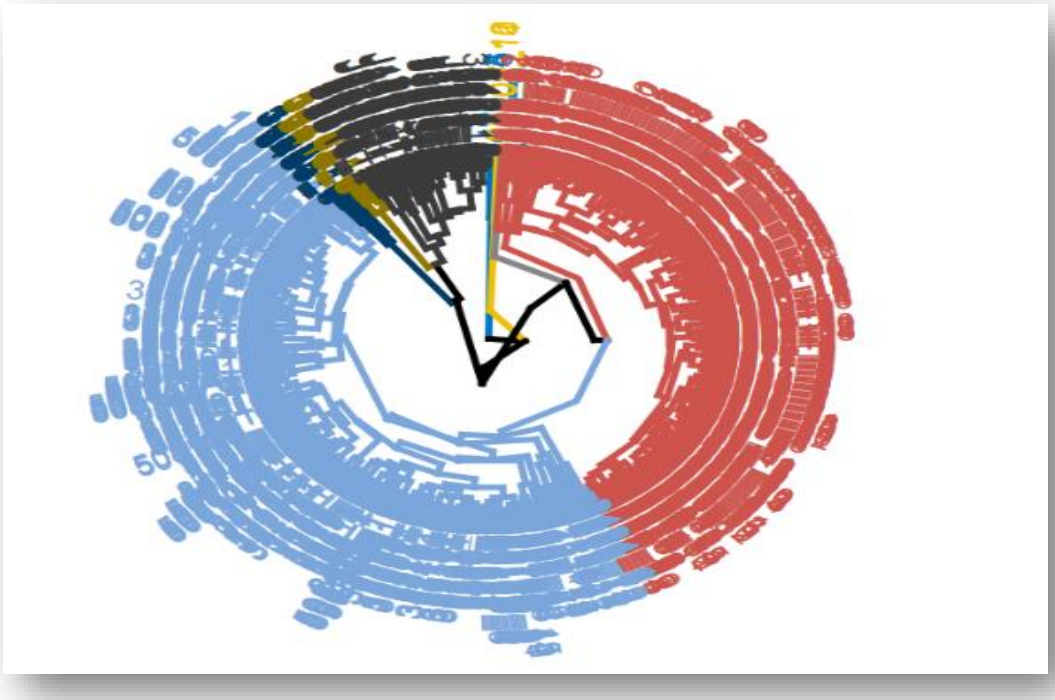
Από όλα τα παραπάνω, παρατηρούμε ότι στο 1ο cluster έχουμε υγιή ποντίκια, τα οποία έλαβαν μεμαντίνη και δε διεγέρθηκαν στη μάθηση. Στο 2ο και 3ο έχουμε υγιή, που έλαβαν ορό και διεγέρθηκαν στη μάθηση. Βέβαια, και σε αυτή την περίπτωση, μόνο 9 ίδιου τύπου ποντίκια βρίσκονται στις συστάδες 2 και 3 (θα μπορούσαμε δηλαδή να θεωρήσουμε αυτές τις δυο ως μία συστάδα), ενώ όλα τα υπόλοιπα βρίσκονται στο 1ο cluster. Άρα, ούτε εδώ είναι καλός ο διαχωρισμός.

- **k=8**

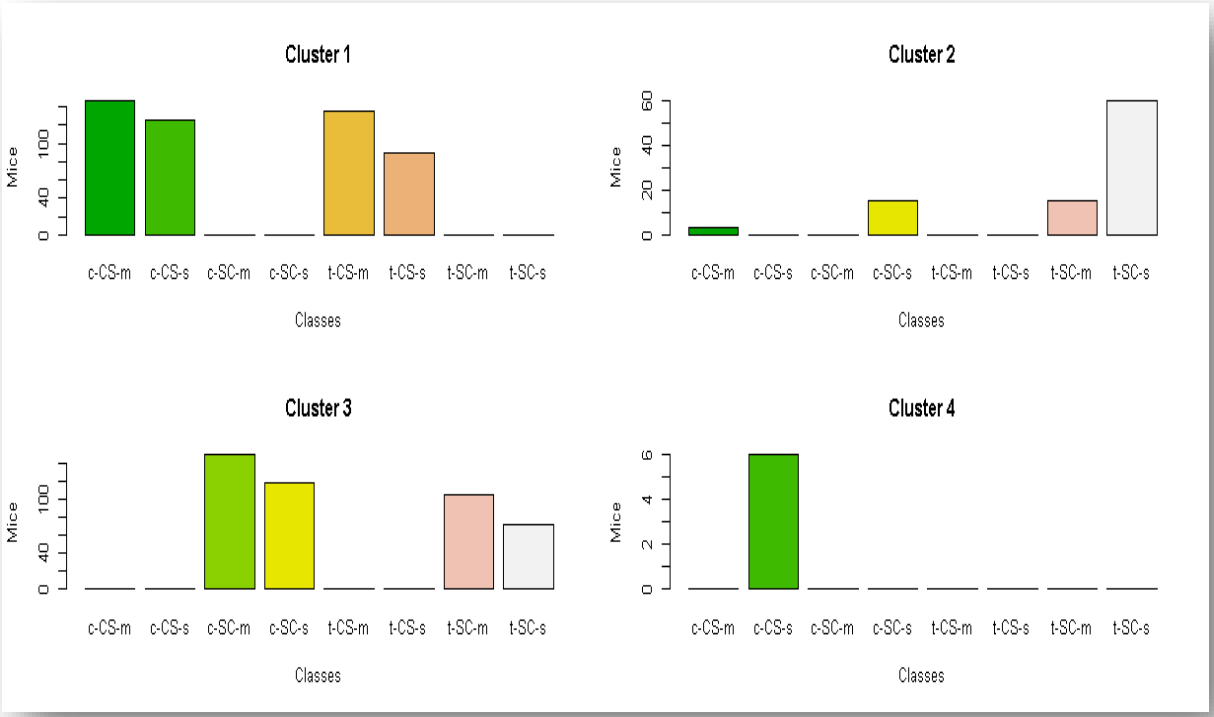
```
> table(grp8_spearman)
  1   2   3   4   5   6   7   8
498  93 446   6   3   1  15  15
```



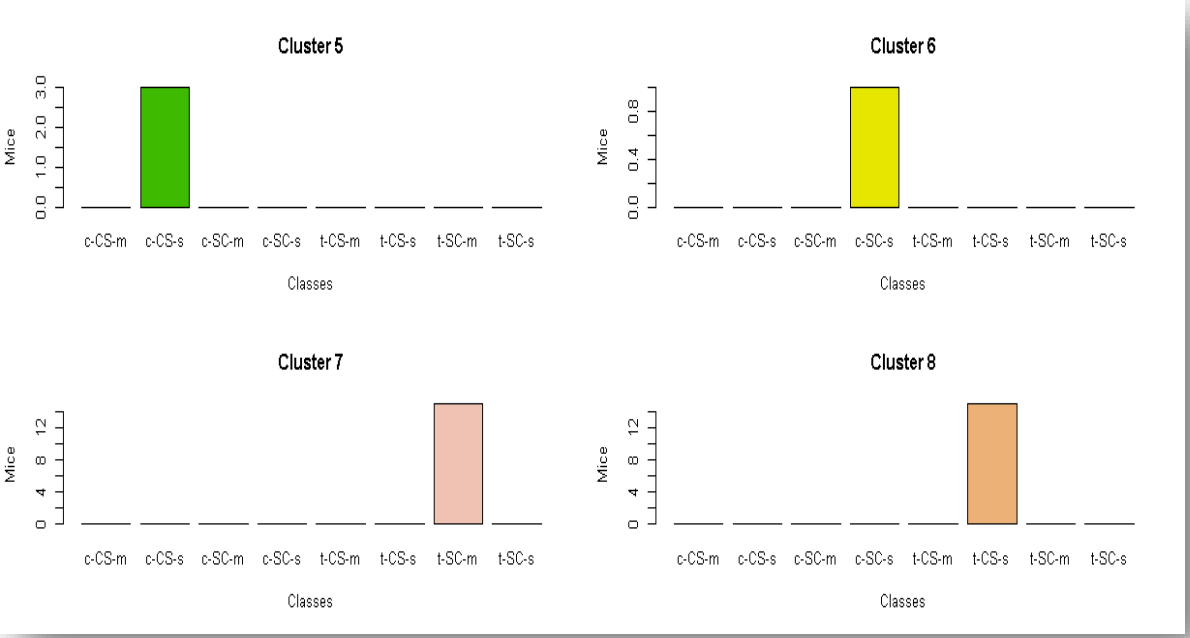
Εικόνα 27 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Spearman για k=8



Εικόνα 28 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Spearman για k=8



Εικόνα 29 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (α)



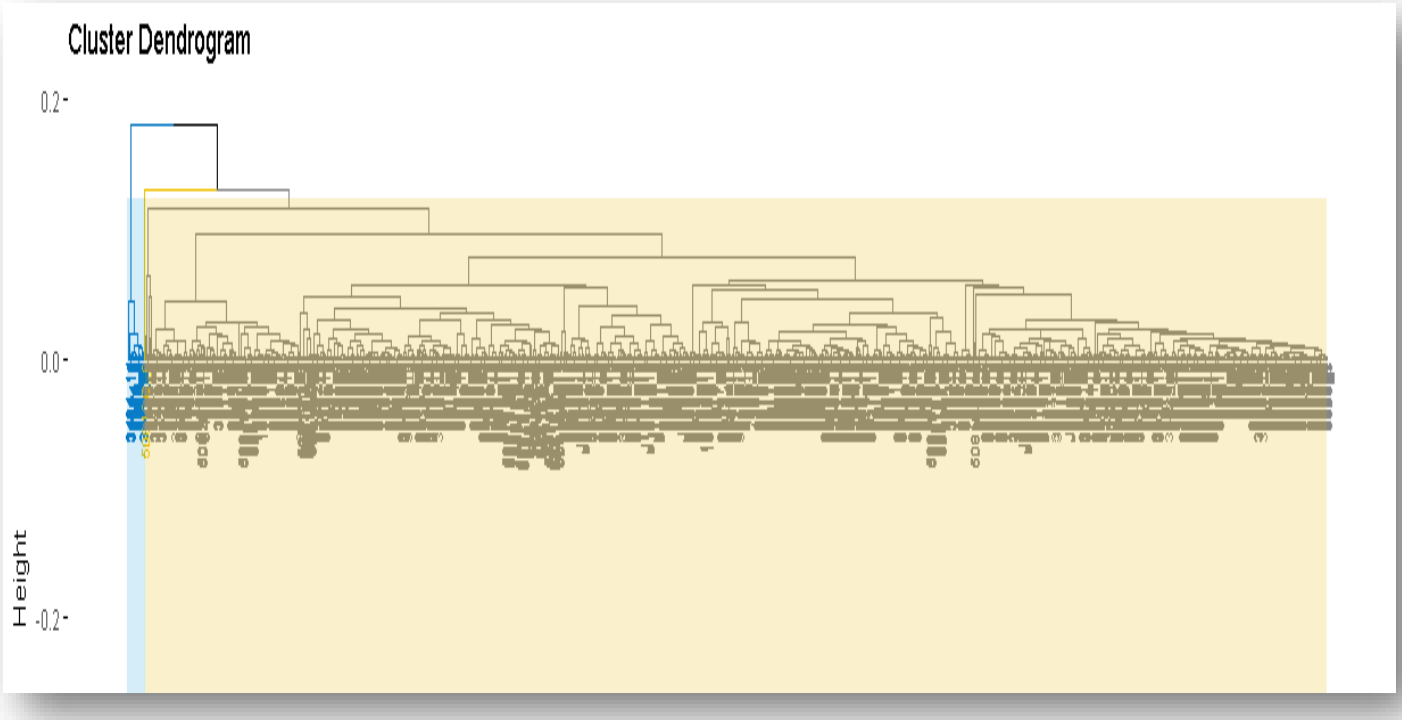
Εικόνα 30 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (β)

Το 1ο cluster έχει ποντίκια, διεγερμένα στη μάθηση, είτε είναι υγιή είτε όχι, είτε έχουν πάρει μεμαντίνη, είτε όχι. Βέβαια, στο 4ο και 5ο cluster υπάρχουν αντίστοιχα 6 και 3 ποντίκια διεγερμένα, υγιή που έχουν πάρει ορό. Τα τρισωμικά, μη διεγερμένα έχουν χωριστεί στο 2ο (έχουν πάρει ορό) και στο 7ο (έχουν πάρει μεμαντίνη). Στο 3ο υπερισχύουν τα υγιή, μη διεγερμένα, που έχουν πάρει μεμαντίνη, ενώ ένα που έχει λάβει αλατούχο ορό βρίσκεται στο 6ο. Τέλος, στο 8ο cluster έχουμε τρισωμικά, διεγερμένα στη μάθηση ποντίκια, τα οποία έχουν λάβει ορό. Από όλα τα παραπάνω βλέπουμε ότι υπάρχει κάποιου είδους διαχωρισμός ως προς τις πραγματικές κλάσεις των δεδομένων.

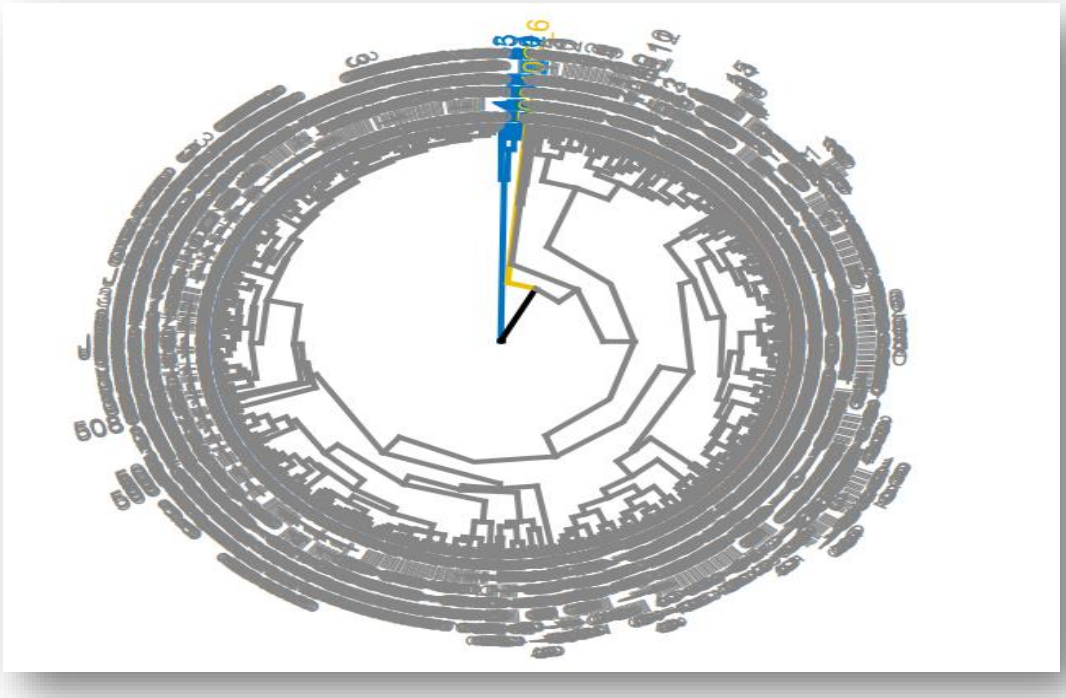
5.2.2 Ιεραρχική Ταξινόμηση με απόσταση PEARSON

- **k=3**

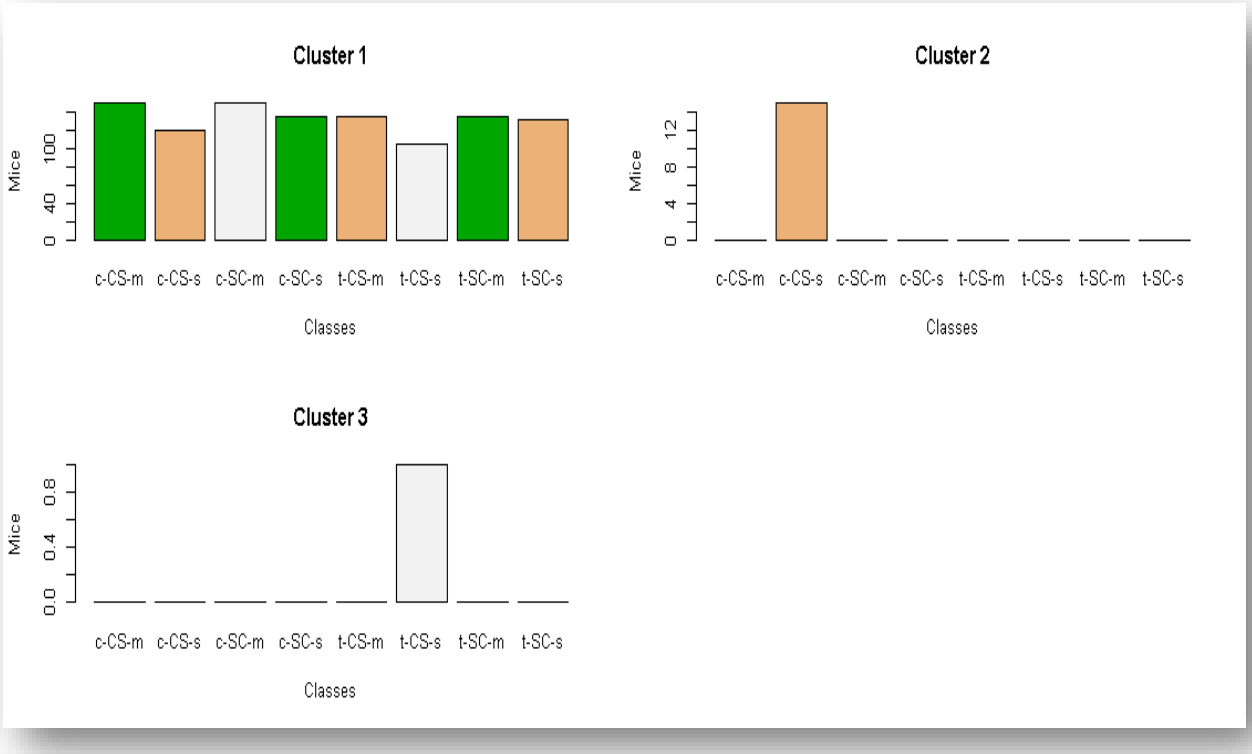
```
> table(grp3_pearson)
   1   2   3
1061  15   1
```



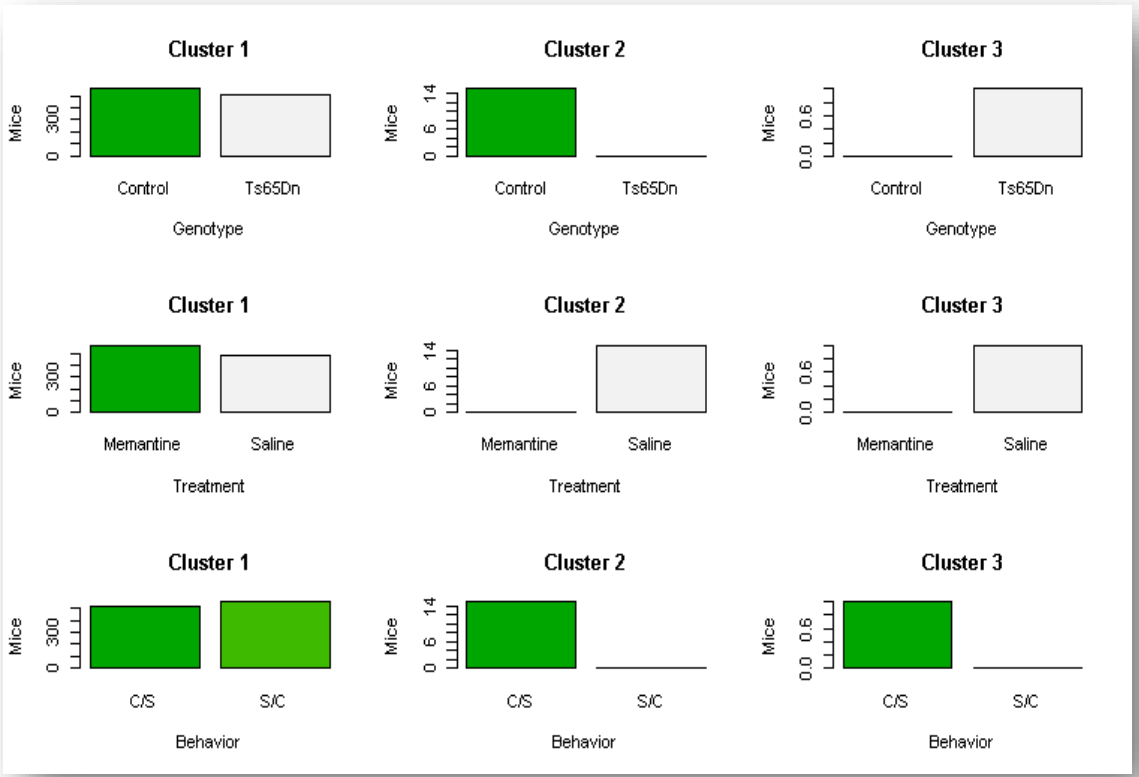
Εικόνα 31 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Pearson για k=3



Εικόνα 32 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Pearson για k=3



Εικόνα 33 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster

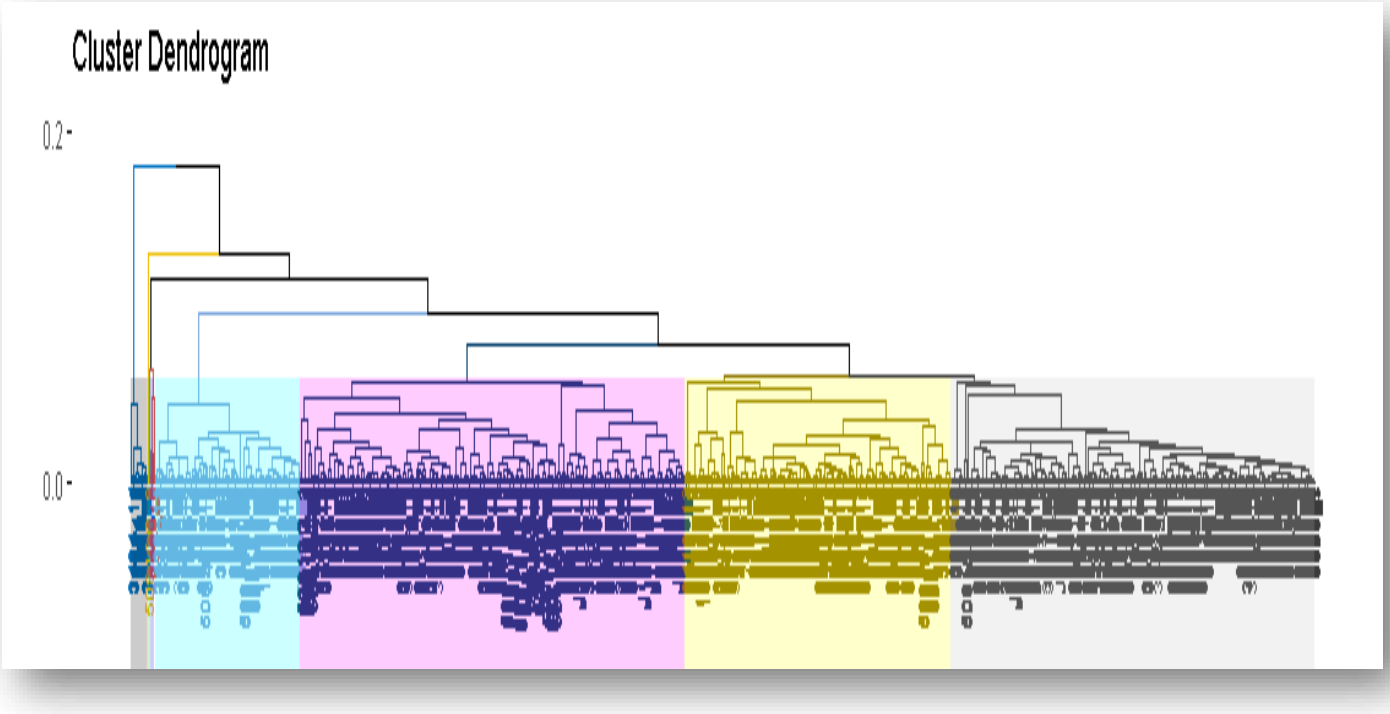


Εικόνα 34 Ποντίκια του κάθε γενότυπου (α), θεραπείας (β) και συμπεριφοράς (γ) στο κάθε cluster

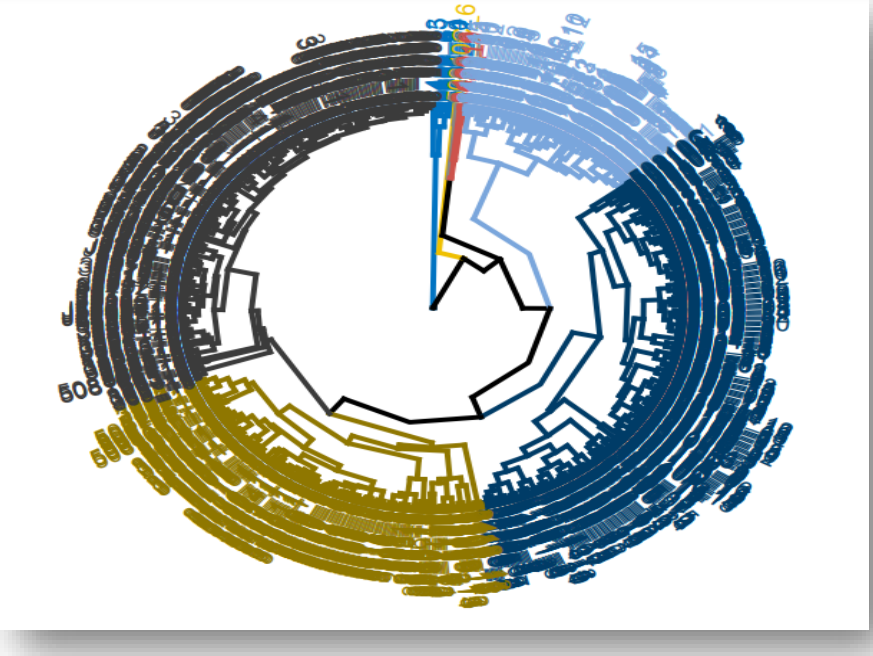
Από όλα τα παραπάνω διαγράμματα, βλέπουμε ότι στο 1ο cluster έχουμε ελάχιστα παραπάνω υγιή, μη διεγερμένα ποντίκια, τα οποία έλαβαν μεμαντίνη, στο 2ο έχουμε υγιή, που διεγέρθηκαν και έλαβαν ορό και τέλος στο 3ο έχουμε τρισωμικά διεγερμένα ποντίκια, στα οποία χορηγήθηκε ορός. Βέβαια, από την εντολή *table(grp3_pearson)* παρατηρούμε ότι το cluster 3 έχει μόνο μια παρατήρηση (θα μπορούσαμε να τη θεωρήσουμε και ακραία παρατήρηση). Επίσης, όλα τα ποντίκια βρίσκονται πάλι στο 1ο cluster, εκτός από 15 υγιή, διεγερμένα, που έλαβαν ορό και βρίσκονται στο 2ο. Όχι τόσο καλός διαχωρισμός.

- **k=8**

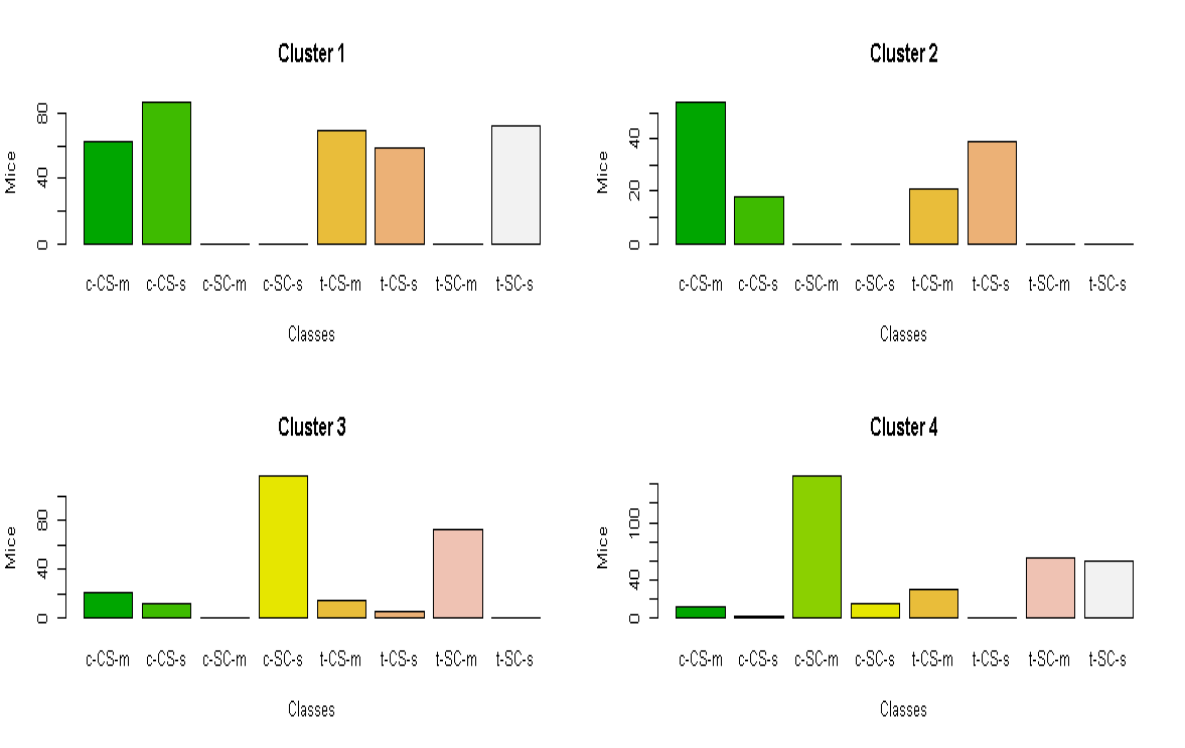
```
> table(grp8_pearson)
  1   2   3   4   5   6   7   8
350 132 242 331   4  15   2   1
```



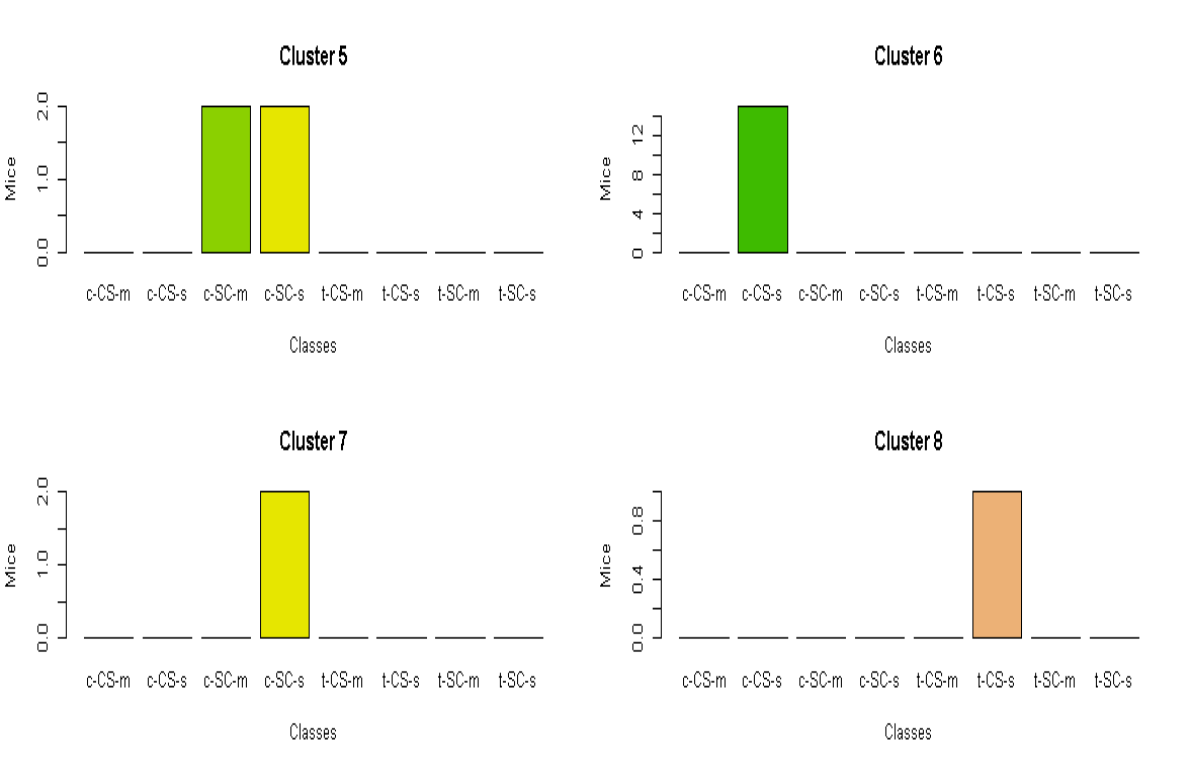
Εικόνα 35 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Pearson για k=8



Εικόνα 36 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Pearson για k=8

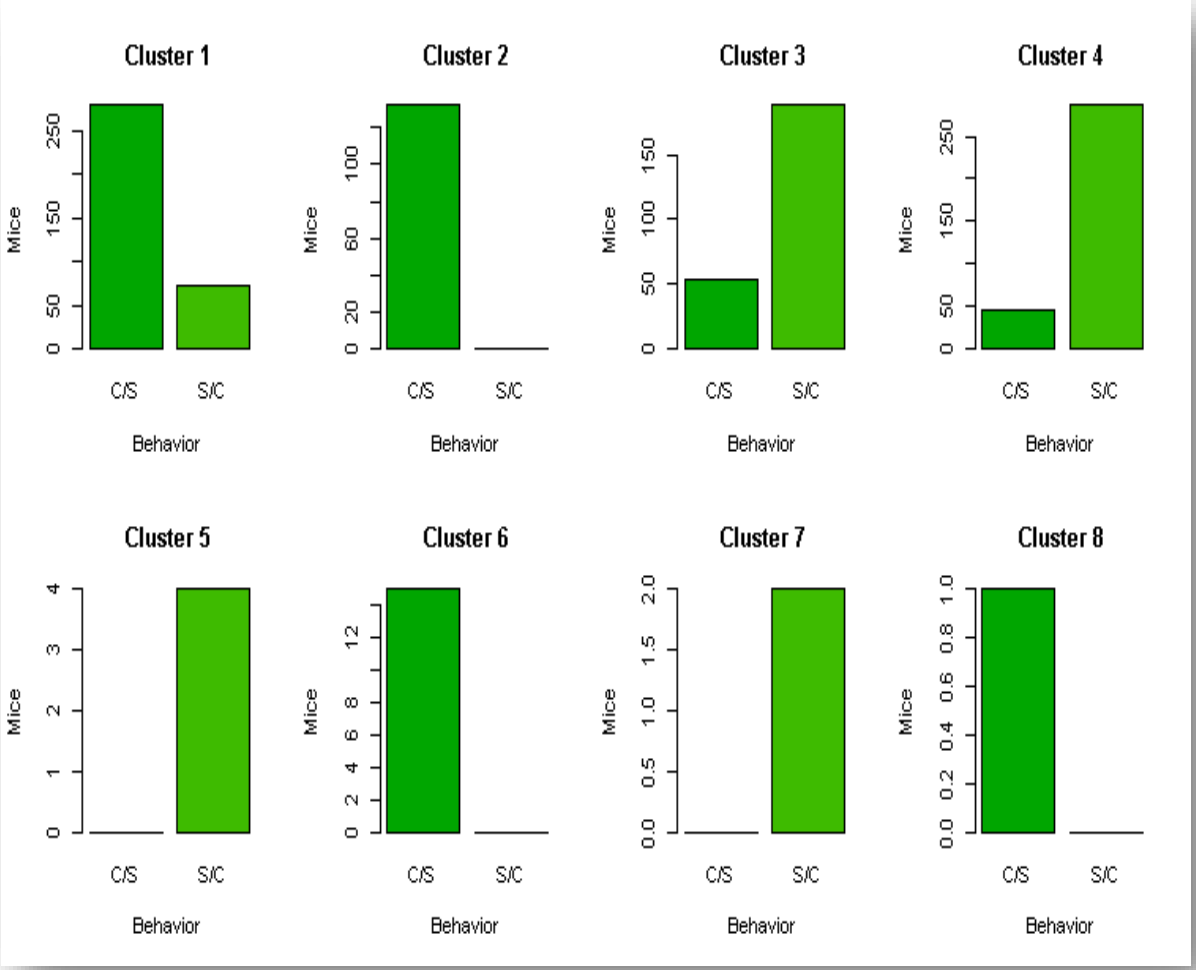


Εικόνα 37 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (α)



Εικόνα 38 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (β)

Παρατηρούμε ότι στο 7ο cluster έχουμε δυο υγιή, μη διεγερμένα στη μάθηση ποντίκια, που έχουν λάβει ορό και στο 8ο μόνο ένα ποντίκι τρισωμικό, διεγερμένο που έλαβε και αυτό ορό. Αυτά τα τρία ποντίκια θα μπορούσαμε να τα θεωρήσουμε ακραίες τιμές. Το ίδιο ισχύει και για το 5ο cluster, το οποίο περιέχει τέσσερα υγιή, μη διεγερμένα ποντίκια. Στο 1ο cluster έχουμε κατά βάση διεγερμένα στη μάθηση ποντίκια, από όλους τους τύπους θεραπείας και γενότυπου, αλλά και κάποια τρισωμικά, μη διεγερμένα, που έλαβαν ορό. Στο 2ο έχουμε επίσης κάποια δεγερμένα ποντίκια, είτε υγιή, τα οποία έλαβαν μεμαντίνη, είτε τρισωμικά, τα οποία έλαβαν ορό. Στο 3ο βρίσκονται ποντίκια μη διεγερμένα, είτε υγιή που έλαβαν ορό, είτε τρισωμικά που έλαβαν μεμαντίνη. Στο 4ο έχουμε κατά κύριο λόγο ποντίκια υγιή, μη διεγερμένα, που έχουν πάρει μεμαντίνη και τέλος, στο 6ο έχουμε δεκαπέντε υγιή, διεγερμένα, που έλαβαν αλατούχο ορό. Παρατηρούμε λοιπόν, ότι δεν υπάρχει κάποιος διαχωρισμός ως προς τις πραγματικές κλάσεις, αλλά κάποιος μικρός διαχωρισμός ως προς τη συμπεριφορά τους.

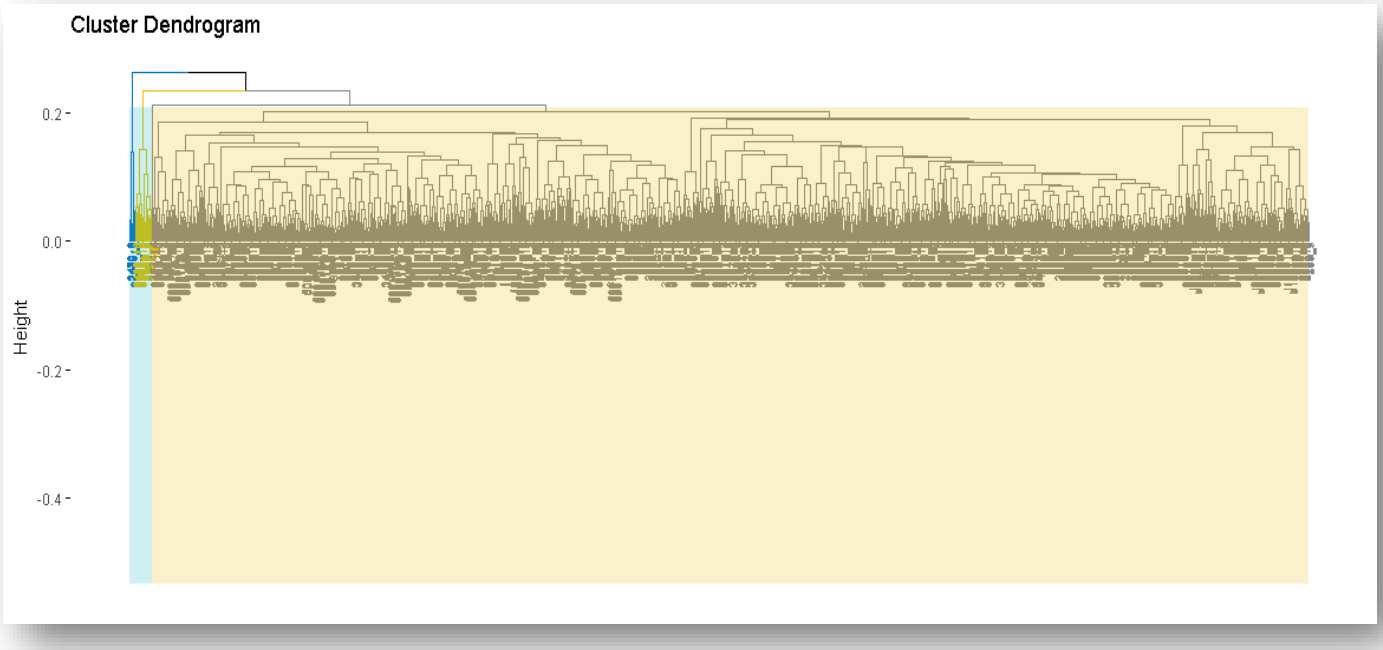


Εικόνα 39 Ποντίκια της κάθε διέγερσης- συμπεριφοράς για μάθηση (behavior) στο κάθε cluster (C/S vs S/C)

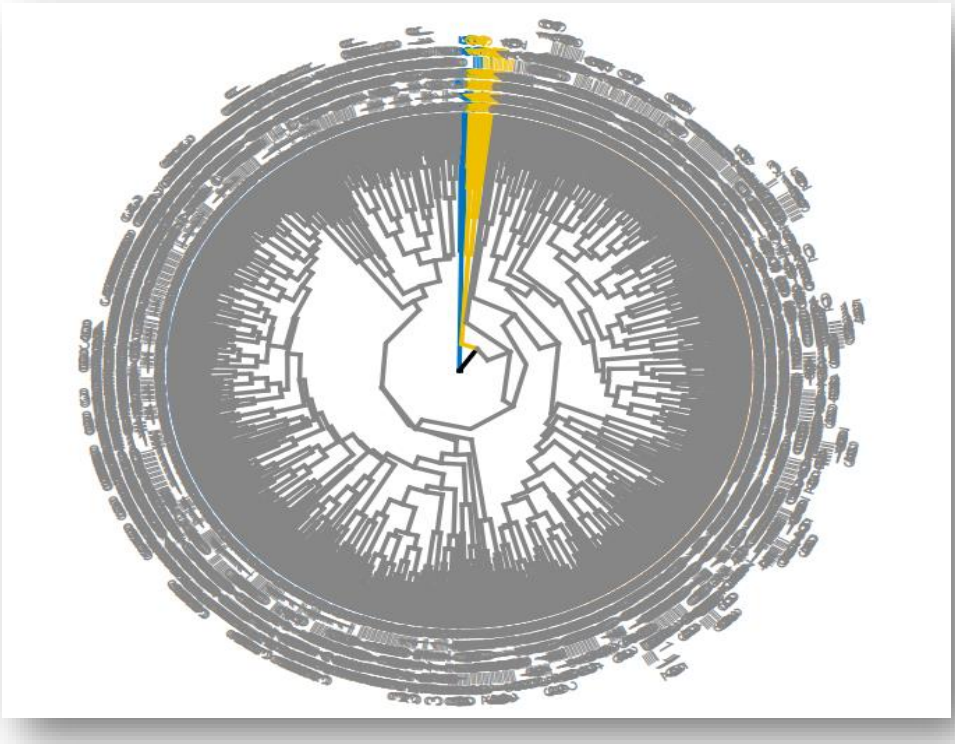
5.2.3 Ιεραρχική Ταξινόμηση με απόσταση KENDALL

- **k=3**

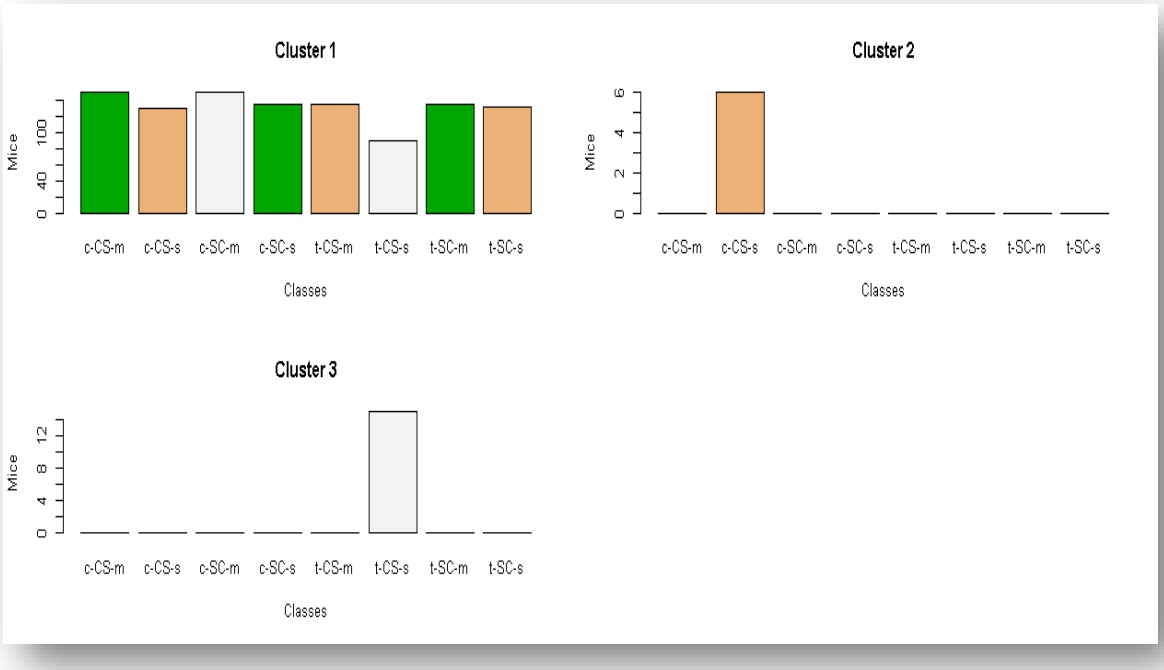
```
> table(grp3_kendall)
   1   2   3
1056   6  15
```



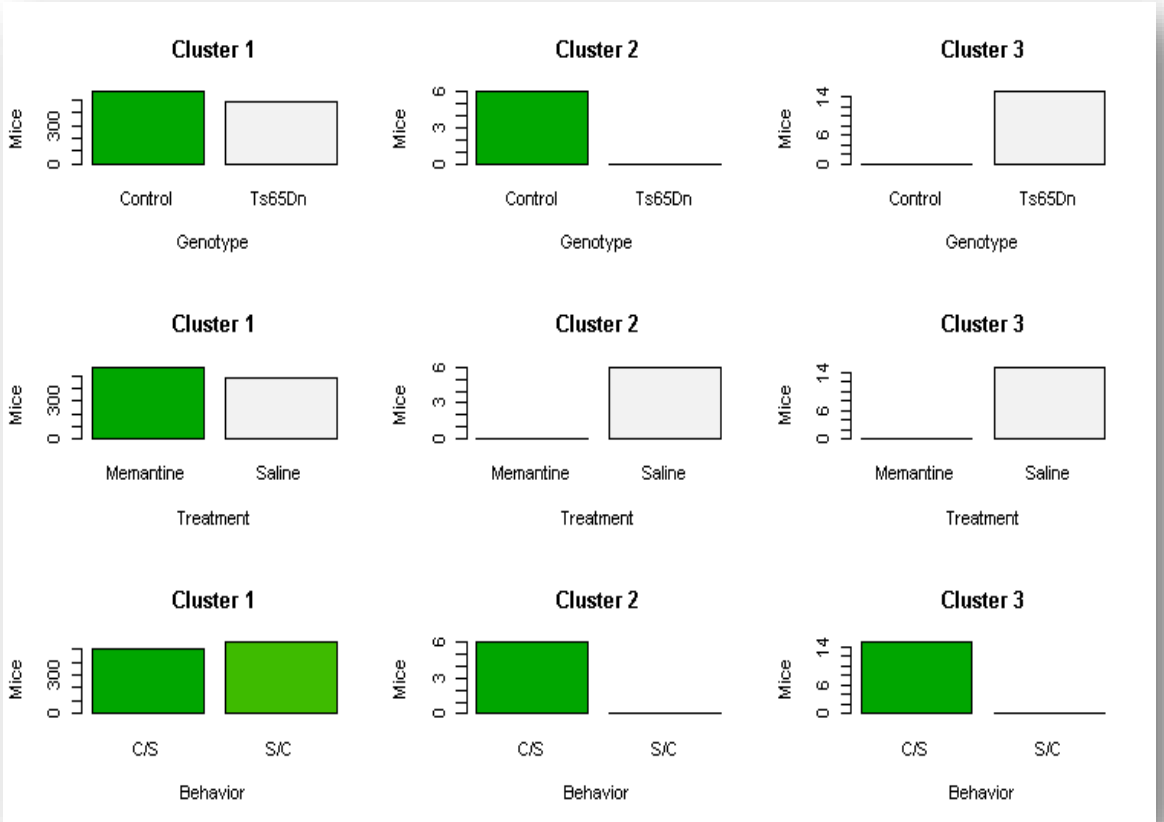
Εικόνα 40 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Kendall για k=3



Εικόνα 41 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Kendall για k=3



Εικόνα 42 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster

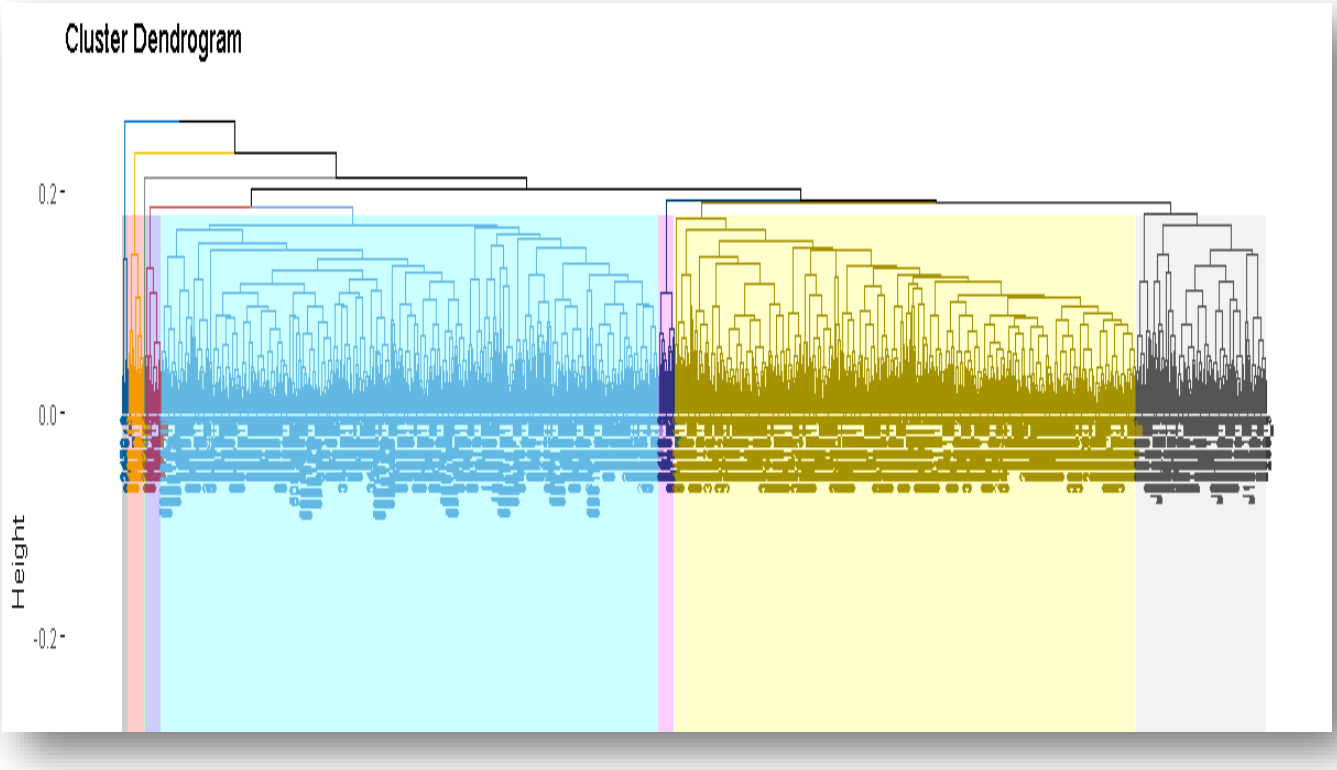


Εικόνα 43 Ποντίκια του κάθε γενότυπου (α), θεραπείας (β) και συμπεριφοράς (γ) στο κάθε cluster

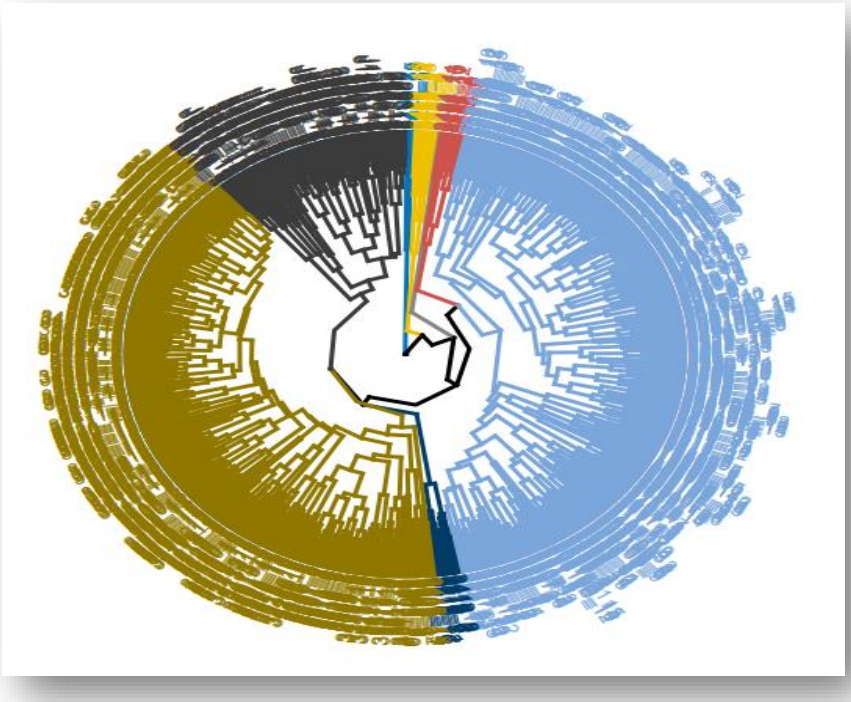
Από όλα τα παραπάνω, βλέπουμε ότι στο 1ο cluster έχουμε υγιή, μη διεγερμένα στη μάθηση ποντίκια, τα οποία έχουν λάβει μεμαντίνη, στο 2ο έχουμε υγιή, διεγερμένα, που έχουν λάβει ορό και στο 3ο έχουμε τρισωμικά, διεγερμένα ποντίκια, στα οποία έχει χορηγηθεί επίσης ορός. Και εδώ παρατηρούμε ότι σχεδόν όλα τα ποντίκια βρίσκονται στο 1ο cluster, μόνο εξι υγιή, διεγερμένα, που έλαβαν ορό στο 2ο και δεκαπέντε στο 3ο τρισωμικά, διεγερμένα, που έλαβαν και αυτά ορό. Συνεπώς, δεν είναι ιδιαίτερα καλός αυτός ο διαχωρισμός.

- **k=8**

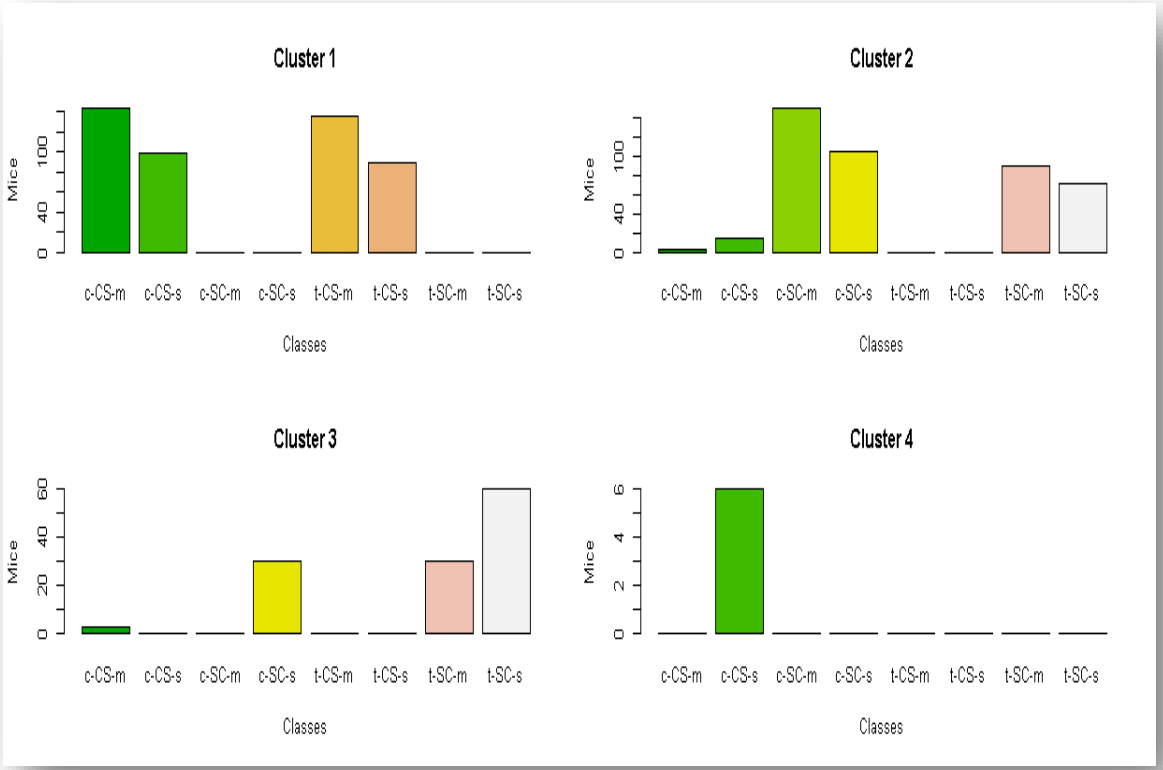
```
> table(grp8_kendall)
  1   2   3   4   5   6   7   8
468 434 123   6  15   1  15  15
```



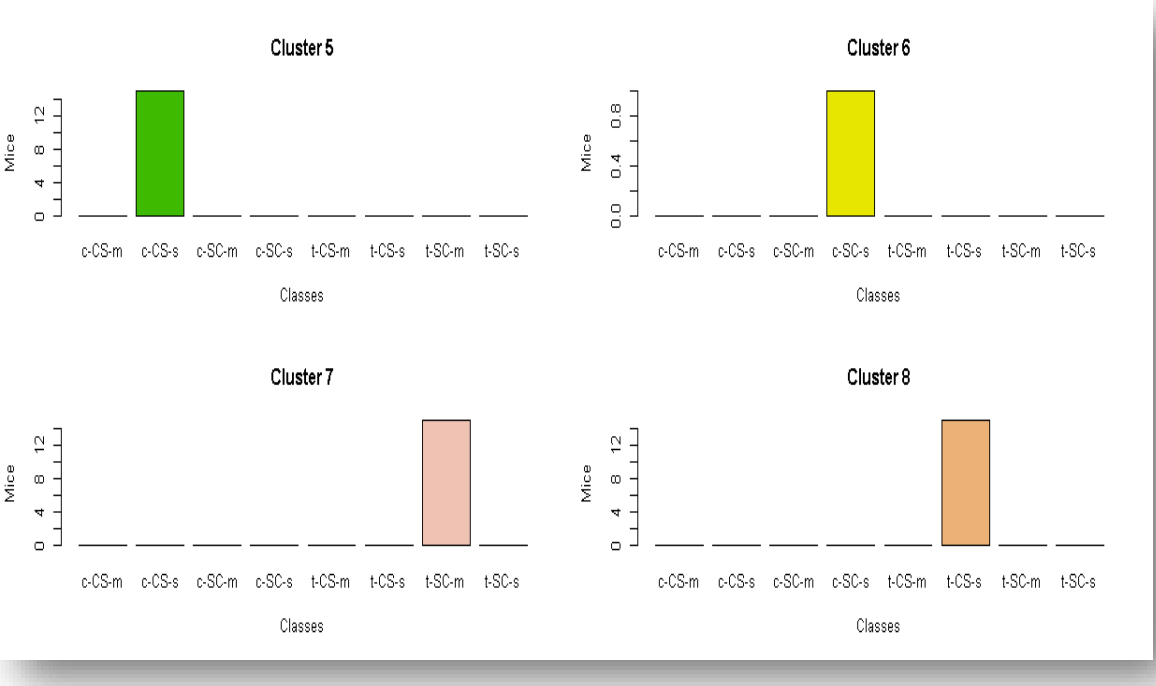
Εικόνα 44 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Kendall για k=8



Εικόνα 45 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την απόσταση Kendall για k=8



Εικόνα 46 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (α)



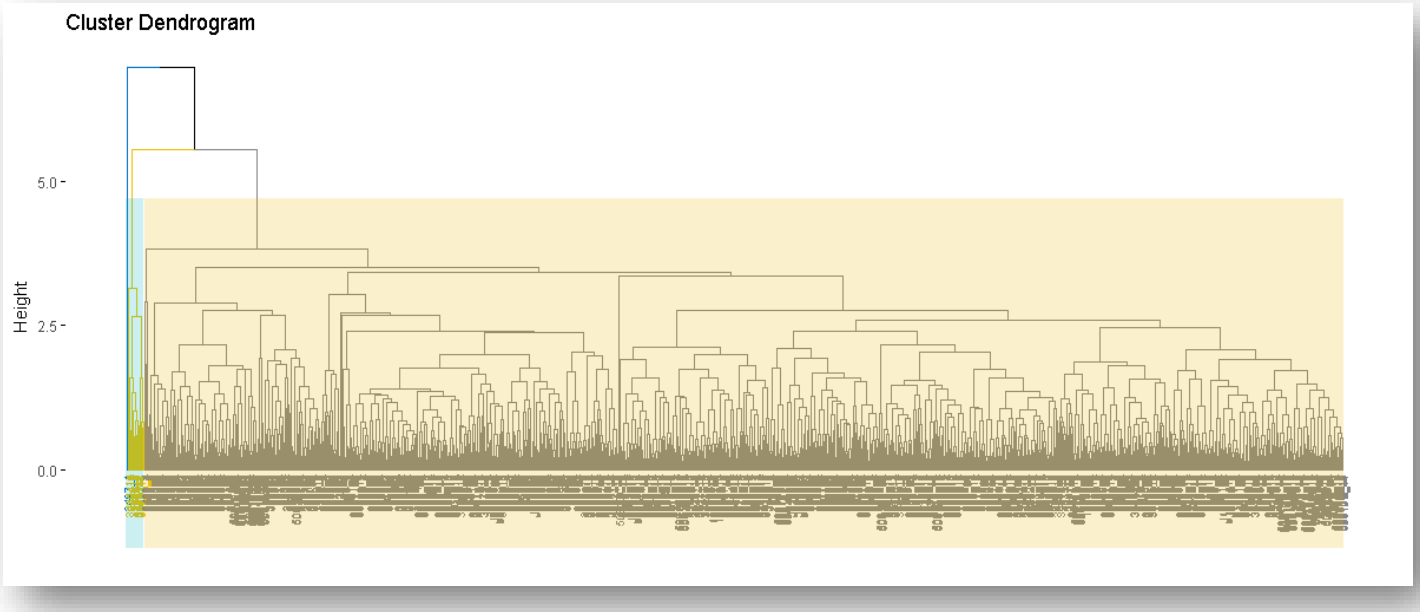
Εικόνα 47 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (β)

Στο 1ο cluster υπάρχουν έντονα υγιή ή μη, διεγερμένα ποντίκια, που έλαβαν μεμαντίνη. Στο 2ο υπερισχύουν τα υγιή, μη διεγερμένα, που έλαβαν είτε μεμαντίνη είτε ορό, στο 3ο τα τρισωμικά, μη διεγερμένα, που έλαβαν ορό, στο 4ο τα υγιή, διεγερμένα στη μάθηση, που πήραν ορό (6 ποντίκια) και στο 5ο επίσης 15 ποντίκια ίδιου τύπου με το 4ο cluster. Στο 6ο cluster έχουμε μόνο ένα ποντίκι, άρα θα το θεωρήσουμε ως ακραία τιμή. Στο 7ο έχουμε τρισωμικά, μη διεγερμένα στη μάθηση ποντίκια, που τους χορηγήθηκε μεμαντίνη και τέλος, στο 8ο έχουμε τρισωμικά, διεγερμένα που τους χορηγήθηκε αλατούχος ορός. Βλέπουμε λοιπόν, ότι υπάρχει ένας διαχωρισμός ως προς τις πραγματικές κλάσεις.

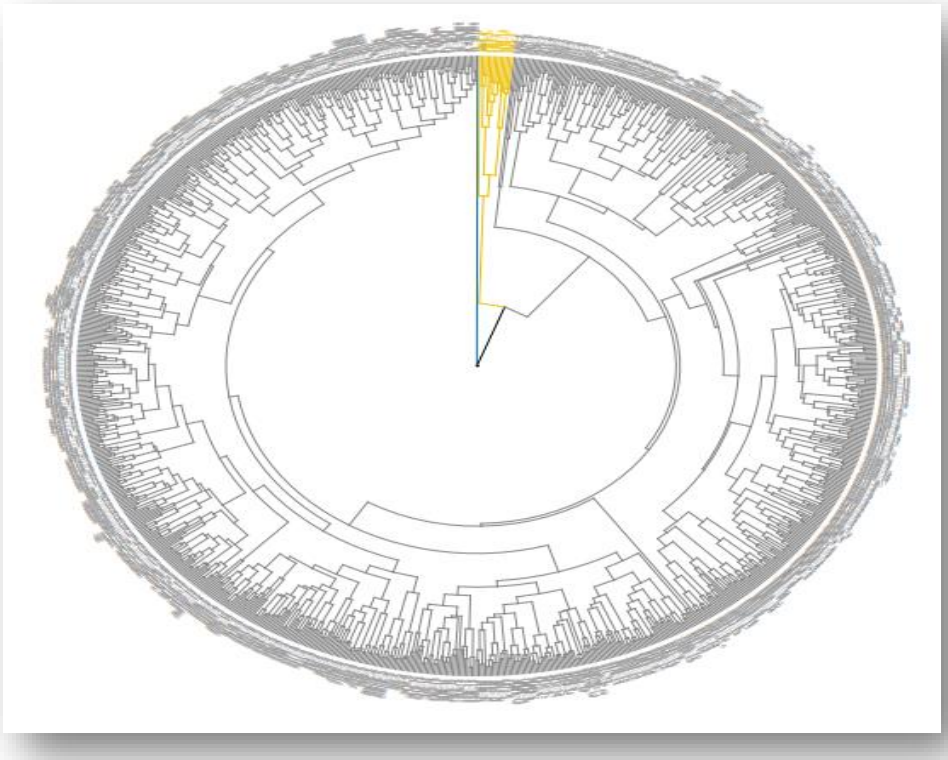
5.2.4 Ιεραρχική Ταξινόμηση με Ευκλείδεια απόσταση

- **k=3**

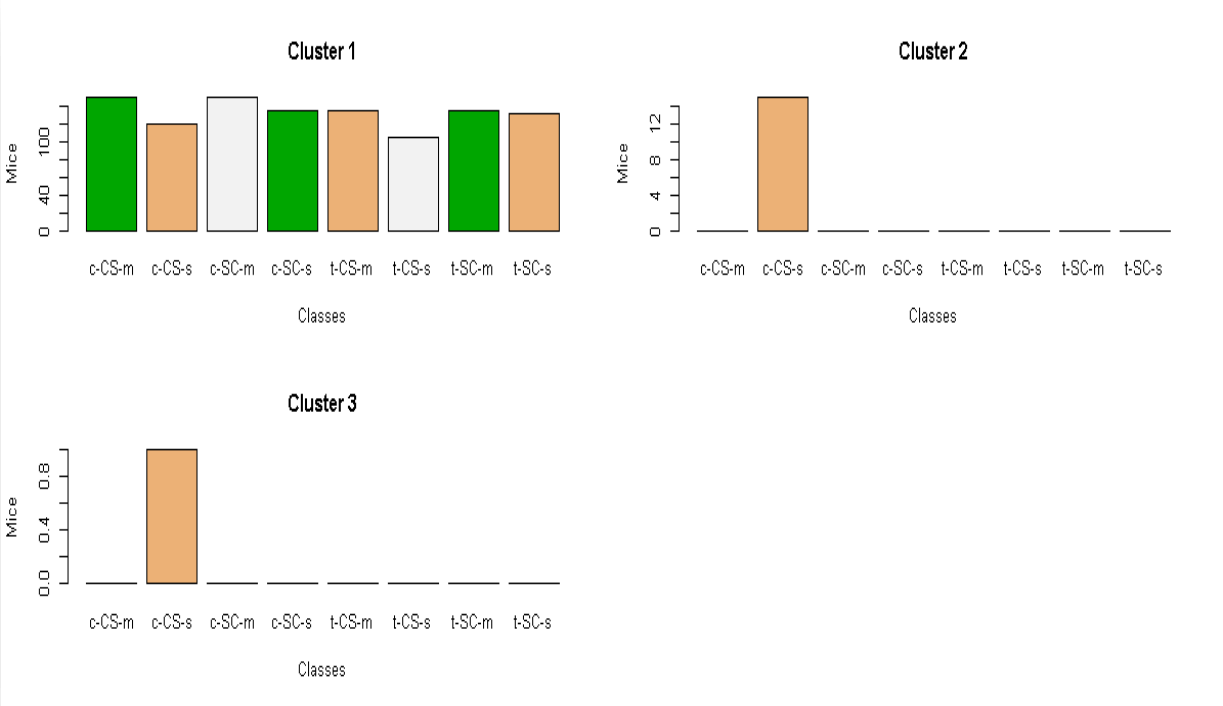
```
> table(grp3_eukleidean)
   1   2   3
1061  15   1
```



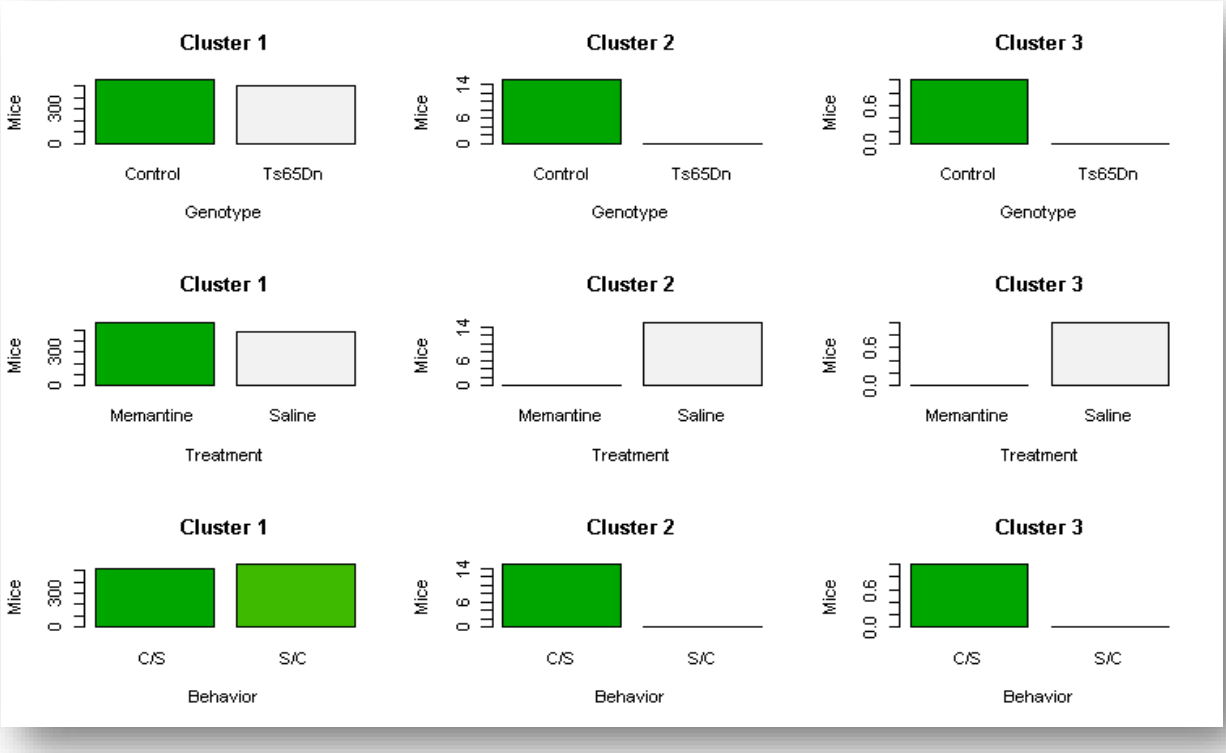
Εικόνα 48 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την Ευκλείδεια απόσταση για k=3



Εικόνα 49 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την Ευκλείδεια απόσταση για k=3



Εικόνα 50 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster

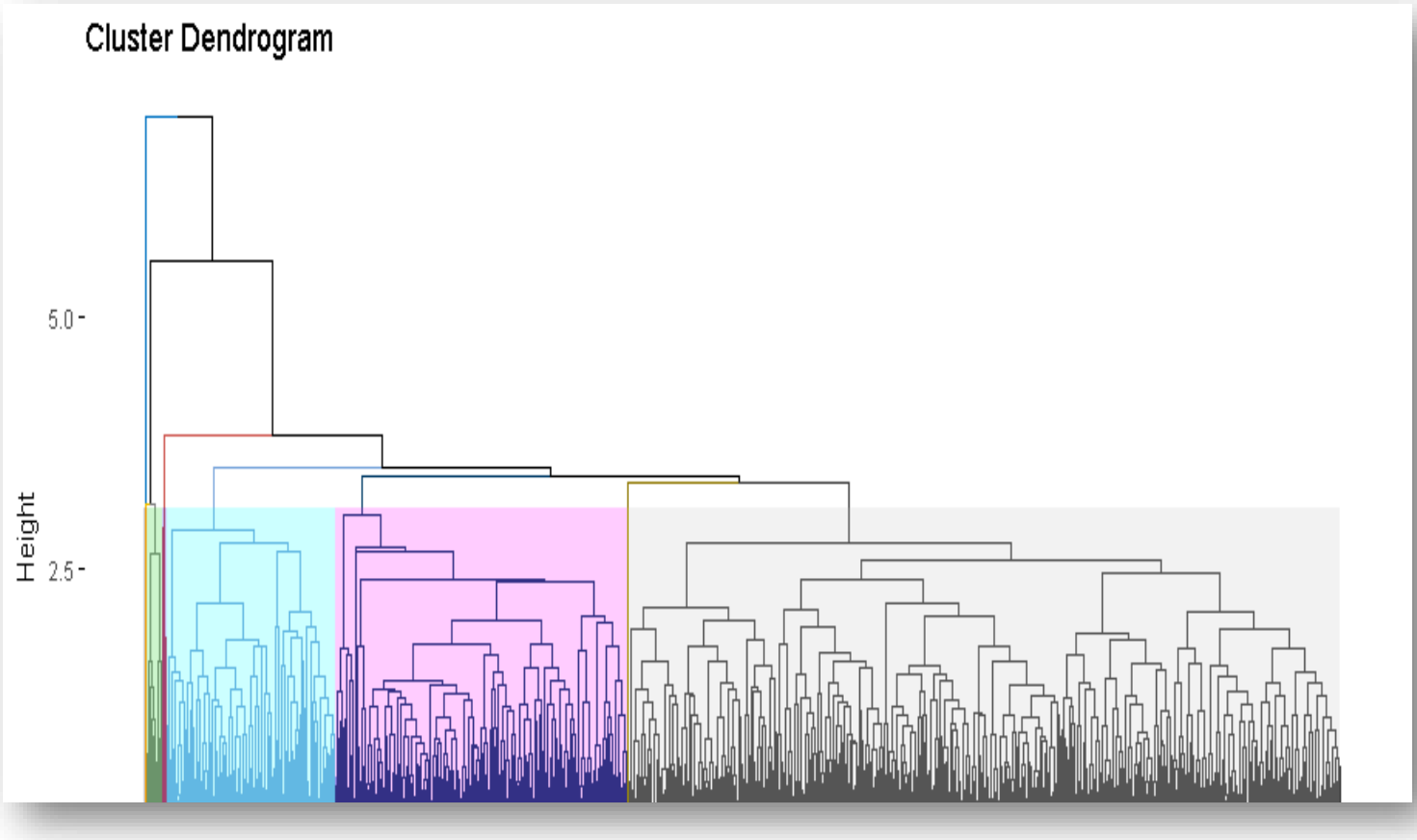


Εικόνα 51 Ποντίκια του κάθε γενότυπου (α), θεραπείας (β) και συμπεριφοράς (γ) στο κάθε cluster

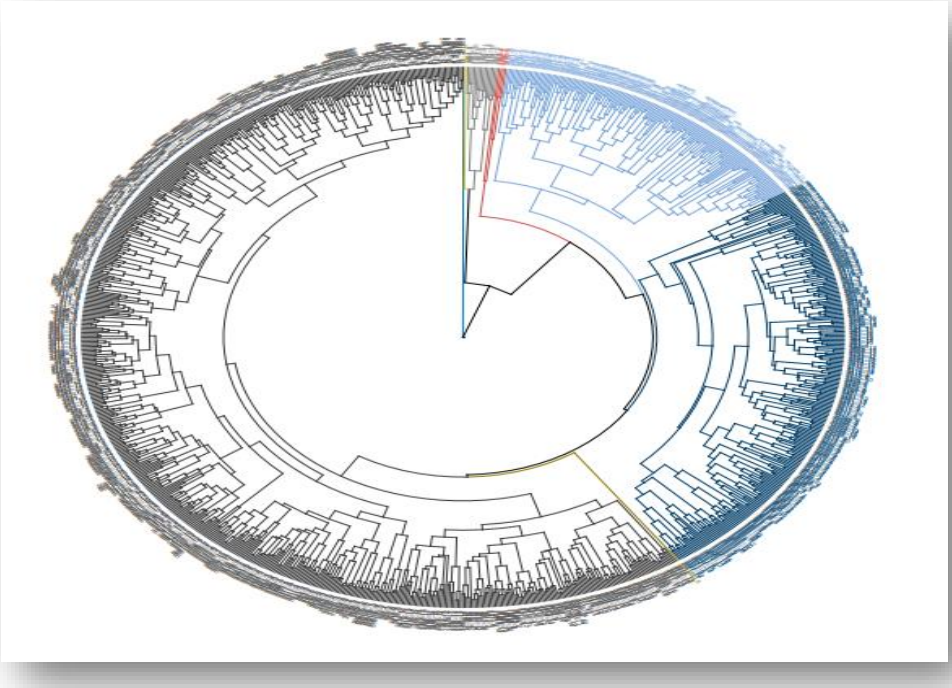
Παρατηρούμε όπι στο 1ο cluster έχουμε υγιή, μη διεγερμένα ποντίκια, που έλαβαν μεμαντίνη, ενώ στο 2ο και στο 3ο έχουμε υγιή, διεγερμένα στη μάθηση, στα οποία χορηγήθηκε ορός. Βέβαια, από την εντολή *table(grp3_eukleidean)* βλέπουμε ότι στο 3ο cluster βρίσκεται μόνο ένα ποντίκι και ότι η πλειοψηφία έχει μαζευτεί στο 1ο, εκτός από 15 που έχουν πάει στο 2ο . Άρα, δεν είναι ιδιαίτερα καλός διαχωρισμός.

- **k=8**

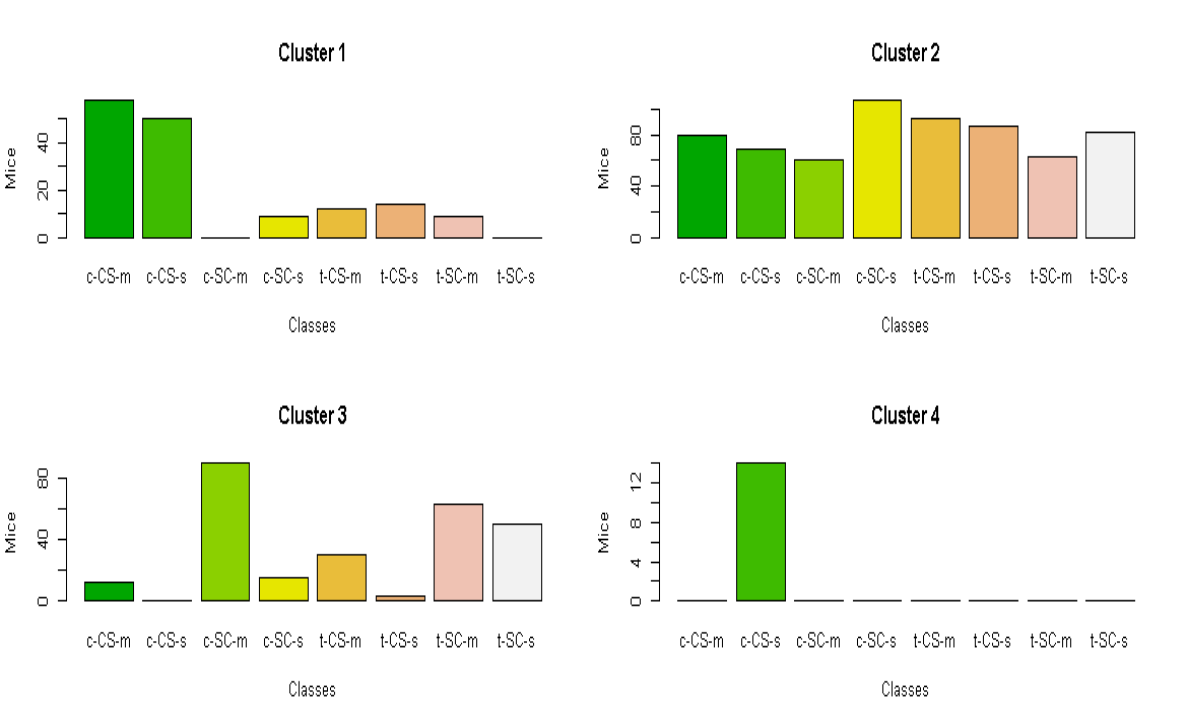
```
> table(grp8_euclidean)
1      2     3    4    5    6    7    8
152 641 263  14   1   1   4   1
```



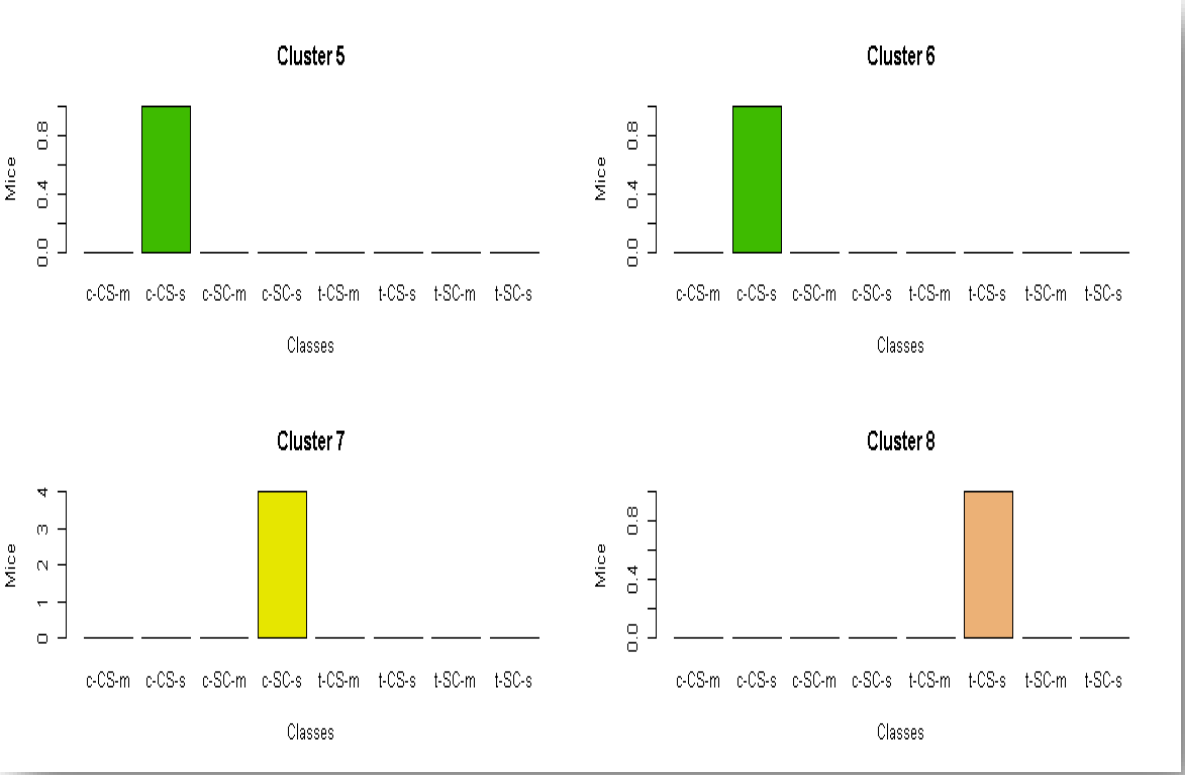
Εικόνα 52 Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την Ευκλείδεια απόσταση για k=8



Εικόνα 53 Κυκλικό Δενδρόγραμμα Ιεραρχικής Ταξινόμησης με την Ευκλείδεια απόσταση για k=8



Εικόνα 54 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (α)



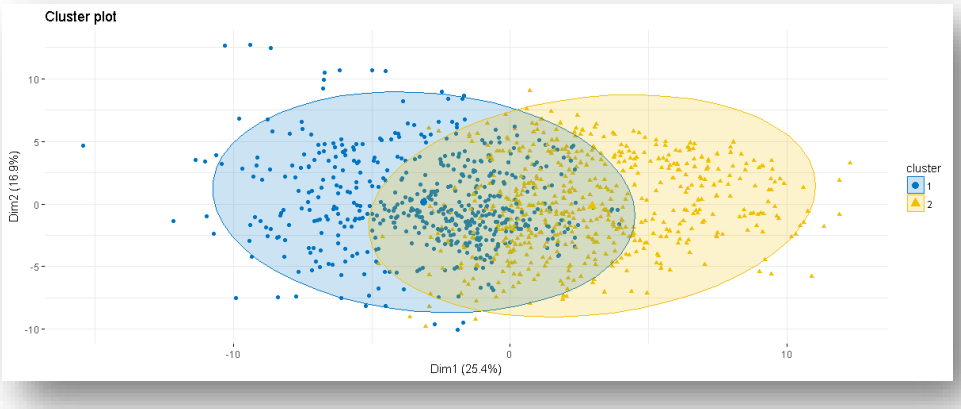
Εικόνα 55 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (β)

Από την εντολή *table(grp8_euclidean)* αλλά και από τα παραπάνω διαγράμματα, τα clusters 5, 6, 7 και 8 μιας και περιέχουν από ένα ή τέσσερα ποντίκια αντιστοίχως, μπορούν να θεωρηθούν ακραίες τιμές και άρα περιττά clusters. Στο 1ο cluster υπάρχουν μόνο υγιή, διεγερμένα ποντίκια, που έλαβαν μεμαντίνη. Στο 2ο υπάρχουν ουσιαστικά ποντίκια από όλες τις πραγματικές κλάσεις. Τέλος, στο 3ο και στο 4ο υπάρχουν τα τρισωμικά, μη διεγερμένα στη μάθηση, που έλαβαν μεμαντίνη και τα διεγερμένα, που έλαβαν αλατούχο ορό αντίστοιχα. Παρατηρούμε λοιπόν, ότι ως προς τις πραγματικές κλάσεις δε γίνεται κάποιου είδους διαχωρισμός, ενώ σε όλα σχεδόν υπερισχύουν τα ποντίκια, στα οποία χορηγήθηκε αλατούχος ορός, συνεπώς, όχι ιδιαίτερα καλός διαχωρισμός, αν και βλέπουμε και εδώ γνωστούς συνδυασμούς κλάσεων.

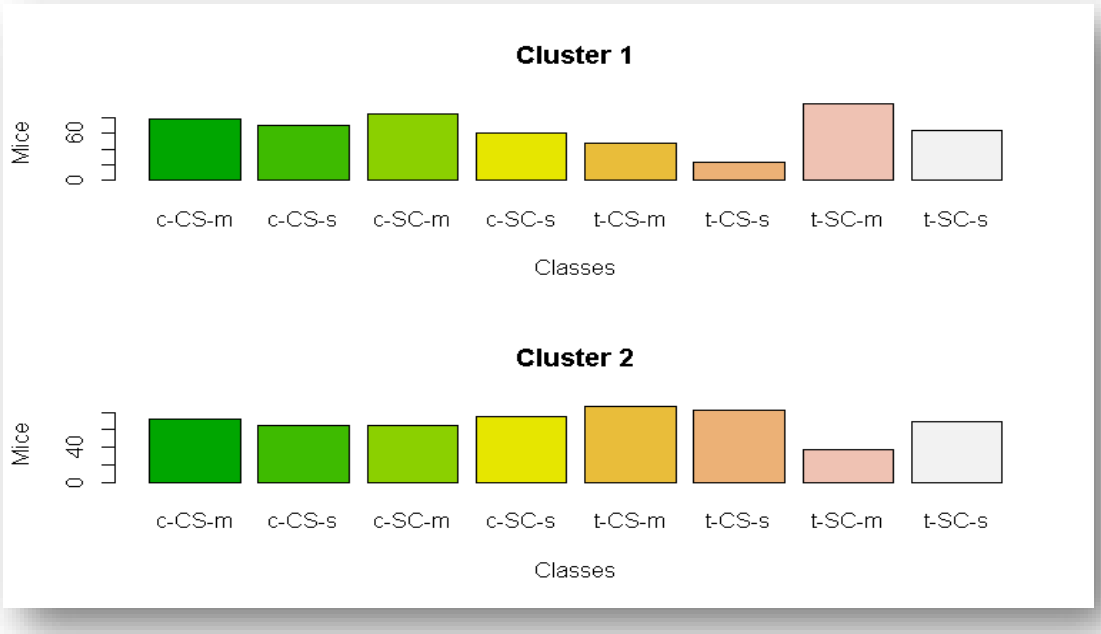
5.2.5 ΑΛΓΟΡΙΘΜΟΣ K-MEANS

- **k=2**

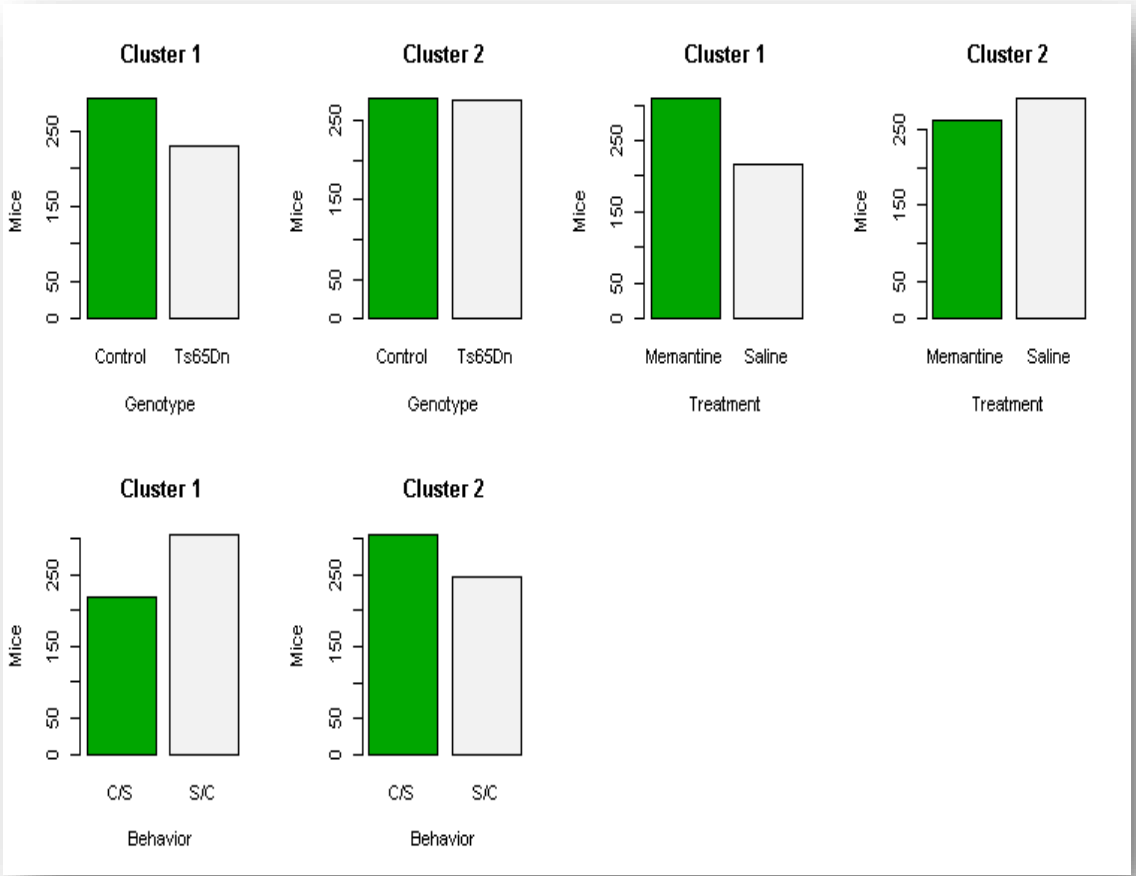
```
## Ο παρακάτω κώδικας είναι αντίστοιχος και για τις υπόλοιπες περιπτώσεις αλγορίθμων ταξινόμησης method
και αριθμού συστάδων k, οπότε θα αναφερθεί μία μόνο φορά.

>set.seed(123)
> km.res2 <- eclust(df, "kmeans", k=2,nstart = 25, graph = FALSE)      ##Αντίστοιχα για k=3,8 και για
                                                                          method= "clara"
>print(km.res2)
> km.res2$size  ## Μέγεθος του κάθε cluster
524 553
>  rownames(df)[km.res2$cluster==1]                                    ## Ποια ποντίκια είναι πχ στο cluster 1
> fviz_cluster(km.res2, geom = "point", ellipse.type = "norm",        ## Γραφική αναπαράσταση των
        palette = "jco", ggtheme = theme_minimal())                      συστάδων

## Η εντολή print(km.res2) τυπώνει τα αποτελέσματα των clusters, δηλώνοντας κυρίως τα κέντρα της κάθε
συστάδας αλλά και σε ποιο cluster έχει μπει η κάθε παρατήρηση.

## Τα plots που ακολουθούν αναπαριστούν πόσες παρατηρήσεις έχουν μπει στο κάθε cluster και ποιές
πραγματικές κλάσεις εμφανίζονται στο καθένα από αυτά.

> par(mfrow=c(2,1))
>plot(classes[km.res2$cluster==1],col=terrain.colors(8))
>title(main="Cluster 1",xlab="Classes",ylab="Mice")
>plot(classes[km.res2$cluster==2],col=terrain.colors(8))
>title(main="Cluster 2",xlab="Classes",ylab="Mice")



## Τέλος, ta barplots δείχνουν με τη σειρά τους ποιος τύπος γενότυπου, φαρμάκου και συμπεριφοράς στη
μάθηση υπάρχει στην κάθε συστάδα αντίστοιχα.
par(mfrow=c(1,2))
plot(genotype[km.res2$cluster==1],col=terrain.colors(2))      ## Αντίστοιχα για treatment και behavior
title(main="Cluster 1",xlab="Genotype",ylab="Mice")
plot(genotype[km.res2$cluster==2],col=terrain.colors(2))      ## Αντίστοιχα για treatment και behavior
title(main="Cluster 2",xlab="Genotype",ylab="Mice")
```



Εικόνα 56 Απεικόνιση των clusters με τον αλγόριθμο K-Means για k=2



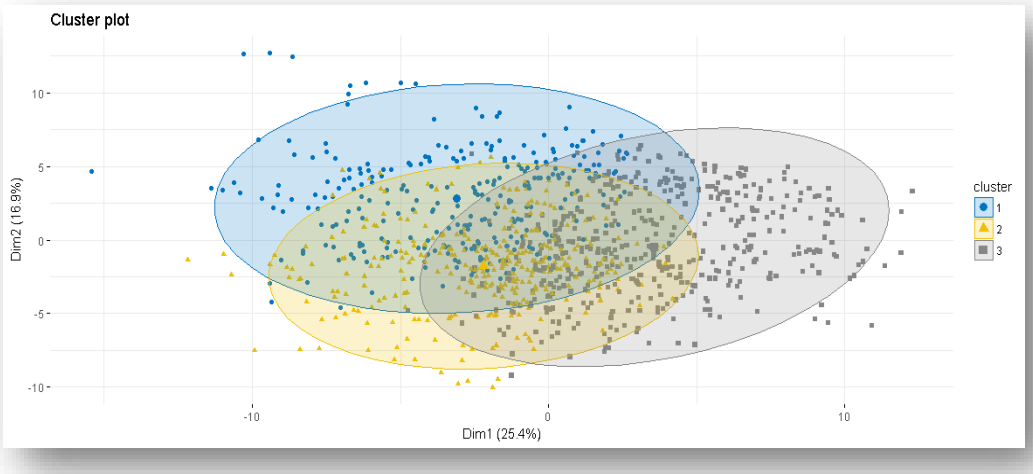
Εικόνα 57 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster



Εικόνα 58 Ποντίκια του κάθε γενότυπου (α), θεραπείας (β) και συμπεριφοράς (γ) στο κάθε cluster

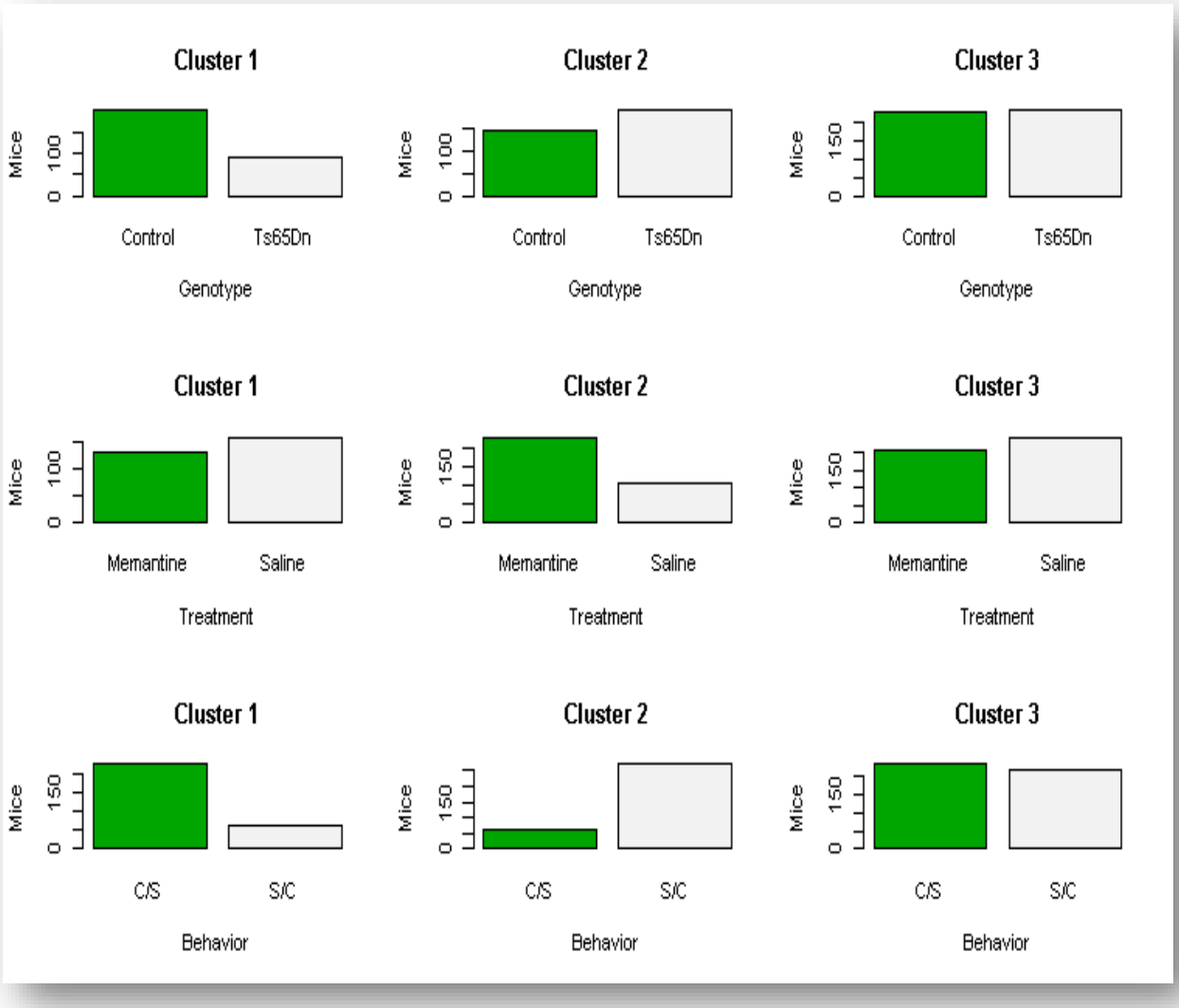
Από τα παραπάνω γραφήματα, χωρίς να υπάρχει έντονος διαχωρισμός, στο 1ο cluster έχουμε υγιή, μη διεγερμένα στη μάθηση ποντίκια, που έλαβαν μεμαντίνη, ενώ στο 2ο έχουμε ελάχιστα περισσότερα τρισωμικά, που έλαβαν ορό και διεγέρθηκαν στη μάθηση.

- **k=3**



Εικόνα 59 Απεικόνιση των clusters με τον αλγόριθμο K-Means για k=3

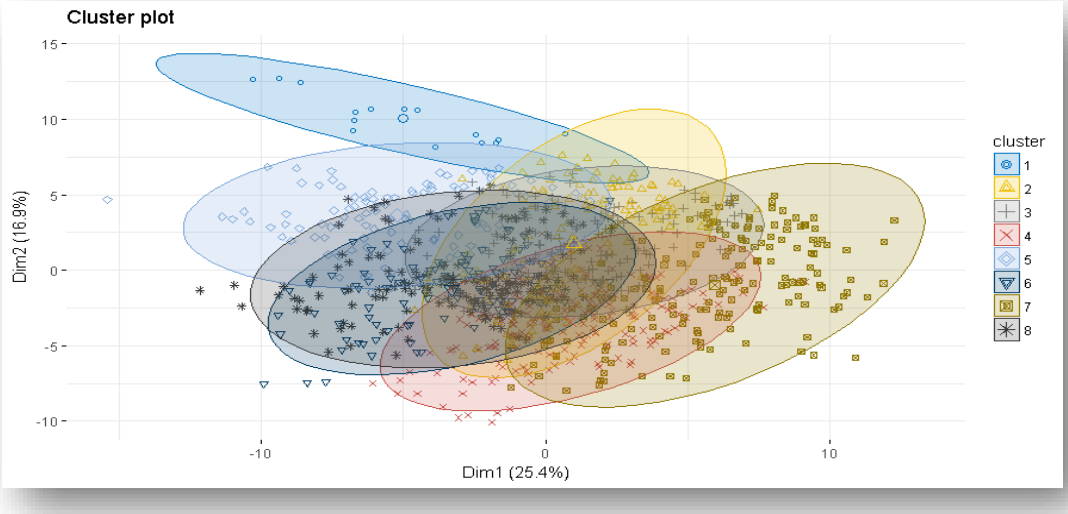
Εικόνα 60 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster



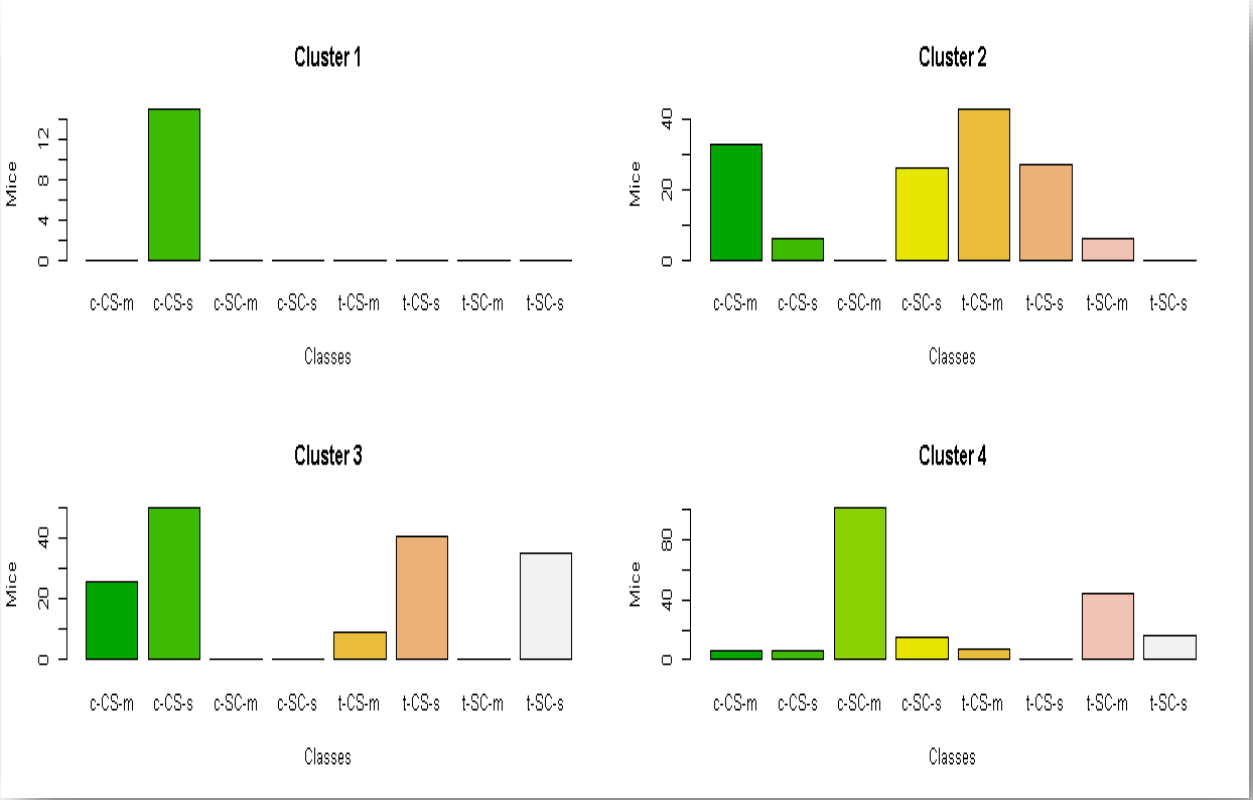
Εικόνα 61 Ποντίκια του κάθε γενότυπου (α), θεραπείας (β) και συμπεριφοράς (γ) στο κάθε cluster

Από τα παραπάνω διαγράμματα, παρατηρούμε ότι στο Cluster 1 είναι τα περισσότερα υγιή, που έλαβαν ορό και διεγέρθηκαν στη μάθηση. Στο 2ο cluster έχουμε ποντίκια τρισωμικά, μη διεγερμένα στη μάθηση, στα οποία χορηγήθηκε μεμαντίνη. Βέβαια, δε γίνεται κάποιος περαιτέρω διαχωρισμός ως προς το γενότυπο.

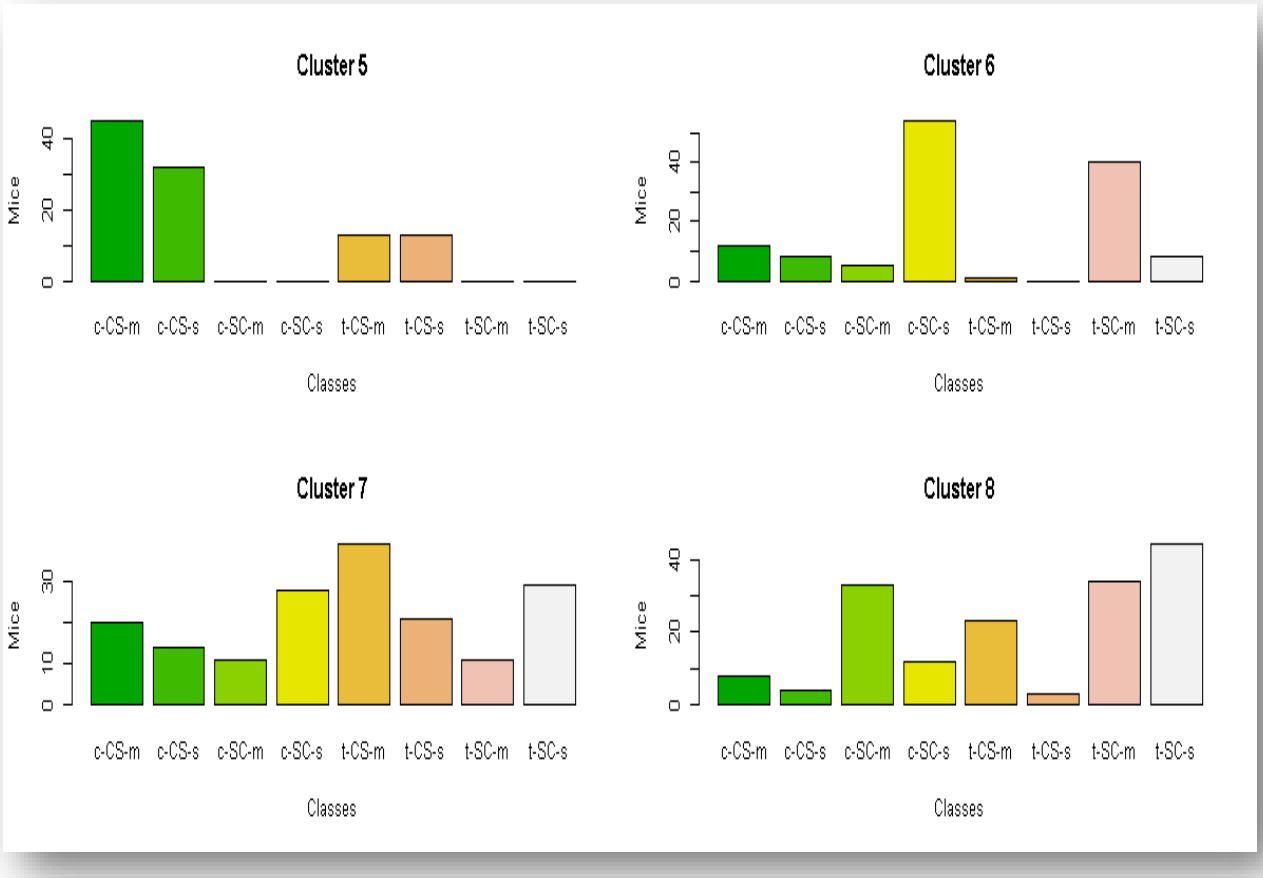
- **k=8**



Εικόνα 62 Απεικόνιση των clusters με τον αλγόριθμο K-Means για k=8

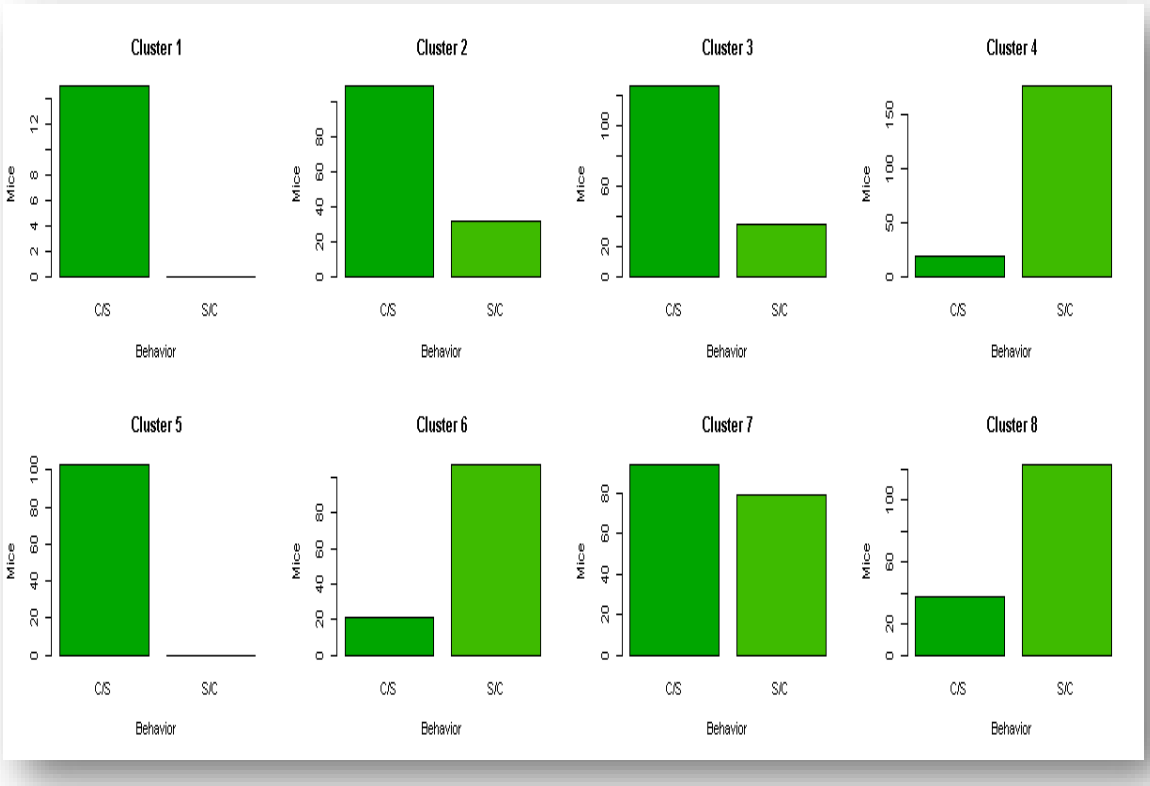


Εικόνα 63 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (α)

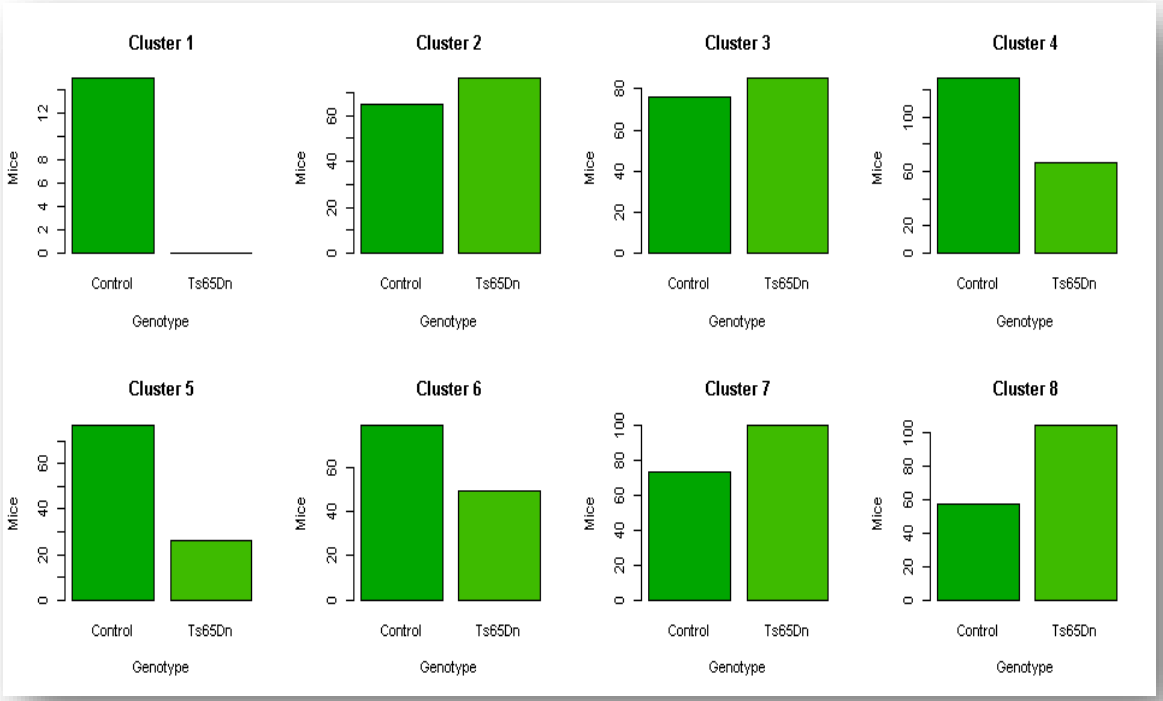


Εικόνα 64 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (β)

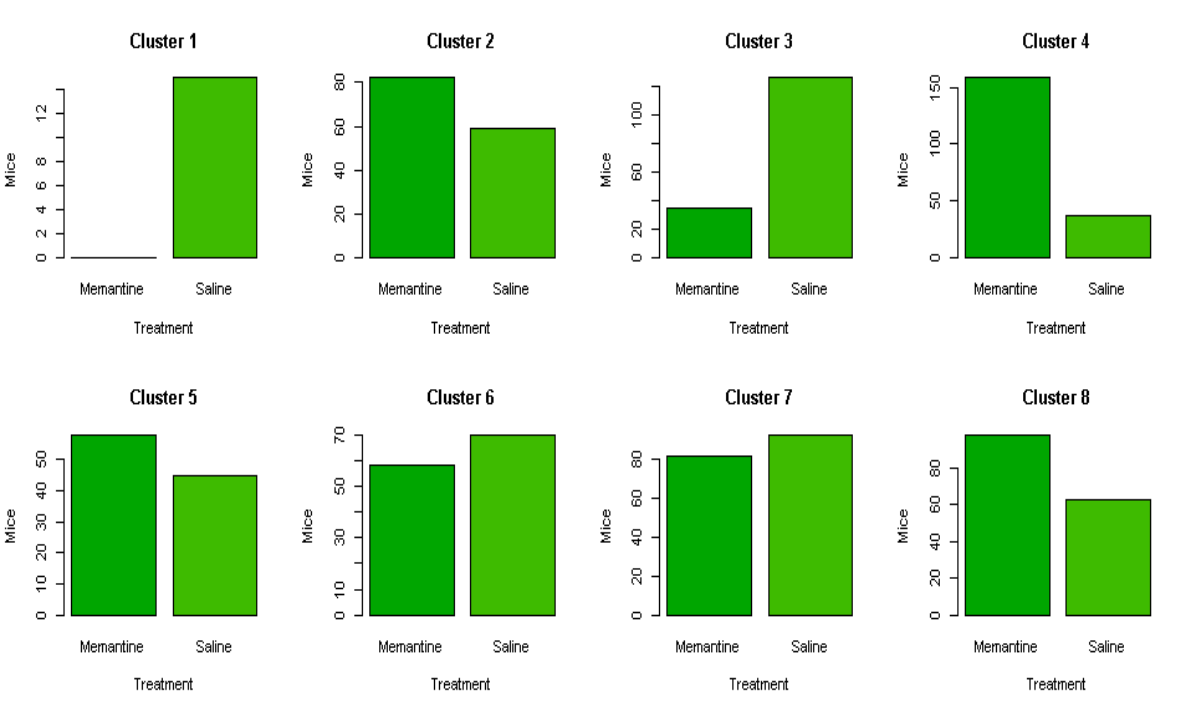
Από τα παραπάνω διαγράμματα, παρατηρούμε ότι στο cluster 1 έχουμε αποκλειστικά υγιή, διεγερμένα στη μάθηση, που έλαβαν ορό. Στο cluster 2 έχουμε έμφαση στα τρισωμικά, διεγερμένα στη μάθηση, που έλαβαν μεμαντίνη. Στο 3ο έχουμε έμφαση στα τρισωμικά, διεγερμένα ποντίκια που έλαβαν ορό, στο 4ο έχουμε τα υγιή, μη διεγερμένα, στα οποία χορηγήθηκε μεμαντίνη και στο 5ο έχουμε έμφαση στα υγιή, διεγερμένα στη μάθηση, τα οποία έλαβαν και αυτά μεμαντίνη. Τέλος, στο 6ο έχουμε κυρίως τα υγιή, μη διεγερμένα, που έλαβαν ορό όπως επίσης και τρισωμικά μη διεγερμένα που έλαβαν μεμαντίνη, ενώ στο 7ο και στο 8ο έχουμε τρισωμικά ποντίκια, με το 7ο να έχει περισσότερα διεγερμένα, που έλαβαν μεμαντίνη και το 8ο περισσότερα μή διεγερμένα, που έλαβαν ορό. Συμπεραίνουμε λοιπόν, ότι παρότι σε κάποιες συστάδες ο διαχωρισμός δεν μην είναι τόσο έντονος, όμως γενικά, τα ποντίκια χωρίζονται στις 8 πραγματικές κλάσεις.



Εικόνα 65 Ποντίκια της κάθε διέγερσης- συμπεριφοράς για μάθηση (behavior) στο κάθε cluster (C/S vs S/C)



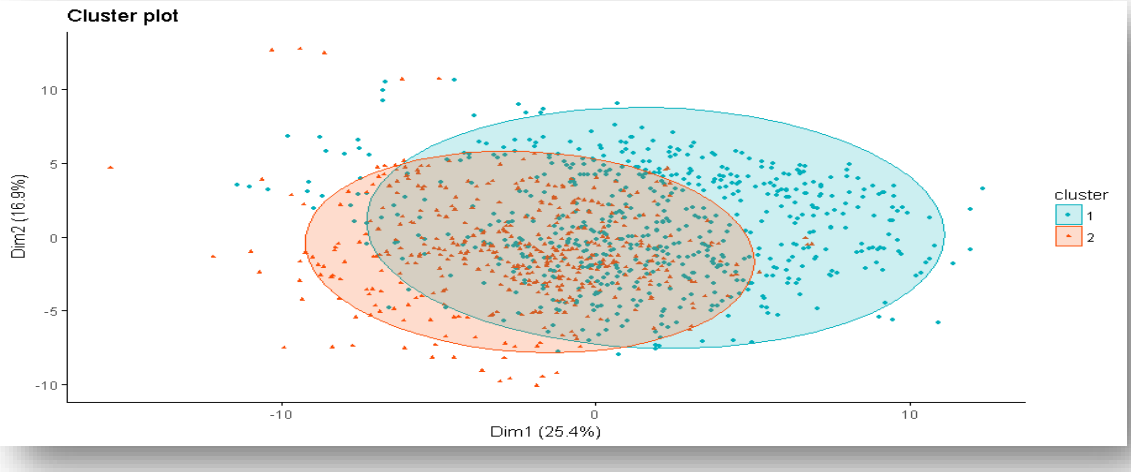
Εικόνα 66 Ποντίκια του κάθε γενότυπου (genotype) στο κάθε cluster (Control vs Ts65Dn)



Εικόνα 67 Ποντίκια του κάθε φαρμάκου (treatment) (Memantine vs Saline)

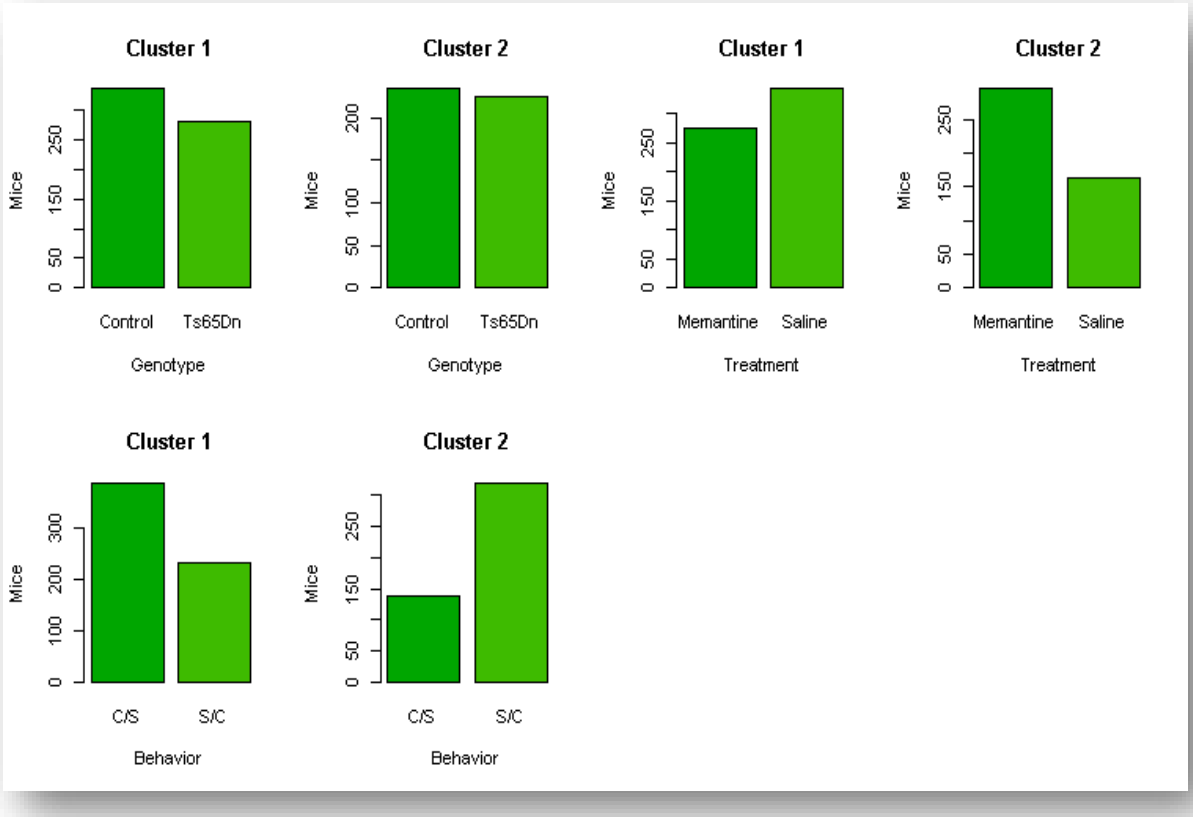
5.2.6 ΑΛΓΟΡΙΘΜΟΣ CLARA

- **k=2**



Εικόνα 68 Απεικόνιση των clusters με τον αλγόριθμο Clara για k=2

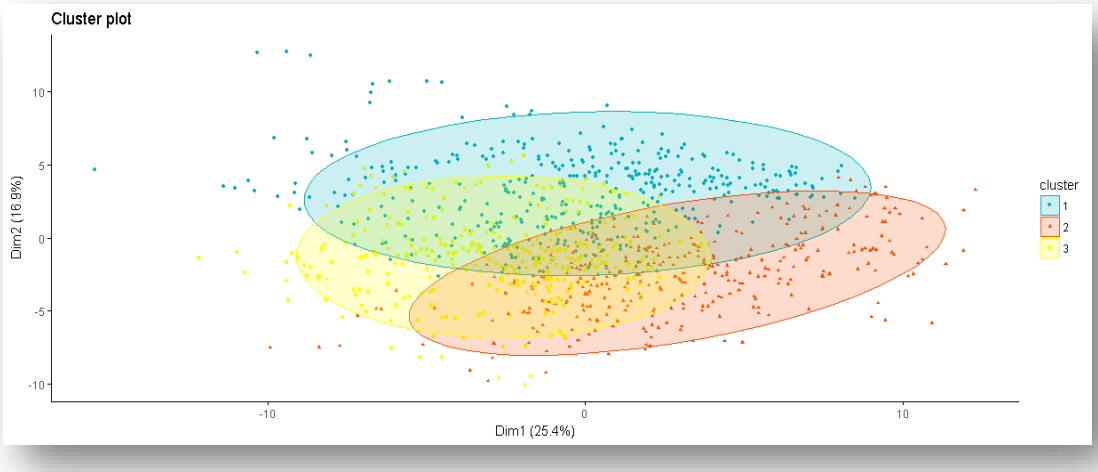
Εικόνα 69 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster



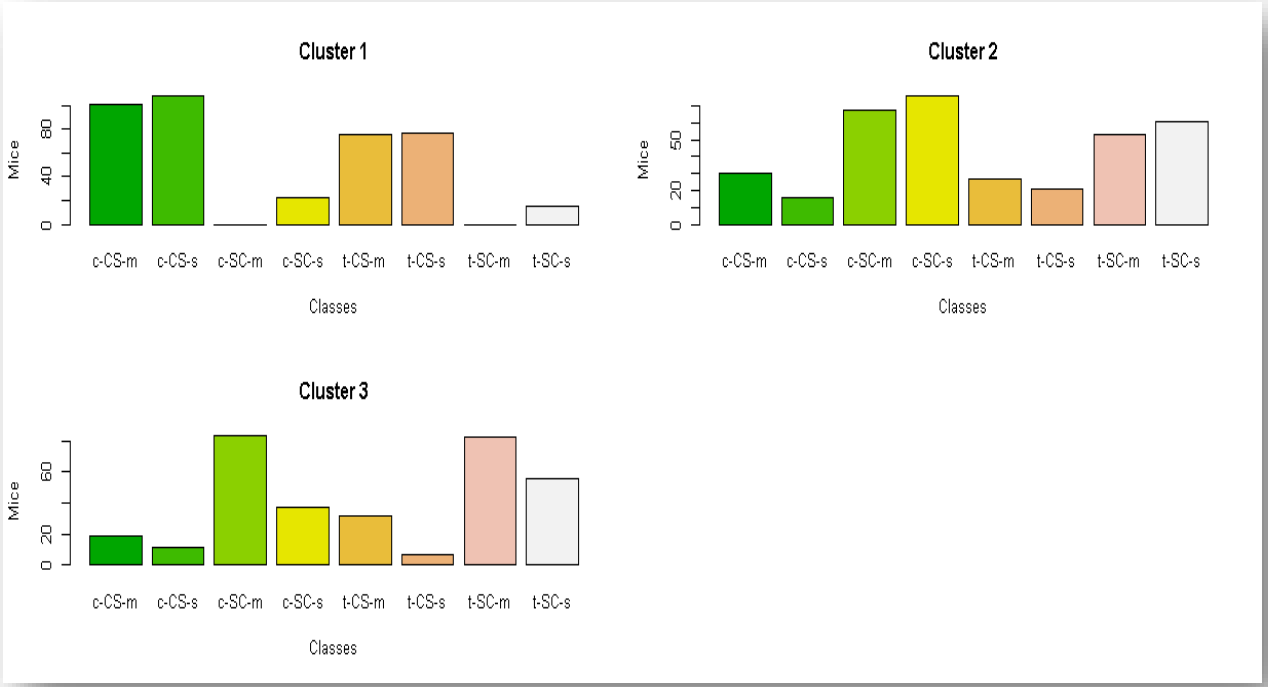
Εικόνα 70 Ποντίκια του κάθε γενότυπου (α), φαρμάκου (β) και συμπεριφοράς (γ) στο κάθε cluster

Από όλα τα παραπάνω, παρατηρούμε ότι στο 1ο cluster έχουμε υγιή, διεγερμένα ποντίκια, στα οποία χορηγήθηκε κυρίως μεμαντίνη αλλά και ορός, ενώ στο 2ο έχουμε υγιή, μη διεγερμένα ποντίκια, που έλαβαν μεμαντίνη. Είναι φανερό ότι δεν υπάρχει κάποιος διαχωρισμός ως προς το γενότυπο.

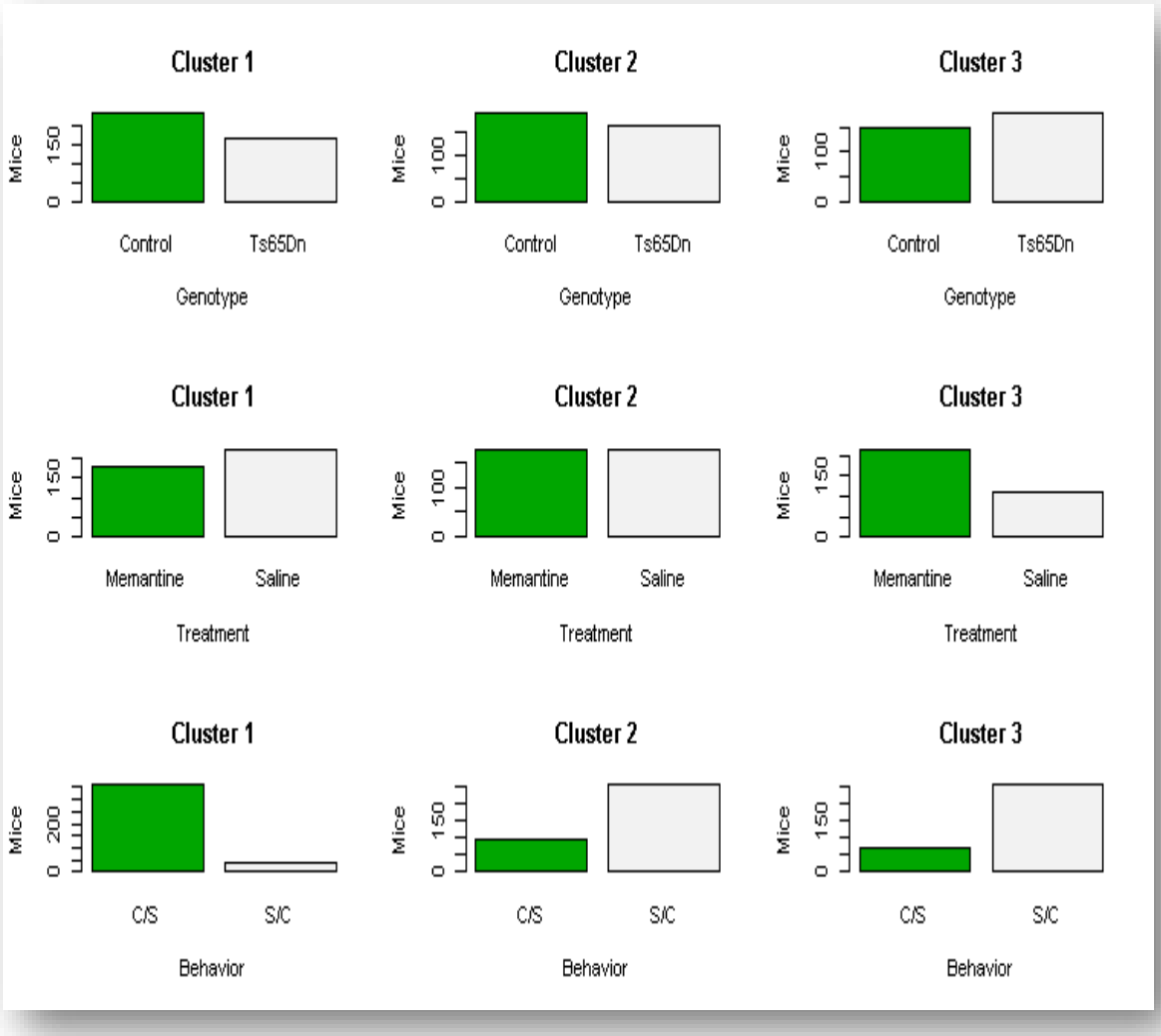
- **k=3**



Εικόνα 71 Απεικόνιση των clusters με τον αλγόριθμο Clara για k=3



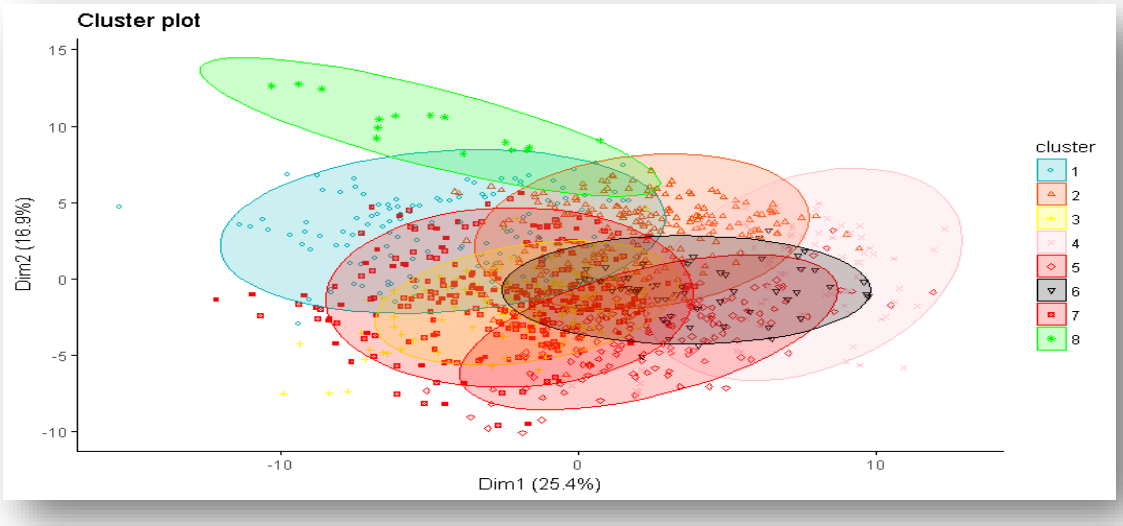
Εικόνα 72 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster



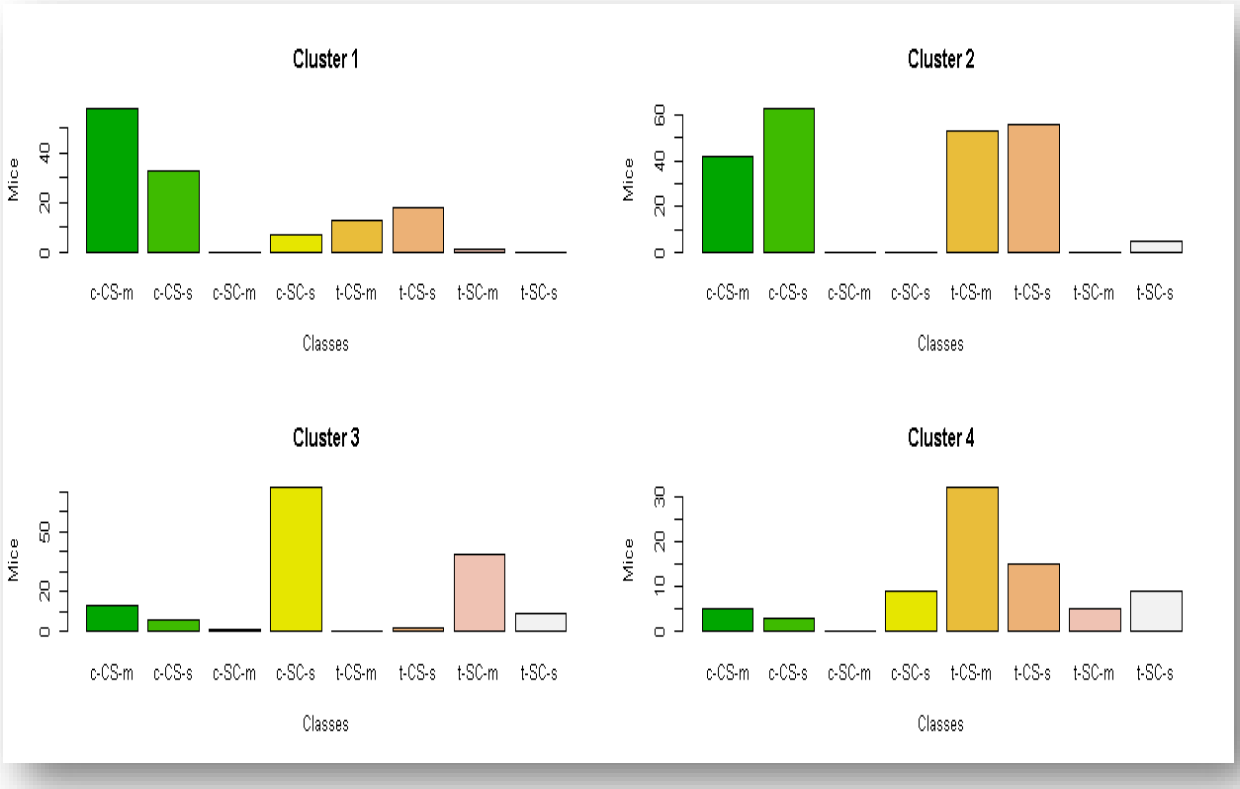
Εικόνα 73 Ποντίκια του κάθε γενότυπου (α), θεραπείας (β) και συμπεριφοράς (γ) στο κάθε cluster

Στο 1ο cluster έχουμε λίγα πιο πολλά ποντίκια υγιή, διεγερμένα για μάθηση, που έλαβαν ορό, Στο 2ο έχουμε υγιή, μη διεγερμένα ποντίκια, από τα οποία τα μισά έλαβαν μεμαντίνη και τα άλλα μισά ορό. Τέλος, στο 3ο υπερισχύουν τα τρισωμικά, μη διεγερμένα που έλαβαν μεμαντίνη. Ως προς το γενότυπο, δε γίνεται κάποιος ιδιαίτερος διαχωρισμός, μιας και οι μπάρες είναι περίπου στο ίδιο ύψος.

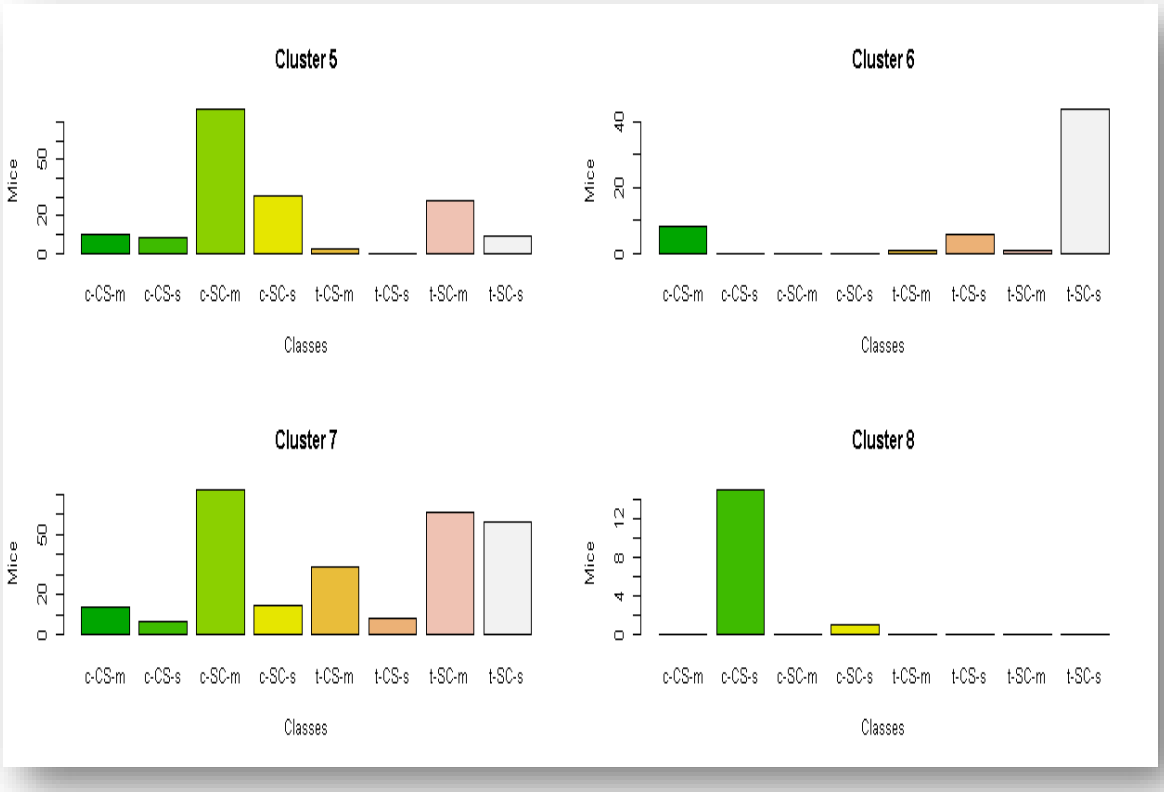
k=8



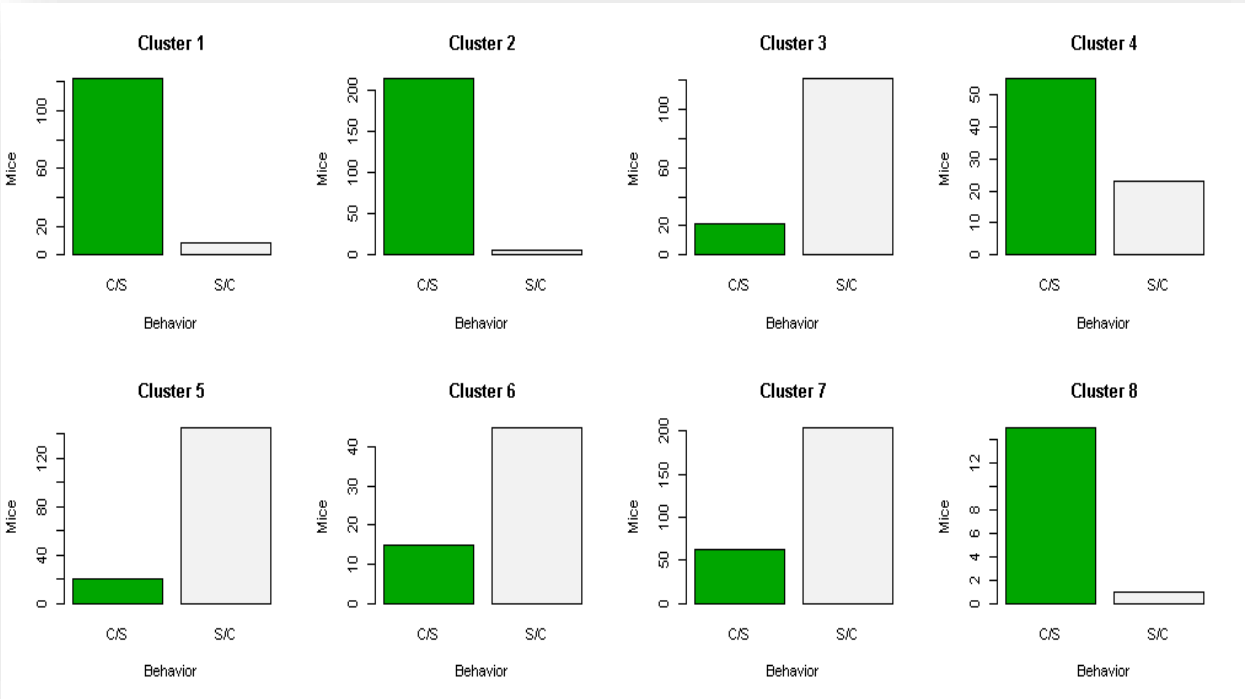
Εικόνα 74 Απεικόνιση των clusters με τον αλγόριθμο Clara για k=8



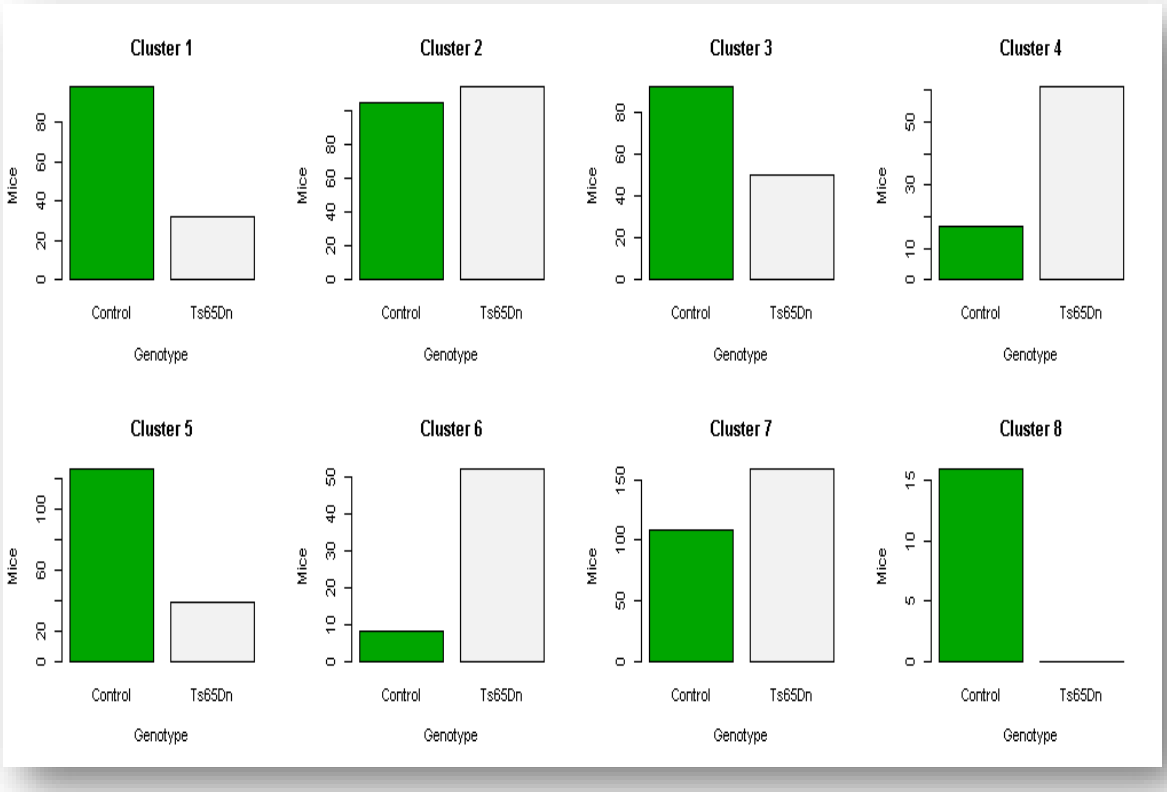
Εικόνα 75 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (α)



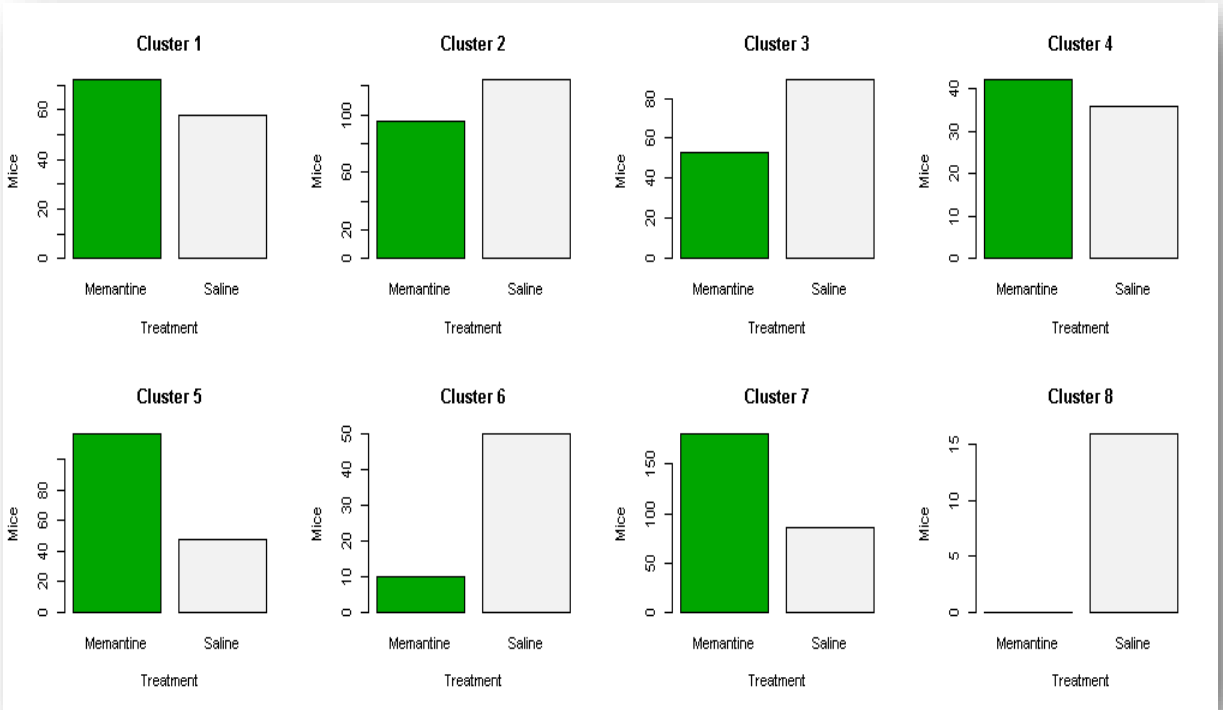
Εικόνα 76 Ποντίκια των πραγματικών κλάσεων που υπάρχουν σε κάθε cluster (β)



Εικόνα 77 Ποντίκια της κάθε συμπεριφοράς-διέγερσης για μάθηση (behavior) στο κάθε cluster (C/S vs S/C)



Εικόνα 78 Ποντίκια του κάθε γενότυπου (genotype) στο κάθε cluster (Control vs Ts65Dn)



Εικόνα 79 Ποντίκια του κάθε φαρμάκου (treatment) στο κάθε cluster (Memantine vs Saline)

Από όλα τα παραπάνω διαγράμματα, παρατηρούμε ότι στο 1ο και στο 8ο cluster έχουμε υγιή, διεγερμένα ποντίκια, με το 1ο να συγκεντρώνει αυτά που έλαβαν μεμαντίνη και το 8ο αυτά που έλαβαν ορό. Στο 2ο και στο 4ο έχουμε τα τρισωμικά, διεγερμένα ποντίκια, με το 2ο να συγκεντρώνει κυρίως αυτά που έλαβαν ορό και το 4ο αυτά που έλαβαν μεμαντίνη. Στο 3ο και στο 5ο έχουμε τα υγιή, μη διεγερμένα στη μάθηση ποντίκια, με αυτά που έλαβαν ορό να βρίσκονται στο 3ο και αυτά που έλαβαν μεμαντίνη στο 5ο. Τέλος, στο 6ο και στο 7ο έχουμε τα τρισωμικά, μη διεγερμένα ποντίκια. Στο 6ο βρίσκονται τα ποντίκια εκείνα, στα οποία χορηγήθηκε ορός και στο 7ο εκείνα στα οποία χορηγήθηκε μεμαντίνη. Εδώ, βλέπουμε λοιπόν, ότι είναι ξεκάθαρος ο διαχωρισμός των ποντικιών στις πραγματικές κλάσεις.

ΣΧΟΛΙΟ

Θα πρέπει να τονιστεί ότι ο προτεινόμενος συνδυασμός μεθόδου ταξινόμησης και αριθμού συστάδων από τον αντίστοιχο αλγόριθμο απέτυχε να μας δώσει κάποιον διαχωρισμό, πράγμα που σημαίνει ότι ο κάθε ερευνητής θα πρέπει να μην επαφίεται σε ένα μόνο αποτέλεσμα και δεδομένο, αλλά να χρησιμοποιεί όλα τα μέσα που διαθέτει για εξακρίβωση των συμπερασμάτων του.

5.3 ΣΥΜΠΕΡΑΣΜΑΤΑ

Αρχικά, η μέθοδος Ιεραρχικής Ταξινόμησης με τις αποστάσεις Spearman και Kendall καθώς και οι μέθοδοι Μη Ιεραρχικής Ταξινόμησης (αλγόριθμοι K-Means και Clara) για k=8 έδωσαν διαχωρισμό ως προς τις πραγματικές κλάσεις των δεδομένων, πράγμα που υποδεικνύει ότι ακόμα και μόνο η γνώση των πρωτεινών μπορεί να οδηγήσει στη δημιουργία των 8 ομάδων.

Έπειτα, βλέποντας συνοπτικά τα αποτελέσματα όλων των μεθόδων για k=2, 3 και 8 παρατηρούμε ότι στους περισσότερους διαχωρισμούς έχουμε τη δημιουργία τριών συστάδων (αριθμός που προτάθηκε και από τους αλγόριθμους Elbow και Silhouette) με τους παρακάτω συνδυασμούς ποντικιών:

- τα υγιή, διεγερμένα ποντίκια, τα οποία είχαν λάβει ορό
- τα μη διεγερμένα, που είχαν λάβει μεμαντίνη (τις περισσότερες φορές μάλιστα τα ποντίκια αυτά ήταν υγιή)
- και τα τρισωμικά, διεγερμένα, στα οποία είχε χορηγηθεί ορός.

Ακόμα και σε ταξινομήσεις, οι οποίες αποτύγχαναν να διαχωρίσουν καλά τα δεδομένα, οι συνδυασμοί αυτοί εμφανίζονταν ακόμα και σαν μικρές συστάδες.

Καταλήγουμε, λοιπόν, στο ότι η γνώση μόνο των πρωτεϊνών μπορεί να οδηγήσει στο διαχωρισμό των ποντικιών στις πραγματικές τους κλάσεις και ότι οι συστάδες που μπορούν πιο έντοντα και πιο εύκολα να διαχωριστούν είναι οι τρεις που αναφέρθηκαν πιο πάνω. Τέλος, παρατηρούμε ότι γενικά υπάρχει πιο έντονη διαφορά στην κατανομή των διεγερμένων και μη διεγερμένων ποντικών μέσα στις ομάδες, με κάποιες φορές ομάδες να περιλαμβάνουν αποκλειστικά ή σχεδόν αποκλειστικά διεγερμένα ποντίκια και άλλες αποκλειστικά ή σχεδόν αποκλειστικά μη διεγερμένα ποντίκια. Φαίνεται λοιπόν ότι η συμπεριφορά των ποντικών (behavior), παίζει ρόλο στην κατασκευή των ομάδων (clusters).

ΞΕΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Bezdek, J. C., & Hathaway, R. J. (n.d.). VAT: a tool for visual assessment of (cluster) tendency. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)* (pp. 2225–2230). IEEE. https://doi.org/10.1109/IJCNN.2002.1007487

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R Package for Cluster Validation. *JSS Journal of Statistical Software*, *25*(4). Retrieved from http://www.jstatsoft.org/

Caccam, M., & Refran, J. (2012). *Cluster analysis*. Retrieved from https://www.slideshare.net/jewelmrefran/cluster-analysis-15529464

Estivill-Castro, V., & Yang, J. (2000). Fast and Robust General Purpose Clustering Algorithms (pp. 208–218). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44533-1_24

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470977811

Han, J., Kamber, M., & Pei, J. (2011). *Data mining : concepts and techniques*. Elsevier Science.

Higuera, C., Gardiner, K. J., & Cios, K. J. (2015). Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome. *PloS One*, *10*(6), e0129126. https://doi.org/10.1371/journal.pone.0129126

Jianchang Mao, J., & Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks*, *7*(1), 16–29. https://doi.org/10.1109/72.478389

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R*. STHDA. Retrieved from file:///C:/Users/mmiller/Downloads/362035585-Practical-Guide-to-Cluster-Analysis-in-R-Unsupervised-Machine-Learning.pdf

Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data, An Introduction to Cluster Analysis. Retrieved from https://leseprobe.buch.de/images-adb/5c/cc/5ccc031f-49c1-452f-a0ac-22babc5e252e.pdf

MacQueen, J. (1967). *Some methods for classification and analysis of*

multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200512992

Murtagh, T. (1984). *Complexities of hierarchic clustering algorithms: State of the art* (Vol. 05). https://doi.org/10.4236/eng.2013.510B113

Norušis, M. (2011). *Οδηγός Ανάλυσης Δεδομένων με το IBM*. Κλειδάριθμος. Retrieved from https://www.politeianet.gr/books/9789604614653-norusis-j-marija-kleidarithmos-odigos-analusis-dedomenon-me-to-ibm-spss-19-cd-rom-213699

Sharma, S. (1996). *Applied multivariate techniques*. J. Wiley. Retrieved from https://dl.acm.org/citation.cfm?id=225519

Sneath, P., & Sokal, R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. Retrieved from http://www.garfield.library.upenn.edu/classics1987/A1987F272800001.pdf

Tan, N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Retrieved from file:///C:/Users/ΑΓΓΕΛΙΔΗΣ/Desktop/ΙΩΝ/ΔΙΠΛΩΜΑΤΙΚΗ/CLUSTER - K_MEANS (παρουσιαση).pdf

Thorndike, A. (1953). SERIOUS RECURRENT INJURIES OF ATHLETES. *Journal of School Health*, *23*(3), 73–78. https://doi.org/10.1111/j.1746-1561.1953.tb06780.x

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Βερύκιος, Β., Καγκλής, Β., & Σταυρόπουλος, Η. (2016). Η επιστήμη των δεδομένων μέσα από τη γλώσσα R. Retrieved from https://repository.kallipos.gr/handle/11419/2965

Ηλιοπούλου, Π. (2015). Πολυμεταβλητές Μέθοδοι Ανάλυσης. In *Γεωγραφική ανάλυση*. ΣΕΑΒ. Retrieved from https://repository.kallipos.gr/bitstream/11419/2065/1/02_chapter_06.pdf

Καράγεωργα, Ι. (2012). *Ανάλυση Συστάδων (Cluster Analysis)*. Retrieved from http://nemertes.lis.upatras.gr/jspui/bitstream/10889/5932/1/Ανάλυση Συστάδων %28Cluster Analysis%29.pdf

Κύρκος, Ε. (2015). Ανάλυση Συστάδων. In *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. ΣΕΑΒ. Retrieved from https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/1238/2/Kef._11.pdf

Μαύρου, Ν. (2012). *Μέθοδοι Πολυμεταβλητής Στατιστικής Ανάλυσης και Εφαρμογές*. Εθνικό Μετσόβιο Πολυτεχνείο.

Μηλιαρέσης, Γ., & Ηλιοπούλου, Π. (2004). Clustering of Zagros Ranges from the Globe DEM representation. *International Journal of Applied Earth Observation and Geoinformation*, *5*(1), 17–28. https://doi.org/10.1016/J.JAG.2003.08.001

Νικολάου, Χ. (2015). *Υπολογιστική Βιολογία*. (Κάλλιπος, Ed.). Ηράκλειο Κρήτης: Κάλλιπος.

Πετρίδης, Δ. (2015a). *Ανάλυση Πολυμεταβλητών Τεχνικών, Εφαρμογές Περιπτώσεων*. ΣΑΕΒ.

Πετρίδης, Δ. (2015b). Ανάλυση Συστάδων. In *Ανάλυση πολυμεταβλητών τεχνικών*. Retrieved from https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/2130/1/06_chapter05.pdf