Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

<u>Διπλωματική εργασία</u>

Μία εμπειρική σύγκριση μεθόδων ανάλυσης κατά συστάδες για κατηγορικά δεδομένα

του

Δεληγιάννη Δημήτριου

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Επιβλέπων Καθηγητής: Άγγελος Μάρκος

Δεκέμβριος 2023

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Άγγελο Μάρκο για την πολύτιμη και συνεχή υποστήριξη που μου έδειξε, για τις συμβουλές και τις γνώσεις που μου προσέφερε και για την συνολική άριστη συνεργασία που είχαμε.

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στη συγκριτική ανάλυση διαφορετικών μεθόδων συσταδοποίησης για κατηγορικά δεδομένα. Συγκεκριμένα, παρουσιάζονται και συγκρίνονται μέθοδοι όπως η K-modes, η Ανάλυση Λανθανουσών Τάξεων (LCA) και η Ανιούσα Ιεραρχική Ταξινόμηση σε δέκα πραγματικά σύνολα δεδομένων από το αποθετήριο UCI. Τα αποτελέσματα της σύγκρισης έδειξαν ότι η LCA υπερτερεί στις περισσότερες περιπτώσεις έναντι των άλλων μεθόδων, ενώ ακολουθούν η μέθοδος K-modes και ορισμένες παραλλαγές της Ανιούσας Ιεραρχικής Ταξινόμησης. Στη συνέχεια, αναλύονται τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου και γίνονται προτάσεις για την αξιοποίηση των μεθόδων αυτών στην συσταδοποίηση κατηγορικών δεδομένων.

Abstract

This thesis compares different clustering methods suitable for categorical data. Specifically, methods such as K-modes, Latent Class Analysis (LCA) and Hierarchical Clustering are presented and compared on ten real datasets obtained from the UCI repository. The results of the experiments showed that LCA outperforms the other methods in most cases, followed by K-modes and some variants of the Hierarchical Clustering. Subsequently, the advantages and disadvantages of each method are presented and specific recommendations are made for the use of these methods in practice.

Περιεχόμενα

1. Εισαγωγή

1.1 Η Ανάλυση σε Συστάδες

Η ανάλυση κατά συστάδες, ή αλλιώς ταξινόμηση ή συσταδοποίηση (clustering), αποτελεί θεμελιώδη τεχνική στον τομέα της στατιστικής ανάλυσης δεδομένων που χρησιμοποιείται για την ομαδοποίηση αντικειμένων ή υποκειμένων ενός συνόλου δεδομένων με βάση τις ομοιότητες μεταξύ τους. Αυτή η μέθοδος δηλαδή περιλαμβάνει την ομαδοποίηση αντικειμένων με βάση τις ομοιότητες τους, δημιουργώντας ομάδες όπου τα μέλη μοιράζονται περισσότερες κοινές ιδιότητες μεταξύ τους από εκείνες που βρίσκονται σε διαφορετικές ομάδες. Η σημασία της υπογραμμίζεται από την ευρεία εφαρμογή της σε διάφορους τομείς, προσφέροντας γνώσεις σχετικά με τις φυσικές ομάδες εντός ενός συνόλου δεδομένων, οι οποίες ενδέχεται να μην είναι άμεσα προφανείς. Το περιεχόμενο αυτής της εργασίας αφορά στη συσταδοποίηση και επικεντρώνονται σε μεθόδους κατάλληλες για κατηγορικά δεδομένα. Η συσταδοποίηση κατηγορικών δεδομένων, λόγω της ιδιαίτερης φύσης της και των εγγενών προκλήσεων που αντιμετωπίζει, έχει γίνει ένα ολοένα και πιο σημαντικό πεδίο μελέτης, προσελκύοντας σημαντικό ερευνητικό ενδιαφέρον.

Τα κατηγορικά (ποιοτικά) δεδομένα διαφέρουν θεμελιωδώς από τα συνεχή (ποσοτικά) δεδομένα καθώς αναφέρονται σε ποιοτικά χαρακτηριστικά των υποκειμένων. Η επικράτησή τους εκτείνεται σε διάφορους τομείς όπως οι κοινωνικές επιστήμες, όπου εμφανίζονται σε έρευνες κοινής γνώμης, στη βιολογία, ιδιαίτερα στην ταξινόμηση γενετικών ακολουθιών, και στην επιστήμη υπολογιστών, ιδίως σε τομείς όπως η μηχανική μάθηση, στην αναγνώριση μοτίβων και στη λήψη αποφάσεων. Οι ποικίλες εφαρμογές της ανάλυσης κατηγορικών δεδομένων, που κυμαίνονται από την αναγνώριση κοινωνικών τάσεων έως την αποκρυπτογράφηση σύνθετων βιολογικών μοτίβων μέσω ταξινόμησης γονιδίων, υπογραμμίζουν τη σημασία τους (Agresti, 2002). Στον τομέα του μάρκετινγκ και της διαχείρισης σχέσεων με τους πελάτες, η κατανόηση των προτιμήσεων των πελατών μέσω της ανάλυσης κατηγορικών δεδομένων, όπως οι επιλογές προϊόντων ή η ανατροφοδότηση υπηρεσιών είναι ζωτικής σημασίας για τις επιχειρηματικές στρατηγικές.

Σε αντίθεση με τα ποσοτικά δεδομένα, όπου η Ευκλείδεια απόσταση μπορεί εύκολα να ποσοτικοποιήσει την ομοιότητα, τα κατηγορικά δεδομένα δεν διαθέτουν εγγενείς αριθμητικές μετρήσεις για τέτοιες συγκρίσεις. Αυτή η ασυμφωνία αποτελεί σημαντική πρόκληση για την ανάλυση σε συστάδες, καθιστώντας αναγκαία την ανάπτυξη εναλλακτικών μέτρων ομοιότητας. Μέτρα όπως η απόσταση Jaccard, η ομοιότητα του συνημίτονου και η απόσταση Hamming έχουν αναπτυχθεί για να αντιμετωπιστεί αυτό το πρόβλημα, παρέχοντας τρόπους ποσοτικοποίησης της ομοιότητας μεταξύ κατηγορικών δεδομένων (Romesburg, 2004).

Πρώιμες προσπάθειες για τη συσταδοποίηση κατηγορικών δεδομένων συχνά βασίζονταν στην προσαρμογή μεθόδων που σχεδιάστηκαν αρχικά για ποσοτικά δεδομένα, όπως η ανάλυση K-means και η Ανιούσα Ιεραρχική Ταξινόμηση. Αυτές οι μέθοδοι, ενώ ήταν πρωτοποριακές από μόνες τους, είχαν περιορισμούς όταν εφαρμόστηκαν απευθείας σε κατηγορικά σύνολα δεδομένων λόγω βασικών διαφορών στην φύση αυτών των δεδομένων (MacQueen, 1967; Sneath & Sokal, 1973).

Οι μέθοδοι συσταδοποίησης μπορούν να χωριστούν σε δύο βασικές κατηγορίες, τις Ιεραρχικές (Hierarchical) και τις Διαμεριστικές (Partitional) μεθόδους. Οι ιεραρχικές μέθοδοι ξεκινούν με ένα σύνολο από ομάδες, όπου κάθε ομάδα αποτελείται από ένα αντικείμενο του αρχικού συνόλου δεδομένων. Στη συνέχεια, οι ομάδες ενώνονται μεταξύ τους με βάση κάποιο κριτήριο εγγύτητας (ή απόστασης, μέχρι να προκύψει μία μοναδική ομάδα που περιλαμβάνει όλα τα αρχικά αντικείμενα. Αντίθετα, οι διαμεριστικές μέθοδοι ξεκινούν με ένα σύνολο από αντικείμενα και έναν προκαθορισμένο αριθμό ομάδων. Στη συνέχεια, τα αντικείμενα κατανέμονται στις ομάδες με βάση κάποιο κριτήριο εγγύτητας από το κέντρο βάρους τους.

Ο πιο δημοφιλής αλγόριθμος διαμεριστικής συσταδοποίησης είναι ο αλγόριθμος K-means. Ο αλγόριθμος K-means ξεκινά με ένα σύνολο από τυχαία επιλεγμένα αντικείμενα, τα οποία αποτελούν τα αρχικά κέντρα των ομάδων. Στη συνέχεια, τα υπόλοιπα αντικείμενα κατανέμονται στις ομάδες με βάση το κοντινότερο κέντρο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να μην αλλάξουν οι ομάδες.

Οι αλγόριθμοι συσταδοποίησης χρησιμοποιούνται σε διάφορους τομείς, όπως η εξόρυξη δεδομένων, η βιοπληροφορική και η μηχανική μάθηση. Οι αλγόριθμοι αυτοί μπορούν να χρησιμοποιηθούν για την ανακάλυψη κρυφών μοτίβων ή ομαδοποιήσεων δεδομένων που δεν είναι εύκολα παρατηρήσιμα.

1.2 Εφαρμογές και χρησιμότητα της Ανάλυσης σε Συστάδες

Η ανάλυση σε συστάδες είναι μια τεχνική πολυμεταβλητής ανάλυσης που χρησιμοποιείται για την ομαδοποίηση αντικειμένων με βάση τις ομοιότητες μεταξύ τους. Η τεχνική αυτή έχει εφαρμογές σε πολλούς διαφορετικούς κλάδους, όπως η βιολογία, η γεωλογία, η κοινωνιολογία και η τεχνολογία.

Μια από τις εφαρμογές της ανάλυσης σε συστάδες είναι η τμηματοποίηση της αγοράς (market segmentation), η οποία χρησιμοποιείται από επιχειρήσεις για να εντοπίσουν διαφορετικές ομάδες αγοραστών με παρόμοιες προτιμήσεις. Η τμηματοποίηση επιτρέπει στις επιχειρήσεις να στοχεύσουν τις προσπάθειές τους σε συγκεκριμένες ομάδες αγοραστών και να εφαρμόσουν διαφορετικές στρατηγικές μάρκετινγκ για κάθε ομάδα. Επίσης μια άλλη εφαρμογή της ανάλυσης σε συστάδες είναι στην ιατρική, όπου χρησιμοποιείται για να συνδέσει ατομικά χαρακτηριστικά με ασθένειες. Η ανάλυση σε συστάδες μπορεί να βοηθήσει στην εύρεση αποτελεσματικών θεραπειών και στην πρόληψη ασθενειών.

Η παρούσα διπλωματική εργασία εστιάζει στην εφαρμογή της ανάλυσης σε συστάδες σε κατηγορικά δεδομένα, δηλαδή δεδομένα που περιγράφονται από ποιοτικές μεταβλητές. Τα ποιοτικά δεδομένα εμφανίζονται συχνά σε πραγματικές εφαρμογές. Η εργασία αυτή παρουσιάζει μια ανασκόπηση των σημαντικότερων μεθόδων ανάλυσης σε συστάδες για ποιοτικά δεδομένα. Η ανασκόπηση αυτή βασίζεται σε μια εκτενή βιβλιογραφική έρευνα και περιλαμβάνει τόσο τις παραδοσιακές όσο και τις νεότερες μεθόδους ανάλυσης σε συστάδες για ποιοτικά δεδομένα.

Συμπερασματικά, η ανάλυση σε συστάδες είναι μια ισχυρή τεχνική που μπορεί να χρησιμοποιηθεί για την εξαγωγή χρήσιμης γνώσης από δεδομένα. Η μελέτη αυτή συμβάλλει στην καλύτερη κατανόηση των δυνατοτήτων της ανάλυσης σε συστάδες για κατηγορικά δεδομένα και δείχνει πώς η τεχνική αυτή μπορεί να εφαρμοστεί σε πραγματικές εφαρμογές.

1.3 Τύποι δεδομένων συσταδοποίησης

Οι τύποι δεδομένων που διαχειρίζονται οι μέθοδοι συσταδοποίησης είναι οι παρακάτω. Ποσοτικές ή συνεχείς μεταβλητές είναι αυτές που μπορούν να πάρουν οποιαδήποτε αριθμητική τιμή, όπως το ύψος, το βάρος, η τιμή ενός προϊόντος, η απόσταση από ένα μέρος σε ένα άλλο. Ποιοτικές μεταβλητές είναι αυτές που εκφράζουν κάποιο ποιοτικό χαρακτηριστικό των αντικειμένων, όπως το επάγγελμα, το φύλο, το χρώμα ματιών. Οι ποιοτικές μεταβλητές μπορούν να διακριθούν σε τρεις κατηγορίες τις Διχοτομικές, τις Ονομαστικές και τις Διατακτικές μεταβλητές. Οι διχοτομικές μεταβλητές μπορούν να πάρουν μόνο δύο τιμές, για παράδειγμα 'άντρας' ή 'γυναίκα'. Οι ονομαστικές μεταβλητές μπορούν να πάρουν πάνω από δύο τιμές, για παράδειγμα οι πιθανές απαντήσεις στην ερώτηση 'ποια είναι η οικογενειακή σας κατάσταση. Καθώς οι διατακτικές μεταβλητές μπορούν να πάρουν τιμές οι οποίες έχουν φυσική διάταξη, για παράδειγμα η συχνότητα εμφάνισης μιας συμπεριφοράς μπορεί να πάρει τις τιμές 'ποτέ', 'σπάνια', 'συχνά', 'πάντα'. Οι μέθοδοι συσταδοποίησης που θα παρουσιάσουμε παρακάτω μπορούν να διαχειριστούν δεδομένα που αποτελούνται μόνο από ποιοτικές μεταβλητές.

1.4 Σκοπός της παρούσας εργασίας

Οι βασικοί στόχοι της παρούσας διπλωματικής εργασίας είναι να παρουσιάσει μια ανασκόπηση τριών μεθόδων συσταδοποίησης για κατηγορικά δεδομένα, να εφαρμόσει και να συγκρίνει αυτές τις μεθόδους σε πραγματικά σύνολα δεδομένων τα οποία αντλήθηκαν από το αποθετήριο μηχανικής μάθησης UCI. Το UCI χρησιμοποιείται από την κοινότητα των επιστημόνων μηχανικής μάθησης για την εμπειρική ανάλυση αλγορίθμων και μεθόδων.

Οι μέθοδοι που θα εξεταστούν είναι η K-modes, η Ανάλυση Λανθανουσών Τάξεων - LCA (Latent Class Analysis) και η Ανιούσα Ιεραρχική Ταξινόμηση (Hierarchical Clustering). Για κάθε μια μέθοδο, θα παρουσιαστεί το στοιχειώδες μαθηματικό υπόβαθρο του αλγορίθμου και η διαδικασία εφαρμογής του. Επιπλέον, θα αναλυθούν τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου.

Η εφαρμογή των μεθόδων σε πραγματικά δεδομένα από το αποθετήριο UCI θα επιτρέψει να αξιολογηθούν τα πλεονεκτήματα και τα μειονεκτήματα των μεθόδων αυτών σε πραγματικές συνθήκες. Συγκεκριμένα, θα εξεταστεί η ικανότητα των μεθόδων να ομαδοποιούν τα δεδομένα με βάση τις ομοιότητες μεταξύ τους και να εξάγουν χρήσιμα και λογικά αποτελέσματα.

Συμπερασματικά, η παρούσα διπλωματική εργασία επιδιώκει να προσφέρει μια ολοκληρωμένη εικόνα των μεθόδων συσταδοποίησης για κατηγορικά δεδομένα και να διερευνήσει την εφαρμογή τους σε ένα πραγματικά σύνολα δεδομένων.

2. Μέθοδοι Ανάλυσης κατά Συστάδες για κατηγορικά δεδομένα

2.1 Η Μέθοδος K-modes

Η μέθοδος K-modes είναι ένας από τους αλγορίθμους μηχανικής μάθησης χωρίς επίβλεψη (unsupervised) που χρησιμοποιείται για να χωρίσει ένα σύνολο δεδομένων σε έναν προκαθορισμένο αριθμό συστάδων (*k*) με βάση την ομοιότητά τους σε κατηγορικά χαρακτηριστικά. Παρουσιάστηκε για πρώτη φορά από τον Huang (1997) ως μια μέθοδος ομαδοποίησης για μεικτού τύπου δεδομένα και βασίζεται στη μέθοδο K-means. Ο αλγόριθμος K-means ξεκινάει με την τυχαία επιλογή *k* παρατηρήσεων από το σύνολο δεδομένων, οι οποίες ονομάζονται κέντρα ή κερκοειδή των *k* συστάδων. Η τιμή του αριθμού των συστάδων *k,* αποφασίζεται από τον χρήστη. Στη συνέχεια, κάθε παρατήρηση ανατίθεται στην πλησιέστερη συστάδα, με βάση την Ευκλείδεια απόστασή της από τα κέντρα. Αυτό σημαίνει ότι οι παρατηρήσεις ανατίθενται στο πλησιέστερο σε αυτές κέντρο. Μετά την αρχική ανάθεση όλων των παρατηρήσεων σε συστάδες, τα κέντρα επανυπολογίζονται ως ο μέσος όρος των παρατηρήσεων της συστάδας τους. Η διαδικασία ανάθεσης των παρατηρήσεων σε κέντρα και επανυπολογισμού των κέντρων επαναλαμβάνεται έως ότου να σταθεροποιηθούν τα κέντρα, δηλαδή μέχρι να μην υπάρχουν περαιτέρω αλλαγές στην ανάθεση των παρατηρήσεων στις συστάδες, ή μέχρι να ολοκληρωθεί ένας προκαθορισμένος αριθμός επαναλήψεων. Λόγω του υπολογισμού μέσων όρων, όπως γίνεται εύκολα αντιληπτό, ο αλγόριθμος K-means μπορεί να εφαρμοστεί μόνο για την ομαδοποίηση παρατηρήσεων που περιγράφονται από συνεχείς μεταβλητές.

Η μέθοδος K-modes χρησιμοποιείται για την ομαδοποίηση παρατηρήσεων που περιγράφονται αποκλειστικά από κατηγορικές μεταβλητές. Ο αλγόριθμος αυτός ομαδοποιεί τις παρατηρήσεις κατά τον ίδιο τρόπο με την K-means με δύο διαφορές: α) χρησιμοποιείται η απόσταση Hamming αντί της Ευκλείδειας απόστασης και β) το κέντρο ή κεντροειδές είναι η επικρατούσα τιμή αντί του μέσου όρου.

Παράδειγμα: παρακάτω εμφανίζεται ένα σύνολο δεδομένων που περιέχει πληροφορίες για τις προτιμήσεις 8 ατόμων (α1, α2, ..., α8) ως προς το χρόνο, το είδος και τον τρόπο άθλησης. Στόχος είναι η ομαδοποίηση των ατόμων σύμφωνα με τις τιμές τους στις παραπάνω μεταβλητές. Ο χρόνος άθλησης, το είδος άθλησης και ο τρόπος άθλησης, είναι όλες κατηγορικές μεταβλητές.

άτομο	**χρόνος άθλησης**	**είδος άθλησης**	**τρόπος άθλησης**
α1	μεσημέρι	TRX	ατομικά
α2	πρωί	άθλημα	ομαδικά
α3	απόγευμα	καλλισθενική	ομαδικά
α4	βράδυ	γιόγκα	ομαδικά
α5	πρωί	TRX	ατομικά
α6	βράδυ	άθλημα	ομαδικά
α7	απόγευμα	καλλισθενική	ατομικά
α8	βράδυ	γιόγκα	ατομικά

Πίνακας 1.1 πίνακας δεδομένων μεγέθους $\boldsymbol{n \times p}$

Αρχικά επιλέγουμε τον αριθμό των ομάδων, έστω ίσο με 3 (k = 3). Στο πρώτο βήμα, επιλέχθηκαν τυχαία οι παρατηρήσεις-αντικείμενα α1, α7, α8 ως τα κέντρα των τριών ομάδων. Ακολουθεί η διαδικασία υπολογισμού της απόστασης των παρατηρήσεων από κάθε ένα από τα κέντρα. Όταν οι τιμές ταυτίζονται, η απόσταση είναι 0, ενώ όταν διαφέρουν η απόσταση είναι 1.

Κέντρα των ομάδων			
α1	μεσημέρι	TRX	ατομικά
α7	απόγευμα	καλλισθενική	ατομικά
α8	βράδυ	γιόγκα	ατομικά

Πίνακας 1.2 πίνακας των κέντρων των ομάδων κεντροειδών

Η σύγκριση του κέντρου α1 με την παρατήρηση α2 δίνει συνολική απόσταση Hamming ίση με 3(1+1+1). Η σύγκριση του κέντρου α1 με την παρατήρηση α3 δίνει συνολικά απόσταση 3(1+1+1). Η σύγκριση του κέντρου α7 με την παρατήρηση α5 δίνει συνολικά απόσταση 2(1+1+0) και ούτω καθεξής. Παρόμοια υπολογίζονται όλες οι αποστάσεις και τοποθετούνται στον παρακάτω πίνακα.

άτομο	**Cluster 1 (α1)**	**Cluster 2 (α7)**	**Cluster 3 (α8)**	**Cluster**
α1	0	2	2	**Cluster 1**
α2	3	3	3	**Cluster 1**
α3	3	1	3	**Cluster 2**
α4	3	3	1	**Cluster 3**
α5	1	2	2	**Cluster 1**
α6	3	3	2	**Cluster 3**
α7	2	0	2	**Cluster 2**
α8	2	2	0	**Cluster 3**

Πίνακας 1.3 πίνακας ανομοιοτήτων

	Cluster 1	Cluster 2	Cluster 3
α1	1	0	0
α2	1	0	0
α3	0	1	0
α4	0	0	1
α5	1	0	0
α6	0	0	1
α7	0	1	0
α8	0	0	1

Πίνακας 1.4 πίνακας διαμέρισης (Y)

Μετά από τον υπολογισμό του πίνακα αποστάσεων οι παρατηρήσεις-αντικείμενα α1,α2 και α5 ανήκουν στην ομάδα 1 (Cluster 1) ενώ οι παρατηρήσεις α3, α7 ανήκουν στην ομάδα 2 (Cluster 2) και οι παρατηρήσεις α4,α6 και α8 ανήκουν στην ομάδα 3 (Cluster 3). Για να ενταχθεί μια παρατήρηση σε μία ομάδα, επιλέγεται πάντα η μικρότερη απόσταση. Σε περίπτωση «ισοπαλίας» (π.χ. βλ. παρατήρηση α2) επιλέγεται τυχαία η ομάδα στην οποία θα τοποθετηθεί η παρατήρηση (π.χ. cluster 1).

Στη συνέχεια υπολογίζεται για κάθε ομάδα το διάνυσμα με τις επικρατέστερες τιμές (mode) και δημιουργούνται με αυτόν τον τρόπο τα νέα κέντρα. Για παράδειγμα για την ομάδα 1 (cluster 1) στην κατηγορία χρόνος άθλησης η επικρατούσα τιμή είναι το πρωί, ενώ για τις κατηγορίες είδος και τρόπος άθλησης οι επικρατούσες τιμές είναι TRX και ατομικά αντίστοιχα. Ομοίως για τις υπόλοιπες ομάδες.

άτομο	**χρόνος άθλησης**	**είδος άθλησης**	**τρόπος άθλησης**
α1	μεσημέρι	TRX	ατομικά
α2	πρωί	άθλημα	ομαδικά
α3	**απόγευμα**	**καλλισθενική**	**ομαδικά**
α4	**βράδυ**	**γιόγκα**	**ομαδικά**
α5	**πρωί**	**TRX**	**ατομικά**
α6	βράδυ	άθλημα	ομαδικά
α7	απόγευμα	καλλισθενική	ατομικά
α8	βράδυ	γιόγκα	ατομικά

Πίνακας 1.5 πίνακας των ομάδων

Αφού δημιουργηθούν τα νέα κέντρα (cluster 1 → α5, cluster 2 → α3 , cluster 3 → α4) ακολουθείται η ίδια διαδικασία σύγκρισης και ομαδοποίησης τοποθετώντας την κάθε παρατήρηση στην κοντινότερη ομάδα (cluster), δηλαδή την ομάδα με τη μικρότερη απόσταση. Στην προκειμένη περίπτωση όλες οι παρατηρήσεις παραμένουν στις ίδιες ομάδες (clusters), οπότε ο αλγόριθμος τερματίζεται, με την τελική ομαδοποίηση να παρουσιάζεται στον παραπάνω πίνακα. Σε περίπτωση που διαμορφωνόταν διαφορετικά η ομαδοποίηση θα συνεχιζόταν η διαδικασία μέχρις ότου να υπάρχει το ίδιο αποτέλεσμα στις δυο τελευταίες ομαδοποιήσεις.

Ακολουθεί ο ορισμός των συμβόλων καθώς και το μαθηματικό υπόβαθρο της μεθόδου.

n	Πλήθος των αντικειμένων ή υποκειμένων
p	Πλήθος των μεταβλητών
X	Πίνακας δεδομένων μεγέθους $n \times p$
k	Αριθμός των ομάδων αντικειμένων ή υποκειμένων
Q_l	Το κεντροειδές της ομάδας l με στοιχεία $q_{l1},\ldots,q_{lk}$, όπου $1 \leq l \leq k$
Y	Πίνακας διαμέρισης, μεγέθους $n \times k$
$d(\cdot,\cdot)$	Μέτρο απόστασης
W	Συνολική απόσταση μέσα σε κάθε συστάδα

Δεδομένου ότι η επιθυμητή διαμέριση είναι ακέραια, δηλαδή κάθε αντικείμενο μπορεί να ανήκει σε μία μόνο ομάδα, η κάθε γραμμή του πίνακα **Y** θα περιλαμβάνει $n - 1$ μηδενικά και ένα μόνο στοιχείο της θα είναι ίσο με 1. Δηλαδή έχουμε $y_{ij} = 1$ όταν το i ανήκει στην ομάδα j (όπου $1 \leq i \leq n,\ 1 \leq j \leq k$). Η αντικειμενική συνάρτηση προς ελαχιστοποίηση από τον αλγόριθμο αντιστοιχεί στην ελαχιστοποίηση των αποστάσεων εντός των ομάδων και δίνεται από:

$$W = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(X_i, Q_l). \quad (2.1)$$

Σημειώνουμε ότι το X_i αναφέρεται στην γραμμή i του **X** και συνεπώς δίνεται από το διάνυσμα μεγέθους p με στοιχεία $x_{il},\ldots,x_{ip}$. Ας εστιάσουμε στον εσωτερικό όρο της συνάρτησης (2.1):

$$W_l = \sum_{i=1}^{n} y_{il} d(X_i, Q_l)\,, 1 \leq l \leq k. \quad (2.2)$$

Ο όρος αυτός αντιστοιχεί στην ελαχιστοποίηση της διασποράς εντός της ομάδας l. Ο στόχος είναι να ελαχιστοποιήσουμε την (1.2), που με τη σειρά της θα οδηγήσει στην ελαχιστοποίηση της (1.1). Ωστόσο, για να μπορέσουμε να ελαχιστοποιήσουμε την παραπάνω συνάρτηση, θα πρέπει να είμαστε σε θέση να ορίσουμε ένα μέτρο απόστασης τόσο για συνεχή όσο και για κατηγορικά δεδομένα. Αυτό μπορεί να επιτευχθεί με το να ορίσουμε μια απόσταση κατάλληλη για δυαδικά δεδομένα. Πιο συγκεκριμένα, ορίζουμε την $d(X_i, Q_l)$ ως:

$$d(X_i, Q_l) = \gamma \sum_{j=p_r+1}^{p} \delta(x_{ij}, q_{lj}). \quad (2.3)$$

Στο σημείο αυτό, διευκρινίζουμε ότι πριν την εφαρμογή του αλγορίθμου οι κατηγορικές μεταβλητές έχουν μετατραπεί σε μεταβλητές 0-1. Το $\delta(x_{ij}, q_{lj})$ είναι η απόσταση Simple Matching, η οποία ισούται με 0 αν και μόνο αν $x_{ij} = q_{lj}$, διαφορετικά ισούται με 1. Τέλος, με γ συμβολίζεται ένας συντελεστής βαρύτητας των μεταβλητών 0-1. Ο συντελεστής αυτός λαμβάνει μία μόνο τιμή, δηλαδή αφορά στο σύνολο των μεταβλητών και όχι σε κάθε μεταβλητή ξεχωριστά. Η τιμή του γ έχει καθοριστική σημασία στον αλγόριθμο, επομένως θα πρέπει να λάβει μια λογική τιμή. Η τιμή του γ προσδιορίζεται από τον χρήστη και στην παρούσα εργασία ορίστηκε ίση με 1 για όλες τις μεταβλητές.

Επιστρέφουμε στη συνάρτηση (2.2), όπου αντικαθιστούμε την απόσταση $d(X_i, Q_l)$ από την (2.3) και η αντικειμενική συνάρτηση δίνεται από:

$$E_l = \underbrace{\gamma \sum_{i=1}^{n} y_{il} \sum_{j=1}^{p} \delta(x_{ij}, q_{lj})}_{:=E_l^c}. \quad (2.4)$$

Για να ελαχιστοποιήσουμε το E_l^c, πρέπει να ορίσουμε ως C_j το σύνολο όλων των μοναδικών τιμών που λαμβάνει η κατηγορική μεταβλητή j (εδώ $p_r + 1 \leq j \leq p$). Υποθέτουμε ότι c_j είναι μία από τις μοναδικές τιμές στο σύνολο C_j και ορίζεται από $P(c_j \in C_j \vee l)$ την πιθανότητα η κατηγορική μεταβλητή j να παίρνει την τιμή c_j στην ομάδα l. Στη συνέχεια μπορούμε να ορίσουμε εκ νέου την E_l^c από την συνάρτηση (2.4) ως:

$$E_l^c = \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=p_r+1}^{p} \delta(x_{ij}, q_{lj}) = \gamma \sum_{j=p_r+1}^{p} n_l \left(1 - P(q_{lj} \in C_j \vee l)\right).$$

Έτσι, προκειμένου να ελαχιστοποιήσουμε την E_l^c, μπορούμε να μεγιστοποιήσουμε την $P(q_{lj} \in C_j \vee l)$. Επομένως θα πρέπει να ισχύει $P(q_{lj} \in C_j | l) \geq P(c_j \in C_j \vee l)$ για $c_j \neq q_{lj}$, που σημαίνει ότι θέλουμε να μεγιστοποιήσουμε την πιθανότητα ότι η κατηγορική τιμή q_{lj} της μεταβλητής j ανήκει στην ομάδα l σε σχέση με όλες τις άλλες τιμές που μπορεί να πάρει η κατηγορική μεταβλητή. Στην ουσία αυτό που καταλαβαίνουμε είναι ότι η q_{lj} πρέπει να είναι η πιο συχνή τιμή της κατηγορικής μεταβλητής j στην ομάδα l. Στην εργασία αυτή, η μέθοδος K-modes εφαρμόστηκε με το πακέτο klaR (Weihs et al., 2017) της R.

2.2 Η Μέθοδος Latent Class Analysis

Η Ανάλυση Λανθανουσών Τάξεων LCA (Latent Class Analysis) είναι μια πιθανοθεωρητική στατιστική μέθοδος που χρησιμοποιείται για τον εντοπισμό διακριτών υποομάδων σε ένα σύνολο παρατηρήσεων που έχουν ορισμένα κοινά χαρακτηριστικά (Hagenaars & McCutcheon, 2002) ή αλλιώς παρόμοια μοτίβα απάντησης (response patterns), όπως μοτίβα απόκρισης σε κλίμακες τύπου Likert. Αυτές οι υποομάδες ονομάζονται λανθάνουσες τάξεις και θεωρείτε ότι αντιπροσωπεύουν αντίστοιχες ομάδες στον πληθυσμό που δεν είναι άμεσα παρατηρήσιμες. Στο πλαίσιο των ερευνών κοινής γνώμης, με την ανάλυση των μοτίβων απάντησης η LCA μπορεί να βοηθήσει τους ερευνητές να κατανοήσουν τους παράγοντες που επηρεάζουν τις απαντήσεις των ατόμων και να εντοπίσουν διαφορετικά τμήματα του πληθυσμού με διαφορετικές ανάγκες ή χαρακτηριστικά.

Η βασική ιδέα πίσω από την LCA είναι ότι οι παρατηρούμενες κατηγορικές μεταβλητές οφείλονται σε μια υποκείμενη λανθάνουσα μεταβλητή, η οποία καθορίζει την πιθανότητα ένα υποκείμενο να λάβει μια συγκεκριμένη τιμή σε μια μεταβλητή. Στόχος της LCA είναι να εκτιμήσει τις παραμέτρους ενός στατιστικού μοντέλου, όπως ο αριθμός των λανθανουσών τάξεων (ομάδων) και την πιθανότητα κάθε υποκείμενο να ανήκει σε κάθε

τάξη. Η διαδικασία εκτίμησης περιλαμβάνει την επίλυση σύνθετων μαθηματικών εξισώσεων όπως επίσης περιλαμβάνει και τη χρήση επαναληπτικών αλγορίθμων.

Η βασική εξίσωση πίσω από την LCA είναι το Θεώρημα του Bayes. Το θεώρημα Bayes αναφέρεται στις δεσμευμένες πιθανότητες δύο συμβάντων. Πιο συγκεκριμένα, η εξίσωση αυτή περιλαμβάνει μια κατηγορική μεταβλητή Y και μια λανθάνουσα μεταβλητή Z με *K* τάξεις. Η πιθανότητα παρατήρησης ενός συγκεκριμένου μοτίβου απόκρισης, y, για μια δεδομένη τάξη, k, μοντελοποιείται με τη χρήση δεσμευμένων πιθανοτήτων, που συμβολίζονται ως P(Y = y | Z = *k*) και οι τιμές τους εκτιμώνται από τα δεδομένα. Η πιθανότητα ένα υποκείμενο να ανήκει σε μια συγκεκριμένη λανθάνουσα τάξη, *k*, δεδομένου του παρατηρούμενου μοτίβου απάντησης, y, μοντελοποιείται ως εξής:

P (Z = *k* | Y = y) = P (Y = y | Z = *k*) * P (Z = *k*) / P (Y = y)

όπου:

P (Z = *k* | Y = y) είναι η post hoc πιθανότητα ένα υποκείμενο να ανήκει στην τάξη *k*, δεδομένου του παρατηρούμενου μοτίβου απόκρισης, y.

P (Y = y | Z = *k*) είναι η πιθανότητα παρατήρησης του μοτίβου απάντησης y δεδομένης της συμμετοχής στην τάξη k, που ονομάζεται επίσης δεσμευμένη πιθανότητα απόκρισης.

P (Z = *k*) είναι η a priori πιθανότητα ένα υποκείμενο να ανήκει στην τάξη *k*, η οποία εκτιμάται από τα δεδομένα.

P (Y = y) είναι η οριακή πιθανότητα παρατήρησης του μοτίβου απάντησης (response pattern) y, η οποία μπορεί να υπολογιστεί ως το σταθμισμένο άθροισμα των υπό όρους πιθανοτήτων σε όλες τις τάξεις.

Για την εύρεση των τιμών των παραμέτρων του μοντέλου που μεγιστοποιούν την πιθανότητα των δεδομένων, χρησιμοποιείται η μέθοδος της Μεγίστης Πιθανοφάνειας (Maximum Likelihood). Αυτή περιλαμβάνει την επίλυση σύνθετων εξισώσεων, συχνά με τη χρήση επαναληπτικών αλγορίθμων, όπως ο αλγόριθμος Expectation-Maximization (EM). Για να προσδιοριστεί ο βέλτιστος αριθμό λανθανουσών τάξεων που πρέπει να

συμπεριληφθούν στο μοντέλο, χρησιμοποιούνται μια σειρά από κριτήρια, όπως το Bayesian Information Criterion (BIC).

Επομένως, η LCA μπορεί να χρησιμοποιηθεί για τον εντοπισμό διαφορετικών τμημάτων ενός πληθυσμού με παρόμοια χαρακτηριστικά ή συμπεριφορές, για την κατανόηση των παραγόντων που διαφοροποιούν τα τμήματα και για την πρόβλεψη της πιθανότητας ενός ατόμου να ανήκει σε ένα συγκεκριμένο τμήμα. Η LCA μπορεί επίσης να χρησιμοποιηθεί για την αξιολόγηση της μεταβλητότητας μέτρησης ή του βαθμού στον οποίο η λανθάνουσα μεταβλητή είναι συνεπής μεταξύ διαφορετικών υποομάδων του πληθυσμού. Τέλος, η LCA είναι ένα ισχυρό εργαλείο για την αποκάλυψη κρυφών δομών σε κατηγορικά δεδομένα και βρίσκει ένα ευρύ φάσμα εφαρμογών σε διαφορετικά επιστημονικά πεδία, όπως στο μάρκετινγκ, τις κοινωνικές επιστήμες, τη δημόσια υγεία, την έρευνα αγοράς και άλλα.

2.3 Ανιούσα Ιεραρχική Ταξινόμηση

Στην ενότητα αυτή θα παρουσιαστεί ο βασικός αλγόριθμος της Ανιούσας Ιεραρχικής Ταξινόμησης, αφού προηγουμένως οριστούν δύο βασικές μετρικές απόστασης για κατηγορικά δεδομένα και δύο κριτήρια συνένωσης των ομάδων κατά τη διαδικασία συγχώνευσής τους. Η Ανιούσα Ιεραρχικής Ταξινόμησης είναι μία τεχνική της πολυμεταβλητής στατιστικής ανάλυσης που χρησιμοποιείται για να ομαδοποιήσει τα δεδομένα σε συστάδες βάσει της ομοιότητας ή της απόστασης μεταξύ τους. Ο στόχος είναι να δημιουργηθούν ομάδες (ή συστάδες) των δεδομένων ώστε τα υποκείμενα εντός κάθε ομάδας να είναι πιο όμοια μεταξύ τους από ό,τι με τα υποκείμενα σε άλλες ομάδες. Η Ιεραρχική Ταξινόμηση διακρίνεται σε δύο βασικές κατηγορίες:

Ανιούσα (Agglomerative): Αυτή είναι η πιο συνηθισμένη προσέγγιση. Ξεκινά με κάθε υποκείμενο να ανήκει σε μία ξεχωριστή ομάδα και στη συνέχεια συγχωνεύει σταδιακά τις ομάδες βάσει της ομοιότητας τους, μέχρι όλα τα υποκείμενα να ανήκουν σε μια μοναδική ομάδα.

Κατιούσα (Divisive): Αρχίζει με όλα τα υποκείμενα σε μία μοναδική ομάδα και σταδιακά τα χωρίζει σε μικρότερες και πιο ομοιογενείς ομάδες. Σε κάθε κατηγορία, η απόφαση για το πώς να συνενωθούν οι συστάδες βασίζεται σε διάφορα κριτήρια, όπως:

α) Το κριτήριο της ελάχιστης απόστασης: Μετρά την απόσταση μεταξύ των πιο κοντινών υποκειμένων δύο συστάδων.

β) Το κριτήριο της μέγιστης απόστασης: Μετρά την απόσταση μεταξύ των πιο απομακρυσμένων υποκειμένων δύο συστάδων.

γ) Μέση απόσταση: Υπολογίζει το μέσο της απόστασης μεταξύ όλων των ζευγών των υποκειμένων από διαφορετικές συστάδες.

δ) Μέθοδος του Ward: Βασίζεται στην ελαχιστοποίηση της αύξησης της συνολικής εσωτερικής διακύμανσης των συστάδων.

Τα αποτελέσματα της Ανιούσας Ιεραρχικής Ταξινόμησης συχνά παρουσιάζονται με τη μορφή ενός δέντρου, γνωστό ως δενδρόγραμμα (dendrogram), το οποίο δείχνει τις ομάδες σε διαφορετικά επίπεδα της ιεραρχίας. Στην περίπτωση των κατηγορικών δεδομένων, θα εστιάσουμε σε δύο μετρικές απόστασης (Manhattan και Gower) και σε τρία κριτήρια συνένωσης ομάδων (κριτήριο του κεντροειδούς, Ward.d και Ward.D2), τα οποία περιγράφονται παρακάτω,

- Η απόσταση Manhattan, γνωστή και ως απόσταση taxicab, είναι ένας τρόπος υπολογισμού της απόστασης μεταξύ υποκειμένων που περιγράφονται από κατηγορικές μεταβλητές. Ονομάζεται έτσι επειδή μιμείται τον τρόπο που ένα ταξί κινείται στους δρόμους μιας πόλης, οι οποίοι σχηματίζουν οικοδομικά τετράγωνα, όπως στο Μανχάταν της Νέας Υόρκης. Η απόσταση Manhattan μεταξύ δύο σημείων (x_1, y_1) και (x_2, y_2) υπολογίζεται ως το άθροισμα των απόλυτων διαφορών των x-συντεταγμένων και των y-συντεταγμένων τους ως εξής :

Απόσταση Manhattan = $|x_1 - x_2| + |y_1 - y_2|$

- Το μέτρο απόστασης του Gower (Gower, 1971) ανάμεσα σε δύο υποκείμενα (διανύσματα) X_i και X_j δίνεται από τη σχέση:

$$d_{Gower}(X_i, X_j) = 1 - \frac{\sum_{k=1}^{p} w_k(X_i, X_j) s_k(X_i, X_j)}{\sum_{k=1}^{p} w_k(X_i, X_j)}, 1 \leq i, j \leq n, i \neq j, (2.5)$$

Όπου w_k είναι το βάρος της μεταβλητής k για τα δύο υποκείμενα X_i *και* X_j και s_k είναι η τιμή ενός μέτρου ομοιότητας μεταξύ των τιμών της μεταβλητής k για τα δυο υποκείμενα. Η τιμή του w_k συνήθως ισούνται με 1 εκτός εάν η τιμή μιας μεταβλητής για ένα ή και τα δύο υποκείμενα απουσιάζει, περίπτωση στην οποία το αντίστοιχο βάρος είναι 0. Επομένως, βλέπουμε ότι το μέτρο απόστασης του Gower μπορεί να διαχειριστεί και τις ελλείπουσες τιμές. Όσον αφορά τον ορισμό του μέτρου ομοιότητας s_k, αυτός εξαρτάται από τον τύπο της μεταβλητής k. Διακρίνουμε τις ακόλουθες περιπτώσεις:

- Για συνεχείς μεταβλητές, $s_k(X_i, X_j) = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$, όπου R_k είναι το δειγματικό εύρος της μεταβλητής k. Στην ουσία το s_k είναι η απόσταση Hamming ή απόσταση Manhattan μεταξύ των x_{ik} και x_{jk}, τυποποιημένη με το εύρος R_k.

- Για κατηγορικές μεταβλητές, $s_k(X_i, X_j) = \mathbb{I}\{x_{ik} = x_{jk}\}$, το s_k ορίζεται ως μια συνάρτηση η οποία ισούται με 1 αν και μόνο αν $x_{ik} = x_{jk}$, αλλιώς ισούται με 0.

Εύκολα αποδεικνύεται ότι το μέτρο από του Gower (2.5) είναι μετρική.

- Με βάση το κριτήριο συνένωσης του κεντροειδούς (centroid) κάθε συστάδα χαρακτηρίζεται από ένα κεντροειδές, το οποίο είναι ο μέσος όρος όλων των σημείων ή υποκειμένων που ανήκουν στη συστάδα. Η απόσταση μεταξύ δύο συστάδων υπολογίζεται βάσει της απόστασης μεταξύ των κεντροειδών τους (π.χ. Manhattan ή Gower). Κατά τη διαδικασία της συγχώνευσης, οι δύο πλησιέστερες συστάδες (βάσει της απόστασης των κεντροειδών τους) συγχωνεύονται σε μία νέα συστάδα. Το κεντροειδές της νέας συστάδας υπολογίζεται ξανά. Το κριτήριο του κεντροειδούς είναι ιδιαίτερα χρήσιμο όταν τα δεδομένα είναι ομοιογενή και όταν οι μεταβλητές είναι σχετικά λίγες. Ωστόσο, μπορεί να μην είναι τόσο αποτελεσματικό σε περιπτώσεις όπου οι συστάδες έχουν διαφορετικά μεγέθη ή σχήματα.

- Τα κριτήρια συνένωσης Ward.d και Ward.D2 αποτελούν δύο παραλλαγές της μεθόδου συνένωσης του Ward.d, η οποία χρησιμοποιείται επιδιώκει να ελαχιστοποιήσει την αύξηση της συνολικής εσωτερικής διακύμανσης των συστάδων κατά τη συνένωση. Οι δύο παραλλαγές διαφοροποιούνται ως προς τον τρόπο υπολογισμού της διακύμανσης. Το κριτήριο Ward.d υπολογίζει την αύξηση της συνολικής εσωτερικής διακύμανσης ως την απόσταση μεταξύ των κεντροειδών των συστάδων που συνενώνονται. Αυτή η μέθοδος επικεντρώνεται στην ελαχιστοποίηση της αύξησης της συνολικής διακύμανσης εντός των συστάδων όταν αυτές συγχωνεύονται. Από την άλλη, το κριτήριο Ward.D2, είναι μια παραλλαγή του Ward.d. Σε αυτή την περίπτωση, η αύξηση της διακύμανσης υπολογίζεται ως το τετράγωνο της απόστασης. Αυτό το κριτήριο τείνει να δίνει μεγαλύτερο βάρος στη συνένωση συστάδων που βρίσκονται πιο κοντά μεταξύ τους, ενισχύοντας την τάση για πιο ομοιογενείς συστάδες.

Και τα δύο κριτήρια είναι σχεδιασμένα για να εντοπίζουν συστάδες που είναι εσωτερικά συνεκτικές και να ελαχιστοποιούν τη διακύμανση εντός αυτών των συστάδων. Ωστόσο, η επιλογή μεταξύ του Ward.d και του Ward.D2 μπορεί να εξαρτηθεί από τα συγκεκριμένα χαρακτηριστικά των δεδομένων και τις απαιτήσεις της ανάλυσης.

3. Εφαρμογή των μεθόδων ανάλυσης κατά συστάδες σε πραγματικά κατηγορικά δεδομένα

Στο κεφάλαιο αυτό εφαρμόζουμε τις μεθόδους ανάλυσης κατά συστάδες που παρουσιάστηκαν στο προηγούμενο κεφάλαιο σε δέκα πραγματικά σύνολα κατηγορικών δεδομένων που προέρχονται από το αποθετήριο UCI. Αρχικά, περιγράφουμε τα χαρακτηριστικά του κάθε συνόλου δεδομένων και στη συνέχεια παρουσιάζουμε τους δείκτες αξιολόγησης της συσταδοποίησης. Οι αναλύσεις πραγματοποίηθηκαν σε R και ο κώδικας παρουσιάζεται στο Παράρτημα Α.

3.1 Περιγραφή των συνόλων δεδομένων

Το σύνολο δεδομένων Breast Cancer Wisconsin από το αποθετήριο UCI περιέχει δεδομένα που προέρχονται από ψηφιοποιημένες εικόνες από αναρρόφηση με λεπτή βελόνα (FNA) μαζών του μαστού. Περιλαμβάνει 569 αντικείμενα και 30 μεταβλητές. Αυτές οι μεταβλητές περιγράφουν χαρακτηριστικά των κυτταρικών πυρήνων που υπάρχουν στις εικόνες και περιλαμβάνουν μετρήσεις όπως ακτίνα, υφή, περίμετρος, εμβαδόν, ομαλότητα, συμπαγής, κοίλο, κοίλα σημεία, συμμετρία και διάσταση φράκταλ. Το σύνολο δεδομένων χρησιμοποιείται για την εργασία ταξινόμησης της πρόβλεψης του κατά πόσον ο καρκίνος είναι καλοήθης ή κακοήθης.

Το σύνολο δεδομένων 'House Votes '84' από το αποθετήριο UCI εμπεριέχει δεδομένα ψηφοφορίας του 1984 του Κογκρέσου των ΗΠΑ. Επί της ουσίας περιλαμβάνει ψήφους για κάθε ένα από τα μέλη της Βουλής των Αντιπροσώπων των Ηνωμένων Πολιτειών σχετικά με τις 16 βασικές ψηφοφορίες που εντόπισε η CQA. Η CQA απαριθμεί εννέα διαφορετικούς τύπους ψήφων: ψήφισε υπέρ, συντάχθηκε υπέρ και ανακοίνωσε υπέρ (αυτοί οι τρεις απλοποιήθηκαν σε ναι), ψήφισε κατά, συντάχθηκε κατά αυτό το σύνολο δεδομένων και ανακοίνωσε κατά (αυτοί οι τρεις απλοποιήθηκαν σε όχι), ψήφισε παρών, ψήφισε παρών για να αποφύγει σύγκρουση συμφερόντων και δεν ψήφισε ή αλλιώς δεν έκανε γνωστή τη θέση του (αυτοί οι τρεις απλοποιήθηκαν σε άγνωστη διάθεση). Περιλαμβάνει 435 αντικείμενα, κάθε μία αντιστοιχίζεται σε ένα μέλος του Κογκρέσου, και περιλαμβάνει 16 βασικές ψηφοφορίες. Ο κύριος στόχος του συνόλου δεδομένων είναι να

ταξινομήσει τα μέλη του Κογκρέσου ως Δημοκρατικούς ή Ρεπουμπλικάνους βάσει των μοτίβων της ψηφοφορίας τους. Συγκεκριμένες ψηφοφορίες που καταγράφονται περιλαμβάνουν θέματα όπως βοήθεια σε βρέφη με αναπηρία, κοινοχρησία κόστους έργου νερού, υιοθέτηση του προϋπολογισμού, πάγωμα των τελών των γιατρών και πολλά άλλα, με απαντήσεις σημειωμένες ως 'y' ναι ή 'n' όχι.

Το σύνολο δεδομένων Car Evaluation (1997) θα μπορούσε να χρησιμοποιηθεί για την αξιολόγηση αυτοκινήτων με βάση διάφορα χαρακτηριστικά, όπως η τιμή αγοράς (buying), το κόστος συντήρησης (maint), ο αριθμός των θυρών (doors), η χωρητικότητα (persons), το μέγεθος του χώρου αποσκευών (lug_boot) και η ασφάλεια (safety). Περιλαμβάνει 1728 αντικείμενα και 6 μεταβλητές. Η μεταβλητή στόχου είναι class, η οποία είναι η αξιολόγηση του αυτοκινήτου (unacc, acc, good, vgood).Το σύνολο δεδομένων περιέχει περίπου ίσες περιπτώσεις για κάθε κατηγορία, εξασφαλίζοντας ισορροπημένη αναπαράσταση.

Το σύνολο δεδομένων Nursery Dataset περιέχει αιτήσεις για παιδικούς σταθμούς και περιλαμβάνει χαρακτηριστικά όπως το επάγγελμα των γονέων (parents), τις συνθήκες του παιδικού σταθμού (has_nurs), τη μορφή του παιδιού(form), τις συνθήκες στέγασης (housing) και την οικονομική κατάσταση (finance), την κοινωνική κατάσταση της οικογένειας (social)και την κατάσταση υγείας του παιδιού (health). Περιλαμβάνει 12960 αντικείμενα και 8 μεταβλητές. Η εργασία ταξινόμησης θα μπορούσε να περιλαμβάνει την κατηγοριοποίηση των αιτήσεων με βάση ορισμένα κριτήρια. Η μεταβλητή στόχου δείχνει εάν το παιδί προτάθηκε για εισαγωγή στο νηπιαγωγείο (recom, not_recom). Όπου η μεταβλητή στόχου είναι κατηγορική και ο στόχος είναι να αναθέσει κάθε περίπτωση σε μια προκαθορισμένη κατηγορία.

Το σύνολο δεδομένων Tic-Tac-Toe Endgame από το αποθετήριο μηχανικής μάθησης UCI περιέχει σύμβολα του αντίστοιχου παιχνιδιού τρίλιζας (X, O ή B, όπου X = ο παίκτης έχει X σημειώσει στο αντίστοιχο τετράγωνο, όπου O = ο παίκτης O έχει σημειώσει στο αντίστοιχο τετράγωνο και όπου B = κενό). Περιέχει μια συλλογή από 958 αντικείμενα και 9 μεταβλητές, κάθε μία από τις οποίες αντιπροσωπεύει μια πιθανή τελική διαμόρφωση πίνακα σε ένα παιχνίδι Tic-Tac-Toe. Το σύνολο δεδομένων περιέχει χαρακτηριστικά όπως

TL (Πάνω αριστερό τετράγωνο), MM (Μεσαίο μεσαίο τετράγωνο), BR (Κάτω δεξί τετράγωνο) και ούτω καθεξής. Κάθε περίπτωση αντιπροσωπεύεται από εννέα χαρακτηριστικά, κάθε ένα από τα οποία αντιστοιχεί στην κατάσταση ενός από τα εννέα τετράγωνα στον πίνακα Tic-Tac-Toe (X, O ή B). Η μεταβλητή στόχου είναι μια δυαδική μεταβλητή που υποδεικνύει εάν η τελική διαμόρφωση του πίνακα αντιπροσωπεύει νίκη για τον παίκτη X ή όχι. Το σύνολο δεδομένων Tic-Tac-Toe Endgame είναι ένας πολύτιμος πόρος για ερευνητές και επαγγελματίες που ενδιαφέρονται να αναπτύξουν και να αξιολογήσουν αλγόριθμους μηχανικής μάθησης για εργασίες ταξινόμησης, ιδιαίτερα στο πλαίσιο της ανάπτυξης παιχνιδιών και της τεχνητής νοημοσύνης

Το ιατρικό σύνολο δεδομένων Lymphography περιέχει πληροφορίες που σχετίζονται με τη λειτουργία του λεμφικού συστήματος, και συγκεκριμένα με τη διάγνωση του καρκίνου του λεμφικού συστήματος. Περιλαμβάνει 148 αντικείμενα και 19 μεταβλητές περιλαμβανόμενη και η μεταβλητή class. Αυτές οι μεταβλητές περιλαμβάνουν διάφορες ιδιότητες του λεμφικού συστήματος και περιγράφουν χαρακτηριστικά των λεμφικών αγγείων όπως το μπλοκάρισμα της παροχής, το μπλοκάρισμα του λεμφικού, τα εξωαγγειακά, την μειωμένη λύμη, τις αλλαγές στη λεμφίδα, την βλάβη στον κόμβο, τις αλλαγές στη δομή και άλλα. Το σύνολο δεδομένων στοχεύει να βοηθήσει στην ανάπτυξη μοντέλων μηχανικής μάθησης για την αυτόματη διάγνωση των λεμφικών νόσων, παρέχοντας πολύτιμες πληροφορίες για τις σχέσεις μεταξύ διαφορετικών χαρακτηριστικών και των συγκεκριμένων τύπων καταστάσεων που υπάρχουν στο λεμφικό σύστημα. Η συνεχιζόμενη έρευνα και ανάπτυξη σε αυτόν τον τομέα έχει τη δυνατότητα να βελτιώσει τις πρακτικές υγειονομικής περίθαλψης και να βελτιώσει τα αποτελέσματα των ασθενών στον τομέα της λεμφολογίας.

Το σύνολο δεδομένων Australian Credit Approval Dataset περιέχει δεδομένα αιτήσεων για πιστωτικές κάρτες. Περιλαμβάνει 690 αντικείμενα και 15 μεταβλητές. Αυτές οι μεταβλητές περιγράφουν χαρακτηριστικά που σχετίζονται με μια σειρά προσωπικών και οικονομικών λεπτομερειών των αιτούντων όπως το όνομα του αιτούντος, η διεύθυνση, πληροφορίες σχετικά με το εισόδημα, την εργασιακή κατάσταση, το πιστωτικό ιστορικό του αιτούντος, πληροφορίες σχετικά με το ζητούμενο ποσό δανείου και άλλα. Η

κατηγοριοποίηση που σχετίζεται με αυτό το σύνολο δεδομένων στοχεύει στην πρόβλεψη εάν μια αίτηση πίστωσης θα εγκριθεί ή θα απορριφθεί. Αυτός ο στόχος είναι κρίσιμος για τους χρηματοπιστωτικούς οργανισμούς προκειμένου να βελτιστοποιήσουν τις διαδικασίες έγκρισης και να λαμβάνουν ενημερωμένες αποφάσεις με βάση το προφίλ του αιτούντος.

Το σύνολο δεδομένων Sponges Dataset περιέχει πληροφορίες σχετικά με σφουγγάρια, μια ποικιλόμορφη ομάδα θαλάσσιων ασπόνδυλων που απαντώνται σχεδόν σε όλους τους θαλάσσιους οικότοπους. Περιλαμβάνει 76 αντικείμενα και 45 μεταβλητές. Το σύνολο δεδομένων περιλαμβάνει πληροφορίες σχετικά με τη μορφολογία (σχήμα, χρώμα), τη φυσιολογία (αναπαραγωγή, μηχανισμοί άμυνας) και την οικολογία (βάθος νερού, κατανομή) των σφουγγαριών, καθώς και τη διανομή τους σε όλο τον κόσμο.. Η συλλογή δεδομένων των σφουγγαριών χρησιμοποιείται κυρίως για ομαδοποίηση, στοχεύει δηλαδή στο να ομαδοποιήσει παρόμοιες περιπτώσεις μαζί. Η ομαδοποίηση είναι χρήσιμη για την ανάλυση της ποικιλομορφίας των σφουγγαριών και την αναγνώριση μοτίβων στη διανομή τους. Επίσης μπορεί να χρησιμοποιηθεί για τη μελέτη των παραγόντων που επηρεάζουν την κατανομή των σφουγγαριών, όπως η θερμοκρασία του νερού, η αλατότητα και ο τύπος του οικοτόπου. Τέλος το σύνολο δεδομένων των Sponges Dataset είναι πολύτιμη πηγή για ερευνητές που μελετούν σφουγγάρια και θαλάσσια οικοσυστήματα.

Το σύνολο δεδομένων Balloons Dataset περιέχει πληροφορίες σχετικά με μπαλόνια. Περιλαμβάνει 16 αντικείμενα και 4 μεταβλητές. Όπως επίσης περιλαμβάνει χαρακτηριστικά των μπαλονιών όπως το μέγεθος (διάμετρος του μπαλονιού σε εκατοστά), το χρώμα (κωδικός χρώματος του μπαλονιού), το σχήμα (σφαίρα, καρδιά), την υφή των μπαλονιών, καθώς και το υλικό (λατέξ, πλαστικό) από το οποίο κατασκευάζονται. Το σύνολο δεδομένων χρησιμοποιείται κυρίως για ταξινόμηση η οποία είναι χρήσιμη για την κατανόηση των χαρακτηριστικών των μπαλονιών και την αναγνώριση μοτίβων μεταξύ τους. Μπορεί να χρησιμοποιηθεί για τη μελέτη των παραγόντων που επηρεάζουν την επιλογή μπαλόνι, όπως η ηλικία, το φύλο και η περίσταση. Όπως επίσης και για την ταυτοποίηση των παραγόντων που επηρεάζουν την ποιότητα των μπαλόνι, όπως η διάρκεια ζωής και η αντοχή.

Το σύνολο δεδομένων Mushrooms Dataset περιέχει πληροφορίες σχετικά με τα μανιτάρια. Περιλαμβάνει 8124 αντικείμενα και 22 μεταβλητές. Επίσης περιλαμβάνει χαρακτηριστικά σχετικά με τη μορφολογία (σχήμα, χρώμα, μέγεθος, οσμή και υφή), τη φυσιολογία (ανάπτυξη, αναπαραγωγή και βιοχημικές ιδιότητες) και την οικολογία (συνήθειες διατροφής και είδη σαπροφυτικών και μυκοριζικών ειδών) των μανιταριών, καθώς και τις φαρμακευτικές τους ιδιότητες(θεραπευτικές ιδιότητες και ενδείξεις χρήσης). Χρησιμοποιείται κυρίως για ταξινόμηση των μανιταριών και στην πρόβλεψη των ιδιοτήτων τους αναγνωρίζοντας μοτίβα μεταξύ τους. Μπορεί να χρησιμοποιηθεί για την ανακάλυψη νέων φαρμακευτικών παραγόντων από μανιτάρια, όπως φάρμακα για τον καρκίνο, τις καρδιακές παθήσεις και τη νόσο του Πάρκινσον. Το σύνολο δεδομένων των μανιταριών είναι μια πολύτιμη πηγή για επιστήμονες και ερευνητές που μελετούν τα μανιτάρια και τις χρήσεις τους. Παρέχει πληθώρα πληροφοριών σχετικά με τα χαρακτηριστικά των μανιταριών, τις παραμέτρους που επηρεάζουν τις ιδιότητές τους και τις πιθανές εφαρμογές τους.

	Αριθμός χαρακτηριστικών	**Αριθμός κλάσεων**	**Είδος δεδομένων**	**Αριθμός περιπτώσεων**
Breast Cancer	*11*	*2*	*Κατηγορικές*	*699*
House Votes 84	*17*	*2*	*Κατηγορικές*	*435*
Car-Evaluation	*6*	*4*	*Κατηγορικές*	*1728*
Nursery	*8*	*2*	*Κατηγορικές*	*12960*
Tic-Tac-Toe	*9*	*2*	*Κατηγορικές*	*958*
Lymphography	*19*	*1-4*	*Κατηγορικές*	*148*
Australian	*15*	*2*	*Κατηγορικές*	*690*
Sponges	*45*	*12*	*Κατηγορικές*	*76*
Balloons	*4*	*2*	*Κατηγορικές*	*16*
Mushrooms	*22*	*2*	*Κατηγορικές*	*8124*

Πίνακας 2.1 πίνακας χαρακτηριστικών συνόλων δεδομένων

3.2 Δείκτες αξιολόγησης

Για την αξιολόγηση της αποτελεσματικότητας των μεθόδων υπολογίστηκαν οι δείκτες Adjusted Rand Index (ARI) και Adjusted Mutual Information (AMI). Πρόκειται για δύο στατιστικά μέτρα που χρησιμοποιούνται για την αξιολόγηση της απόδοσης αλγορίθμων ομαδοποίησης, συγκεκριμένα για τη σύγκριση της ομαδοποίησης που προκύπτει από τον αλγόριθμο με μια δεδομένη ομαδοποίηση αναφοράς (ground truth). Οι δύο δείκτες λαμβάνουν τιμές στο διάστημα μεταξύ -1 και 1, όπου 1 σημαίνει τέλεια συμφωνία μεταξύ των ομαδοποιήσεων και 0 ή αρνητικές τιμές υποδηλώνουν καμία ή τυχαία συμφωνία.

Adjusted Rand Index (ARI). Ο ARI (Hubert & Arabie, 1985) είναι μια διορθωμένη εκδοχή του δείκτη RI (Rand Index). Κατά τα σύγκριση δύο ομαδοποιήσεων, ο δείκτης RI είναι το ποσοστό των παρατηρήσεων που τοποθετούνται στην ίδια ομάδα στις δύο ομαδοποιήσεις. Ο ARI αποτελεί μια διόρθωση του Rand Index για την πιθανότητα τυχαίας συμφωνίας. Η μαθηματική του έκφραση είναι:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \quad (3.1)$$

όπου n_{ij} είναι ο αριθμός των παρατηρήσεων που βρίσκονται τόσο στην ομάδα i της πρώτης ομαδοποίησης όσο και στην ομάδα j της δεύτερης, a_i είναι ο αριθμός των παρατηρήσεων στην ομάδα i της πρώτης ομαδοποίησης, b_j είναι ο αριθμός των παρατηρήσεων στην ομάδα j της δεύτερης ομαδοποίησης, και n είναι ο συνολικός αριθμός των παρατηρήσεων.

Adjusted Mutual Information (AMI): Ο δείκτης AMI (Vinh, Epps, & Bailey, 2009) είναι μια διορθωμένη εκδοχή του δείκτη Αμοιβαίας Πληροφορίας (Mutual Information) για την πιθανότητα τυχαίας συμφωνίας. Ο δείκτης MI αντιστοιχεί στην ποσότητα της πληροφορίας που είναι κοινή ανάμεσα στις δύο ομαδοποιήσεις. Η μαθηματική του έκφραση είναι:

$$\text{AMI} = \frac{MI - E[MI]}{\max\big(H(U), H(V)\big) - E[MI]} \quad (3.2)$$

όπου MI είναι η Αμοιβαία Πληροφορία (Mutual Information), και H(U), H(V) είναι οι εντροπίες των δύο διαμερισμών.

Η αμοιβαία Πληροφορία (MI) υπολογίζεται ώς :

$$\mathrm{MI}(U,V) = \sum_{i=1}^{R} \sum_{j}^{C} P(i,j) \log\left(\frac{P(i,j)}{P(i)P(j)}\right) \quad (3.3)$$

όπου:

- P(i,j) είναι η συνδυαστική πιθανότητα του στοιχίου i να βρίσκεται στη συστάδα j .
- P(i), P'(j) είναι οι περιθωριακές πιθανότητες των συστάδων.

Οι δύο αυτοί δείκτες επιλέχθηκαν για την αξιολόγηση της απόδοσης των αλγορίθμων ομαδοποίησης διότι εξετάζουν διαφορετικές πτυχές της συμφωνίας μεταξύ δύο ομαδοποιήσεων. Συγκεκριμένα, ο δείκτης ARI επηρεάζεται από την ύπαρξη ομάδων μικρού μεγέθους, κάτι που δε συμβαίνει με τον δείκτη AMI. Αντίθετα, ο δείκτης AMI είναι επηρεάζεται από τον αριθμό των ομάδων και καταλήγει σε υψηλότερες τιμές από αυτές του ARI, όταν υπάρχουν στα δεδομένα πάνω από 4 ή 5 ομάδες.

4. Αποτελέσματα

Τα αποτελέσματα της εφαρμογής των διαφορετικών μεθόδων ομαδοποίησης, όπως αποτυπώνονται στις τιμές του δείκτη ARI, παρουσιάζονται στον Πίνακα 3.1. Η μέθοδος LCA είχε την καλύτερη επίδοση στα έξι από τα δέκα σύνολα δεδομένων (breast, votes, nursery, tictactoe, lymphography και balloons – σε δύο από αυτά εντόπισε με απόλυτη επιτυχία τις ομάδες), η K-modes σε δύο σύνολα δεδομένων (australian και mushroom), ενώ η Ανιούσα Ιεραρχική Ταξινόμηση σε άλλα δύο (car και sponges, για τους συνδυασμούς Gower-Centroid και Gower-Ward.D2, αντίστοιχα). Ωστόσο, αξίζει να σημειωθεί ότι για το σύνολο δεδομένων car, όλες οι μέθοδοι είχαν πολύ χαμηλή επίδοση.

Με βάση τη μέση επίδοση των μεθόδων σε όλα τα σύνολα δεδομένων, παρατηρείται η η ξεκάθαρη υπεροχή της LCA (mean ARI = 0.539), ενώ ακολουθούν με

μικρή διαφορά μεταξύ τους η K-modes και ο συνδυασμός Gower+Ward.D2 (mean ARI = 0.369 και 0.343, αντίστοιχα). Αντίστοιχες επιδόσεις είχαν και οι συνδυασμοί Manhattan+Ward.D2 (mean ARI = 0.330), Manhattan+Ward.D (0.311) και Gower+Ward.D (0.305). Επομένως, όταν η μέθοδος ομαδοποίησης είναι η Ανιούσα Ιεραρχική Ταξινόμηση, φαίνεται να υπερέχουν τα κριτήρια συνένωσης Ward.D και Ward.D2, ανεξάρτητα από τη μετρική απόστασης. Επιπρόσθετα, αξίζει να σημειωθεί ότι σε αρκετές περιπτώσεις η τιμή του δείκτη ARI είναι αρνητική ή πολύ κοντά στο 0, που σημαίνει ότι οι συγκεκριμένες ομαδοποιήσεις δεν απέχουν πολύ από την τυχαία ομαδοποίηση.

Στο Διάγραμμα 1.1 (Heatmap), το οποίο αντιστοιχεί στις τιμές του Πίνακα 3.1, μπορούμε να διακρίνουμε περισσότερες λεπτομέρειες, όπως για τις διαφορές μεταξύ των συνόλων δεδομένων. Συγκεκριμένα, υπάρχουν σύνολα δεδομένων για τα οποία οι περισσότερες μέθοδοι έχουν σταθερά καλή, μέτρια (votes, sponges) ή χαμηλή απόδοση (cars, australian). Επιπλέον, υπάρχουν σύνολα στα οποία η LCA φαίνεται να είναι η μοναδική μέθοδος που δίνει καλά αποτελέσματα (nursery, tic tac toe, balloons).

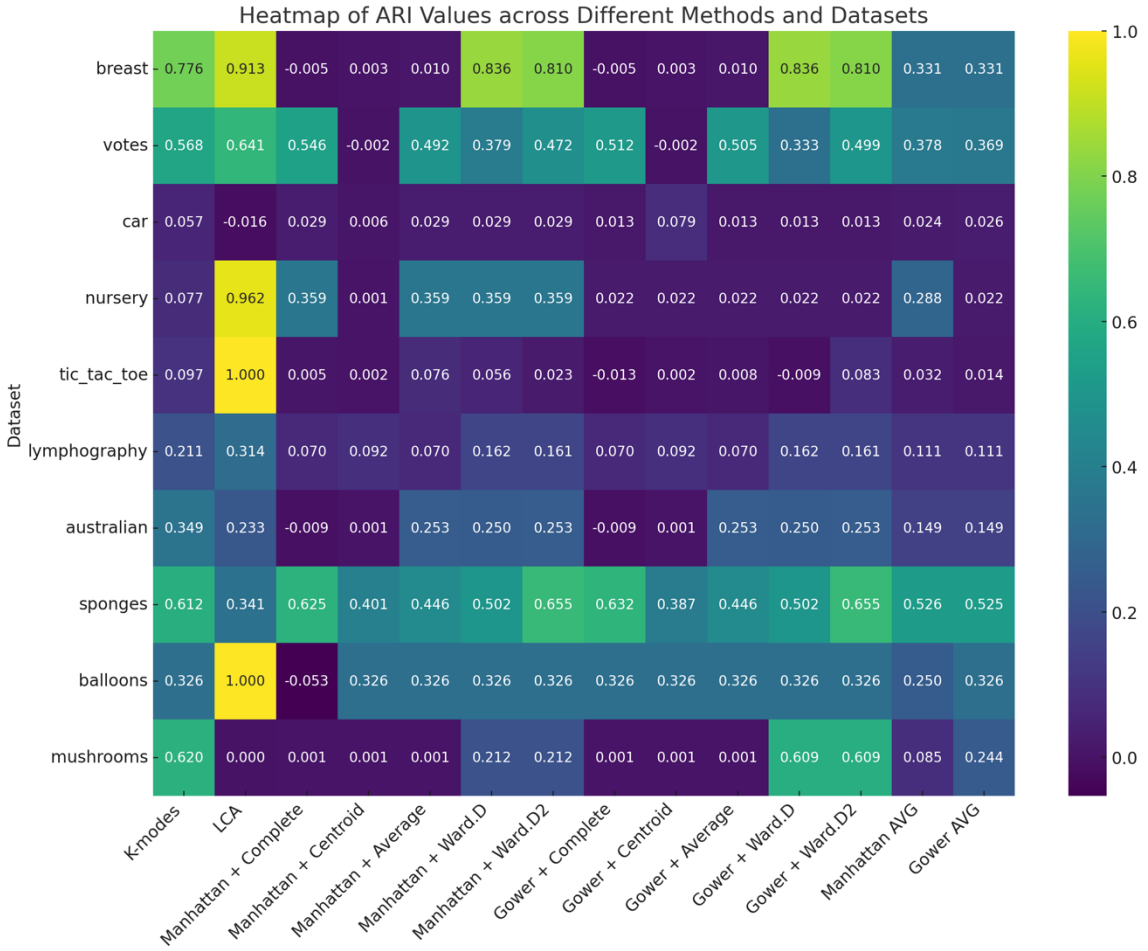
Στον πίνακα 3.1 αποτυπώνονται οι τιμές του δείκτη ARI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων)

Δεδομένα	K-modes	LCA	M+Cmp	M+Cntr	M+Avg	M+Ward.D	M+Ward.D2
breast	0.776	0.913	-0.005	0.003	0.010	0.836	0.810
votes	0.568	0.641	0.546	-0.002	0.492	0.379	0.472
car	0.057	-0.016	0.029	0.006	0.029	0.029	0.029
nursery	0.077	0.962	0.359	0.001	0.359	0.359	0.359
tictactoe	0.097	1.000	0.005	0.002	0.076	0.056	0.023
lymphogr	0.211	0.314	0.070	0.092	0.070	0.162	0.161
australian	0.349	0.233	-0.009	0.001	0.253	0.250	0.253
sponges	0.612	0.341	0.625	0.401	0.446	0.502	0.655
balloons	0.326	1.000	-0.053	0.326	0.326	0.326	0.326
mushroom	0.620	**0.000**	0.001	0.001	0.001	0.212	0.212
Μέσος όρος	**0.369**	**0.539**	**0.157**	**0.083**	**0.206**	**0.311**	**0.330**

Πίνακας 3.1 (μέρος α') Τιμές του δείκτη ARI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων). Σημειώση: M = Manhattan, G = Gower. Avg = Average, Cmp = Complete, Cntr = Centroid

Δεδομένα	G+Cmp	G+Cntr	G+Avg	G+Ward.D	G+Ward.D2	M+Avg	G+Avg
breast	-0.005	0.003	0.010	0.836	0.810	0.331	0.331
votes	0.512	-0.002	0.505	0.333	0.499	0.378	0.369
car	0.013	0.079	0.013	0.013	0.013	0.024	0.026
nursery	0.022	0.022	0.022	0.022	0.022	0.288	0.022
tic_tac_toe	-0.013	0.002	0.008	-0.009	0.083	0.032	0.014
lymphogr	0.070	0.092	0.070	0.162	0.161	0.111	0.111
australian	-0.009	0.001	0.253	0.250	0.253	0.149	0.149
sponges	0.632	0.387	0.446	0.502	0.655	0.526	0.525
balloons	0.326	0.326	0.326	0.326	0.326	0.250	0.326
mushroom	0.001	0.001	0.001	0.609	0.609	0.085	0.244
Μέσος όρος	**0.155**	**0.091**	**0.165**	**0.305**	**0.343**	**0.217**	**0.212**

Πίνακας 3.1 (μέρος β') Τιμές του δείκτη ARI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων). Σημειώση: M = Manhattan, G = Gower. Avg = Average, Cmp = Complete, Cntr = Centroid



Διάγραμμα 1.1 Heatmap με τις τιμές του δείκτη ARI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων)

Στο Διάγραμμα 1.1 (Heatmap) αποτυπώνονται οι τιμές του δείκτη ARI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων) .

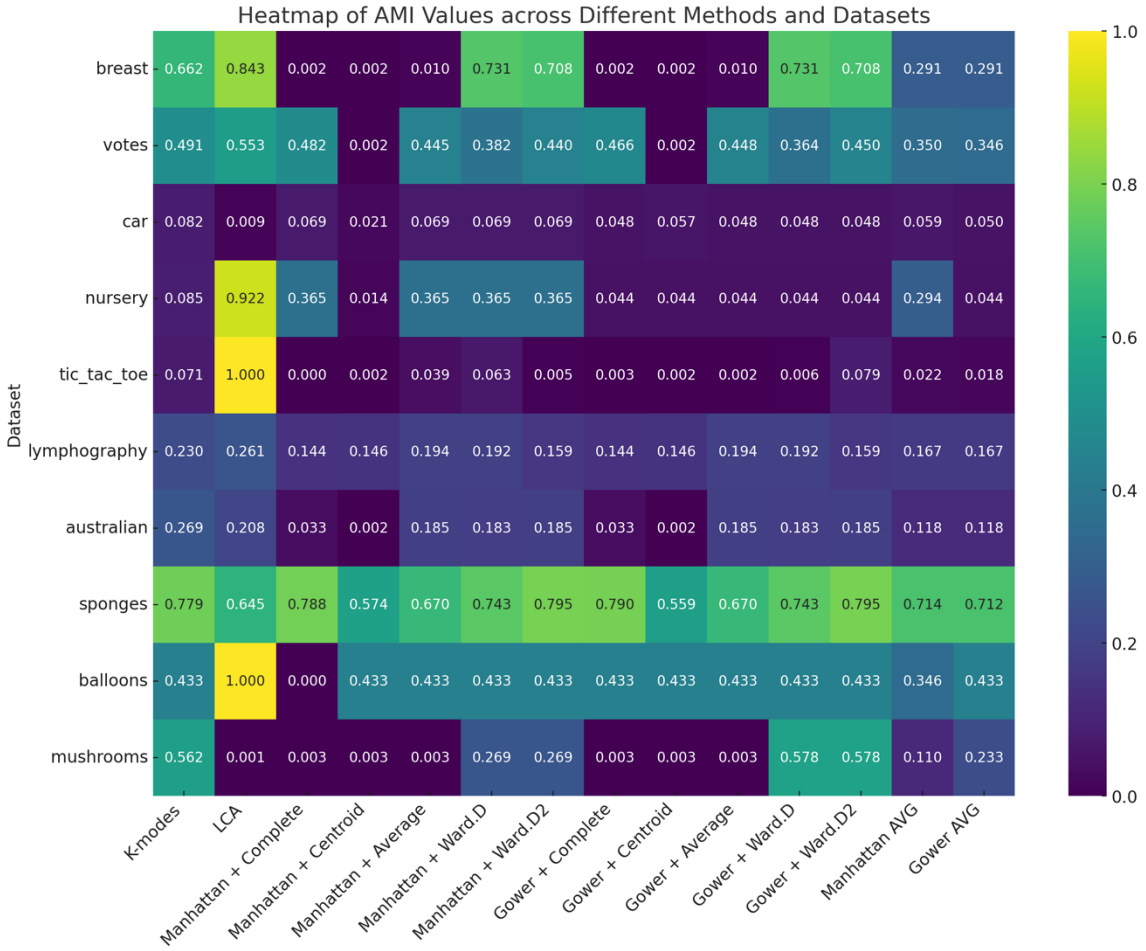
Όπως φαίνεται στον πίνακα 3.2 και το Διάγραμμα 1.2, τα αποτελέσματα με τον δείκτη AMI συμφωνούν σε σημαντικό βαθμό με αυτά στα οποία κατέληξε ο δείκτης ARI. Οι δύο καλύτερες μέθοδοι παραμένουν οι ίδιες, δηλαδή οι LCA και K-modes, ενώ μικρές διαφορές εντοπίζονται στην απόδοση των συνδυασμών της ανιούσας ιεραρχικής ταξινόμησης.

Δεδομένα	K-modes	LCA	M+Cmp	M+Cntr	M+Avg	M+Ward.D	M+Ward.D2
breast	0.662	**0.843**	0.002	0.002	0.010	0.731	0.708
votes	0.491	**0.553**	0.482	0.002	0.445	0.382	0.440
car	0.082	0.009	0.069	0.021	0.069	0.069	0.069
nursery	0.085	**0.922**	0.365	0.014	0.365	0.365	0.365
tictactoe	0.071	**1.000**	0.000	0.002	0.039	0.063	0.005
lymphogr	0.230	**0.261**	0.144	0.146	0.194	0.192	0.159
australian	**0.269**	0.208	0.033	0.002	0.185	0.183	0.185
sponges	**0.779**	0.645	0.788	0.574	0.670	0.743	0.795
balloons	0.433	**1.000**	0.000	0.433	0.433	0.433	0.433
mushroom	0.562	0.001	0.003	0.003	0.003	0.269	0.269
Μέσος όρος	**0.366**	**0.544**	**0.189**	**0.120**	**0.241**	**0.343**	**0.343**

Πίνακας 3.2 (μέρος α') Τιμές του δείκτη AMI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων). Σημειώση: M = Manhattan, G = Gower. Avg = Average, Cmp = Complete, Cntr = Centroid

Δεδομένα	G+Cmp	G+Cntr	G+Avg	G+Ward.D	G+Ward.D2	M+Avg	G+Avg
breast	0.002	0.002	0.010	0.731	0.708	0.291	0.291
votes	0.466	0.002	0.448	0.364	0.450	0.350	0.346
car	0.048	0.057	0.048	0.048	0.048	**0.059**	0.050
nursery	0.044	0.044	0.044	0.044	0.044	0.294	0.044
tic_tac_toe	0.003	0.002	0.002	0.006	0.079	0.022	0.018
lymphogr	0.144	0.146	0.194	0.192	0.159	0.167	0.167
australian	0.033	0.002	0.185	0.183	0.185	0.118	0.118
sponges	0.790	0.559	0.670	0.743	0.795	0.714	0.712
balloons	0.433	0.433	0.433	0.433	0.433	0.346	0.433
mushroom	0.003	0.003	0.003	0.578	**0.578**	0.110	0.233
Μέσος όρος	**0.197**	**0.125**	**0.204**	**0.332**	**0.348**	**0.247**	**0.241**

Πίνακας 3.2 (μέρος β') Τιμές του δείκτη AMI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων). Σημειώση: M = Manhattan, G = Gower. Avg = Average, Cmp = Complete, Cntr = Centroid



Διάγραμμα.1.2 Heatmap με τις τιμές του δείκτη AMI για τη σύγκριση της πραγματικής ομαδοποίησης με την ομαδοποίηση που προέκυψε από τους αλγόριθμους K-modes, LCA και Ανιούσα Ιεραρχική Ταξινόμηση (για 12 διαφορετικούς συνδυασμούς μέτρων απόστασης και κριτηρίων συνένωσης των ομάδων)

5. Συμπεράσματα

Η παρούσα εργασία επικεντρώθηκε στην εξέταση και ανάλυση διαφόρων μεθόδων ανάλυσης κατά συστάδες σε κατηγορικά δεδομένα. Από την ενδελεχή μελέτη των διαφόρων προσεγγίσεων, έγινε σαφές ότι, παρά την πληθώρα των διαθέσιμων μεθόδων, κάθε μία παρουσιάζει τις δικές της ιδιαιτερότητες και περιορισμούς. Η LCA είχε την καλύτερη επίδοση στα περισσότερα από τα σύνολα δεδομένων στα οποία εφαρμόστηκε, ενώ είχε και 100% ακρίβεια σε δύο από αυτά. Η μέθοδος K-modes ήταν η καλύτερη σε δύο από τα σύνολα δεδομένων, ενώ η Ανιούσα Ιεραρχική Ταξινόμηση σε άλλα δύο, με τη χρήση συνδυασμών Gower-Centroid και Gower-Ward.D2. Κατά συνέπεια, η LCA είχε την υψηλότερη μέση απόδοση. Ακολουθούν η K-modes και ο συνδυασμός Gower+Ward.D2 με μικρή διαφορά μεταξύ τους. Άλλοι συνδυασμοί όπως οι Manhattan+Ward.D2, Manhattan+Ward.D και Gower+Ward.D έχουν συγκρίσιμες επιδόσεις. Σε πολλές περιπτώσεις, ο δείκτης ARI ήταν αρνητικός ή πολύ κοντά στο μηδέν, κάτι που δείχνει ότι κάποιες ομαδοποιήσεις δεν ήταν καλύτερες από τυχαίες.

Αν υποθέσουμε ότι τα σύνολα δεδομένων που χρησιμοποιήσαμε στο εμπειρικό μέρος της εργασίας μας αποτελούν τυχαίο δείγμα όλων των πιθανών (άπειρων) πραγματικών συνόλων δεδομένων τα οποία υπόκεινται σε ανάλυση σε συστάδες, τότε φαίνεται οι LCA και K-modes να αποτελούν καλή επιλογή για κατηγορικά δεδομένα. Η Ανιούσα Ιεραρχική Ταξινόμηση με την απόσταση Gower αναμένεται να είναι αποτελεσματική όταν σκοπός είναι να αναδειχθεί ταυτόχρονα και η ιεραρχία των λύσεων της ομαδοποίησης. Ωστόσο, η υπόθεση που κάναμε παραπάνω είναι μάλλον αυστηρή, διότι δεν είμαστε βέβαιοι ότι τα σύνολα δεδομένων του αποθετηρίου UCI είναι αντιπροσωπευτικά αυτών που αναλύονται συνήθως στην πράξη. Ένα άλλο ζήτημα που ενδέχεται να επηρέασε τα αποτελέσματα των υπό σύγκριση μεθόδων είναι το γεγονός ότι, σύμφωνα με τους ερευνητές που τα έκαναν διαθέσιμα στο αποθετήριο, τα δεδομένα είναι κατάλληλα για προβλήματα κατηγοριοποίησης (classification). Στα προβλήματα κατηγοριοποίησης, μία από τις μεταβλητές μας ορίζεται ως εξαρτημένη και οι υπόλοιπες ως ανεξάρτητες. Αυτή η διάκριση, ωστόσο, παύει να υπάρχει στα προβλήματα ανάλυσης σε συστάδες. Σε ένα σύνολο δεδομένων ενδέχεται να υπάρχουν πάνω από μία «σωστές» ή ερμηνεύσιμες λύσεις ομαδοποίησης. Με αυτήν την έννοια, η σύγκριση του αποτελέσματος

των υπό σύγκριση μεθόδων με την μία και μοναδική «πραγματική» λύση, ίσως να αδικεί ορισμένες μεθόδους, οι οποίες ενδέχεται επίσης να καταλήγουν σε διαφορετικές αλλά ερμηνεύσιμες λύσεις.

Συμπερασματικά, η επιλογή της κατάλληλης μεθόδου εξαρτάται σημαντικά από τη φύση των δεδομένων και τον στόχο της ανάλυσης. Επιπρόσθετα, αναδείχθηκε η ανάγκη για περαιτέρω βελτίωση των υφιστάμενων μεθόδων, καθώς και για την ανάπτυξη νέων προσεγγίσεων που θα μπορούσαν να αντιμετωπίσουν τις προκλήσεις των κατηγορικών δεδομένων με ακόμη μεγαλύτερη ακρίβεια και αποτελεσματικότητα.

Με βάση τα ευρήματα και τις παρατηρήσεις αυτής της μελέτης, προκύπτουν σημαντικές προτάσεις για μελλοντική έρευνα στον τομέα της ανάλυσης κατά συστάδες για κατηγορικά δεδομένα. Καταρχάς, υπάρχει ανάγκη για ανάπτυξη νέων μεθόδων που θα ξεπερνούν τους περιορισμούς των υφιστάμενων τεχνικών, ιδιαίτερα σε ό,τι αφορά την απόδοση και την ευελιξία με διαφορετικά σύνολα δεδομένων. Επιπλέον, η έρευνα θα μπορούσε να επικεντρωθεί στην εφαρμογή αυτών των μεθόδων σε πιο πολύπλοκα και μεγάλα σύνολα δεδομένων, καθώς και στην εξέταση της εφαρμογής τους σε διαφορετικά επιστημονικά πεδία. Τέλος, θα ήταν σκόπιμο να διερευνηθεί η συνδυαστική χρήση ποιοτικών και ποσοτικών μεθόδων ανάλυσης, προκειμένου να επιτευχθεί μια πιο ολοκληρωμένη και πολυδιάστατη προσέγγιση στην κατανόηση και την ανάλυση κατηγορικών δεδομένων.

Βιβλιογραφία

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience.

Aha, D. (1991). *The Tic-Tac-Toe Endgame Dataset. Machine Learning Repository*. University of California, Irvine.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732-764.

Guha, S., Rastogi, R., & Shim, K. (1999). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, *25*(5), 345-366.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283-304.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193-218.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, No. 14, pp. 281-297). University of California Press.

Romesburg, H. C. (2004). *Cluster Analysis for Researchers*. Lulu Press.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical Taxonomy*. W.H. Freeman.

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). *Data Analysis and Decision Support*, 335-343, Springer-Verlag, Berlin.

Vinh, N. X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837-2854.

Παράρτημα Α. Κώδικας σε R

```
install.packages("klaR")
install.packages("mclust")
install.packages("poLCA")
install.packages("cluster")
install.packages("fastDummies")
install.packages("aricode")

library(klaR)
library(mclust)
library(poLCA)
library(cluster)
library(fastDummies)
library(aricode)

##### ------breast-cancer-wisconsin------ #####

###---1---###

# Load breast data set
breast <- read.csv("breast-cancer-wisconsin.csv",header=TRUE,)

# Convert all variables to factors
breast <- data.frame(breast)
colnames <- names(breast)
breast[,colnames] <- lapply(breast[,colnames],factor)

#Remove column id
breast <- breast[,-1]

# 1. Apply K-modes
outkmodes <- kmodes(breast[,-10],modes = 2,iter.max = 100)

#Adjusted Rand Index & AMI score
ARIvalue <- adjustedRandIndex(outkmodes$cluster,breast[,10])
AMIvalue <- AMI(outkmodes$cluster,breast[,10])
ARI <- ARIvalue
AMIscore <-AMIvalue

print(ARI)
print(AMIscore)

# 2.  Apply LCA
f <- cbind(clump_thickness,uniformity_cell_size,
```

```
        uniformity_cell_shape,marginal_adhesion,
        single_epi_cell_size,bare_nuclei,
        bland_chroma,normal_nucleo,
        mitoses)~1
M0 <- poLCA(f,breast,nclass=2) # log-likelihood: -543.6498

table(M0$predclass,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,breast[,10])
AMI(M0$predclass,breast[,10])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
man.dist <- daisy(breast[,-10], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])


# 3.2 Manhattan + Centroid
man.dist <- daisy(breast[,-10], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])

# 3.3 Manhattan + Average
man.dist <- daisy(breast[,-10], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
```

```
AMI(memb,breast[,10])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(breast[,-10], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(breast[,-10], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])

# 3.6 Gower + Complete
man.dist <- daisy(breast[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])


# 3.7 Gower + Centroid
man.dist <- daisy(breast[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])
```

```
# 3.8 Gower + Average
man.dist <- daisy(breast[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])

# 3.9 Gower + Ward.D
man.dist <- daisy(breast[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])

# 3.10 Gower + Ward.D2
man.dist <- daisy(breast[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,breast[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,breast[,10])
AMI(memb,breast[,10])

##### ------house_votes_84------ #####

###---2---###

# Load votes data set
votes <- read.csv("house-votes-84.csv",header=TRUE)

# Convert all variables to factors
votes <- data.frame(votes)
colnames <- names(votes)
votes[,colnames] <- lapply(votes[,colnames],factor)
```

```
# 1. Apply K-modes
ARIvalue = NULL
AMIvalue = NULL
for (i in 1:100) {
  outkmodes <- kmodes(votes[,-17],modes = 2,iter.max = 100)

  #Adjusted Rand Index & AMI score
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,votes[,17])
  AMIvalue[i] <- AMI(outkmodes$cluster,votes[,17])
}

ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)

# 2.  LCA

f <- cbind(handicapped_infants,water_project_cost_sharing,

adoption_of_the_budget_resolution,physician_fee_freeze,el_salvador_aid,religious_grou
ps_in_schools,
        anti_satellite_test_ban,aid_to_nicaraguan_contras,mx_missile,immigration,
        synfuels_corporation_cutback,education_spending,superfund_right_to_sue,crime,
        duty_free_exports,export_administration_act_south_africa,Class_Name)~1

M0 <- poLCA(f,votes,nclass=2) # log-likelihood: -543.6498

table(M0$predclass,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,votes[,17])
AMI(M0$predclass,votes[,17])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
votes_dummy <- dummy_cols(votes[,-17])
man.dist <- daisy(data.matrix(votes_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
```

```
AMI(memb,votes[,17])


# 3.2 Manhattan + Centroid
man.dist <- daisy(data.matrix(votes_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

# 3.3 Manhattan + Average
man.dist <- daisy(data.matrix(votes_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(data.matrix(votes_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(data.matrix(votes_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])
```

```
# 3.6 Gower + Complete
man.dist <- daisy(votes[,-17], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

# 3.7 Gower + Centroid
man.dist <- daisy(votes[,-17], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

# 3.8 Gower + Average
man.dist <- daisy(votes[,-17], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

# 3.9 Gower + Ward.D
man.dist <- daisy(votes[,-17], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

# 3.10 Gower + Ward.D2
```

```
man.dist <- daisy(votes[,-17], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,votes[,17])

#Adjusted Rand Index
adjustedRandIndex(memb,votes[,17])
AMI(memb,votes[,17])

##### ------Car-Evaluation------ #####

###---3---###
# Load car data set
car <- read.csv("Car-Evaluation.csv",header=TRUE)

# Convert all variables to factors
car <- data.frame(car)
colnames <- names(car)
car[,colnames] <- lapply(car[,colnames],factor)

# 1. Apply K-modes
ARIvalue = NULL
AMIvalue = NULL
for (i in 1:100) {
  outkmodes <- kmodes(car[,-7],modes = 4,iter.max = 100)

  #Adjusted Rand Index & AMI score
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,car[,7])
  AMIvalue[i] <- AMI(outkmodes$cluster,car[,7])
}

ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)

# 2. Apply LCA
f <- cbind(buying,maint,doors,persons,lug_boot,safety)~1
M0 <- poLCA(f,car,nclass=4) # log-likelihood: -543.6498

table(M0$predclass,car[,7])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,car[,7])
AMI(M0$predclass,car[,7])
```

```
# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
car_dummy <- dummy_cols(car[,-7])
man.dist <- daisy(data.matrix(car_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])

# 3.2 Manhattan + Centroid
man.dist <- daisy(data.matrix(car_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])

# 3.3 Manhattan + Average
man.dist <- daisy(data.matrix(car_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(data.matrix(car_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
```

```
AMI(memb,car[,7])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(data.matrix(car_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])

# 3.6 Gower + Complete
man.dist <- daisy(car[,-7], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])

# 3.7 Gower + Centroid
man.dist <- daisy(car[,-7], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])

# 3.8 Gower + Average
man.dist <- daisy(car[,-7], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])
```

```
# 3.9 Gower + Ward.D
man.dist <- daisy(car[,-7], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])

# 3.10 Gower + Ward.D2
man.dist <- daisy(car[,-7], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 4)

table(memb,car[,7])

#Adjusted Rand Index
adjustedRandIndex(memb,car[,7])
AMI(memb,car[,7])



##### ------Nursery------ #####

###---4---###

# Load nursery data set
nursery <- read.csv("nursery.csv",header=TRUE)

# Convert all variables to factos
nursery <- data.frame(nursery)
colnames <- names(nursery)
nursery[,colnames] <- lapply(nursery[,colnames],factor)

# 1. Apply K-modes
ARIvalue = NULL
AMIvalue = NULL
for (i in 1:100) {
  outkmodes <- kmodes(nursery[,-9],modes = 5,iter.max = 100)

  #Adjusted Rand Index & AMI score
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,nursery[,9])
  AMIvalue[i] <- AMI(outkmodes$cluster,nursery[,9])
}
```

```
ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)


# 2. Apply LCA

f <- cbind(parents,has_nurs,form,children,housing,finance,social,health,class)~1
M0 <- poLCA(f,nursery,nclass=5)# log-likelihood: -543.6498

table(M0$predclass,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,nursery[,9])
AMI(M0$predclass,nursery[,9])


# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
nursery_dummy <- dummy_cols(nursery[,-9])
man.dist <- daisy(data.matrix(nursery_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.2 Manhattan + Centroid
man.dist <- daisy(data.matrix(nursery_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.3 Manhattan + Average
man.dist <- daisy(data.matrix(nursery_dummy), metric = c("manhattan"))
```

```
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(data.matrix(nursery_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(data.matrix(nursery_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.6 Gower + Complete
man.dist <- daisy(nursery[,-9], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])


# 3.7 Gower + Centroid
man.dist <- daisy(nursery[,-9], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
```

```
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.8 Gower + Average
man.dist <- daisy(nursery[,-9], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.9 Gower + Ward.D
man.dist <- daisy(nursery[,-9], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])

# 3.10 Gower + Ward.D2
man.dist <- daisy(nursery[,-9], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 5)

table(memb,nursery[,9])

#Adjusted Rand Index
adjustedRandIndex(memb,nursery[,9])
AMI(memb,nursery[,9])



##### ------tic_tac_toe------ #####
```

```
###---5---###
# Load tic_tac_toe data set
titato <- read.csv("tic_tac_toe.csv",header=TRUE)

# Convert all variables to factors
titato <- data.frame(titato)
colnames <- names(titato)
titato[,colnames] <- lapply(titato[,colnames],factor)

# 1. Apply K-modes
ARIvalue = NULL
AMIvalue <- NULL
for (i in 1:100) {
  outkmodes <- kmodes(titato[,-10],modes = 2,iter.max = 100)

  #Adjusted Rand Index & AMI score
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,titato[,10])
  AMIvalue[i] <- AMI(outkmodes$cluster,titato[,10])
}

ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)


# 2.  Apply LCA

f <- cbind(top_left_square,top_mid_square,top_right_square,
          middle_left_square,middle_middle_square,middle_right_square,
          bottom_left_square,bottom_middle_square,bottom_right_square,
          Class)~1

M0 <- poLCA(f,titato,nclass=2) # log-likelihood: -543.6498

table(M0$predclass,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,titato[,10])
AMI(M0$predclass,titato[,10])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete

titato_dummy <- dummy_cols(titato[,-10])
```

```
man.dist <- daisy(data.matrix(titato_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])


# 3.2 Manhattan + Centroid
man.dist <- daisy(data.matrix(titato_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

# 3.3 Manhattan + Average
man.dist <- daisy(data.matrix(titato_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(data.matrix(titato_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(data.matrix(titato_dummy), metric = c("manhattan"))
```

```
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

# 3.6 Gower + Complete
man.dist <- daisy(titato[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])


# 3.7 Gower + Centroid
man.dist <- daisy(titato[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

# 3.8 Gower + Average
man.dist <- daisy(titato[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

# 3.9 Gower + Ward.D
man.dist <- daisy(titato[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
```

```
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

# 3.10 Gower + Ward.D2
man.dist <- daisy(titato[,-10], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,titato[,10])

#Adjusted Rand Index
adjustedRandIndex(memb,titato[,10])
AMI(memb,titato[,10])

##### ------lymphography------ #####

###---6---###



# Load lymphography data set
lympho <- read.csv("lymphography.csv",header=TRUE)


# Convert all variables to factor
lympho <- data.frame(lympho)
colnames <- names(lympho)
lympho[,colnames] <- lapply(lympho[,colnames],factor)

# 1. Apply K-modes
ARIvalue = NULL
AMIvalue = NULL
for (i in 1:100) {
  outkmodes <- outkmodes <- kmodes(lympho[,-19],modes = 4,iter.max = 100)

  #Adjusted Rand Index & AMI score
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,lympho[,19])
  AMIvalue[i] <- AMI(outkmodes$cluster,lympho[,19])
}

ARI <- max(ARIvalue)
```

```
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)

# 2.  Apply LCA

f <- cbind(lymphatics,block._of_affere,bl_of_lymph_c,bl_of_lymph_s,by_pass,
        extravasates,regeneration_of,early_uptake_in,lym_nodes_dimin,
        lym_nodes_enlar,changes_in_lym,defect_in_node,changes_in_node,
        changes_in_stru,special_forms,dislocation_of,exclusion_of_no,
        no_of_nodes_in,class)~1

M0 <- poLCA(f,lympho,nclass=4) # log-likelihood: -543.6498

table(M0$predclass,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,lympho[,19])
AMI(M0$predclass,lympho[,19])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
man.dist <- daisy(lympho[,-19], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])


# 3.2 Manhattan + Centroid
man.dist <- daisy(lympho[,-19], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])

# 3.3 Manhattan + Average
```

```
man.dist <- daisy(lympho[,-19], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(lympho[,-19], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(lympho[,-19], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])

# 3.6 Gower + Complete
man.dist <- daisy(lympho[,-19], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])


# 3.7 Gower + Centroid
man.dist <- daisy(lympho[,-19], metric = c("gower"))
```

```
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])

# 3.8 Gower + Average
man.dist <- daisy(lympho[,-19], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])

# 3.9 Gower + Ward.D
man.dist <- daisy(lympho[,-19], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])

# 3.10 Gower + Ward.D2
man.dist <- daisy(lympho[,-19], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 4)

table(memb,lympho[,19])

#Adjusted Rand Index
adjustedRandIndex(memb,lympho[,19])
AMI(memb,lympho[,19])



##### ------australian------ #####
```

```
###---7---###



# Load australia data set
aus <- read.csv("australian.csv",header=TRUE,)

# Convert all variables to factor
aus <- data.frame(aus)
colnames <- names(aus)
aus[,colnames] <- lapply(aus[,colnames],factor)


# 1. Apply K-modes
ARIvalue = NULL
AMIvalue <- NULL
for (i in 1:100) {
  outkmodes <- outkmodes <- kmodes(aus[,-15],modes = 2,iter.max = 100)

  #Adjusted Rand Index & AMI score
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,aus[,15])
  AMIvalue[i] <- AMI(outkmodes$cluster,aus[,15])
}

ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)

# 2.  Apply LCA

f <- cbind(A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12,A13,A14,A15)~1
M0 <- poLCA(f,aus,nclass=2) # log-likelihood: -543.6498

table(M0$predclass,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,aus[,15])
AMI(M0$predclass,aus[,15])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
man.dist <- daisy(aus[,-15], metric = c("manhattan"))
```

```
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])


# 3.2 Manhattan + Centroid
man.dist <- daisy(aus[,-15], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])

# 3.3 Manhattan + Average
man.dist <- daisy(aus[,-15], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(aus[,-15], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(aus[,-15], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
```

```
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])

# 3.6 Gower + Complete
man.dist <- daisy(aus[,-15], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])


# 3.7 Gower + Centroid
man.dist <- daisy(aus[,-15], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])

# 3.8 Gower + Average
man.dist <- daisy(aus[,-15], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])

# 3.9 Gower + Ward.D
man.dist <- daisy(aus[,-15], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)
```

```
table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])

# 3.10 Gower + Ward.D2
man.dist <- daisy(aus[,-15], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,aus[,15])

#Adjusted Rand Index
adjustedRandIndex(memb,aus[,15])
AMI(memb,aus[,15])



##### ------sponges------ #####

###---8---###




# Load sponges data set
sponges <- read.csv("sponges.csv",header=TRUE)

sponges<- sponges[,-47]

# Convert all variables to factors
sponges <- data.frame(sponges)
colnames <- names(sponges)
sponges[,colnames] <- lapply(sponges[,colnames],factor)

# 1. Apply K-modes
ARIvalue = NULL
AMIvalue = NULL
for (i in 1:100) {
  outkmodes <- kmodes(sponges[,-1],modes = 12,iter.max = 100)

  #Adjusted Rand Index & AMI score
```

```
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,sponges[,1])
  AMIvalue[i] <- AMI(outkmodes$cluster,sponges[,1])
}

ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)

# 2.Apply LCA

f <-
cbind(A_CAPAS_DEL_CORTEX,A_CAPA_INTERNA_DEL_CORTEX,A_CORTEX,
A_CORTEX_FIBROSO,
A_CORTEX_SOLO_DE_ESPICULAS_TANGENCIALES,A_CUERPOS_EXTRANOS
_EN_EL_CORTEX,
A_GROSOR_DEL_CORTEX,A_HACES_DE_ESPICULAS_PRINCIPALES_EN_POM
PON_EN_EL_CORTEX,
A_TILOSTILOS_ADICIONALES_COANOSOMA,B.NUMERO_DE_TIPOS_DE_ME
GASCLERAS,
C_TIPO_ESPICULA_PRINCIPAL_DIACTINA_TUBERCULADA,C_TIPO_ESPICUL
A_PRINCIPAL_ESTILO,
C_TIPO_ESPICULA_PRINCIPAL_ESTILOS_2_TAMANOS,C_TIPO_ESPICULA_PR
INCIPAL_ESTILO_TILOSTILO,
C_TIPO_ESPICULA_PRINCIPAL_ESTRONGILOXA,C_TIPO_ESPICULA_PRINCIP
AL_OXAS,
C_TIPO_ESPICULA_PRINCIPAL_TILOSTILO,D_ESPICULA_PRINCIPAL_ESTILO
,
D_ESPICULA_PRINCIPAL_TILOSTILO,D_FORMA_BASE_TILOSTILO_PRINCIP
AL,
E_DISPOSICION_MEGASCLERAS_ECTOSOMICAS_EN_EL_ECTOSOMA,E_FOR
MA_BASE_TILOSTILO_ECTOSOMICO,
E_FORMA_MEGASCLERA_ECTOSOMICA,E_TIPO_MEGASCLERA_ECTOSOMI
CA,F_TIPO_DE_EXOSTILO,
G_FORMA_MEGASCLERA_INTERMEDIARIA,G_TIPO_MEGASCLERA_INTERM
EDIARIA,
H_LONGITUD_MEGASCLERAS,I_MICROSCLERAS,I_TIPO_MICROSCLERA,J_A
STER,
J_DIAMETRO_ESFERASTER,J_TIPO_DE_ASTER,J_TIPO_DE_DIPLASTER,J_TIP
O_DE_ESFERASTER,
K_FORMA_FINAL,L_NUMERO_DE_PAPILAS,L_PAPILAS,M_COLOR,N_SUPERF
ICIE,
O_DISPOSICION_ESPICULAR_ESQUELETO,P_ALOJA_CANGREJO_ERMITANO,
P_PERFORANTE,
        P_PSEUDORAICES,P_SUSTRATO)~1
```

```
M0 <- poLCA(f,sponges,nclass=12)# log-likelihood: -543.6498
table(M0$predclass,sponges[,1])
#Adjusted Rand Index
adjustedRandIndex(M0$predclass,sponges[,1])
AMI(M0$predclass,sponges[,1])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
sponges_dummy <- dummy_cols(sponges[,-1])
man.dist <- daisy(data.matrix(sponges_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 12)

table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.2 Manhattan + Centroid
man.dist <- daisy(data.matrix(sponges_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 12)

table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.3 Manhattan + Average
man.dist <- daisy(sponges[,-1], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 12)

table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(sponges[,-1], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 12)
```

```
table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(sponges[,-1], metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 12)

table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.6 Gower + Complete
man.dist <- daisy(sponges[,-1], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 12)

table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])


# 3.7 Gower + Centroid
man.dist <- daisy(sponges[,-1], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 12)

table(memb,nursery[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.8 Gower + Average
man.dist <- daisy(sponges[,-1], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 12)
```

```
table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.9 Gower + Ward.D
man.dist <- daisy(sponges[,-1], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 12)

table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

# 3.10 Gower + Ward.D2
man.dist <- daisy(sponges[,-1], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 12)

table(memb,sponges[,1])

#Adjusted Rand Index
adjustedRandIndex(memb,sponges[,1])
AMI(memb,sponges[,1])

##### ------ballons------ #####

###---9---###
# Load balloons data set
adult <- read.csv("adult+stretch - adult+stretch.csv",header=TRUE)

# Convert all variables to factors
adult <- data.frame(adult)
colnames <- names(adult)
adult[,colnames] <- lapply(adult[,colnames],factor)

# 1. Apply K-modes
ARIvalue = NULL
AMIvalue = NULL
for (i in 1:100) {
  outkmodes <- kmodes(adult[,-5],modes = 2,iter.max = 100)

  #Adjusted Rand Index & AMI score
```

```
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,adult[,5])
  AMIvalue[i] <- AMI(outkmodes$cluster,adult[,5])
}

ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)

# 2. Εφαρμογή της LCA

f <- cbind(COLOR,SIZE,ACT,AGE,INFLATED)~1
M0 <- poLCA(f,adult,nclass=2)# log-likelihood: -543.6498

table(M0$predclass,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,adult[,5])
AMI(M0$predclass,adult[,5])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
adult_dummy <- dummy_cols(adult[,-5])
man.dist <- daisy(data.matrix(adult_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])


# 3.2 Manhattan + Centroid
man.dist <- daisy(data.matrix(adult_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])
```

```
# 3.3 Manhattan + Average
man.dist <- daisy(data.matrix(adult_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(data.matrix(adult_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(data.matrix(adult_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

# 3.6 Gower + Complete
man.dist <- daisy(adult[,-5], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

# 3.7 Gower + Centroid
man.dist <- daisy(adult[,-5], metric = c("gower"))
```

```
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

# 3.8 Gower + Average
man.dist <- daisy(adult[,-5], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

# 3.9 Gower + Ward.D
man.dist <- daisy(adult[,-5], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

# 3.10 Gower + Ward.D2
man.dist <- daisy(adult[,-5], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,adult[,5])

#Adjusted Rand Index
adjustedRandIndex(memb,adult[,5])
AMI(memb,adult[,5])

##### ------mushrooms------ #####

###---10---###
```

```
# Load mushrooms data set
mushrooms <- read.csv("agaricus-lepiota - agaricus-lepiota.csv",header=TRUE)

# Convert all variables to factors
mushrooms <- data.frame(mushrooms)
colnames <- names(mushrooms)
mushrooms[,colnames] <- lapply(mushrooms[,colnames],factor)

# 1. Apply K-modes
ARIvalue = NULL
AMIvalue = NULL
for (i in 1:100) {
  outkmodes <- kmodes(mushrooms[,-23],modes = 2,iter.max = 100)

  #Adjusted Rand Index & AMI score
  ARIvalue[i] <- adjustedRandIndex(outkmodes$cluster,mushrooms[,23])
  AMIvalue[i] <- AMI(outkmodes$cluster,mushrooms[,23])
}

ARI <- max(ARIvalue)
AMIscore <- max(AMIvalue)
print(ARI)
print(AMIscore)

# 2.Apply LCA

f <- cbind(cap_shape,cap_surface,cap_color,bruises,odor,gill_attachment,
        gill_spacing,gill_size,gill_color,stalk_shape,stalk_root,stalk_surface_above_ring,
        stalk_surface_below_ring,stalk_color_above_ring,stalk_color_below_ring,
        vevil_type,vevil_color,ring_number,ring_type,spore_print_color,population,
        habitat,class)~1
M0 <- poLCA(f,mushrooms,nclass=2)# log-likelihood: -543.6498

table(M0$predclass,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(M0$predclass,mushrooms[,23])
AMI(M0$predclass,mushrooms[,23])

# 3. Εφαρμογή Ιεραρχικής Ταξινόμησης

# 3.1 Manhattan + Complete
mushrooms_dummy <- dummy_cols(mushrooms[,-23])
man.dist <- daisy(data.matrix(mushrooms_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)
```

```
table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.2 Manhattan + Centroid
man.dist <- daisy(data.matrix(mushrooms_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.3 Manhattan + Average
man.dist <- daisy(data.matrix(mushrooms_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.4 Manhattan + Ward.D
man.dist <- daisy(data.matrix(mushrooms_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.5 Manhattan + Ward.D2
man.dist <- daisy(data.matrix(mushrooms_dummy), metric = c("manhattan"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])
```

```
#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.6 Gower + Complete
man.dist <- daisy(mushrooms[,-23], metric = c("gower"))
mytree <- hclust(man.dist,method = "complete")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.7 Gower + Centroid
man.dist <- daisy(mushrooms[,-23], metric = c("gower"))
mytree <- hclust(man.dist,method = "centroid")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.8 Gower + Average
man.dist <- daisy(mushrooms[,-23], metric = c("gower"))
mytree <- hclust(man.dist,method = "average")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.9 Gower + Ward.D
man.dist <- daisy(mushrooms[,-23], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
```

```
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])

# 3.10 Gower + Ward.D2
man.dist <- daisy(mushrooms[,-23], metric = c("gower"))
mytree <- hclust(man.dist,method = "ward.D2")
memb <- cutree(mytree, k = 2)

table(memb,mushrooms[,23])

#Adjusted Rand Index
adjustedRandIndex(memb,mushrooms[,23])
AMI(memb,mushrooms[,23])
```