Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Τεχνικές ομαδοποίησης μικτών δεδομένων

Κωνσταντίνα Π. Αντωνοπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Σεπτέμβριος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Καθηγητής Τήνιος Πλάτων
- Αναπληρωτής Καθηγητής Πελέκης Νικόλαος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics
Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Mixed data clustering techniques

By

Konstantina P. Antonopoulou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2023

ὡς ἐν ἔργῳ γε παντὶ τὸ μὲν καλὸν ἐκ πολλῶν οἷον ἀριθμῶν εἰς ἕνα καιρὸν ἡκόντων ὑπὸ συμμετρίας τινὸς καὶ ἁρμονίας ἐπιτελεῖται, τὸ δ΄ αἰσχρὸν ἐξ ἑνὸς τοῦ τυχόντος ἐλλείποντος ἢ προσόντος ἀτόπως εὐθὺς ἑτοίμην ἔχει τὴν γένεσιν…

(Ηθικά, Πλούταρχος)

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλαν στην εκπόνησή της.

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της διπλωματικής εργασίας μου, κ. Κούτρα Μάρκο, ο οποίος είναι ένας εξαίρετος επιστήμονας, για τη γνώση του και τις συμβουλές του, κατά την διάρκεια των σπουδών μου τόσο στο Μεταπτυχιακό Πρόγραμμα, όσο και στο προπτυχιακό, καθώς και την περίοδο της συγγραφής της παρούσας διπλωματικής εργασίας.

Παράλληλα, θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής κ. Τήνιο Πλάτωνα και κ. Πελέκη Νικόλαο για την συμμετοχή τους στην επιτροπή αξιολόγησης της διπλωματικής εργασίας μου.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στην οικογένεια και τους φίλους μου για την ανεκτίμητη στήριξη τους, την πίστη τους στις δυνατότητες μου και την ενθάρρυνση τους τόσο στις δύσκολες στιγμές όσο και στις καλές.

Περίληψη

Στη Στατιστική, οι τεχνικές ομαδοποίησης (*Cluster Analysis*) αποσκοπούν στην δημιουργία ομοιογενών ομάδων, ενώ παράλληλα κάθε ομάδα (ή συστάδα) να είναι όσο το δυνατόν πιο διαφορετική από μία άλλη. Κατά συνέπεια, εξετάζοντας ξεχωριστά κάθε ομάδα μπορούμε να επιτύχουμε πιο εύκολη και αποδοτικότερη επεξεργασία για τα δεδομένα που διαθέτουμε. Σε πολλούς χώρους όπως η υγεία, τα χρηματοοικονομικά, το μάρκετινγκ κ.ά, εμφανίζεται η ανάγκη μελέτης πολυδιάστατων μικτών δεδομένων, δηλαδή δεδομένων που περιλαμβάνουν τόσο αριθμητικά όσο και κατηγορικά χαρακτηριστικά. Συνήθως, η εφαρμογή τεχνικών ομαδοποίησης σε μικτά σύνολα δεδομένα, γίνεται με σκοπό την εύρεση δομών και ομαδοποίηση παρόμοιων αντικειμένων για περαιτέρω ανάλυση.

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι να παρουσιαστούν και να εφαρμοστούν τεχνικές ανάλυσης μικτών δεδομένων σε προσομοιωμένα δεδομένα και παράλληλα να παρουσιαστούν αναλύσεις σε πραγματικά δεδομένα. Η αξιολόγηση των τεχνικών αυτών θα γίνει με βάση τα αποτελέσματα των τεχνικών που θα εφαρμοστούν στα προσομοιωμένα δεδομένα, αλλά και μέσω βιβλιογραφικής ανασκόπησης των συγκεκριμένων τεχνικών σε πραγματικά δεδομένα.

Abstract

In Statistics, clustering analysis techniques aim to create homogeneous groups, while at the same time ensuring that groups are as different as possible from each other. Consequently, by examining each group separately we can achieve easier and more efficient processing for the available data. Mixed data sets, i.e., data that include both numerical and categorical characteristics, arise in various fields such as healthcare, finance, marketing, and others. Typically, the application of cluster Analysis techniques to mixed data sets is exploited in order to identify structures and group similar objects for further analysis.

The aim of this thesis is to present and apply mixed data clustering techniques to simulated data and perform statistical analyses on real data. An evaluation of these techniques will also be performed using the results obtained from the applied techniques on simulated data, as well as through a bibliographic review of these techniques on real data.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Μία από τις πρώτες ενέργειες που πρέπει να κάνει ένας ερευνητής για οποιαδήποτε ανάλυση δεδομένων που θέλει να διενεργήσει, είναι να συγκεντρώσει τα απαραίτητα δεδομένα που θα χρειαστεί. Η πλειονότητα των δεδομένων, ανεξαρτήτως του πεδίου που γίνεται η ανάλυση, είναι πολυμεταβλητά, δηλαδή δεδομένα που περιέχουν δύο ή περισσότερες μεταβλητές. Σε μερικές περιπτώσεις, ο ερευνητής κρίνει πως θα πρέπει να μελετήσει κάθε μεταβλητή ξεχωριστά. Όμως, τις περισσότερες φορές η ανάλυση γίνεται για όλες τις μεταβλητές ταυτόχρονα έτσι ώστε να υπάρξει κατανόηση της δομής των δεδομένων και της εξάρτησης των μεταβλητών. Για τον λόγο αυτό, η χρήση μίας ή περισσότερων μεθόδων της πολυμεταβλητής ανάλυσης είναι ωφέλιμη, ίσως και αναγκαία.

Η πολυμεταβλητή ανάλυση αναφέρεται σε μεθόδους όπου ο ερευνητής προσπαθεί να κατανοήσει, περιγράψει και εξερευνήσει πολυδιάστατα δεδομένα. Μία από τις κύριες εφαρμογές της πολυμεταβλητής ανάλυσης είναι η ανάλυση κατά συστάδες ή ανάλυση σε ομάδες (*Cluster Analysis*). Η ανάλυση σε συστάδες είναι μία στατιστική μέθοδος που έχει σαν στόχο να κατατάξει σε ομάδες τις παρατηρήσεις ενός συνόλου δεδομένων κάνοντας χρήση της πληροφορίας που υπάρχει στις διαθέσιμες μεταβλητές.

Η ανάλυση κατά συστάδες ανήκει σε μία ευρύτερη ομάδα τεχνικών, αυτή της μη επιβλεπόμενης μάθησης (*unsupervised learning*). Στη μη επιβλεπόμενη μάθηση, γίνεται ανάλυση στα δεδομένα χωρίς να υπάρχει εκ των προτέρων γνώση κάποιου είδους ταμπέλας/ετικέτας που να τα διαχωρίζει, δηλαδή δεν υπάρχει κάποια μεταβλητή στόχος, όπως γίνεται στην επιβλεπόμενη μάθηση (Hennig και Liao (2013)). Ο κύριος σκοπός της μη επιβλεπόμενης μάθησης είναι η εύρεση μοτίβων/ομοιοτήτων ανάμεσα στα δεδομένα οι οποίες τα ταξινομούν σε ομάδες.

Η εύρεση ομοιοτήτων ανάμεσα στα δεδομένα εξαρτάται τόσο από τα ίδια τα δεδομένα που έχουν συλλεχθεί, όσο και από το πρόβλημα που καλείται να αναλύσει ο ερευνητής. Για την εύρεση ομοιοτήτων στα δεδομένα, στην βιβλιογραφία έχουν προταθεί αρκετά μέτρα απόστασης και ομοιότητας. Ένα μέτρο απόστασης παίρνει μικρές τιμές όταν τα δεδομένα μοιάζουν μεταξύ τους και μεγάλες τιμές όταν είναι ανόμοια. Αντίθετα, στα μέτρα ομοιότητας αναμένουμε μεγάλες τιμές όταν οι παρατηρήσεις μοιάζουν μεταξύ τους και μικρές τιμές όταν είναι ανόμοιες. Συνεπώς, κάνοντας χρήση κάποιου μέτρου απόστασης ή ομοιότητας, μπορεί να γίνει η ταξινόμηση των παρατηρήσεων σε ομάδες.

Για το πρόβλημα της ομαδοποίησης, έχουν προταθεί πολλές προσεγγίσεις. Αρχικά, υπάρχουν οι εμπειρικοί τρόποι, που εφαρμόζονται με την χρήση ποικίλων γραφικών παραστάσεων των πολυδιάστατων παρατηρήσεων. Ακολούθως, έχουν αναπτυχθεί τεχνικές όπου έχουν μαθηματική ή στατιστική υπόσταση. Οι πιο διαδεδομένες χωρίζονται σε ιεραρχικές και μη ιεραρχικές. Οι ιεραρχικές μέθοδοι ξεκινάνε με κάθε παρατήρηση να είναι μία ομάδα και καταλήγουν σε μία ομάδα που περιέχει όλες τις παρατηρήσεις ή αντίστροφα. Στις μη ιεραρχικές μεθόδους, υπάρχει η εκ των προτέρων γνώση των αριθμών των ομάδων που θα δημιουργηθούν και γίνεται χρήση κάποιου επαναληπτικού αλγόριθμου που τοποθετεί τις παρατηρήσεις σε ομάδες. Επιπροσθέτως, υπάρχουν και κάποιες λιγότερο δημοφιλείς τεχνικές, όπως είναι η ομαδοποίηση με βάση την πυκνότητα, η Gaussian ομαδοποίηση κ.ά.

Ως επί των πλείστο, πριν την ανάλυση κατά συστάδες εφαρμόζεται η ανάλυση κύριων συνιστωσών (*principal components analysis*). Η συγκεκριμένη τεχνική αποσκοπεί στην μείωση διαστάσεων του συνόλου δεδομένων, ειδικά όταν υπάρχουν υπερβολικά πολλές μεταβλητές. Η βασική ιδέα είναι η δημιουργία ενός καινούριου πλήθους ασυσχέτιστων μεταβλητών, οι οποίες είναι γραμμικός συνδυασμός των αρχικών, και περιέχουν όσο το δυνατόν περισσότερη πληροφορία από τις αρχικές.

Όλα τα παραπάνω μπορούν πλέον να εφαρμοσθούν σήμερα με μεγάλη ευκολία σε αριθμητικά δεδομένα. Όμως, η άνθηση της τεχνολογίας δημιούργησε την ανάγκη ανάπτυξης περισσότερων τεχνικών που μπορούν να εφαρμοσθούν σε μικτά δεδομένα. Τα μικτά δεδομένα απαρτίζονται από άτομα/αντικείμενα που περιλαμβάνουν πληροφορία, τόσο σε αριθμητικές μεταβλητές όσο και σε κατηγορικές. Στην εποχή μας, τέτοιου είδους σύνολα δεδομένων είναι πιο συνηθισμένα από τα αριθμητικά σύνολα δεδομένων, σχεδόν σε όλους τους τομείς επιστημών που μπορεί να χρησιμοποιηθεί η ανάλυση κατά συστάδες. Στην παρούσα διπλωματική εργασία θα ασχοληθούμε με την μελέτη αλγορίθμων ομαδοποίησης που έχουν εφαρμογή σε μικτά δεδομένα.

Πιο αναλυτικά, στο επόμενο κεφάλαιο περιγράφονται τα είδη των δεδομένων, η προπαρασκευή αυτών και παρουσιάζονται κάποια dataset μικτών δεδομένων. Στο 3^{ο} κεφάλαιο γίνεται μία ανασκόπηση της ανάλυσης κατά συστάδες, η εφαρμογή της σε αριθμητικά δεδομένα και η παρουσίαση των μεθόδων που αφορούν μικτά δεδομένα. Το 4^{ο} κεφάλαιο ακολουθεί με την εφαρμογή και σύγκριση των μεθόδων σε προσομοιωμένα δεδομένα, ενώ στο επόμενο κεφάλαιο παρουσιάζονται οι εφαρμογές των μεθόδων σε πραγματικά δεδομένα, μέσω βιβλιογραφικής ανασκόπησης. Στο τελευταίο κεφάλαιο αναλύονται και ερμηνεύονται τα αποτελέσματα και γίνονται προτάσεις για επέκταση του υλικού που παρουσιάστηκε στην παρούσα διπλωματική.

ΚΕΦΑΛΑΙΟ 2

Μικτά δεδομένα

2.1 Κατηγορίες Δεδομένων

Η συγκέντρωση, επεξεργασία και αποθήκευση απεριόριστων δεδομένων με σκοπό την ανακάλυψη τάσεων, προτύπων και απαντήσεων σε ερωτήματα της καθημερινότητας καθιστά την εξοικείωση μαζί τους αναγκαία. Η μετατροπή των δεδομένων σε πληροφορία και της πληροφορίας σε γνώση προς αξιοποίηση, προϋποθέτει την ορθή χρήση τεχνικών μηχανικής μάθησης (περιοχή της τεχνητής νοημοσύνης, η οποία διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν μοτίβα από τα ίδια τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά), αντίστοιχες με το είδος των δεδομένων που έχουν συλλεχθεί.

Σύμφωνα με τους Everitt και Dunn (1991),οι βασικές κατηγορίες δεδομένων που συλλέγονται προς επεξεργασία είναι οι εξής:

- Κατηγορικά δεδομένα (*categorical/nominal*): Τα δεδομένα αυτά προέρχονται από μεταβλητές, οι τιμές των οποίων εκφράζουν διακριτές κατηγορίες ή τάξεις. Εξ ονόματος λοιπόν, στη συγκεκριμένη κατηγορία δε θα μπορούσε να υπάρχει κάποια διάταξη κατά σειρά μεγέθους, συνεπώς αποκλείονται τα αριθμητικά στοιχεία. Η αριθμητική αντιστοίχιση μπορεί να προκύψει με συμβολικό χαρακτήρα για την καλύτερη επεξεργασία τους στον υπολογιστή, χωρίς να έχει νόημα η εκτέλεση αριθμητικών πράξεων με αυτά. Κάποια ενδεικτικά αντιπροσωπευτικά δείγματα κατηγορικών μεταβλητών αποτελούν το φύλο, το χρώμα ματιών, η ομάδα αίματος, το θρήσκευμα, η οικογενειακή κατάσταση, το αν ένα άτομο πάσχει ή όχι από κάποια ασθένεια κ.ά.

- Διατάξιμα δεδομένα (*ordinal*): Τα διατάξιμα δεδομένα προέρχονται από μεταβλητές που εκφράζουν κάποιο ποιοτικό χαρακτηριστικό, το οποίο όμως υπόκειται σε διάταξη. Η φυσική διάταξη που προκύπτει δεν υπονοεί απαραίτητα τη γνώση ύπαρξης απόστασης ανάμεσα στις κατηγορίες, ούτε την ύπαρξη απόστασης από ένα σταθερό και ανεξάρτητο σημείο αναφοράς. Χαρακτηριστικά παραδείγματα διατάξιμων δεδομένων αποτελούν το επίπεδο εκπαίδευσης ενός ατόμου (π.χ. πρωτοβάθμια, δευτεροβάθμια, τριτοβάθμια), η στάση απέναντι στην οικονομική πολιτική της κυβέρνησης (πολύ αρνητική, αρνητική, αδιάφορη, θετική, πολύ θετική) ο αυτοπροσδιορισμός της ψυχικής του υγείας (π.χ. κακή, μέτρια, καλή, εξαιρετική) κ.ά. Όπως και στα κατηγορικά δεδομένα, η αριθμητική αντιστοίχιση χρησιμοποιείται για λόγους διευκόλυνσης στην αξιοποίηση της πληροφορίας.

- Ποσοτικά δεδομένα με αυθαίρετο μηδέν (*interval*): Τα δεδομένα αυτά προέρχονται από ποσοτικές μεταβλητές και εκφράζουν κάποιο ποσοτικό, δηλαδή αριθμητικό χαρακτηριστικό. Το μηδέν έχει ορισθεί αυθαίρετα ως αρχή της κλίμακας μέτρησης και δεν εκφράζει την αρχή των εννοιών των μεταβλητών. Οι ιδιότητές τους περιλαμβάνουν την ταξινόμηση κατά σειρά μεγέθους και τον προσδιορισμό της μεταξύ τους απόστασης. Ανάμεσα στα στοιχεία του συγκεκριμένου συνόλου, μπορεί να εφαρμοστεί η πράξη της πρόσθεσης αλλά όχι του πολλαπλασιασμού. Επιπλέον, ενώ το διάστημα ανάμεσα σε δύο τιμές της μεταβλητής ορίζει τη διαφορά τους, η αναλογική τους έκφραση δεν έχει κάποια αξία. Ένα χαρακτηριστικό

παράδειγμα για το προηγούμενο διαφαίνεται στα δεδομένα που εκφράζουν θερμοκρασία. Η θερμοκρασία των 20 βαθμών Κελσίου δε νοείται δύο φορές μεγαλύτερη από τη θερμοκρασία των 10 βαθμών Κελσίου. Αντιπροσωπευτικά παραδείγματα ποσοτικών δεδομένων με αυθαίρετο μηδέν αποτελούν ο δείκτης νοημοσύνης, η θερμοκρασία και άλλα φυσικά μεγέθη που μετριούνται σε ποικίλες κλίμακες (πχ. Κελσίου, Φαρενάιτ), η μέτρηση στοιχείων της προσωπικότητας - με κλίμακα μονάδας μέτρησης όχι σταθερή και κοινά αποδεκτή από όλους στο χώρο των κοινωνικών επιστημών κ.ά.

- Ποσοτικά δεδομένα με απόλυτο μηδέν (*ratio*): Τα ποσοτικά δεδομένα με απόλυτο μηδέν μπορεί να θεωρηθούν ως μια υποπερίπτωση της προηγούμενης κατηγορίας. Η διαφορά τους εντοπίζεται στο γεγονός ότι το μηδέν της κλίμακας μέτρησης είναι ένα αντικειμενικό σημείο αναφοράς που εκφράζει την πραγματική κατάσταση της μεταβλητής. Η μηδενική, δηλαδή, τιμή του χαρακτηριστικού υποδηλώνει την απουσία τιμής στο στοιχείο αυτό, όπως για παράδειγμα στοιχείο με μηδενικό βάρος ή ύψος. Όλες οι ιδιότητες των ποσοτικών δεδομένων με αυθαίρετο μηδέν εφαρμόζονται και εδώ, ενώ επιπλέον ο λόγος μεταξύ των διαφορετικών τιμών του χαρακτηριστικού έχει σημασία. Ενδεικτικά παραδείγματα για τη συγκεκριμένη κατηγορία περιλαμβάνουν το βάρος, το μήκος, την ηλικία, την ταχύτητα ενός σώματος κ.ά.

Στα περισσότερα σύνολα δεδομένων που εξετάζονται η συνηθέστερη σύνθεση που παρουσιάζεται είναι η ύπαρξη μικτών δεδομένων, δεδομένα δηλαδή που προέρχονται συνήθως από όλες τις προηγούμενες κατηγορίες ή κάποιες από αυτές. Ένα τέτοιο παράδειγμα παρουσιάζεται παρακάτω:

Individual	Sex	Age (yrs)	IQ	Depression	Health	Weight (lbs)
1	Male	21	120	Yes	Very good	150
2	Male	43	NK	No	Very good	160
3	Male	22	135	No	Average	135
4	Male	86	150	No	Very poor	140
5	Male	60	92	Yes	Good	110
6	Female	16	130	Yes	Good	110
7	Female	NK	150	Yes	Very good	120
8	Female	43	NK	Yes	Average	120
9	Female	22	84	No	Average	105
10	Female	80	70	No	Good	100

Note: NK = not known

Πίνακας 1: Υποθετικό παράδειγμα πίνακα δεδομένων για δέκα άτομα από τους Everitt και Dunn (1991).

Είναι εύκολο να παρατηρήσει κανείς, πως ο Πίνακας 1, απαρτίζεται από διάφορες μεταβλητές που εκφράζουν είτε ποσοτικές πληροφορίες είτε ποιοτικές πληροφορίες. Όπως αναφέρθηκε προηγουμένως, μπορεί να υπάρξει αριθμητική αντιστοίχιση για τις ποιοτικές μεταβλητές π.χ. Sex=1 για άντρες και Sex=2 για γυναίκες, Health=1 όταν η υγεία είναι πολύ κακή και Health=5 όταν η υγεία είναι πολύ καλή κ.τ.λ. Ο κάθε ερευνητής θα πρέπει να είναι πολύ προσεκτικός με την εκάστοτε αντιστοίχιση, για να μην εκφράζει διαφορετική πληροφορία. Τέλος, στον Πίνακα 1, υπάρχουν ελλειπούσες τιμές (*missing values*), όπου συμβολίζονται με "NK"(*Not Known*). Γενικά, αυτό μπορεί να συμβεί είτε από ανθρώπινο λάθος του ερευνητή είτε επειδή το άτομο ξέχασε ή αρνήθηκε να παρέχει την συγκεκριμένη πληροφορία. Αυτό είναι ένα ζήτημα που θα αναλυθεί στη συνέχεια.

2.2 Προπαρασκευή δεδομένων

Η προπαρασκευή των δεδομένων είναι ένα σημαντικό βήμα στη διαδικασία ανάλυσης δεδομένων και είναι ζωτικής σημασίας για τη διασφάλιση της ακρίβειας και της αξιοπιστίας αυτών. Εξασφαλίζει τη συνέπεια μεταξύ των διαφόρων πηγών στοιχείων ή πληροφοριών, βοηθώντας στη συγχώνευση δεδομένων από διαφορετικές πηγές προέλευσης σε ένα ενιαίο σύνολο δεδομένων, διασφαλίζοντας ότι αυτά είναι αξιόπιστα και ακριβή και μπορούν να χρησιμοποιηθούν μαζί. Η προπαρασκευή των δεδομένων περιλαμβάνει διάφορες επιμέρους ενέργειες όπως καθαρισμός δεδομένων, αντιμετώπιση ακραίων τιμών (*outliers*), χειρισμός ελλειπουσών τιμών, μετασχηματισμός δεδομένων κ.ά.

2.2.1. Καθαρισμός δεδομένων

Ο καθαρισμός των δεδομένων αποτελεί σημαντικό βήμα στη διαδικασία ανάλυσης δεδομένων. Είναι η διαδικασία εντοπισμού και διόρθωσης σφαλμάτων, ασυνεπειών και τιμών που λείπουν σε ένα σύνολο δεδομένων. Ανεξάρτητα από το πόσο προσεκτικά έχουν εισαχθεί τα δεδομένα, τα σφάλματα είναι αναπόφευκτα. Αυτό θα μπορούσε να οδηγήσει σε εσφαλμένη κωδικοποίηση, έλλειψη δεδομένων κ.ά.

Κάθε ερευνητής για να διενεργήσει μια ανάλυση, θα πρέπει να διαχειριστεί έναν μεγάλο όγκο δεδομένων. Τα μικτά δεδομένα συνήθως αναφέρονται σε στοιχεία που περιλαμβάνουν τόσο αριθμητικές όσο και κατηγορικές μεταβλητές. Είναι σημαντικό να προσδιοριστούν οι τύποι δεδομένων στο σύνολο δεδομένων, πριν από τη διεξαγωγή της οποιασδήποτε ανάλυσης. Μια ανάλυση για να είναι αξιόπιστη, τα δεδομένα πρέπει να είναι ποιοτικά και να διακρίνονται από ακρίβεια, εγκυρότητα, συνέπεια και ομοιομορφία. Επίσης, βασικό χαρακτηριστικό της ανάλυσης είναι οι τεχνικές που θα χρησιμοποιήσει για την εύρεση και αντιμετώπιση των προβληματικών δεδομένων.

Σύμφωνα με τους Ουζούνη και Νακάκη (2011), η εγκυρότητα των δεδομένων αναφέρεται στο βαθμό στον οποίο τα δεδομένα αντιπροσωπεύουν με ακρίβεια τα πραγματικά στοιχεία που προορίζονται να μετρήσουν. Η αξιοπιστία μπορεί να ερμηνευθεί με τους όρους «σταθερότητα» και «εσωτερική συνοχή» που αφορά στις παραμέτρους τις οποίες θα πρέπει να εξεταστούν από τον ερευνητή. Η αξιοπιστία είναι ο συσχετισμός μιας μεταβλητής, ενός παράγοντα ή ενός μοντέλου με κάτι υποθετικό, που μετρά αληθινά αυτό που επιθυμεί να μετρηθεί. Δίνει δηλαδή μετρήσεις που είναι απαλλαγμένες από σφάλματα και παρέχουν συνεπή αποτελέσματα.

Η συνέπεια των δεδομένων αναφέρεται στην ορθότητα και την ακρίβεια των δεδομένων, πράγμα που σημαίνει ότι τα δεδομένα είναι απαλλαγμένα από σφάλματα και αντανακλούν την πραγματικότητα όσο το δυνατόν καλύτερα. Η συνέπεια των δεδομένων είναι σημαντική, επειδή τα ασυνεπή δεδομένα μπορεί να οδηγήσουν σε εσφαλμένες αποφάσεις, παραπλανητική ανάλυση και αναξιόπιστες πληροφορίες.

Σημαντικό βήμα στον καθαρισμό των δεδομένων, είναι η κατάργηση των διπλότυπων εγγραφών. Αυτές μπορούν να προκύψουν όταν τα ίδια δεδομένα εισάγονται πολλές φορές. Η κατάργηση των διπλότυπων δεδομένων, μας εξασφαλίζει ότι η κάθε παρατήρηση είναι μοναδική και αποφεύγεται η στρέβλωση της ανάλυσης. Η διαδικασία κατάργησης διπλότυπων δεδομένων περιλαμβάνει την αναγνώριση και την κατάργηση γραμμών ή εγγραφών που έχουν διπλότυπες τιμές σε μία ή περισσότερες στήλες. Τα βήματα για την κατάργηση των διπλότυπων είναι τα εξής: προσδιορίζουμε τις στήλες που περιέχουν διπλότυπες τιμές, ταξινομούμε το σύνολο δεδομένων ως προς τις στήλες με πιθανά διπλότυπα. Η ταξινόμηση του συνόλου

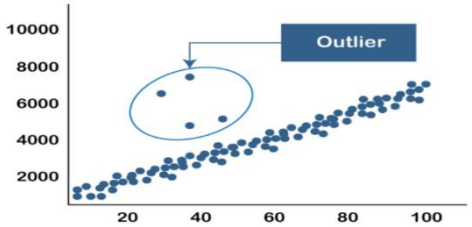
δεδομένων με αυτόν τον τρόπο, βοηθά στην ομαδοποίηση παρόμοιων εγγραφών. Αφού εντοπίσουμε τις διπλότυπες τιμές, τις καταργούμε από το σύνολο δεδομένων.

2.2.2. Εύρεση και αντιμετώπιση ακραίων τιμών

Οι ακραίες τιμές είναι παρατηρήσεις, που διαφέρουν σημαντικά από άλλες. Συνήθως μπορεί να προκληθούν από σφάλματα στη συλλογή δεδομένων ή τη μέτρηση ή ακόμη να είναι και ορθές παρατηρήσεις, που απέχουν όμως πολύ από το μέσο όρο. Ο ερευνητής θα πρέπει να είναι ιδιαίτερα προσεκτικός στην αντιμετώπισή τους, γιατί αυτές οι τιμές μπορούν να έχουν σημαντικό αντίκτυπο στην ανάλυση δεδομένων και τη στατιστική μοντελοποίηση. Οι ακραίες τιμές μπορούν να στρεβλώσουν τα αποτελέσματα της ανάλυσης, να δώσουν μη αξιόπιστες προβλέψεις, να επηρεάσουν την ακρίβεια των στατιστικών μέτρων, όπως τους μέσους όρους, τις τυπικές αποκλίσεις κ.ά. Επειδή υπάρχουν διαφορετικές αιτίες για την ύπαρξη ακραίων τιμών, οι τεχνικές ανίχνευσης και αντιμετώπισης τους, βασίζονται στον τομέα εφαρμογής της κάθε ανάλυσης.

Υπάρχουν διάφορες προσεγγίσεις για την αντιμετώπιση των ακραίων τιμών (Johnson και Wichern (2007)). Η πιο εύκολη θα ήταν, να αφαιρεθούν από το σύνολο δεδομένων. Ωστόσο, αυτό θα πρέπει να γίνεται μόνο μετά από προσεκτική εξέταση, καθώς η κατάργησή τους μπορεί να καταργήσει σημαντικές πληροφορίες από αυτό.

Από τις απλούστερες μεθόδους εύρεσης ακραίων τύπων, είναι η χρήση διαφόρων γραφικών παραστάσεων, σε μία ή και περισσότερες διαστάσεις. Με αυτό τον τρόπο, απεικονίζονται οι τιμές που διαφοροποιούνται σε σχέση με τα υπόλοιπα σύνολα δεδομένων. Στην Εικόνα 1 παρατίθεται ένα γραφικό παράδειγμα ακραίων τιμών.

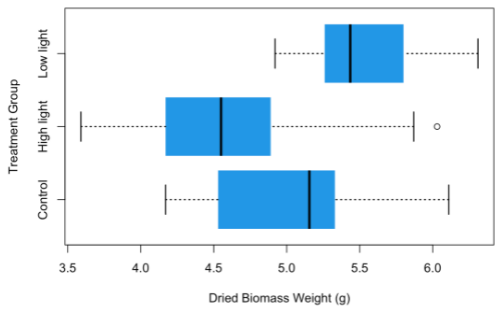


Εικόνα 1: Δείγμα 4 ακραίων τιμών.

(https://www.tutorialandexample.com/outlier-analysis-in-data-mining#)

Όπως φαίνεται σε αυτό το γράφημα, οι ακραίες τιμές είναι σημεία που βρίσκονται έξω από το μοτίβο της κατανομής.

Ένας ακόμη γραφικός τρόπος παρουσίασης είναι και το θηκόγραμμα (*box-plot*). Είναι ένας πολύ απλός, αλλά αποτελεσματικός τρόπος να απεικονίσουμε τις ακραίες τιμές (βλ. Εικόνα 2).



Εικόνα 2: θηκόγραμμα 1 μεταβλητής με 3 επίπεδα.

(https://www.r-bloggers.com/2022/04/how-to-make-a-boxplot-in-r/)

Ο αλγόριθμος DBSCAN έχει πλεονεκτήματα έναντι άλλων, όπως η ικανότητα χειρισμού ομάδων διαφορετικών σχημάτων και μεγεθών καθώς και την ικανότητα αναγνώρισης θορύβου. Έχει δηλαδή την ικανότητα να ομαδοποιεί παρατηρήσεις που είναι κοντά η μία στην άλλη σε μια περιοχή υψηλής πυκνότητας, ενώ ταυτοποιεί παρατηρήσεις που βρίσκονται σε περιοχές χαμηλής πυκνότητας, ως θόρυβος. Συνεπώς, ο ερευνητής μπορεί να τον χρησιμοποιήσει για εύρεση ακραίων τιμών.

Η μέθοδος Z-score, είναι ένας απλός και αποτελεσματικός τρόπος να ανιχνεύσουμε ακραίες τιμές σε ένα σύνολο δεδομένων. Αυτή είναι μία παραμετρική μέθοδος, καθώς βασίζεται στη χρήση του μέσου όρου $\bar{x}$ και της τυπικής απόκλισης s της μεταβλητής. Ο τύπος για τον υπολογισμό του Z-score, είναι:

$$Z = (x - \bar{x})/s$$

Οι παρατηρήσεις με τιμή μεγαλύτερη από ένα συγκεκριμένο όριο (συνήθως 2 ή 3) θεωρούνται ακραίες τιμές. Αυτό συμβαίνει, επειδή οι παρατηρήσεις που απέχουν περισσότερο από 2 ή 3 τυπικές αποκλίσεις από τη μέση τιμή, θεωρούνται εξαιρετικά απίθανο να παρατηρηθούν όταν τα δεδομένα προέρχονται από μία κανονική κατανομή. Ωστόσο, προϋποθέτει ότι το σύνολο δεδομένων ακολουθεί μια κανονική κατανομή και μπορεί να μη λειτουργεί καλά για τα σύνολα δεδομένων από άλλες κατανομές. Επιπλέον, μπορεί να μην είναι αποτελεσματική για σύνολα δεδομένων με μικρό μέγεθος δείγματος, καθώς η μέση τιμή και η τυπική απόκλιση, μπορεί να μην είναι αντιπροσωπευτικές του πλήθους του δείγματος.

Οι μέθοδοι ανίχνευσης ακραίων τιμών για κατηγορικά δεδομένα, είναι διαφορετικές από εκείνες των αριθμητικών, καθώς αυτά δεν μπορούν εύκολα να τυποποιηθούν ή να μετρηθούν, από την άποψη της απόστασης. Η μέθοδος που βασίζεται στη συχνότητα για την ανίχνευση ακραίων τιμών σε κατηγορικά δεδομένα περιλαμβάνει την εξέταση της συχνότητας κάθε κατηγορίας στο σύνολο δεδομένων και τον προσδιορισμό των κατηγοριών. Ακολουθεί ένα παράδειγμα του τρόπου με τον οποίο μπορεί να εφαρμοστεί αυτή η μέθοδος:

Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων που περιέχει πληροφορίες σχετικά με τα αγαπημένα χρώματα των ανθρώπων σε μια συγκεκριμένη περιοχή, το οποίο έχει τις ακόλουθες κατηγορίες: κόκκινο, μπλε, πράσινο, κίτρινο και μαύρο. Η συχνότητα κάθε κατηγορίας έχει ως εξής: Κόκκινο=10, Μπλε=25, Πράσινο=15, Κίτρινο=5 και Μαύρο=2.

Σε αυτήν την περίπτωση, μπορούμε να δούμε ότι οι κατηγορίες "κίτρινο" και "μαύρο" εμφανίζονται πολύ σπάνια, ενώ οι κατηγορίες "μπλε" και "πράσινο" εμφανίζονται πιο συχνά. Επομένως, μπορούμε να θεωρήσουμε τις κατηγορίες "κίτρινο" και "μαύρο" ως πιθανές ακραίες τιμές.

Ωστόσο, είναι σημαντικό να λαμβάνεται υπόψη το εύρος των δεδομένων πριν από τη λήψη οποιωνδήποτε αποφάσεων με βάση τις προσδιοριζόμενες ακραίες τιμές. Σε αυτό το παράδειγμα, οι κατηγορίες "κίτρινο" και "μαύρο" μπορεί να μην είναι απαραίτητα ακραίες αν είναι αντιπροσωπευτικές μιας συγκεκριμένης ομάδας.

Ως εκ τούτου, είναι σημαντικό να αξιολογηθούν προσεκτικά τα αποτελέσματα του εντοπισμού ακραίων τιμών και να εξεταστεί το πλαίσιο των δεδομένων πριν από τη λήψη οποιωνδήποτε αποφάσεων με βάση τις προσδιοριζόμενες ακραίες τιμές.

2.2.3. Εύρεση και αντιμετώπιση ελλειπουσών τιμών

Οι τιμές που λείπουν από ένα σύνολο δεδομένων είναι ένα κοινό πρόβλημα στην ανάλυση δεδομένων. Αυτές μπορεί να οφείλονται σε διάφορους λόγους, όπως σφάλματα καταχώρισης δεδομένων, καταστροφή δεδομένων, απώλεια δεδομένων ή απλά επειδή ορισμένες παρατηρήσεις δεν συλλέχθηκαν.

Οι ελλείπουσες τιμές μπορούν να δημιουργήσουν προβλήματα στην ανάλυση δεδομένων. Μερικά από αυτά είναι (Enders (2010)):

- Μεροληψία: Οι ελλείπουσες τιμές μπορούν να εισάγουν μεροληψία στην ανάλυση αν δεν αντιμετωπίζονται σωστά. Τα ελλείποντα στοιχεία ενδέχεται να μην είναι τυχαία και να σχετίζονται με το αποτέλεσμα ή άλλες μεταβλητές, με αποτέλεσμα μεροληπτικές εκτιμήσεις και εσφαλμένα συμπεράσματα.
- Μειωμένο μέγεθος δείγματος: Όταν υπάρχουν ελλείπουσες τιμές, το μέγεθος του δείγματος μειώνεται, το οποίο μπορεί να μειώσει τη στατιστική ισχύ και την ακρίβεια της ανάλυσης. Αυτό μπορεί να καταστήσει πιο δύσκολο να εντοπιστούν σημαντικές επιδράσεις ή σχέσεις.
- Ασυμμετρικές κατανομές: Οι ελλείπουσες τιμές μπορούν να επηρεάσουν τη μορφή της κατανομής των μεταβλητών, ειδικά αν δεν λείπουν τυχαία. Αυτό μπορεί να επηρεάσει την εγκυρότητα των στατιστικών δοκιμών και των παραδοχών που γίνονται στην ανάλυση.
- Απώλεια πληροφοριών: Η διαγραφή ή η αγνόηση των τιμών που λείπουν χωρίς να ληφθούν υπόψη οι λόγοι που συμβαίνει αυτό μπορεί να οδηγήσει σε απώλεια πολύτιμων πληροφοριών. Αυτό μπορεί να μειώσει την ακρίβεια και την αξιοπιστία της ανάλυσης.
- Εσφαλμένος καταλογισμός: Αν οι ελλείπουσες τιμές καταλογίζονται χωρίς κατάλληλη εξέταση των δεδομένων και του μηχανισμού ελλείψεων, οι υποθετικές τιμές ενδέχεται να μην αντιπροσωπεύουν επακριβώς τις αληθείς τιμές. Αυτό μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα και ερμηνείες.
- Αυξημένη πολυπλοκότητα: Η αντιμετώπιση των τιμών που λείπουν απαιτεί πρόσθετα βήματα στην προπαρασκευή και ανάλυση δεδομένων, τα οποία μπορούν να αυξήσουν την πολυπλοκότητα και τον χρόνο που απαιτείται για την ανάλυση.

Οι μέθοδοι αντιμετώπισης των ελλειπουσών τιμών, διαφέρουν αναλόγως της κατηγορίας δεδομένων που επεξεργαζόμαστε. Υπάρχουν διαφορετικές προσεγγίσεις που ακολουθούμε στα δείγματα με αριθμητικά δεδομένα και διαφορετικές στα δείγματα των κατηγορικών δεδομένων. Παρακάτω θα αναφέρουμε τεχνικές που αφορούν τις προσεγγίσεις των μικτών δεδομένων.

- Διαγραφή τιμών που λείπουν: Αν τα δεδομένα που λείπουν είναι ελάχιστα ή λείπουν εντελώς τυχαία, μπορούν να διαγραφούν οι γραμμές με τις τιμές που λείπουν. Ωστόσο, η προσέγγιση αυτή μπορεί να οδηγήσει σε μείωση του μεγέθους του δείγματος και πιθανή απώλεια πολύτιμων πληροφοριών.
- Απουσία κατηγορίας: Αντιστοιχίζεται μια ξεχωριστή κατηγορία για να αντιπροσωπεύσει τις τιμές που λείπουν. Είναι σημαντικό να εξεταστεί προσεκτικά ο τρόπος με τον οποίο θα ερμηνευτεί αυτή η κατηγορία που λείπει στην ανάλυση. Ο ερευνητής ξεκινάει με τον προσδιορισμό των τιμών που λείπουν από τα δεδομένων. Οι τιμές που λείπουν αντιπροσωπεύονται συνήθως από κενά κελιά, «NA», «N/A» (*Not Available*), «N/K» ή άλλες ενδείξεις που καθορίζονται στο σύνολο δεδομένων.
- Χρήση μέσου όρου/διαμέσου: Μια κοινή προσέγγιση είναι η αντικατάσταση των ελλειπουσών αριθμητικών τιμών με τη μέση τιμή ή τη διάμεσο της αντίστοιχης μεταβλητής. Αυτή η μέθοδος υποθέτει ότι οι τιμές που λείπουν είναι παρόμοιες με τις υπάρχουσες τιμές

όσον αφορά την κεντρική τους τάση. Μπορεί, όμως, να υποτιμήσει τη μεταβλητότητα και να στρεβλώσει την κατανομή των δεδομένων.
- Προηγμένες Τεχνικές Απόδοσης: Υπάρχουν και πιο εξελιγμένες μέθοδοι υπολογισμού που έχουν σχεδιαστεί ειδικά για κατηγορικά δεδομένα, όπως η regression imputation, η hot-deck imputation ή η multiple imputation (βλ. Enders (2010)). Αυτές οι μέθοδοι λαμβάνουν υπόψη τις σχέσεις μεταξύ των μεταβλητών και καταγράφουν την αβεβαιότητα που σχετίζεται με τις ελλείπουσες τιμές, παρέχοντας πιο ακριβείς υπολογισμούς.

2.2.4. Μετασχηματισμός δεδομένων

Ο μετασχηματισμός των δεδομένων, περιλαμβάνει την κλιμάκωση των δεδομένων, σε μια κοινή κλίμακα για να διασφαλιστεί ότι όλες οι μεταβλητές είναι εξίσου σημαντικές στην ανάλυση. Αυτό είναι σημαντικό, όταν ασχολούμαστε με μεταβλητές που έχουν διαφορετικές μονάδες και κλίμακες. Ανατρέχοντας στη βιβλιογραφία (Saxena (2023)), εντοπίζουμε πολλές τεχνικές μετασχηματισμού δεδομένων, όπως: την δυαδική κωδικοποίηση, την κωδικοποίηση one-hot, την τυποποίηση (γνωστή ως z-score), την κανονικοποίηση, τον μετασχηματισμός Box-Cox κ.ά.

Η κωδικοποίηση one-hot (*one-hot encoding*) είναι μια μέθοδος που χρησιμοποιείται συχνά για την μετατροπή κατηγορικών μεταβλητών σε δυαδική αναπαράσταση με διανύσματα. Κάθε διάνυσμα έχει το ίδιο μήκος με τον αριθμό των μοναδικών κατηγοριών της μεταβλητής. Στο δυαδικό διάνυσμα, μόνο ένα στοιχείο είναι 1 ενώ τα υπόλοιπα είναι 0, δείχνοντας την παρουσία ή απουσία μιας συγκεκριμένης κατηγορίας. Στη συνέχεια, εξηγούμε πως λειτουργεί η κωδικοποίηση one-hot:

- Αναγνωρίζουμε την κατηγορική μεταβλητή που χρειάζεται να κωδικοποιηθεί. Για παράδειγμα, ας υποθέσουμε ότι έχουμε μια κατηγορική μεταβλητή με την ονομασία «Χρώμα» με τις κατηγορίες: «Κόκκινο», «Μπλε», και «Πράσινο».
- Δημιουργούμε ένα δυαδικό διάνυσμα για κάθε κατηγορία. Το μήκος κάθε διανύσματος είναι ίσο με τον συνολικό αριθμό των μοναδικών κατηγοριών (τρεις σε αυτήν την περίπτωση). Δηλαδή, θα δημιουργήσουμε τρία δυαδικά διανύσματα:
    - Διάνυσμα για το «Κόκκινο»: [1, 0, 0]
    - Διάνυσμα για το «Πράσινο»: [0, 1, 0]
    - Διάνυσμα για το «Μπλε»: [0, 0, 1]
- Αντικαθιστούμε την αρχική κατηγορική μεταβλητή με τα δυαδικά διανύσματα. Κάθε παρατήρηση που είχε «Κόκκινο» στην αρχική μεταβλητή θα έχει τώρα [1, 0, 0] ως αναπαράστασή της, κ.ο.κ.

Η δυαδική κωδικοποίηση (*binary encoding*) είναι μια τεχνική που χρησιμοποιείται για το μετασχηματισμό των κατηγορικών μεταβλητών σε μια αριθμητική αναπαράσταση χρησιμοποιώντας δυαδικά ψηφία, η οποία βοηθάει στη διατήρηση των πληροφοριών. Η διαδικασία περιλαμβάνει τη δημιουργία δυαδικών χαρακτηριστικών για κάθε μοναδική κατηγορία στην αρχική μεταβλητή. Πρακτικά η δυαδική κωδικοποίηση και η κωδικοποίηση one-hot είναι ισοδύναμες. Ακολούθως εξηγούμε πως λειτουργεί η δυαδική κωδικοποίηση:

- Προσδιόρισε τις μοναδικές κατηγορίες στην κατηγορική μεταβλητή. Για παράδειγμα, ας θεωρήσουμε μια κατηγορική μεταβλητή «Χρώμα» με τρεις κατηγορίες: «Κόκκινο», «Μπλε», και «Πράσινο».

- Καταχώρησε σε κάθε κατηγορία μια μοναδική ακέραια τιμή. Στο συγκεκριμένο παράδειγμα, έχουμε τρεις κατηγορίες, μπορούμε να εκχωρήσουμε τις τιμές ως εξής: «Κόκκινο» ως 1, το «Μπλε» ως 2 και το «Πράσινο» ως 3.
- Μετέτρεψε κάθε ακέραιο στη δυαδική του αναπαράσταση. Για παράδειγμα, η δυαδική αναπαράσταση για το 1 είναι 001, για το 2 είναι 010 και για το 3 είναι 011.
- Δημιούργησε δυαδικά χαρακτηριστικά με βάση τις δυαδικές αναπαραστάσεις που έχουμε αποκτήσει στο προηγούμενο βήμα. Κάθε δυαδικό χαρακτηριστικό αντιπροσωπεύει εάν μια συγκεκριμένη κατηγορία είναι παρούσα ή όχι. Για παράδειγμα, δημιουργήσαμε τρία δυαδικά χαρακτηριστικά: «Κόκκινο» (001), «Μπλε» (010) και «Πράσινο» (011). Εάν η αρχική κατηγορία είναι το «Κόκκινο», το δυαδικό χαρακτηριστικό για το «Κόκκινο» θα έχει τιμή 1, ενώ τα δυαδικά χαρακτηριστικά για το «Μπλε» και το «Πράσινο» θα έχουν τιμές 0.

Η κλιμάκωση χαρακτηριστικών (*Feature Scaling*) είναι μια τεχνική προπαρασκευής δεδομένων που χρησιμοποιείται για τη μεταφορά όλων των μεταβλητών σε παρόμοια κλίμακα. Είναι σημαντικό επειδή πολλοί αλγόριθμοι μηχανικής μάθησης είναι ευαίσθητοι στην κλίμακα των μεταβλητών και έχοντας χαρακτηριστικά σε διαφορετικές κλίμακες μπορεί να επηρεαστεί η απόδοση και η ακρίβεια των μοντέλων. Δύο συνηθισμένες μέθοδοι για την κλιμάκωση χαρακτηριστικών είναι οι παρακάτω:

- Τυποποίηση (Z-score): Η τυποποίηση μετασχηματίζει τα δεδομένα έτσι ώστε να έχει μέσο όρο 0 και τυπική απόκλιση 1. Αυτό επιτυγχάνεται αφαιρώντας τη μέση τιμή του χαρακτηριστικού από κάθε παρατήρηση και στη συνέχεια διαιρώντας με την τυπική απόκλιση. Ο τύπος για την τυποποίηση είναι: $Z = (\ x - \bar{x})/s$.
- Κανονικοποίηση: Η κανονικοποίηση κλιμακώνει τα δεδομένα σε ένα σταθερό εύρος, συνήθως μεταξύ 0 και 1. Επιτυγχάνεται με αφαίρεση της ελάχιστης τιμής του χαρακτηριστικού από κάθε παρατήρηση και στη συνέχεια διαίρεση με το εύρος του δείγματος. Ο τύπος για την κανονικοποίηση είναι: $x_{scaled} = (x - min)/(max - min)$.

Η επιλογή μεταξύ τυποποίησης και κανονικοποίησης εξαρτάται από τις συγκεκριμένες απαιτήσεις του συνόλου δεδομένων και τον αλγόριθμο που χρησιμοποιείται. Γενικά, η τυποποίηση είναι πιο αποτελεσματική σε ακραίες τιμές και λειτουργεί καλά με αλγορίθμους που υποθέτουν κανονικά κατανεμημένα δεδομένα. Η κανονικοποίηση είναι κατάλληλη για δεδομένα με γνωστά όρια και αλγορίθμους που βασίζονται σε υπολογισμούς απόστασης.

Λαμβάνοντας το όνομά της, από τους στατιστικούς George Box και David Cox, οι οποίοι εισήγαγαν τη μέθοδο το 1964, ο μετασχηματισμός Box-Cox είναι μια τεχνική που χρησιμοποιείται για το μετασχηματισμό μη κανονικών δεδομένων σε περίπου κανονικά κατανεμημένα δεδομένα (Box και Cox (1964)). Χρησιμοποιείται σε αριθμητικά δεδομένα για να αντιμετωπίσει τις ασύμμετρες ή μη κανονικές κατανομές δεδομένων, οι οποίες μπορούν να παραβιάσουν τις υποθέσεις πολλών στατιστικών μοντέλων. Ο συγκεκριμένος μετασχηματισμός των δεδομένων, βοηθάει να γίνουν τα δεδομένα πιο συμμετρικά και κατάλληλα για ανάλυση με χρήση τεχνικών που υποθέτουν την κανονικότητα, όπως η γραμμική παλινδρόμηση. Ο τύπος μετασχηματισμού Box-Cox είναι ο εξής:

$$x' = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ ln\ (x), & \lambda = 0 \end{cases}.$$

Είναι σημαντικό να σημειωθεί ότι ο μετασχηματισμός Box-Cox υποθέτει ότι οι τιμές των δεδομένων είναι θετικές. Γενικά θα λέγαμε ότι είναι ένα χρήσιμο εργαλείο, στην ανάλυση και

τη μοντελοποίηση, καθώς μπορεί να βελτιώσει την ακρίβεια και την ερμηνευσιμότητα των στατιστικών μοντέλων.

2.3 Περιπτώσεις – dataset μικτών δεδομένων

Στη σημερινή εποχή και σε ένα διαρκώς εξελισσόμενο περιβάλλον επιχειρήσεων και προκλήσεων, η ορθή χρήση της γνώσης, παράγει απτή επιχειρηματική αξία και όχι μόνο. Επιχειρήσεις και οργανισμοί, σε μεγαλύτερο ή μικρότερο βαθμό, διαχειρίζονται τις πληροφορίες και τις γνώσεις που αποκομίζουν από αυτή, βελτιώνοντας την αποτελεσματικότητά τους, μειώνοντας τα κόστη, αυξάνοντας την κερδοφορία και την καινοτομία τους. Είναι επομένως προφανές ότι η ανάλυση δεδομένων έχει βαρύνουσα σημασία και αποτελεί πολύτιμο επιχειρηματικό εργαλείο.

Μικτά δεδομένα μπορούν να βρεθούν σε διάφορους τομείς και κλάδους, που πραγματοποιούνται αναλύσεις δεδομένων και ειδικότερα ανάλυση κατά συστάδες. Βασικά πεδία εφαρμογών βρίσκουμε στα παρακάτω (βλ. Κύρκος (2015)):

- Πωλήσεις και διαφήμιση
- Τράπεζες
- Ασφαλιστικές Εταιρείες
- Ηλεκτρονικό εμπόριο
- Χρηματιστήριο
- Τηλεπικοινωνίες
- Τομέας Υγείας
- Περιβαλλοντική – Μετεωρολογική Επιστήμη
- Μεταφορά και Logistics.

Η αναγνώριση της καταναλωτικής συμπεριφοράς των πελατών, παίζει βασικό ρόλο στην αύξηση των πωλήσεων. Η ανάλυση μικτών δεδομένων, βρίσκει άμεση εφαρμογή στην αναγνώριση και κατανόηση της καταναλωτικής συμπεριφοράς ενός ατόμου, εντοπίζοντας προϊόντα ή συνδυασμούς τους, που πωλούνται συχνά μαζί (π.χ. μπύρες, αναψυκτικά, προϊόντα υγιεινής για ενήλικες και βρέφη). Επιπροσθέτως, επιμερίζοντας τα χαρακτηριστικά των προϊόντων, οι εταιρείες κατανοούν ακόμα καλύτερα τι ωθούν τους καταναλωτές στις αγορές τους.

Οι τραπεζικοί οργανισμοί, είναι κατεξοχήν οι κατάλληλοι υποψήφιοι, για την εφαρμογή αναλύσεων μικτών δεδομένων. Βάσει των κανονιστικών τους διατάξεων, είναι υποχρεωμένοι να διατηρούν αναλυτικά στοιχεία των πελατών τους και δεδομένα για όλες τις συναλλαγές αυτών. Αναλύοντας τα παραπάνω, έχουν τη δυνατότητα να αντιμετωπίζουν ζητήματα ξεπλύματος χρήματος καθώς και να δημιουργούν προφίλ των πελατών τους, δημιουργώντας με αυτό τον τρόπο, αναλύσεις τμηματοποίησης της αγοράς. Οι διασταυρούμενες πληροφορίες σε μια τράπεζα, παρέχουν τη δυνατότητα ταυτόχρονης πώλησης προϊόντων π.χ. δάνεια, πιστωτικές κάρτες κ.ά. Προσελκύουν πελάτες για να έχουν μια πιο μακροχρόνια και σταθερή σχέση με την τράπεζα.

Η ανάλυση των μικτών δεδομένων, βρίσκει παραπλήσιες εφαρμογές και στον κλάδο των ασφαλειών. Μείωση του κόστους, αύξηση των πωλήσεων και των κερδών τους, διατήρηση του υπάρχοντος πελατολογίου και ανάπτυξη νέου. Ανανεώνοντας και αναπτύσσοντας τα προϊόντα τους, διαχειρίζονται την επισφάλεια, που συνήθως αποτελεί το μεγαλύτερο ρίσκο του κλάδου. Αναλόγως του τομέα ασφάλισης, υγείας ή υπηρεσιών (οχημάτων, κατοικίας κ.ά.), οι

κατηγορίες των δεδομένων είναι πολλαπλές. Παραδείγματος χάριν, σε ένα ασφαλιστήριο υγείας η ηλικία και το ιατρικό ιστορικό του ατόμου είναι αλληλένδετα χαρακτηριστικά. Αντίστοιχα, σε μία ασφάλεια οχήματος, η χρονολογία κατασκευής, ο κυβισμός και το είδος του οχήματος παίζουν σημαντικό ρόλο στο είδος των προτεινόμενων υπηρεσιών από την ασφαλιστική. Όταν τα δεδομένα ψηφιοποιούνται και αναλύονται, δίνουν τη δυνατότητα της αυτόματης απόρριψης ασυνεπών πληροφοριών ή σφαλμάτων, επιτρέποντας σε αυτούς τους οργανισμούς να έχουν μόνο συνεπείς και σωστές πληροφορίες, κάνοντάς τους γρήγορους και ευέλικτους στις λήψεις αποφάσεων και ενεργειών.

Το Ηλεκτρονικό Εμπόριο έχει ιδιαίτερα χαρακτηριστικά, δεδομένου ότι οι συναλλαγές πραγματοποιούνται ηλεκτρονικά. Συνέπεια αυτού είναι, η παραγωγή και η καταγραφή μεγάλων ποσοτήτων δεδομένων. Η ανάλυση αυτών προσφέρει πληθώρα πληροφοριών, όπως, το προφίλ του χρήστη, την καταναλωτική του συμπεριφορά και τις προτιμήσεις του, αφού είναι καταγεγραμμένη με λεπτομέρεια η ηλεκτρονική αγορά του προϊόντος. Προηγμένα συστήματα ηλεκτρονικού εμπορίου δίνουν τη δυνατότητα της παρακολούθησης της διαδρομής περιήγησης του χρήστη, πόσες παρόμοιες επιχειρήσεις και ποιες επισκέφθηκε καθώς και το πόση ώρα παρέμεινε σε κάποιο διαδικτυακό κατάστημα.

Η διαδικασία της χρηματοπιστωτικής διαμεσολάβησης (Χρηματιστήριο), θεωρείται ως ο σημαντικότερος τομέας για την ανάπτυξη της οικονομίας μιας χώρας. Η πρόβλεψη της διακύμανσης των μετοχών και των δεικτών, η τιμή μιας μετοχής του ίδιου προϊόντος ή επιχείρησης μπορεί να διαφέρει από χώρα σε χώρα. Τα χρηματιστήρια συναντιούνται σε κάθε χώρα ή ήπειρο, συνδέονται μεταξύ τους, αναπτύσσουν σχέσεις και αλληλοεπηρεάζονται. Όλα τα παραπάνω, επεξεργάζονται και ανάλογα με τις διακυμάνσεις της αγοράς της κάθε χώρας, ανακατατάσσονται και αναδιαμορφώνονται, για να μπορούν να έχουν εφαρμογή στην οικονομία της κάθε χώρας.

Οι επιχειρήσεις στον χώρο των τηλεπικοινωνίων, βρίσκονται στην κορυφαία κατάταξη στον χρήση τεχνολογιών πληροφορικής. Με αποτέλεσμα, οι όγκοι καταγραφής δεδομένων είναι τεράστιοι. Στοιχεία πελάτων, προσωπικά δεδομένα, ηλικία, διεύθυνση, οικογενειακή κατάσταση, στοιχεία κλήσεων, διάρκεια και κόστος κλήσης, είναι κάποια από τα δεδομένα που μπορούν να καταγραφούν. Τα τεχνικά στοιχεία της λειτουργείας των τηλεπικοινωνιών, παρέχουν τη δυνατότητα, να παράγουν ταυτόχρονα και τεχνικά δεδομένα: αντιμετώπιση προβλημάτων δικτύου, πωλήσεις και διαφήμιση προϊόντων και υπηρεσιών καθώς και αντιμετώπιση φαινομένων απάτης.

Στον τομέα της Υγείας και την Περίθαλψη, τα μικτά δεδομένα, βρίσκουν πλήρη εφαρμογή. Τα ιατρικά δεδομένα προσφέρουν ένα μείγμα αριθμητικών και κατηγορικών μεταβλητών, δεδομένα κειμένου (συνταγογραφήσεις) και δεδομένα χρονολογικών σειρών. Όλα αυτά αποτυπώνονται σε προφίλ ασθενών, εργαστηριακές τιμές εξετάσεων τους, δημογραφική τους κατάσταση και αρχεία ασθενείας κατά χρονολογική σειρά.

Τα περιβαλλοντικά δεδομένα, περιλαμβάνουν διάφορους τύπους πληροφοριών, σχεδόν όμοιες με εκείνες που χρησιμοποιούνται στην Μετεωρολογική Επιστήμη. Αριθμητικές μετρήσεις, όπως, θερμοκρασία, επίπεδα ρύπανσης, ποσοστά υγρασίας, καθώς και κατηγορικές μεταβλητές ή χωρικά δεδομένα, γεωγραφικές συντεταγμένες και περιβαλλοντικές εκθέσεις. Εστιάζοντας στις αλληλεπιδράσεις μεταξύ του εδάφους, του νερού, του αέρα στη βιόσφαιρα και των ζωντανών οργανισμών της, οι επιστήμονες ερευνούν πως αυτές οι σχέσεις συντελούν στις περιβαλλοντικές αλλαγές σε διάφορες χωρικές και χρονικές κλίμακες.

Σημαντικός κλάδος χρήσης μικτών δεδομένων, είναι και αυτός της μεταφοράς και της εφοδιαστικής αλυσίδας (*logistics*). Το κόστος μεταφοράς, ο χρόνος μεταφοράς και οι εκλυόμενες εκπομπές διοξειδίου του άνθρακα κατά τη μεταφορά είναι δεδομένα που μπορούν να αναλυθούν σε: αεροπορική, ακτοπλοϊκή ή οδική μεταφορά, μέρες παράδοσης ή άφιξης σε τελικό προορισμό, τύποι οχημάτων κ.ά, λαμβάνοντας υπόψιν και γεωγραφικές παραμέτρους. Η διαχείριση της εφοδιαστικής αλυσίδας, περιλαμβάνει πληθώρα κατηγοριών προς συλλογή και ανάλυση. Προμήθειες, παραγωγή, αποθήκευση, μεταφορά και πωλήσεις εντός των επιχειρήσεων, αλλά και μεταξύ αυτών. Η διαθεσιμότητα, η δυναμικότητα και η συνέπεια, είναι μετρήσιμα ποιοτικά στοιχεία και μπορούν να καταγραφούν και να αξιολογηθούν. Αντικειμενικό σκοπό, αποτελεί η αύξηση της συνολικής κερδοφορίας κατά μήκος της αλυσίδας, η οποία βέβαια συνεπάγεται και την αύξηση της κερδοφορίας όλων των εταίρων της.

Συνοψίζοντας, παρουσιάζονται δύο παραδείγματα που αφορούν πραγματικά δεδομένα. Στο πρώτο παράδειγμα, τα δεδομένα αφορούν τηλεφωνική καμπάνια ενός πορτογαλικού τραπεζικού ιδρύματος. Τα δεδομένα είναι μικτά αφού περιέχουν τόσο ποσοτικές (balance,duration) όσο και κατηγορικές μεταβλητές (education,job). Εμφανίζονται οι πρώτες έξι γραμμές του dataset:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
1:	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown
2:	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown
3:	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown
4:	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown
5:	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown
6:	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown

(Ανακτήθηκε από: https://archive.ics.uci.edu/dataset/222/bank+marketing)

Στο δεύτερο παράδειγμα, περιλαμβάνεται μία μίξη πληροφοριών ηλεκτρονικού εμπορίου και μεταφοράς προϊόντων. Τα δεδομένα είναι μικτά αφού περιέχουν τόσο ποσοτικές (Delivery_person_Age, Delivery_person_Ratings) όσο και κατηγορικές μεταβλητές (Delivery_person_ID, Road_traffic_density). Εμφανίζονται οι πρώτες έξι γραμμές του dataset:

	ID	Delivery_person_ID	Delivery_person_Age	Delivery_person_Ratings	Restaurant_latitude	Restaurant_longitude	Delivery_location_latitude	Delivery_location_longitude	Order_Date	Time_Orderd	Time_Order_picked	Weatherconditions	Road_traffic_density	Vehicle_condition	Type_of_order	Type_of_vehicle	multiple_deliveries	Festival	City	Time_taken(min)
1:	0x4607	INDORES13DEL02	37	4.9	22.74505	75.89247	22.76505	75.91247	19-03-2022	11:30:00	11:45:00	conditions Sunny	High	2	Snack	motorcycle	0	No	Urban	(min) 24
2:	0xb379	BANGRES18DEL02	34	4.5	12.91304	77.68324	13.04304	77.81324	25-03-2022	19:45:00	19:50:00	conditions Stormy	Jam	2	Snack	scooter	1	No	Metropolitian	(min) 33
3:	0x5d6d	BANGRES19DEL01	23	4.4	12.91426	77.67840	12.92426	77.68840	19-03-2022	08:30:00	08:45:00	conditions Sandstorms	Low	0	Drinks	motorcycle	1	No	Urban	(min) 26
4:	0x7a6a	COIMBRES13DEL02	38	4.7	11.00367	76.97649	11.05367	77.02649	05-04-2022	18:00:00	18:10:00	conditions Sunny	Medium	0	Buffet	motorcycle	1	No	Metropolitian	(min) 21
5:	0x70a2	CHENRES12DEL01	32	4.6	12.97279	80.24998	13.01279	80.28998	26-03-2022	13:30:00	13:45:00	conditions Cloudy	High	1	Snack	scooter	1	No	Metropolitian	(min) 30
6:	0x9bb4	HYDRES09DEL03	22	4.8	17.43167	78.40832	17.46167	78.43832	11-03-2022	21:20:00	21:30:00	conditions Cloudy	Jam	0	Buffet	motorcycle	1	No	Urban	(min) 26

(Ανακτήθηκε από: https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset)

ΚΕΦΑΛΑΙΟ 3

Ανάλυση κατά συστάδες

3.1 Τί είναι η ανάλυση κατά συστάδες

Η ανάλυση κατά συστάδες ή ανάλυση σε ομάδες είναι μία στατιστική τεχνική που έχει ως στόχο να κατανείμει σε ομάδες, τις παρατηρήσεις από ένα σύνολο δεδομένων, σύμφωνα με την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Η συγκεκριμένη ανάλυση μπορεί να επιτευχθεί μέσω της ομοιογένειας που υπάρχει μεταξύ των παρατηρήσεων. Πιο συγκεκριμένα, μία επιτυχημένη ομαδοποίηση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις σε κάθε ομάδα που έχει δημιουργηθεί, να είναι όσο γίνεται πιο όμοιες μεταξύ τους. Από την άλλη μεριά, οι παρατηρήσεις που υπάρχουν σε διαφορετικές ομάδες, θα πρέπει να διαφέρουν όσο το δυνατόν περισσότερο (Κούτρας (2020)).

Ο διαμερισμός ατόμων ή αντικειμένων σε ομάδες είναι μία διαδικασία που ο άνθρωπος μαθαίνει από μικρή ηλικία. Αυτό είναι εύκολα αντιληπτό εάν αναλογιστεί κανείς ότι ένα παιδί μπορεί να διαχωρίσει το αυτοκίνητο από το ποδήλατο, το τρένο από το αεροπλάνο, την τηλεόραση από το ραδιόφωνο, τις γυναίκες από τους άνδρες κ.ά. Συνεπώς, είναι μία ρουτίνα στην καθημερινότητα του ανθρώπου, που γίνεται αυτόματα, ενώ επιπρόσθετα υπάρχει και η τάση να δημιουργεί επιπλέον υποομάδες αυτών των ομάδων.

Η ανάπτυξη των τεχνικών που αφορά την ανάλυση κατά συστάδες ξεκίνησε την δεκαετία του 1960, εντούτοις δεν υπήρχαν τα υπολογιστικά μέσα για την εφαρμογή τους. Η τεχνολογία έχει σημειώσει τεράστια πρόοδο τις τελευταίες δεκαετίες, όσον αφορά τα στατιστικά πακέτα, την ανάπτυξη μικροϋπολογιστών κ.λ.π. Κατά συνέπεια, η εφαρμογή των συγκεκριμένων τεχνικών μπορεί να γίνει με μικρό υπολογιστικό κόστος και σε δεδομένα που έχουν πληθώρα χαρακτηριστικών.

Με την πάροδο των χρόνων, όλο και περισσότερες επιστήμες κάνουν χρήση των τεχνικών της ανάλυσης κατά συστάδες. Ακολούθως, θα δώσουμε κάποια παραδείγματα επιστημών που έχει καταλήξει χρήσιμη η ομαδοποίηση δεδομένων:

- Στις Ιατρικές επιστήμες, η ανάλυση κατά συστάδες βοηθάει στην διάγνωση, θεραπεία, αλλά ακόμα και στην εύρεση βέλτιστων μεθόδων θεραπείας.
- Στις Κοινωνικές επιστήμες, ο ερευνητής ενδιαφέρεται να ομαδοποιήσει άτομα ή κοινωνίες ή ακόμα και μοτίβα συμπεριφορών.
- Στις Βίο-επιστήμες, η ανάλυση σε συστάδες οδηγεί στον καθορισμό των ειδών που συναντάμε στη φύση.
- Στις Οικονομικές επιστήμες, λόγου χάρη, στο μάρκετινγκ καθίσταται χρήσιμη η ομαδοποίηση σε έναν μεγάλο αριθμό δυνητικών πελατών, οι οποίοι προσεγγίζονται στη συνέχεια σύμφωνα με τις ανάγκες τους σε συγκεκριμένη περιοχή προϊόντων.
- Στην επιστήμη των Υπολογιστών, η ανάλυση κατά συστάδες θεωρείται απαραίτητη για την εφαρμογή των μεθόδων της τεχνητής νοημοσύνης, της ρομποτικής, των νευρωνικών δικτύων, αυτών που χρησιμοποιούνται για την εξέταση δαχτυλικών αποτυπωμάτων, αναγνώριση προσώπου ή φωνής κ.ά. Ένας σημαντικός πυλώνας της συγκεκριμένης επιστήμης, που παίζει ρόλο η ομαδοποίηση δεδομένων, είναι το διαδίκτυο. Η συμπεριφορά

κάθε χρήστη του διαδικτύου ομαδοποιείται ανάλογα με τις σελίδες που επισκέπτεται, τον χρόνο παραμονής του και τις αναζητήσεις που κάνει.

Σε γενικές γραμμές, η ομαδοποίηση δεδομένων είναι ο διαμερισμός ατόμων ή αντικειμένων σύμφωνα με ένα σύνολο κανόνων, το οποίο θα πρέπει να κρίνεται ως προς την χρησιμότητα του. Για παράδειγμα, εάν θέλουμε να ομαδοποιήσουμε ένα σύνολο δεδομένων που απαρτίζεται από βιβλία, θα ήταν πιο χρήσιμο να γίνει με βάσει το είδος (π.χ. βιογραφία, μυθιστόρημα), τον συγγραφέα και την χρονολογία, παρά με το χρώμα που έχει το εξώφυλλο κάθε βιβλίου. Τελικά, ο εκάστοτε ερευνητής θα πρέπει να αναγνωρίζει εάν η ομαδοποίηση που προέκυψε έχει κάποιο νόημα ή προέκυψε από τη γεωμετρία των δεδομένων, π.χ. να είναι αποτέλεσμα από μίξη ετερογενών δεδομένων.

Προηγουμένως, αναφέραμε πως μία επιτυχημένη ομαδοποίηση θα καταλήξει σε ομάδες, όπου οι παρατηρήσεις σε καθεμία είναι όσο το δυνατόν πιο όμοιες. Η ομοιότητα των παρατηρήσεων καθορίζεται από τη φύση του προβλήματος του εκάστοτε ερευνητή και από τα δεδομένα που έχει συλλέξει για την επίλυση του προβλήματος. Για τον λόγο αυτό, έχει προταθεί η χρήση ενός μέτρου ομοιότητας ή ενός μέτρου απόστασης. Ένα μέτρο απόστασης, θα πρέπει να δίνει μία πολύ μικρή τιμή για τις παρατηρήσεις που μοιάζουν μεταξύ τους. Αντίθετα, ένα μέτρο ομοιότητας θα πρέπει να δίνει μεγάλη τιμή για παρατηρήσεις που μοιάζουν μεταξύ τους. Αντίστοιχα, για ανόμοιες παρατηρήσεις θα πρέπει να δίνει αντίθετα αποτελέσματα.

Σε σχέση με τα μέτρα απόστασης, ας υποθέσουμε πως έχουμε $\mathbf{x}_i$ και $\mathbf{x}_j$ δύο άτομα/αντικείμενα του συνόλου δεδομένων. Ένα μέτρο απόστασης συμβολίζεται ως εξής: $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ και ικανοποιεί τις παρακάτω τρεις ιδιότητες:

$$\begin{aligned} &\text{I}_1.\ d_{ij} \geq 0\ \forall\ i,j \text{ και } d_{ij} = 0 \Leftrightarrow i = j, \\ &\text{I}_2.\ d_{ij} \leq d_{is} + d_{sj} \text{ (τριγωνική ανισότητα)}, \\ &\text{I}_3.\ d_{ij} = d_{ji} \quad \text{(συμμετρική ιδιότητα)}. \end{aligned}$$

Στην πράξη, η δεύτερη ιδιότητα (τριγωνική) δεν ικανοποιείται σε όλα τα μέτρα απόστασης και η τρίτη ιδιότητα (συμμετρική) μπορεί να μην γίνει απαιτητή. Ας υποθέσουμε πως έχουμε p μεταβλητές και ας συμβολίσουμε με s_r τη διακύμανση της r μεταβλητής και w_r ένα μη αρνητικό βάρος για αυτήν. Τα πιο γνωστά μέτρα απόστασης είναι τα εξής:

- Ευκλείδεια απόσταση: $d_{ij} = \sqrt{\sum_{r=1}^{p}(x_{ir} - x_{jr})^2}$
- Σταθμισμένη Ευκλείδεια απόσταση: $d_{ij} = \sqrt{\sum_{r=1}^{p} w_r(x_{ir} - x_{jr})^2}$
- Απόσταση του Pearson: $d_{ij} = \sqrt{\sum_{r=1}^{p}\left(\frac{x_{ir} - x_{jr}}{s_r}\right)^2}$
- Απόσταση Manhattan: $d_{ij} = \sum_{r=1}^{p}|x_{ir} - x_{jr}|$

Η πιο δημοφιλής απόσταση είναι η Ευκλείδεια. Πολλές φορές χρησιμοποιείται το τετράγωνο της Ευκλείδειας απόστασης, το οποίο φυσικά δεν είναι μέτρο απόστασης αφού δεν ικανοποιεί τη I_2. Αξίζει να σημειωθεί πως η απόσταση του Pearson είναι μία ειδική περίπτωση της σταθμισμένης Ευκλείδειας απόστασης. Τέλος, η απόσταση Manhattan θα μπορούσαμε να πούμε πως μοιάζει με την Ευκλείδεια, με μόνη διαφορά ότι χρησιμοποιούμε απόλυτες αποκλίσεις, αντί για τετραγωνικές. Επίσης, δίνει παρόμοια αποτελέσματα με την Ευκλείδεια,

πέρα από τις περιπτώσεις που έχουμε ακραίες παρατηρήσεις, όπου δίνει πιο ανθεκτικά αποτελέσματα.

Στην συνέχεια, θα παρουσιάσουμε κάποια μέτρα ομοιότητας που μπορούν να χρησιμοποιηθούν εναλλακτικά των μέτρων αποστάσεων. Υπολογίζοντας κάποιο κατάλληλο μέτρο ομοιότητας για το σύνολο δεδομένων που εξετάζεται, μπορούν να ταξινομηθούν οι παρατηρήσεις σε ίδια ομάδα (εάν η τιμή είναι μεγάλη) ή σε διαφορετική ομάδα (εάν η τιμή είναι μικρή). Ένα μέτρο ομοιότητας συμβολίζεται συνήθως ως: $s_{ij} = s(x_i, x_j)$ και ικανοποιεί τις παρακάτω τρεις ιδιότητες:

$$\text{I}_1.\ s_{ij} \geq 0\ \forall\ i,j\ \kappa\alpha\iota\ i = j \Rightarrow s_{ij} = 0,$$
$$\text{I}_2.\ s_{ij} \leq 1,$$
$$\text{I}_3.\ s_{ij} = s_{ji}.$$

Για ποσοτικά δεδομένα, το πιο γνωστό μέτρο απόστασης είναι η απόλυτη τιμή του δειγματικού συντελεστή συσχέτισης και ορίζεται ως εξής:

$$s_{ij} = \frac{|\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i\cdot})(x_{jr} - \bar{x}_{j\cdot})|}{\left(\sum_{r=1}^{p}(x_{ir} - \bar{x}_{i\cdot})^2 \sum_{r=1}^{p}(x_{jr} - \bar{x}_{j\cdot})^2\right)^{1/2}},$$

όπου $\bar{x}_{i\cdot} = \frac{1}{p}\sum_{r=1}^{p} x_{ir}$ και $\bar{x}_{j\cdot} = \frac{1}{p}\sum_{r=1}^{p} x_{jr}$.

Για μικτά δεδομένα, όσον αφορά την ομοιότητα, προτάθηκε από τον Gower το εξής μέτρο ομοιότητας:

$$s_{ij} = \frac{\sum_{r=1}^{p} w_{ij}(r) s_{ij}(r)}{\sum_{r=1}^{p} w_{ij}(r)}.$$

Εάν η μεταβλητή *r* είναι συνεχής, ορίζεται το εύρος της ως $R_r = max\ x_{ir} - min\ x_{ir}$, για κάθε $i = 1,2,\ldots,p$ και θέτουμε $s_{ij}(r) = 1 - \frac{|x_{ir} - x_{jr}|}{R_r}$. Εάν η μεταβλητή *r* είναι διακριτή, θέτουμε $s_{ij}(r) = 1$ εάν $x_{ir} = x_{jr}$ και $s_{ij}(r) = 0$ εάν $x_{ir} \neq x_{jr}$. Τέλος, ανάλογα με το εάν η σύγκριση στην *r* μεταβλητή έχει νόημα ή όχι, τα βάρη θα παίρνουν την τιμή 1 ή 0.

Αξίζει να σημειωθεί ότι στην περίπτωση που έχουμε δημιουργήσει ένα μέτρο απόστασης, τότε ένα αντίστοιχο μέτρο ομοιότητας (και αντίστροφα) προκύπτει από τον τύπο:

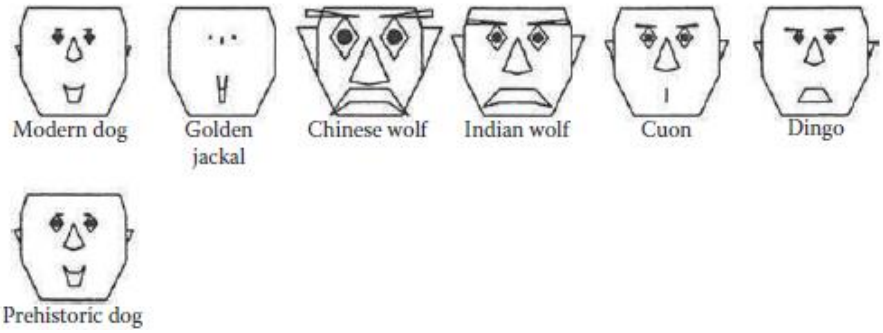
$$s_{ij} = \frac{1}{1 + d_{ij}}.$$

3.2 Τεχνικές ομαδοποίησης για αριθμητικά δεδομένα

Το πρόβλημα της ομαδοποίησης, ειδικά για αριθμητικά δεδομένα, μπορεί να προσεγγιστεί με ποικίλους τρόπους. Οι εμπειρικοί τρόποι βασίζονται στις γραφικές αναπαραστάσεις των παρατηρήσεων. Από την άλλη μεριά, έχουν αναπτυχθεί μέθοδοι, που έχουν μαθηματική ή στατιστική βάση. Οι συγκεκριμένες θεωρούνται κατά βάση πιο αποδοτικές από τις εμπειρικές και μπορούν να χωριστούν σε ιεραρχικές, μη ιεραρχικές, αυτές που βασίζονται στην πυκνότητα κ.ά.

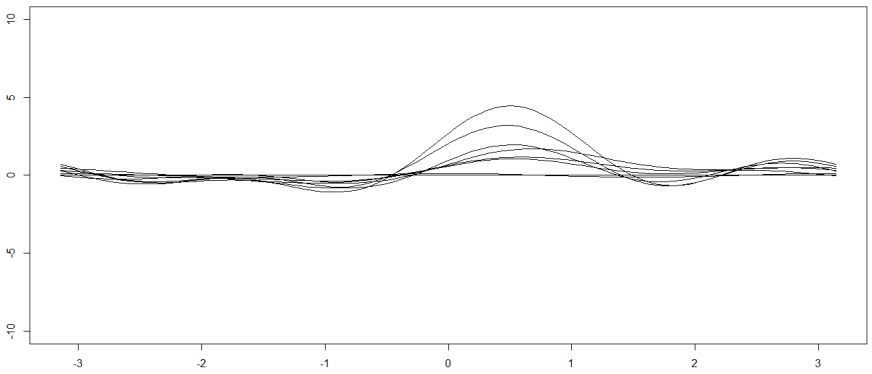
3.2.1. Εμπειρικές μέθοδοι ομαδοποίησης για αριθμητικά δεδομένα

Στις εμπειρικές τεχνικές, χρησιμοποιούνται γραφικές παραστάσεις για την αναπαράσταση των πολυδιάστατων δεδομένων. Σε αυτές τις τεχνικές συμπεριλαμβάνονται τα sun ray plots, Glyphs, Chernoff faces όπως και οι καμπύλες Andrews. Στα Chernoff faces κάθε πρόσωπο είναι ένα αντικείμενο/άτομο που μελετάμε και κάθε χαρακτηριστικό του προσώπου όπως τα μάτια, το στόμα κ.λ.π. αντιπροσωπεύει τις τιμές των μεταβλητών. Έτσι, διαφοροποιώντας το σχήμα, τον προσανατολισμό, το μήκος κ.λ.π. των χαρακτηριστικών του προσώπου, μπορεί να γίνει ομαδοποίηση των αντικειμένων/ατόμων που μελετάμε. Το παρακάτω σχήμα δείχνει έξι ράτσες σκύλων όπου θέλουμε με βάση κάποιες μετρήσεις κάτω γνάθου να αναγνωρίσουμε ποια από αυτές είναι πιο όμοια με τον προϊστορικό σκύλο. Φαίνεται πως ο προϊστορικός σκύλος μοιάζει περισσότερο με το μοντέρνο σκύλο και λιγότερο με τον κινέζικο λύκο.



Σχήμα 1: Γραφική αναπαράσταση μετρήσεων κάτω γνάθου σε διαφορετικές ράτσες σκύλων-Chernoff faces από τους Manly και Navarro Alberto (2017).

Αντίστοιχα, για το ίδιο παράδειγμα με τις ράτσες σκύλων, δημιουργήσαμε τις καμπύλες Andrews μέσω του προγράμματος RStudio. Κάθε ράτσα σκύλου απεικονίζεται από μία καμπύλη. Φαίνεται ότι τέσσερις ράτσες σκύλων μοιάζουν περισσότερο συγκριτικά με τις υπόλοιπες.



Σχήμα 2: Γραφική αναπαράσταση μετρήσεων κάτω γνάθου σε διαφορετικές ράτσες σκύλων-Καμπύλες Andrews

3.2.2. Μη ιεραρχικές μέθοδοι ομαδοποίησης για αριθμητικά δεδομένα

Στις μη ιεραρχικές μεθόδους, ο αριθμός των ομάδων είναι γνωστός εκ των προτέρων. Ένα άλλο χαρακτηριστικό, είναι ότι γίνεται χρήση της έννοιας του κέντρου (*centroid*), το οποίο είναι το διάνυσμα των μέσων τιμών των παρατηρήσεων για κάθε μεταβλητή που υπάρχουν στην ομάδα. Στις συγκεκριμένες μεθόδους, χρησιμοποιείται ένας επαναληπτικός αλγόριθμος με τον οποίο γίνεται η τοποθέτηση των παρατηρήσεων στις ομάδες, σύμφωνα με το ποια ομάδα είναι πιο κοντά στην κάθε παρατήρηση. Η τοποθέτηση της κάθε παρατήρησης γίνεται με βάση την απόστασή της από το κέντρο της ομάδας. Σε κάθε επανάληψη ή βήμα του εκάστοτε αλγόριθμου, υπολογίζεται εκ νέου το κέντρο της ομάδας και ξανά γίνεται ταξινόμηση των παρατηρήσεων στην πιο κοντινή ομάδα.

Από τις πιο γνωστές μη ιεραρχικές μεθόδους, είναι η μέθοδος *k*-means. Το κέντρο βάρους για κάθε ομάδα υπολογίζεται κάθε φορά με την τοποθέτηση ενός νέου αντικειμένου/ατόμου. Ο *k*-means μπορεί να περιγραφεί αλγοριθμικά με τα επόμενα βήματα:

Βήμα 1ο: Από ένα σύνολο δεδομένων, καθόρισε *k* μητρικά σημεία για κάθε ομάδα.

Βήμα 2ο: Κατάταξε κάθε αντικείμενο/άτομο στην ομάδα με την οποία το αντικείμενο/άτομο έχει την μικρότερη απόσταση από το κέντρο της. Κάθε φορά που γίνεται τοποθέτηση αντικειμένου/ατόμου επανυπολόγισε το κέντρο της ομάδας.

Βήμα 3ο: Όταν έχει ολοκληρωθεί το δεύτερο βήμα, τα κέντρα που έχουν δημιουργηθεί θεωρούνται ως μητρικά σημεία και εκτέλεσε μία τελική σάρωση στα δεδομένα τοποθετώντας κάθε αντικείμενο/άτομο στο κοντινότερο μητρικό σημείο.

Μία εναλλακτική προσέγγιση του *k*-means είναι να γίνεται η τοποθέτηση όλων των αντικειμένων/ατόμων στις ομάδες και μετά υπολογίζονται τα κέντρα. Αυτή η προσέγγιση μπορεί να περιγραφθεί αλγοριθμικά με τα επόμενα βήματα:

Βήμα 1ο: Διαμέρισε το σύνολο δεδομένων σε *k* ομάδες και υπολόγισε τα κέντρα.

Βήμα 2ο: Υπολόγισε τις αποστάσεις κάθε αντικειμένου/ατόμου από το κέντρο κάθε ομάδας. Κατάταξε κάθε αντικείμενο/άτομο στην ομάδα που έχει το κέντρο με την μικρότερη απόσταση. Υπολόγισε εκ νέου τα κέντρα των ομάδων που επηρεάσθηκαν από την καινούρια ταξινόμηση.

Βήμα 3ο: Επανάλαβε το δεύτερο βήμα μέχρι να υπάρξει επανάληψη που δεν θα διαφέρουν οι ομάδες από την προηγούμενη επανάληψη.

Ένα από τα μεγαλύτερα πλεονεκτήματα του *k*-means είναι η ταχύτητα που ολοκληρώνεται, αφού δεν χρειάζεται πολλές επαναλήψεις. Επιπλέον, δεν έχει μεγάλο υπολογιστικό κόστος επειδή δεν κρατάει στην μνήμη του υπολογιστή πολλά στοιχεία. Από την άλλη μεριά, οι ομάδες που προκύπτουν επηρεάζονται σε μεγάλο βαθμό από τα αρχικά μητρικά σημεία, το οποίο μπορεί να οδηγήσει σε λάθος αποτελέσματα. Επίσης, επηρεάζεται από ακραίες παρατηρήσεις, οι οποίες μπορεί να δημιουργήσουν ομάδες με διάσπαρτες παρατηρήσεις.

Σε γενικές γραμμές, ο αλγόριθμος *k*-means καταλήγει σε ομαδοποίηση των δεδομένων που είναι πολύ κοντά στην τελική λύση, σε μόλις λίγες επαναλήψεις. Έχει παρατηρηθεί πως αυτό είναι ωφέλιμο για μεγάλα σύνολα δεδομένων. Στις περιπτώσεις που χρειάζονται αρκετές επαναλήψεις, έχει παρατηρηθεί πως στις τελευταίες, οι διαφοροποιήσεις είναι συνήθως ελάχιστες και προκύπτουν σε παρατηρήσεις που βρίσκονται στα όρια μεταξύ δύο ομάδων. Τέλος, τις περισσότερες φορές, η ομαδοποίηση που καταλήγει ο αλγόριθμος, δημιουργεί ισοπληθείς ομάδες.

3.2.3. Ιεραρχικές μέθοδοι ομαδοποίησης για αριθμητικά δεδομένα

Οι ιεραρχικές μέθοδοι ομαδοποίησης χωρίζονται σε δύο κατηγορίες: στις συσσωρευτικές (*agglomerative*) και στις διαιρετικές μεθόδους (*divisive*). Στις συσσωρευτικές μεθόδους, κάθε αντικείμενο/άτομο αντιστοιχεί σε μία ομάδα και με συνεχόμενες συγχωνεύσεις, ομαδοποιούνται σε μία ομάδα. Αντίθετα, οι διαιρετικές μέθοδοι ξεκινάνε με μία ομάδα που εμπεριέχει όλα τα αντικείμενα/άτομα και διαιρώντας την, καταλήγουν σε ομάδες που αποτελούνται από ένα μόνο αντικείμενο/άτομο. Θα παρουσιάσουμε αποκλειστικά τις συσσωρευτικές μεθόδους, διότι είναι οι πιο δημοφιλείς.

Αλγοριθμικά, οι συσσωρευτικές μέθοδοι, μπορούν να περιγραφθούν ως εξής:

Βήμα 1ο: Ανάθεσε κάθε αντικείμενο/άτομο σε μία ομάδα και δημιούργησε τον πίνακα αποστάσεων όλων των αντικειμένων/ατόμων μεταξύ τους.

Βήμα 2ο: Εντόπισε στον πίνακα των αποστάσεων το κοντινότερο ζευγάρι ομάδων.

Βήμα 3ο: Συνένωσε το κοντινότερο ζευγάρι ομάδων σε μία καινούρια ομάδα. Επανυπολόγισε τον πίνακα αποστάσεων, διαγράφοντας τις γραμμές και τις στήλες που αφορούσαν τις ομάδες που ενώθηκαν και προσθέτοντας μία γραμμή και μία στήλη που θα περιέχει τις αποστάσεις για την καινούρια ομάδα από τις υπόλοιπες ομάδες.

Βήμα 4ο: Επανάλαβε το δεύτερο και τρίτο βήμα, μέχρι να δημιουργηθεί μία ομάδα που περιέχει όλες τις παρατηρήσεις. Σε κάθε βήμα, κατέγραψε τις λεπτομέρειες που αφορούν τις συγχωνεύσεις.

Αξίζει να σημειωθεί πως αντί για πίνακα αποστάσεων, ο εκάστοτε ερευνητής μπορεί να χρησιμοποιήσει εναλλακτικά έναν πίνακα με μέτρα ομοιότητας.

Ενώ στην Ενότητα 3.1 έχουμε ορίσει μέτρα απόστασης μεταξύ στοιχείων, παρατηρούμε πως χρειαζόμαστε μέτρα αποστάσεων μεταξύ ομάδων. Στη συνέχεια, θα ορίσουμε μερικές διαδεδομένες τεχνικές για τον καθορισμό των αποστάσεων μεταξύ ομάδων. Η πιο γνωστή τεχνική είναι η μέθοδος της απλής συνένωσης, όπου η απόσταση μεταξύ δύο ομάδων ορίζεται ως η απόσταση των κοντινότερων σημείων ανάμεσα σε αυτές τις ομάδες. Αντίστοιχα, η μέθοδος της πλήρης συνένωσης ορίζει την απόσταση μεταξύ δύο ομάδων ως η απόσταση των πιο μακρινών σημείων ανάμεσα σε αυτές τις ομάδες. Επιπροσθέτως, υπάρχει η μέθοδος της μέσης συνένωσης, όπου η απόσταση μεταξύ δύο ομάδων υπολογίζεται ως ο μέσος όρος των αποστάσεων μεταξύ όλων των ζευγών παρατηρήσεων ανάμεσα σε αυτές τις ομάδες. Τέλος, υπάρχει η μέθοδος των σταθμισμένων μέσων, όπου η απόσταση είναι ο μέσος όρος των αποστάσεων όλων των σημείων της μίας ομάδας με την άλλη.

Ένα από τα μεγαλύτερα πλεονεκτήματα των ιεραρχικών μεθόδων είναι πως δεν χρειάζεται να προκαθοριστεί ο αριθμός των ομάδων που θα δημιουργηθούν. Όμως, οι συσσωρευτικές μέθοδοι απαιτούν πολύ χρόνο και υπολογιστική ισχύ, ειδικά για μεγάλα σύνολα δεδομένων. Επίσης, η σύνθεση των ομάδων δεν μπορεί να αναστραφεί, το οποίο σημαίνει ότι οι παρατηρήσεις που ενώνονται σε αρχικά βήματα, δεν μπορούν να χωριστούν σε επόμενα.

Γενικά, οι ιεραρχικές μέθοδοι δεν καταλήγουν σε συγκεκριμένο πλήθος ομάδων και καλείται ο εκάστοτε ερευνητής να επιλέξει πόσες ομάδες θα κρατήσει, μέσω δενδρογράμματος ή κάποιας άλλης τεχνικής. Συνήθως, γίνεται χρήση ενός δενδρογράμματος, όπου φαίνονται γραφικά οι ομάδες που δημιουργούνται σε όλα τα στάδια, όπως και οι αποστάσεις που έγιναν οι συνενώσεις.

3.2.4. DBSCAN για αριθμητικά δεδομένα

Σε αυτήν την ενότητα, θα αναλύσουμε τον αλγόριθμο DBSCAN (*Density Based Spatial Clustering of Applications with Noise*), ο οποίος είναι από τους πιο βασικούς αλγορίθμους που βασίζονται στην πυκνότητα των δεδομένων (Ester, Kriegel, Sander και Xu (1996)). Ο DBSCAN είναι ικανός να καταλήξει σε δημιουργία ομάδων οι οποίες έχουν αυθαίρετα σχήματα, δηλαδή μπορεί να διαχειριστεί ομάδες διαφορετικών μεγεθών/σχημάτων. Το κεντρικό νόημα του συγκεκριμένου αλγορίθμου είναι ότι για κάθε σημείο μίας ομάδας, πρέπει να υπάρχει μία γειτονιά (με δεδομένη ακτίνα), μέσα στην οποία θα πρέπει να περιέχεται ένας ελάχιστος αριθμός σημείων. Αυτό σημαίνει ότι η πυκνότητα της γειτονιάς θα πρέπει να υπερβαίνει ένα ορισμένο όριο.

Ακολούθως, θα ορίσουμε κάποιες έννοιες που χρειάζονται για την ανάλυση του αλγορίθμου:

- MinPts: ελάχιστος αριθμός σημείων μέσα στην ομάδα
- Eps: για κάθε σημείο της ομάδας θα πρέπει να υπάρχει ένα άλλο σημείο της ομάδας με απόσταση μικρότερη του Eps
- Eps-γειτονιά: τα σημεία μέσα σε Eps απόσταση από ένα σημείο
- Πυρήνας: ένα σημείο με αρκετά πυκνή Eps-γειτονιά, δηλαδή τέτοιο ώστε πληθυσμός ≥ MinPts
- Σύνορο: ένα σημείο το οποίο δεν έχει ικανή Eps-γειτονιά ώστε να ονομαστεί πυρήνας, αλλά ανήκει στην Eps-γειτονιά κάποιου πυρήνα
- Θόρυβος: ένα σημείο το οποίο δεν είναι ούτε πυρήνας ούτε σύνορο

Οι έννοιες MinPts και Eps είναι απαραίτητο να καθοριστούν πριν ξεκινήσει ο αλγόριθμος, δηλαδή δίνονται σαν είσοδοι στον αλγόριθμο. Τα βήματα του DBSCAN είναι τα εξής:

Βήμα 1^{ο}: Επέλεξε ένα τυχαίο αρχικό σημείο και δημιούργησε την Eps-γειτονιά του.

Βήμα 2^{ο}: Εφόσον έχει δημιουργηθεί μία επαρκής γειτονιά γύρω από το σημείο, συνέχισε την διαδικασία και όρισε το σημείο σαν διαβασμένο. Διαφορετικά, όρισε το σημείο σαν ακραία τιμή.

Βήμα 3^{ο}: Εάν ένα σημείο είναι μέρος της κατηγορίας, τότε όλα τα σημεία της Eps-γειτονιάς του τοποθετούνται στην κατηγορία και επανάλαβε τα προηγούμενα βήματα για κάθε σημείο της γειτονιάς, μέχρι να μην υπάρχουν καινούρια σημεία που να μπορούν να μπουν σε αυτήν την κατηγορία.

Βήμα 4^{ο}: Επέλεξε ένα καινούριο σημείο που δεν έχει ορισθεί σαν διαβασμένο και επανάλαβε όλα τα προηγούμενα βήματα για να δημιουργηθεί καινούρια κατηγορία. Ο αλγόριθμος ολοκληρώνεται όταν όλα τα σημεία έχουν ορισθεί σαν διαβασμένα.

Από τα βασικότερα πλεονεκτήματα του DBSCAN είναι πως δεν απαιτείται ο εκ των προτέρων καθορισμός των ομάδων. Επίσης, έχει καλή αντοχή στον θόρυβο και τις ακραίες τιμές. Οι ακραίες τιμές, είναι εύκολα κατανοητό, πως προσδιορίζονται κατά την διάρκεια της διαδικασίας της ομαδοποίησης, και είναι αυτές που θα έχουν σαν αποτέλεσμα τη δημιουργία μίας ομάδας. Στον αντίποδα, ο DBSCAN δεν μπορεί να οδηγήσει σε καλή ομαδοποίηση όταν διαχειρίζεται δεδομένα με μεγάλες διαφορές στις πυκνότητες, αφού οι MinPts και Eps δεν θα μπορούν να ορισθούν κατάλληλα. Επιπροσθέτως, δεν αποδίδει καλά σε σύνολα δεδομένων που έχουν μεγάλες διαστάσεις.

3.3 Μέθοδος *k*-means για μικτά δεδομένα

Ο κλασσικός αλγόριθμος *k*-means δεν μπορεί να εφαρμοστεί σε μικτά δεδομένα, καθώς τα τελευταία περιλαμβάνουν κατηγορικές μεταβλητές και δεν υπάρχει δυνατότητα υπολογισμού του μέτρου απόστασης. Για να εφαρμοστεί ο αλγόριθμος *k*-means σε μικτά δεδομένα, θα πρέπει να γίνει κατάλληλη τροποποίηση, ώστε να ενσωματωθούν οι κατηγορικές μεταβλητές στη μετρική που θα χρησιμοποιηθεί. Θα αναφέρουμε στη συνέχεια, μια πρόταση η οποία έγινε από τους Ahmad και Dey (2007) για την εφαρμογή του αλγορίθμου *k*-means σε μικτά δεδομένα.

Οι Ahmad και Dey (2007) πρότειναν έναν βελτιωμένο αλγόριθμο *k*-means, μεταβάλλοντας τη συνάρτηση κόστους του Huang (1997), η οποία ελαχιστοποιείται για ομαδοποίηση μικτών δεδομένων, και παρουσιάζοντας ένα καινούριο μέτρο απόστασης. Αναφορικά με το μέτρο απόστασης, βασίστηκαν στην έννοια της συνύπαρξης (*co-occurrence*) των παρατηρήσεων, δηλαδή μίας ποσότητας η οποία εκφράζει πόσο στενά συνδεδεμένες είναι οι παρατηρήσεις μεταξύ τους.

Σύμφωνα με τον Huang (1997), η συνάρτηση κόστους περιλαμβάνει δύο μέρη, ένα μέρος για τις αριθμητικές μεταβλητές και ένα μέρος για τις κατηγορικές μεταβλητές. Υποθέτουμε πως διαθέτουμε παρατηρήσεις για n άτομα $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\mathrm{n}$ και p μεταβλητές, εκ των οποίων οι p_r είναι αριθμητικές μεταβλητές και οι p_c κατηγορικές μεταβλητές. Επομένως ισχύει ότι $p = p_r + p_c$. Επιπλέον, ας θεωρήσουμε k το σύνολο των ομάδων που επιθυμούμε να δημιουργήσουμε με τα αντίστοιχα κέντρα τους $C_1, C_2, \dots, C_k$.

Έστω ότι θα υπολογίσουμε την συνάρτηση κόστους που αφορά ένα άτομο $\mathbf{x}_\mathrm{i}$ με την κοντινότερη ομάδα *j*, μέσω της απόστασης που θα αναλυθεί στη συνέχεια. Ας συμβολίσουμε με x_i^r τις παρατηρήσεις των αριθμητικών μεταβλητών και με x_i^c τις παρατηρήσεις των κατηγορικών μεταβλητών. Το κέντρο της ομάδας *j* ορίζεται ως $C_j = (C_{j1}, C_{j2}, \dots, C_{jp})$. Ας συμβολίσουμε με C_j^r το μέσο όρο για κάθε αριθμητική μεταβλητή και με C_j^c την επικρατέστερη τιμή της κάθε κατηγορικής μεταβλητής. Επιπροσθέτως, για τις κατηγορικές μεταβλητές ως γ_j το βάρος της *j* ομάδας, όπου δηλώνεται από τον εκάστοτε ερευνητή, και ορίζεται ότι η απόσταση $d(l,q)=0$ για $l=q$ και $d(l,q)=1$ για $l \neq q$. Τότε η συνάρτηση κόστους για την ομάδα *j* ορίζεται ως εξής:

$$\zeta_j = \sum_{i=1}^{n}\left(\sum_{t=1}^{p_r}\left(x_{it}^r - C_{jt}^r\right)^2 + \gamma_j \sum_{t=1}^{p_c} d(x_{it}^c, C_{jt}^c)\right).$$

Όμως, λόγω του γεγονότος ότι για τις κατηγορικές μεταβλητές χρησιμοποιείται η επικρατέστερη τιμή, παρατηρήθηκε πως υπήρχε απώλεια πληροφορίας, καθώς καμία άλλη τιμή δεν συμβάλει στον υπολογισμό της απόστασης. Συμπληρωματικά, σε όλες τις αριθμητικές μεταβλητές δεν ορίζεται κάποιο βάρος και το βάρος για τις κατηγορικές μεταβλητές δίνεται από τον εκάστοτε ερευνητή. Κατά συνέπεια, δεν είναι ρεαλιστικό εφόσον κάθε αριθμητική μεταβλητή μπορεί να μην έχει την ίδια επίδραση στην ομαδοποίηση και ο ερευνητής μπορεί να δώσει λάθος τιμή στο βάρος. Τέλος, έχει παρατηρηθεί ότι η απόσταση $d(l,q)=1$ για $l \neq q$ δεν είναι ρεαλιστική, αφού για διαφορετικά ζεύγη l,q θα μπορούσαν να χρησιμοποιηθούν διαφορετικές τιμές ανάλογα από τις σχετικές συχνότητες των ζευγών εντός μίας κατηγορίας, διότι οι σχετικές συχνότητες μπορεί να παρέχουν πληροφορίες σχετικά με τη σημαντικότητα του συγκεκριμένου ζεύγους κατά τη διαφοροποίηση μεταξύ ποικίλων περιπτώσεων σε αυτήν την κατηγορία.

Για αυτόν τον λόγο, οι Ahmad και Dey (2007) τροποποίησαν την προαναφερόμενη συνάρτηση κόστους, προσθέτοντας την w_i που είναι η σημαντικότητα της *i*-στης αριθμητικής μεταβλητής. Σχετικά με το κέντρο της ομάδας, για τις αριθμητικές μεταβλητές δεν αλλάζει

κάτι. Αντίθετα, για τις κατηγορικές μεταβλητές, ορίζουν το κέντρο της ομάδας έτσι ώστε να αντιπροσωπεύεται από την ποσοστιαία κατανομή των τιμών της μεταβλητής που υπάρχουν μέσα στην ομάδα. Έτσι, για κάθε κατηγορική μεταβλητή, όρισαν την $\Omega(x_i^c, C_j^c)$ ως την συνάρτηση της παρατηρηθείσας τιμής και την ποσοστιαία κατανομή όλων των τιμών που βρίσκονται στη ομάδα, έννοια η οποία θα αναλυθεί αργότερα. Τελικά, η συνάρτηση κόστους για την ομάδα *j* ορίζεται ως εξής:

$$\zeta_j = \sum_{i=1}^{n}(\sum_{t=1}^{p_r}\left(w_t\left(x_{it}^r - C_{jt}^r\right)\right)^2 + \sum_{t=1}^{p_c}\Omega(x_{it}^c, C_{jt}^c)^2).$$

Αξίζει να σημειωθεί πως το βάρος για τις αριθμητικές μεταβλητές στη συνάρτηση κόστους, υπολογίζεται από τα ίδια τα δεδομένα και δεν δίνεται από τον εκάστοτε ερευνητή.

Στη συνέχεια, θα αναλύσουμε το μέτρο απόστασης που θα εφαρμοσθεί. Το μέτρο απόστασης που χρησιμοποιείται για την συνάρτηση κόστους σε σχέση με τις αριθμητικές μεταβλητές, είναι η σταθμισμένη Ευκλείδεια απόσταση μεταξύ της *i*-στης παρατήρησης και του κέντρου της ομάδας που εξετάζουμε. Επίσης, γίνεται κανονικοποίηση των αριθμητικών μεταβλητών. Όσον αφορά τις κατηγορικές μεταβλητές, είναι πιο δύσκολο να βρεθεί το μέτρο απόστασης, αφού οι παρατηρήσεις δεν αντιστοιχούν σε μετρήσιμα μεγέθη. Επιπροσθέτως, θα πρέπει να λάβουμε υπόψιν τη σημαντικότητα που έχει η κάθε κατηγορική μεταβλητή, όπως και την ομοιογένεια των δεδομένων. Για να υπολογίσουμε το μέτρο απόστασης για τις κατηγορικές μεταβλητές, χρειαζόμαστε τους κάτωθι ορισμούς.

Ας υποθέσουμε ότι έχουμε δύο τιμές **x** και **y** από μία κατηγορική μεταβλητή A_i σε ένα σύνολο δεδομένων. Για αυτές τις δύο τιμές, λαμβάνουμε υπόψιν τη συνολική τους κατανομή (*overall distribution*) σε συνδυασμό με τη συνύπαρξη που υπάρχει με τις τιμές των υπόλοιπων μεταβλητών. Έστω ότι έχουμε ένα υποσύνολο **w** και το συμπληρωματικό αυτού ~**w** από μία δεύτερη κατηγορική μεταβλητή A_j. Θα χρησιμοποιήσουμε το σύμβολο $P_i(\mathbf{w}|\mathbf{y})$ για να δηλώσουμε τη δεσμευμένη πιθανότητα του ενδεχομένου **w** της A_j δοθέντος ότι στην A_i το στοιχείο έχει τιμή **x**. Αντίστοιχα, το $P_i(\sim\mathbf{w}|\mathbf{y})$ ορίζεται ως η δεσμευμένη πιθανότητα του ενδεχομένου ~**w** της A_j δοθέντος ότι στην A_i το στοιχείο έχει τιμή **y**. Τότε, η απόσταση μεταξύ του ζεύγους των τιμών **x** και **y** της A_i σε σχέση με την κατηγορική μεταβλητή A_j και ένα συγκεκριμένο υποσύνολο **w**, με βάση τις δεσμευμένες πιθανότητες συνύπαρξης των τιμών, είναι:

$$d_{\mathbf{w}}^{i}(\mathbf{x}, \mathbf{y}) = P_i(\mathbf{w}\,|\mathbf{x}) + P_i(\sim\mathbf{w}|\mathbf{y}).$$

Ας συμβολίσουμε με $|A_i|$ τον πληθάριθμο του συνόλου A_i, ο αριθμός των δυνατών τιμών για το **w** θα είναι $2^{|A_i|}$.

Ακολούθως, έστω ότι έχουμε ένα σύνολο **ω** το οποίο είναι ένα υποσύνολο του **w**. Η απόσταση μεταξύ των διανυσμάτων των **x** και **y** της A_i σε σχέση με την κατηγορική μεταβλητή A_j, με βάση τις δεσμευμένες πιθανότητες συνύπαρξης ενός συνόλου τιμών της A_j,ορίζεται ως:

$$d^{ij}(\mathbf{x}, \mathbf{y}) = P_i(\boldsymbol{\omega}\,|\mathbf{x}) + P_i(\sim\boldsymbol{\omega}\,|\mathbf{y}),$$

όπου το **ω** μεγιστοποιεί την ποσότητα $P_i(\boldsymbol{\omega}\,|\mathbf{x}) + P_i(\sim\boldsymbol{\omega}\,|\mathbf{y})$. Εφόσον έχουμε $2^{|A_i|}$ δυνατές τιμές για το **w**, η ποσότητα $P_i(\boldsymbol{\omega}\,|\mathbf{x}) + P_i(\sim\boldsymbol{\omega}\,|\mathbf{y})$ μπορεί να πάρει τιμές από 1 έως 2. Για να παραμείνει η απόσταση $d^{ij}(\mathbf{x}, \mathbf{y})$ μεταξύ 0 και 1, χρησιμοποιείται ο εξής τύπος:

$$d^{ij}(\mathbf{x}, \mathbf{y}) = P_i(\boldsymbol{\omega}\,|\mathbf{x}) + P_i(\sim\boldsymbol{\omega}\,|\mathbf{y}) - 1.$$

Για καλύτερη κατανόηση, θα παρουσιάσουμε ένα παράδειγμα. Έστω ότι διαθέτουμε το παρακάτω σύνολο δεδομένων:

	Director	Actor	Genre
t1 (Godfather II)	Scorsese	De Niro	Crime
t2 (Good Fellas)	Coppola	De Niro	Crime
t3 (vertigo)	Hitchcock	Stewart	Thriller
t4 (N by NW)	Hitchcock	Grant	Thriller
t5 (Bishop's Wife)	Koster	Grant	Comedy
t6 (Harvey)	Koster	Stewart	Comedy

Πίνακας 2: Σύνολο δεδομένων έξι ταινιών από τους Ahmad και Dey (2007)

Το σύνολο δεδομένων απαρτίζεται από έξι ταινίες με τρείς κατηγορικές μεταβλητές: Σκηνοθέτης, Ηθοποιός και Είδος. Ενδεικτικά, θα υπολογιστεί η απόσταση ανάμεσα στη κατηγορία De Niro και Stewart δοθείσης της μεταβλητής Σκηνοθέτης. Εφαρμόζοντας τον τύπο της απόστασης, έχουμε το εξής αποτέλεσμα:

$$d^{Actor,Director}(De\ Niro, Stewart) = P(Scorsese|De\ Niro) + P(Coppola|De\ Niro) + P(Hitchcock|Stewart) + P(Koster|Stewart) - 1 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} - 1 = 1.$$

Παρατηρούμε πως, οι τιμές των δεσμευμένων πιθανοτήτων για κάθε Σκηνοθέτη δεδομένου του De Niro ή του Stewart είναι όλες 0.5, πράγμα που σημαίνει ότι κάθε Σκηνοθέτης είναι εξίσου πιθανό να συνδέεται είτε με De Niro ή Stewart. Δεδομένου ότι δεν υπάρχει Σκηνοθέτης που είναι πιθανότερο να συνδέεται με τον έναν ηθοποιό από τον άλλο, η απόσταση μεταξύ του De Niro και Stewart σε σχέση με τον Σκηνοθέτη είναι 1, που δείχνει ότι δεν υπάρχει κοινός Σκηνοθέτης μεταξύ τους.

Για να υπολογιστεί η απόσταση του ζεύγους των παρατηρήσεων x και y σε σχέση με μία αριθμητική μεταβλητή, θα πρέπει πρώτα να γίνει η διακριτοποίηση της. Θα πρέπει να σημειωθεί πως η διακριτοποίηση των αριθμητικών μεταβλητών, χρησιμοποιείται μόνο για τον υπολογισμό των αποστάσεών τους με μία κατηγορική μεταβλητή και για τον υπολογισμό της σημαντικότητας των μεταβλητών. Για ένα σύνολο με p μεταβλητές (κατηγορικές και κατηγοριοποιημένες αριθμητικές μεταβλητές), η απόσταση για δύο παρατηρήσεις x και y για κάθε κατηγορική μεταβλητή A_i, δίνεται από τον τύπο:

$$d(x,y) = \frac{1}{p-1}\sum d^{ij}(x,y),$$

όπου η άθροιση γίνεται για $j \epsilon \{1,\ldots,p\}$ με $i \neq j$.

Θα πρέπει να αναφερθεί, πως για την απόσταση $d(x,y)$ ισχύουν οι τρεις ιδιότητες, όπως έχουν περιγραφεί στην Ενότητα 3.1.

Όσον αφορά τη σημαντικότητα της κάθε κατηγορικής μεταβλητής, αυτή καθορίζεται από το πόσο καλά είναι διαχωρισμένο κάθε ζευγάρι παρατηρήσεων από τις υπόλοιπες κατηγορικές μεταβλητές. Εφόσον έχει οριστεί το μέτρο απόστασης ανάμεσα σε ζευγάρια παρατηρήσεων στις κατηγορικές μεταβλητές, συμπεραίνεται ότι η σημαντικότητα κάθε μεταβλητής αφομοιώνεται από το αντίστοιχο μέτρο απόστασης.

Στα προηγούμενα αναφέραμε τι γίνεται για τις κατηγορικές μεταβλητές, στη συνέχεια θα δώσουμε αντίστοιχες έννοιες για τις αριθμητικές. Ορίζοντας προηγουμένως τη συνάρτηση κόστους και το μέτρο απόστασης για τις αριθμητικές μεταβλητές, εφαρμόσθηκε κανονικοποίηση και χρησιμοποιήθηκε σταθμισμένη Ευκλείδεια απόσταση. Για να υπολογίσουμε τη σημαντικότητα των αριθμητικών μεταβλητών, θα πρέπει πρώτα να γίνει η

διακριτοποίηση τους. Αρχικά, θα ορίσουμε τον ίδιο αριθμό *S* διαστημάτων για όλες τις αριθμητικές μεταβλητές. Στη συνέχεια, αναθέτουμε μία $u[1], u[2], \ldots, u[S]$ κατηγορική τιμή για κάθε διάστημα. Τέλος, υπολογίζουμε την απόσταση $d(u[r], u[s])$ για κάθε ζεύγος τιμών $u[r], u[s]$ των αριθμητικών μεταβλητών που διακριτοποιήθηκαν, όπως την υπολογίσαμε και για τις κατηγορικές μεταβλητές. Η σημαντικότητα w_i μιας αριθμητικής μεταβλητής A_i υπολογίζεται ως ο μέσος όρος της $d(u[r], u[s])$ για όλα τα $u[r] \neq u[s]$:

$$w_i = \sum_{k=1}^{S} \sum_{j>k}^{S} \frac{d(u[k], u[j])}{S(S-1)/2}.$$

Σε αυτό το σημείο, έχουμε ορίσει το μέτρο απόστασης και την σημαντικότητα τόσο για τις αριθμητικές, όσο και για τις κατηγορικές μεταβλητές. Στην συνέχεια, θα αναπτύξουμε την ανάλυση που αφορά το κέντρο της ομάδας.

Στην περίπτωση του κλασσικού αλγόριθμου *k*-means, ο υπολογισμός του κέντρου της ομάδας δεν γίνεται να συμπεριλάβει όλες τις πληροφορίες από ένα σύνολο μικτών δεδομένων. Για το λόγο αυτό, οι Ahmad και Dey (2007) πρότειναν ένα τροποποιημένο κέντρο ομάδας μέσω του μέτρου απόστασης, το οποίο υπολογίζεται κατά την διάρκεια της ομαδοποίησης. Όσον αφορά τις αριθμητικές μεταβλητές χρησιμοποιείται ο κανονικοποιημένος μέσος όρος (*normalized mean*). Όμως, αυτό δεν μπορεί να υπολογιστεί για τις κατηγορικές. Καθώς η απόσταση μεταξύ δύο κατηγορικών τιμών καθορίζεται από την αντίστοιχη κατανομή τους εντός του συνόλου δεδομένων, κάθε απόσταση θα ποικίλει μεταξύ διαφορετικών ζευγών. Άρα, εάν μία παρατήρηση *r* είναι πιο κοντά σε μία παρατήρηση *s* από μία άλλη παρατήρηση *t*, τότε $d(r, s) < d(r, t)$ και τελικά επιτυγχάνεται καλύτερη ομαδοποίηση για τις παρατηρήσεις *r,s* από τις παρατηρήσεις *r,t*.

Ας υποθέσουμε ότι έχουμε N_C πλήθος ατόμων σε μία ομάδα *C*. Έστω ότι $N_{i,k,C}$ δηλώνει το σύνολο των στοιχείων στη ομάδα, που έχει η *k*-στη τιμή της *i*-στης μεταβλητής, θεωρώντας ότι η *i*-στη μεταβλητή έχει m_i διαφορετικές τιμές. Επομένως το κέντρο της ομάδας των κατηγορικών μεταβλητών μπορεί να οριστεί ως εξής:

$$(1/N_C)[(N_{1,1,C}, N_{1,2,C}, \ldots, N_{1,m_1,C}), (N_{2,1,C}, N_{2,2,C}, \ldots, N_{1,m_2,C}), \ldots, (N_{p,1,C}, N_{p,2,C}, \ldots, N_{p,m_p,C})].$$

Προχωρώντας, όσον αφορά την απόσταση ανάμεσα σε ένα άτομο και το κέντρο μίας ομάδας, αυτή υπολογίζεται από το άθροισμα των αποστάσεων τόσο των αριθμητικών, όσο και των κατηγορικών μεταβλητών. Όπως έχουμε αναφέρει, για τις αριθμητικές μεταβλητές χρησιμοποιείται η Ευκλείδεια απόσταση μεταξύ της τιμής και του κανονικοποιημένου μέσου όρου του κέντρου.

Για τις κατηγορικές μεταβλητές, προτείνεται η παρακάτω που λαμβάνει υπόψη την αναλογική παρουσία κάθε κατηγορικής τιμής στην ομάδα. Έστω $A_{i,k}$ η *k*-στη τιμή της A_i κατηγορικής μεταβλητής και ότι η A_i έχει m_i διαφορετικές τιμές. Υποθέτουμε ότι διαθέτουμε ένα άτομο μόνο με κατηγορικές παρατηρήσεις *X*. Τότε, η απόσταση από την ομάδα *C* ορίζεται ως εξής:

$$\Omega(X, C) = \left(\frac{N_{i,1,C}}{N_C}\right) d\left(X, A_{i,1}\right) + \left(\frac{N_{i,2,C}}{N_C}\right) d\left(X, A_{i,2}\right) + \cdots + \left(\frac{N_{i,m_i,C}}{N_C}\right) d\left(X, A_{i,m_i}\right).$$

Με βάση την πρώτη ιδιότητα που αναφέραμε παραπάνω, ισχύει ότι $d(x, y) \leq 1$. Συνεπώς, αντίστοιχα ισχύει ότι $\Omega(X, C) \leq 1$. Τελικά, η απόσταση μεταξύ ενός ατόμου και του κέντρου της ομάδας C_j υπολογίζεται ως εξής:

$$d(x_i, C_j) = \sum_{t=1}^{p_r}(w_t(x_{it}^r - C_{jt}^r))^2 + \sum_{t=1}^{p_c}(\Omega(x_{it}^C, C_{jt}^C))^2.$$

Εν τέλει, έχοντας ορίσει όλες τις απαραίτητες αποστάσεις που χρειάζονται, ο τροποποιημένος αλγόριθμος *k*-means μπορεί να περιγραφεί αλγοριθμικά με τα επόμενα βήματα:

Βήμα 1ο: Εκχώρησε με τυχαίο τρόπο, όλα τα άτομα σε έναν *k* προκαθορισμένο αριθμό ομάδων.

Βήμα 2ο: Για κάθε κατηγορική μεταβλητή, υπολόγισε την απόσταση $d(r, s)$ που αφορά τις κατηγορικές τιμές *r* και *s*. Για κάθε αριθμητική μεταβλητή, υπολόγισε τη σημαντικότητα της μεταβλητής. Διαμέρισε τα άτομα τυχαία σε διαφορετικές ομάδες.

Βήμα 3ο: Υπολόγισε τα κέντρα για τις ομάδες $C_1, C_2, \ldots, C_k$.

Βήμα 4ο:Κάθε άτομο ταξινομείται στην ομάδα με το κοντινότερο κέντρο χρησιμοποιώντας την απόσταση $d(x_i, C_j)$.

Βήμα 5ο: Εάν τα νέα κέντρα δεν διαφέρουν ή έχεις φτάσει στον προκαθορισμένο αριθμό επαναλήψεων σταμάτα, αλλιώς επανάλαβε τα βήματα 3 και 4.

Αξιοσημείωτο είναι, πως κατά τους Ahmad και Dey (2007), το υπολογιστικό κόστος του παραπάνω αλγορίθμου είναι γραμμικό ως προς το συνολικό πλήθος των ατόμων.

Τέλος, θα αναφέρουμε κάποια μέτρα αξιολόγησης της ομαδοποίησης. Ας υποθέσουμε ότι διαθέτουμε a_i άτομα που ταξινομήθηκαν σε μία ομάδα που αντιστοιχεί στην πραγματική τους κατηγορία, b_i άτομα που ταξινομήθηκαν σε μία ομάδα που δεν αντιστοιχεί στην πραγματική τους κατηγορία και c_i άτομα που έχουν απορριφθεί λανθασμένα από μία ομάδα που αντιστοιχεί στην πραγματική τους κατηγορία. Τα μέτρα αξιολόγησης για την *i*-στη κατηγορία ορίζονται ως εξής:

- Ακρίβεια (*precision):*
$$p_i = a_i/(a_i + b_i)$$
- Ευαισθησία (*recall):*
$$r_i = a_i/(a_i + c_i)$$

Οι Ahmad και Dey (2007) καθόρισαν την απόδοση των αλγορίθμων ομαδοποίησης ως μικροακρίβεια (*micro-p*) και μικροευαισθησία (*micro-r*):

$$micro - p = micro - r = (\sum_{i=1}^{C} a_i)/n.$$

Η μικροακρίβεια και η μικροευαισθησία είναι ισοδύναμες, αφού ισχύει ότι: $\sum_{i=1}^{C} a_i + b_i = \sum_{i=1}^{C} a_i + c_i = n$.

3.4 Μέθοδος KAMILA για μικτά δεδομένα

Λόγω της ανερχόμενης τάσης ύπαρξης μικτών συνόλων δεδομένων με μεγάλο όγκο (*Big Data*) σε πολλούς τομείς, δημιουργήθηκε η ανάγκη για την ανάπτυξη νέων μεθόδων για την διαχείριση των μικτών δεδομένων. Οι Foss, Markatou, Ray και Heching (2016) ανέπτυξαν τον αλγόριθμο KAMILA, ο οποίος βασίζεται σε μία εκτιμήτρια πυκνότητας με την μέθοδο του πυρήνα (*kernel density estimator*), η οποία υπολογίζεται από τα ίδια τα δεδομένα. Ως εκ τούτου η μέθοδος KAMILA είναι ημι-παραμετρική.

Ειδικότερα, ο αλγόριθμος KAMILA για την αποφυγή απώλειας πληροφοριών, εστιάζει στην αρχική μορφή των μεταβλητών, χωρίς κάποιου είδους παραμετροποίηση. Επίσης, όπως στα Gaussian μικτά πολυωνυμικά μοντέλα, παρέχει ισότιμη συνεισφορά τόσο των αριθμητικών, όσο και των κατηγορικών μεταβλητών. Επιπλέον, αποφεύγει τις περιοριστικές παραμετρικές υποθέσεις, όπως ο *k*-means, γενικεύοντας τη μορφή των ομάδων σε μια ευρεία κατηγορία ελλειπτικών κατανομών. Τέλος, δεν ζητάτε από τον εκάστοτε ερευνητή να χρησιμοποιήσει κάποιο σύστημα κωδικοποίησης (*coding scheme*). Συνεπώς, ο αλγόριθμος KAMILA εμπεριέχει τα καλύτερα χαρακτηριστικά των *k*-means και των Gaussian μικτών πολυωνυμικών μοντέλων.

Ωστόσο, έχει παρατηρηθεί, πως στα πολυμεταβλητά δεδομένα, ο υπολογισμός της εκτιμήτριας με την μέθοδο του πυρήνα δημιουργεί πρόβλημα όσον αφορά τον χρόνο υπολογισμού για υψηλών διαστάσεων δεδομένα (*high-dimensional data*) και την υπερπροσαρμογή του δείγματος (*overfitting*). Με τον αλγόριθμο KAMILA, αυτού του είδους τα προβλήματα μπορεί να εξαλειφθούν, εκμεταλλευόμενοι κάποιες ιδιότητες από την σφαιρική κατανομή των ομάδων (*spherically distributed clusters*), όπως θα δούμε στη συνέχεια. Με αυτόν τον τρόπο, ο υπολογισμός της εκτιμήτριας με την μέθοδο του πυρήνα γίνεται με μεγαλύτερη ακρίβεια και πιο γρήγορα. Είναι χρήσιμο να σημειωθεί ότι η σφαιρική κατανομή αναφέρεται σε δεδομένα όπου οι πυκνότητες είναι ακτινωτά συμμετρικές (*radially symmetric*) γύρω από το διάνυσμα μέσου όρου. Αυτό σημαίνει ότι η πυκνότητα εξαρτάται μόνο από την απόσταση του δείγματος από το κέντρο της κατανομής.

Η μέθοδος KAMILA μπορεί να εφαρμοστεί και για ελλειπτικές ομάδες, όπως θα δούμε αργότερα. Όμως, ο εκάστοτε ερευνητής θα πρέπει να προσδιορίσει εκ των προτέρων εάν θα γίνει η χρήση του αλγορίθμου KAMILA για σφαιρικές ή ελλειπτικές ομάδες. Αυτό μπορεί να επιτευχθεί μέσω μίας αρχικής διερεύνησης των δεδομένων και τους στόχους που υπάρχουν για την ομαδοποίηση.

Στη συνέχεια θα δοθούν οι ορισμοί που θα χρειαστούν για τη μέθοδο KAMILA. Έστω ότι $\mathbf{V}_1,\mathbf{V}_2,\ldots,\mathbf{V}_N$ είναι ένα δείγμα από ανεξάρτητες και ισόνομες μεταξύ τους μεταβλητές από ένα $P\times 1$ σύνολο συνεχών, τυχαίων διανυσμάτων, δηλαδή $\mathbf{V}_i = (V_{i1}, V_{i2}, \ldots, V_{ip}, \ldots, V_{iP})^T$, που ακολουθούν μίξη κατανομών με πυκνότητα *h*:

$$f_{\mathbf{V}}(\mathbf{v}) = \sum_{g=1}^{G} \pi_g h\,(\mathbf{v}; \boldsymbol{\mu}_g),$$

όπου *G* δηλώνει τον αριθμό των ομάδων στην μίξη, $\boldsymbol{\mu}_g$ δηλώνει το $P\times 1$ κέντρο της *g*-στής ομάδας του $\mathbf{V}_i$ και π_g δηλώνει την εκ των προτέρων πιθανότητα να επιλεγεί μία παρατήρηση από τον **g**-στο πληθυσμό. Έστω ότι $\mathbf{W}_1,\mathbf{W}_2,\ldots,\mathbf{W}_N$ είναι ένα δείγμα από ανεξάρτητες και ισόνομες μεταξύ τους μεταβλητές από ένα $Q\times 1$ σύνολο διακριτών, τυχαίων διανυσμάτων, όπου κάθε στοιχείο είναι μία μίξη από πολυωνυμικές τυχαίες μεταβλητές, τέτοιο ώστε $\mathbf{W}_i = (W_{i1}, W_{i2}, \ldots, W_{iq}, \ldots, W_{iQ})^T$, με $W_{iq}\epsilon\{1, \ldots, l, \ldots, L_q\}$:

$$f_{\mathbf{W}}(\mathbf{w}) = \sum_{g=1}^{G} \pi_g \prod_{q=1}^{Q} m(w_q; \boldsymbol{\theta}_{gq}),$$

όπου $m(w; \boldsymbol{\theta}) = \prod_{l=1}^{L_q} \theta_l^{I\{w=l\}}$ δηλώνει την πολυωνυμική συνάρτηση μάζας πιθανότητας, $I\{\cdot\}$ δηλώνει την δείκτρια συνάρτηση και $\boldsymbol{\theta}_{gq}$ δηλώνει το $L_q\times 1$ παραμετρικό διάνυσμα της πολυωνυμικής συνάρτησης μάζας πιθανότητας που αντιστοιχεί στην *q*-στη τυχαία μεταβλητή της *g*-στής ομάδας. Υποθέτουμε ότι οι W_{iq} και $W_{iq'}$ είναι υπό όρους ανεξάρτητοι πληθυσμοί $\forall\, q \neq q'$.

Έστω ότι $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_N$ είναι ένα ανεξάρτητο και ισόνομο δείγμα από $\mathbf{X}_{i_{(P+Q)\times 1}} = (\mathbf{V}_i^T, \mathbf{W}_i^T)^T$, όπου το $\mathbf{V}_i$ είναι υπό όρους ανεξάρτητο δείγμα από το $\mathbf{W}_i$. Στην t-οστή επανάληψη, το $\hat{\mu}_g^{(t)}$ δηλώνει τον εκτιμητή του κέντρου της ομάδας που αφορά τον πληθυσμό g και το $\hat{\theta}_{gq}^{(t)}$ δηλώνει τον εκτιμητή για τις παραμέτρους της πολυωνυμικής κατανομής που αφορά την q-οστή τυχαία διακριτή μεταβλητή από τον πληθυσμό g.

Στην ομαδοποίηση μικτών δεδομένων, οι κατηγορικές μεταβλητές ενδέχεται να μην είναι ανεξάρτητες μεταξύ τους. Για να αντιμετωπιστεί αυτό το πρόβλημα, προτείνεται μια μέθοδος όπου αυτές οι μεταβλητές αντικαθίστανται από μια νέα κατηγορική μεταβλητή που περιλαμβάνει όλους τους πιθανούς συνδυασμούς επιπέδων των αρχικών μεταβλητών. Για παράδειγμα, ας υποθέσουμε ότι υπάρχουν δύο κατηγορικές μεταβλητές $\mathbf{W}_{i1}$ και $\mathbf{W}_{i2}$, που δεν είναι υπό όρους ανεξάρτητες, με $\mathbf{L}_1$ και $\mathbf{L}_2$ επίπεδα αντίστοιχα. Τότε θα αντικατασταθούν από μια νέα μεταβλητή $\mathbf{W}_{*i}$ με $\mathbf{L}_1\times\mathbf{L}_2$ επίπεδα, όπου κάθε επίπεδο θα αντιπροσωπεύει έναν μοναδικό συνδυασμό επιπέδων από τις αρχικές μεταβλητές. Αυτό επιτρέπει μία πιο ακριβή αναπαράσταση των σχέσεων μεταξύ των κατηγορικών μεταβλητών στη διαδικασία ομαδοποίησης.

Όσον αφορά τις συνεχείς ομάδες, για τον αλγόριθμο KAMILA, υπολογίζεται μία μονομεταβλητή εκτιμήτρια πυκνότητας με την μέθοδο του πυρήνα. Για να εκτιμηθούν οι πυκνότητες των συνεχών μεταβλητών, χρησιμοποιείται η μέθοδος μετασχηματισμού, η οποία είναι ένα γενικό πλαίσιο για την εκτίμηση των πυκνοτήτων μεταβλητών που έχουν μετασχηματιστεί με κάποια γνωστή συνάρτηση. Δηλαδή, ο υπολογισμός της εκτιμήτριας πυκνότητας μίας συνεχής μεταβλητής X, με την μέθοδο μετασχηματισμού κάποιας γνωστής συνάρτησης, γίνεται μέσω του τύπου:

$$\hat{f}(x) = \hat{g}(t(x))t'(x),$$

όπου t δηλώνει την παραγωγίσιμη συνάρτηση (π.χ. $\log(x), \sqrt{x}$) που χρησιμοποιούμε, g δηλώνει την συνάρτηση πυκνότητας πιθανότητας της $t(x)$ με την εκτιμήτρια της μεθόδου του πυρήνα $\hat{g}$ και $t'(x)$ δηλώνει την παράγωγο της $t(x)$.

Προχωρώντας, θα δώσουμε μία χρήσιμη πρόταση που αποδείχθηκε από τους Foss, Markatou, Ray και Heching (2016). Έστω ότι έχουμε ένα $\mathbf{V} = (V_1, V_2, \ldots, V_p)^T$ τυχαίο διάνυσμα που ακολουθεί μία σφαιρικά συμμετρική κατανομή με κέντρο στην αρχή, δηλαδή η κατανομή είναι ανεπηρέαστη από περιστροφές γύρω από την αρχή. Άρα, η κατανομή του διανύσματος εξαρτάται μόνο από το μέγεθος ή το μήκος του και όχι από την κατεύθυνσή του. Έστω ότι $f_R\left(\sqrt{\mathbf{v}^T\mathbf{v}}\right)$ είναι συνάρτηση πυκνότητας πιθανότητας της Ευκλείδειας απόστασης. Για να υπολογιστεί η εκτιμήτρια με την μέθοδο του πυρήνα, ορίζεται η συνάρτηση:

$$f_{\mathbf{V}}(\mathbf{v}) = \frac{f_R\left(\sqrt{\mathbf{v}^T\mathbf{v}}\right)\Gamma\left(\frac{p}{2}+1\right)}{p\sqrt{\mathbf{v}^T\mathbf{v}}^{p-1}\pi^{p/2}}, \tag{3.4.1}$$

όπου $\sqrt{\mathbf{v}^T\mathbf{v}} \in [0,\infty)$. Για τις συναρτήσεις πυκνότητας πιθανότητας των σφαιρικών ομάδων, θέτουμε τη t ως την Ευκλείδεια απόσταση και κάνοντας αντικατάσταση της f_R με την μονομεταβλητή εκτιμήτρια πυκνότητας με την μέθοδο του πυρήνα $\hat{f}_R$, τότε έχουμε την τεχνική για την εκτίμηση πυκνότητας $\hat{f}_V$.

Ο αλγόριθμος KAMILA αρχικοποιείται πολλές φορές για να αποφευχθεί να κολλήσει σε κάποιο τοπικό βέλτιστο, κάτι το οποίο σημαίνει ότι έχει συγκλίνει σε μια μη βέλτιστη λύση και

δεν μπορεί να βρει μια καλύτερη λύση με περαιτέρω επαναλήψεις. Κάθε αρχικοποίηση ξεκινά με ένα διαφορετικό σύνολο αρχικών τιμών για τις παραμέτρους. Ο αλγόριθμος τρέχει επαναληπτικά μέχρι να πληρούνται συγκεκριμένα κριτήρια διακοπής.

Το κριτήριο διακοπής βασίζεται σε δύο συνθήκες. Η πρώτη συνθήκη είναι να φτάσει σε ένα προκαθορισμένο μέγιστο αριθμό επαναλήψεων. Αυτό είναι ένα μέτρο ασφαλείας για να διασφαλιστεί ότι ο αλγόριθμος δεν θα τρέξει απεριόριστα. Η δεύτερη συνθήκη είναι ότι η σύνθεση του πληθυσμού δεν αλλάζει από την προηγούμενη επανάληψη. Αυτό σημαίνει ότι ο αλγόριθμος σύγκλινε σε μια σταθερή λύση και πιθανότατα οι επόμενες επαναλήψεις δεν θα βελτιώσουν τα αποτελέσματα.

Αρχικοποιώντας τον αλγόριθμο πολλαπλές φορές και χρησιμοποιώντας το κριτήριο σύγκλισης, ο αλγόριθμος KAMILA μπορεί να βρει μία βέλτιστη τιμή ή μια κοντινή σε βέλτιστη λύση. Ο αριθμός των αρχικοποιήσεων και ο μέγιστος αριθμός επαναλήψεων είναι παράμετροι που μπορούν να ρυθμιστούν για να βελτιωθεί η απόδοση του αλγορίθμου. Κάθε επανάληψη του αλγόριθμου KAMILA, γίνεται σε δύο στάδια: το στάδιο της διαχώρισης και το στάδιο της εκτίμησης.

Στην αρχή της κάθε επανάληψης, εκτιμάται η παράμετρος $\hat{\mu}_{gp}^{(0)}$ που δηλώνει την τυχαία λήψη από μία ομοιόμορφη κατανομή με όρια ίσα με την ελάχιστη και μέγιστη τιμή της p-στης αριθμητικής μεταβλητής. Επίσης, εκτιμάται η παράμετρος $\hat{\theta}_{gq}^{(0)}$ που δηλώνει την λήψη από την Dirichlet κατανομή με παραμέτρους ίσες με την μονάδα. Για τον υπολογισμό της Ευκλείδειας απόστασης, ας ορίσουμε το προαιρετικό βάρος ως ξ. Θα προχωρήσουμε με τον ορισμό της Ευκλείδειας απόστασης από την i-στη παρατήρηση και την $\hat{\mu}_g^{(t)}$:

$$d_{ig}^{(t)} = \sqrt{\sum_{p=1}^{P}[\xi_p(v_{ip} - \hat{\mu}_{gp}^{(t)})]^2}.$$

Επίσης, η ελάχιστη απόσταση για την i-στη παρατήρηση υπολογίζεται ως: $r_i^{(t)} = min_g(d_{ig}^{(t)})$. Η πυκνότητα του πυρήνα για τις ελάχιστες αποστάσεις υπολογίζεται ως εξής:

$$\hat{f}_R^{(t)}(r) = \frac{1}{Nh^{(t)}}\sum_{l=1}^{N} k(\frac{r - r_l^{(t)}}{h^{(t)}}),$$

όπου $k(\cdot)$ είναι η συνάρτηση του πυρήνα και $h^{(t)}$ είναι η παράμετρος που ορίζει το εύρος της συνάρτησης πυρήνα στην t-στη επανάληψη. Στον αλγόριθμο KAMILA η $k(\cdot)$ είναι η Gaussian συνάρτηση και $h = 0.9An^{-1/5}$, όπου $A = min(\hat{\sigma}, \frac{\hat{q}}{1.34})$, $\hat{\sigma}$ δηλώνει την δειγματική τυπική απόκλιση και $\hat{q}$ δηλώνει το δειγματικό ενδοτεταρτημοριακό εύρος. Η συνάρτηση $\hat{f}_R^{(t)}(r)$ χρησιμοποιείται για την εκτίμηση της (3.4.1).

Ακολούθως, υπολογίζεται η λογαριθμική πιθανότητα του i-στου παρατηρηθέντος κατηγορικού διανύσματος, υποθέτοντας ότι οι Q κατηγορικές μεταβλητές μέσα στην g-στη ομάδα είναι ανεξάρτητες: $\log\left(c_{ig}^{(t)}\right) = \sum_{q=1}^{Q} \xi_q \cdot \log\left(m(w_{iq}; \hat{\theta}_{gq}^{(t)})\right)$, όπου ξ_q είναι ένα προαιρετικό βάρος που αντιστοιχεί στη μεταβλητή q, το οποίο μπορεί να χρησιμοποιηθεί για τη συνεισφορά κάθε μεταβλητής στη συνολική λογαριθμική πιθανότητα.

Όπως έχουμε αναφέρει παραπάνω, για τον αλγόριθμο KAMILA η χρήση βαρών είναι προαιρετική. Στην συγκεκριμένη περίπτωση, η χρήση βαρών γίνεται για να υπάρξει συμβατότητα με άλλες στρατηγικές.

Η κάτωθι ποσότητα ορίζεται για την ανάθεση ατόμων μέσα στην ομάδα:

$$H_i^{(t)}(g) = \log\left[\hat{f}_V^{(t)}(d_{ig}^{(t)})\right] + \log\left[c_{ig}^{(t)}\right],$$

με την *i*-στη παρατήρηση να έχει ανατεθεί στον πληθυσμό g που μεγιστοποιεί την ποσότητα $H_i^{(t)}(g)$. Όπως έχουμε αναφέρει ήδη, η (3.4.1) δημιουργείται για τις ελάχιστες αποστάσεις, αλλά στην συνέχεια υπολογίζεται με το $d_{ig}^{(t)}$ για όλο τον πληθυσμό g.

Μετά το στάδιο της διαχώρισης των *N* παρατηρήσεων στην *t*-στη επανάληψη, στο στάδιο της εκτίμησης υπολογίζονται οι εκτιμημένες παράμετροι $\hat{\mu}_{gp}^{(t+1)}$ και $\hat{\theta}_{gp}^{(t+1)}$ για όλα τα *g,p,q*:

$$\hat{\mu}_{gp}^{(t+1)} = \frac{\sum_{i\in\Omega_g^{(t)}} v_i}{|\Omega_g^{(t)}|} \text{ και } \hat{\theta}_{gq}^{(t+1)} = \frac{\sum_{i\in\Omega_g^{(t)}} I\{w_{iq}=l\}}{|\Omega_g^{(t)}|},$$

όπου $\Omega_g^{(t)}$ δηλώνει το σύνολο των δεικτών που έχουν ανατεθεί στον πληθυσμό g στην επανάληψη *t* και $I\{\cdot\}$ δηλώνει την δείκτρια συνάρτηση.

Τέλος, για κάθε επανάληψη, υπολογίζεται η αντικειμενική συνάρτηση (*objective function*):

$$\sum_{i=1}^{N} max_g\{H_i^{(final)}(g)\}.$$

Έχοντας ορίσει όλες τις απαραίτητες συναρτήσεις που χρειάζονται, ο αλγόριθμος KAMILA μπορεί να περιγραφεί αλγοριθμικά με τα επόμενα βήματα:

Βήμα 1ο: Κάνε την αρχικοποίηση για τις παραμέτρους $\hat{\mu}_{gp}^{(0)}$ και $\hat{\theta}_{gq}^{(0)}$ $\forall$ *g, q, p*.

Βήμα 2ο: Εκτέλεσε το στάδιο της διαχώρισης, υπολογίζοντας τα: $d_{ig}^{(t)}, r_i^{(t)}, \hat{f}_V^{(t)}, c_{ig}^{(t)}, H_i^{(t)}(g)$.

Βήμα 3ο: Εκτέλεσε το στάδιο της εκτίμησης, υπολογίζοντας τις $\hat{\mu}_{gp}^{(t+1)}$ και $\hat{\theta}_{gq}^{(t+1)}$ παραμέτρους.

Βήμα 4ο: Επανάλαβε τα βήματα 2 και 3, μέχρι να συγκλίνει σε ένα αποτέλεσμα ή να έχει φτάσει στον μέγιστο αριθμό επαναλήψεων.

Βήμα 5ο: Υπολόγισε την αντικειμενική συνάρτηση. Ο αλγόριθμος KAMILA εξάγει το αποτέλεσμα του διαχωρισμού, που μεγιστοποιεί την αντικειμενική συνάρτηση.

Είναι ιδιαίτερα σημαντικό να αναφερθεί ότι σε άλλες μεθόδους έχουμε δει πως πρέπει να υπολογίζεται η απόσταση μεταξύ δύο σημείων, κάτι το οποίο δεν είναι εφαρμόσιμο ακόμα στην μέθοδο KAMILA.

Τέλος, ο αλγόριθμος KAMILA μπορεί να εφαρμοσθεί και για ελλειπτικές ομάδες σε ένα σύνολο δεδομένων. Για να ξεκινήσουμε, το σύνολο δεδομένων αποτελείται από έναν *V* πίνακα συνεχών δεδομένων με διαστάσεις *N*×*P*, όπου το *N* δηλώνει το αριθμό των παρατηρήσεων και το *P* δηλώνει τον αριθμό των συνεχών μεταβλητών.

Το πρώτο βήμα στη γενίκευση του KAMILA για ελλειπτικές ομάδες είναι η διαχώριση του συνολικού αθροίσματος τετραγώνων και του πίνακα των διανυσματικών γινομένων των συνεχών μεταβλητών. Αυτή η διαδικασία διαχώρισης παρέχει μια εκτίμηση του πίνακα συνδιακύμανσης των ομάδων και συμβολίζεται ως $\hat{\Sigma}$. Ο πίνακας συνδιακύμανσης αποτυπώνει τις σχέσεις και τη μεταβλητότητα μεταξύ των συνεχών μεταβλητών εντός κάθε ομάδας.

Μόλις εκτιμηθεί ο πίνακας συνδιακύμανσης, τα συνεχή δεδομένα μετασχηματίζονται κλιμακωτά (*rescaled*). Ο σκοπός του κλιμακωτού μετασχηματισμού είναι να μετατρέψει τα δεδομένα έτσι ώστε κάθε ομάδα να έχει προσεγγιστικά έναν ταυτοτικό πίνακα συνδιακύμανσης. Αυτό επιτυγχάνεται παίρνοντας τον κλιμακωτά μετασχηματισμένο πίνακα δεδομένων $V^* = V\hat{\Sigma}^{-1/2}$.

Μετά το βήμα του κλιμακωτού μετασχηματισμού, ο αλγόριθμος KAMILA μπορεί να εφαρμοστεί στο μετασχηματισμένο σύνολο δεδομένων. Ο αλγόριθμος KAMILA σχεδιάστηκε για την αναγνώριση ομάδων στα δεδομένα με βάση την ομοιότητά τους. Εφαρμόζοντας τον KAMILA στα μετασχηματισμένα δεδομένα με προσεγγιστικούς ταυτοτικούς πίνακες συνδιακύμανσης, ο αλγόριθμος μπορεί να διαχειριστεί αποτελεσματικά τις ελλειπτικές ομάδες.

Τέλος, προτείνονται τα ακόλουθα τρία μέτρα αξιολόγησης της ομαδοποίησης: καθαρότητα (*purity*) (Manning, Raghavan και Schutze (2008)), μάκρο-ακρίβεια (*macro-precision*) και μάκρο-ευαισθησία (*macro-recall*) (Modha και Spangler (2003)). Ας υποθέσουμε ότι διαθέτουμε συνολικά k ομάδες και c κατηγορίες. Επίσης, ας υποθέσουμε ότι διαθέτουμε a_i άτομα που αντιστοιχούν σε σωστές κατηγορίες, b_i άτομα που αντιστοιχούν σε λανθασμένες κατηγορίες και c_i άτομα που έχουν απορριφθεί λανθασμένα από μία σωστή κατηγορία. Τα μέτρα αξιολόγησης ορίζονται ως εξής:

- Καθαρότητα*:*

$$purity = \frac{1}{n}\sum_{i} max_j \, |k_i \cap c_j|$$

- Μάκρο-ακρίβεια*:*

$$p = \frac{1}{c}\sum_{i=1}^{c} \frac{a_i}{(a_i + b_i)}$$

- Μάκρο-ευαισθησία:

$$r = \frac{1}{c}\sum_{i=1}^{c} \frac{a_i}{(a_i + c_i)}$$

3.5 Μέθοδος DBSCAN για μικτά δεδομένα

Ο αλγόριθμος DBSCAN είναι από τους βασικότερους μη-παραμετρικούς αλγορίθμους που βασίζονται στην πυκνότητα των δεδομένων και όπως έχουμε αναφέρει προηγουμένως, έχει την ικανότητα να δημιουργεί ομάδες οι οποίες έχουν αυθαίρετα σχήματα.

Συγκριτικά με τους μη ιεραρχικούς και ιεραρχικούς αλγορίθμους, ο DBSCAN μπορεί πιο εύκολα να αναγνωρίζει μοτίβα στα δεδομένα. Κατά συνέπεια, είναι πιο προσαρμοστικός, χωρίς να είναι απαραίτητο ο εκάστοτε ερευνητής να έχει εκ των προτέρων γνώση των δεδομένων. Συμπληρωματικά, μπορεί γρήγορα να ομαδοποιήσει μικρά σύνολα δεδομένων, αλλά δεν ανταποκρίνεται καλά σε μεγάλα σύνολα δεδομένων. Αξίζει να σημειωθεί πως ο υπολογισμός του μέτρου ομοιότητας ή απόστασης είναι το πιο χρονοβόρο, αλλά ταυτόχρονα και το πιο κρίσιμο στάδιο.

Όπως έχουμε αναφέρει σε προηγούμενο κεφάλαιο, πριν ξεκινήσει ο αλγόριθμος απαιτείται ο καθορισμός των MinPts και Eps. Έχει παρατηρηθεί πως εάν δοθεί μία πολύ μικρή τιμή στο Eps, τότε πολλά άτομα θα χαρακτηριστούν ως θόρυβος και οι φυσικές ομάδες θα χωριστούν σε περισσότερες. Από την άλλη μεριά, εάν δοθεί μία πολύ μεγάλη τιμή στο Eps, πολλά σημεία που είναι χαρακτηρισμένα σαν θόρυβος, θα ταξινομηθούν σε ομάδες και οι φυσικές ομάδες θα συγχωνευθούν.

Αντίστοιχα, δίνοντας μία πολύ μικρή τιμή στο MinPts, πολλά σημεία θα οριστούν σαν πυρήνες, με αποτέλεσμα ο θόρυβος να συγχωνευθεί σε ομάδες. Δίνοντας μία πολύ μεγάλη τιμή στο MinPts, θα μειωθεί δραματικά ο αριθμός των πυρήνων, με αποτέλεσμα να καταλήξει σε πολύ μικρότερο αριθμό ομάδων. Συνεπώς, η σωστή επιλογή αυτών των δύο είναι πολύ σημαντική.

Όταν δεν παρέχεται το Eps, χρησιμοποιείται μία προσέγγιση για τον προσδιορισμό της τιμής του. Δημιουργείται ένα διάγραμμα k-dist για να βοηθήσει τον χρήστη στην επιλογή μιας κατάλληλης τιμής Eps από την κατακόρυφη θέση που αντιστοιχεί στο γόνατο του σχεδίου. Η διαδικασία περιγράφεται με τα ακόλουθα βήματα:

Βήμα 1ο: Όρισε ένα πιθανό εύρος τιμών για τα MinPts

Βήμα 2ο: Για κάθε παρατήρηση στον πίνακα απόστασης, ταξινόμησε τις αποστάσεις με αύξουσα σειρά και αποθήκευσε το υποσύνολο MinPts των αποστάσεων.

Βήμα 3ο: Ταξινόμησε το υποσύνολο των αποστάσεων σε φθίνουσα σειρά και σχεδίασε την ακολουθία των αποστάσεων.

Ο DBSCAN είναι σχεδιασμένος να μπορεί να διαχειριστεί αριθμητικά δεδομένα. Όμως, μετατρέποντας τα κατηγορικά δεδομένα, οι Liu, Yang και He (2017) πρότειναν έναν τροποποιημένο DBSCAN για μικτά δεδομένα. Αυτός ο αλγόριθμος ονομάστηκε EPDCA και βασίζεται στην πυκνότητα εντάσσοντας τις έννοιες της εντροπίας και της κατανομής πιθανότητας (*entropy and probability distribution*). Στον EPDCA δεν διαφέρουν τα αλγοριθμικά βήματα από αυτά του DBSCAN, όπως έχουν ήδη αναλυθεί.

Γενικά, η εντροπία μετράει την αβεβαιότητα ή την τυχαιότητα μίας μεταβλητής ή μίας κατανομής πιθανότητας σε ένα σύνολο δεδομένων. Ποσοτικοποιεί τη μέση ποσότητα πληροφορίας που χρειάζεται, για να περιγραφθεί ένα γεγονός. Στον EPDCA, η εντροπία χρησιμοποιείται ως κριτήριο για την αξιολόγηση της ομαδοποίησης. Για να προχωρήσουμε στον ορισμό της εντροπίας, θα πρέπει πρώτα να ορίσουμε την σχετική συχνότητα για τα κατηγορικά δεδομένα.

Υποθέτουμε πως διαθέτουμε p_c κατηγορικές μεταβλητές, από ένα σύνολο δεδομένων με n άτομα. Έστω ότι η $f(x^c)$ συμβολίζει το πλήθος των φορών που θα εμφανιστεί η τιμή x^c σε μία κατηγορική μεταβλητή. Τότε η σχετική συχνότητα συμβολίζεται ως εξής:

$$p(x^c) = \frac{f(x^c)}{n}.$$

Εάν στην κατηγορική μεταβλητή δεν υπάρχει η τιμή x^c, τότε η $f(x^c)$ παίρνει την τιμή μηδέν.

Ας συμβολίσουμε με n_i το σύνολο των ατόμων που υπάρχουν μέσα σε μία ομάδα k_i και με n_{ti} το σύνολο των ατόμων που ανήκουν σε μία κατηγορία t από συνολικά K κατηγορίες. Τότε η εντροπία της ομάδας k_i υπολογίζεται από τον τύπο:

$$E(k_i) = -\frac{1}{log\ (n)} \sum_{t=1}^{K} \frac{n_{ti}}{n_i} log\ (\frac{n_{ti}}{n_i}).$$

Όταν η τιμή της εντροπίας είναι κοντά στην μονάδα, τότε η ομάδα είναι ομοιόμορφα κατανεμημένη με διαφορετικές κατηγορίες. Από την άλλη μεριά, όταν η τιμή της εντροπίας είναι κοντά στο μηδέν, αυτό μας υποδηλώνει ότι η ομάδα απαρτίζεται από άτομα που είναι μία καθαρή κατηγορία. Σε γενικές γραμμές, όσο πιο μικρή είναι η τιμή της εντροπίας, τόσο καλύτερη ομαδοποίηση έχει γίνει.

Αντίστοιχα, η βασική συνάρτηση της εντροπίας μίας κατηγορικής τιμής σε μία ομάδα k, ορίζεται ως εξής:

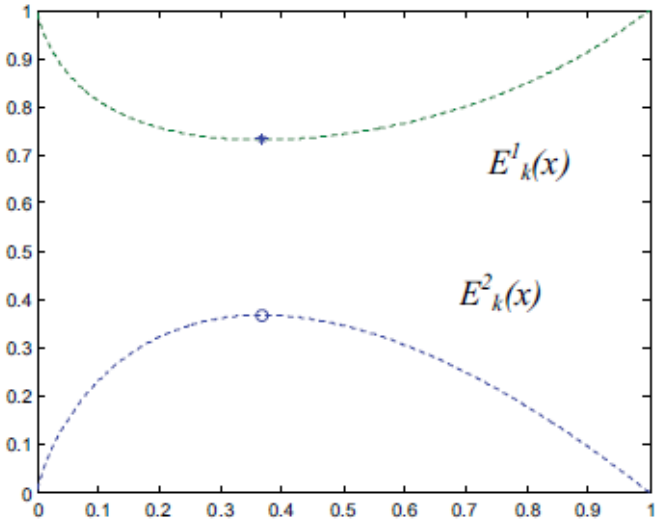
$$E_k^1(x^c) = -p(x^c) log_2(p(x^c)).$$

Οι Liu, Yang και He (2017) πρότειναν μία αναθεωρημένη συνάρτηση της παραπάνω. Λαμβάνει υπόψη την πιθανότητα μιας συγκεκριμένης τιμής και αναθέτει μεγαλύτερη σημαντικότητα στις μεταβλητές με χαμηλή πιθανότητα. Η συγκεκριμένη συνάρτηση ορίζεται ως εξής:

$$E_k^2(x^c) = \frac{1}{1-p(x^c) log_2(p(x^c))}.$$

Η εντροπία είναι ένα αποτελεσματικό μέτρο της αβεβαιότητας για τις κατηγορικές μεταβλητές, που μπορεί να υπολογιστεί εύκολα. Επιπροσθέτως, το μέτρο της εντροπίας δεν χρειάζεται να υπολογισθεί από την αρχή όταν προστίθενται νέα άτομα στο σύνολο δεδομένων. Αυτό είναι πολύ σημαντικό, ειδικά στα μεγάλα μικτά σύνολα δεδομένων που μεταβάλλονται συνεχώς.

Μία σύγκριση των παραπάνω τύπων παρουσιάζεται στην κάτωθι εικόνα. Στον οριζόντιο άξονα παρουσιάζονται οι τιμές του x^c και στον κάθετο άξονα παρουσιάζονται οι τιμές των $E_k^1(x^c)$ και $E_k^2(x^c)$:



Εικόνα 3: Συνάρτηση εντροπίας και αναθεωρημένη συνάρτηση εντροπίας από τους Liu, Yang και He (2017).

Η εντροπία μπορεί να χρησιμοποιηθεί για να μετρήσει την αβεβαιότητα που αφορά την ανάθεση ενός ατόμου σε μία ομάδα, με βάση την απόστασή του από την μέση εσωτερική απόσταση της ομάδας. Η μέση εσωτερική απόσταση ορίζεται ως η μέση απόσταση μεταξύ των ατόμων μέσα σε μία ομάδα.

Από την Εικόνα 3, παρατηρείται πως όταν η απόσταση μεταξύ ενός ατόμου και της μέσης εσωτερικής απόστασης μίας ομάδας συγκλίνει, το άτομο είναι λιγότερο βέβαιο ότι ανήκει σε αυτή την ομάδα και η τιμή εντροπίας του είναι μεγαλύτερη. Αυτό συμβαίνει επειδή το άτομο είναι πιο κοντά στα όρια της ομάδας και θα μπορούσε ενδεχομένως να ανήκει σε διαφορετική ομάδα. Από την άλλη μεριά, όταν οι αποκλίσεις στην απόσταση μεταξύ ενός ατόμου και της μέσης εσωτερικής απόστασης μίας ομάδας είναι πολύ μικρές ή μεγάλες, τότε η τιμή της εντροπίας θα μειωθεί. Αυτό συμβαίνει επειδή το άτομο είναι πιο μακριά από τα όρια της ομάδας και είναι πιο πιθανό να ανήκει στην συγκεκριμένη ομάδα.

Όπως αναφέραμε προηγουμένως, για να εφαρμοσθεί ο EPDCA, θα πρέπει να γίνει κατάλληλη μετατροπή στα κατηγορικά δεδομένα. Με τον υπολογισμό της εντροπίας, σε συνδυασμό με την κατανομή πιθανότητας, μπορεί να επιτευχθεί η ταυτόχρονη επεξεργασία τόσο στα αριθμητικά, όσο και στα κατηγορικά δεδομένα. Αξίζει να σημειωθεί πως αυτό μπορεί να συμβεί χωρίς να χαθεί σημαντικό μέρος της πληροφορίας και χωρίς να προκύψουν δυσκολίες από την μετατροπή των διαφορετικών τύπων δεδομένων.

Ας συμβολίσουμε με x_{ij} το επίπεδο j της i μεταβλητής. Ακολούθως, θα παρουσιαστούν κάποιοι εύχρηστοι τύποι για την μετατροπή των δεδομένων.

- Μέσω της βασικής συνάρτησης εντροπίας (*Basic Entropy*) (BE): μετασχηματισμός της πιθανότητας μίας τιμής μεταβλητής σε εντροπία:
$$x'_{ij} = -p(x_{ij}) ln\,(p(x_{ij}))$$
- Μέσω της αναθεωρημένης συνάρτησης εντροπίας (*Single Entropy*) (SE): η πιθανότητα μίας τιμής μεταβλητής χρησιμοποιείται ως βάρος:
$$x'_{ij} = \frac{p(x_{ij})}{1 - p(x_{ij}) ln\,(p(x_{ij}))}$$
- Μέσω της αναθεωρημένης συνάρτησης εντροπίας με κοινή πληροφορία (*Single Entropy with Mutual statistic*) (SEM): η αντίστροφη συνάρτηση της πιθανότητας της συσχετισμένης μεταβλητής χρησιμοποιείται ως βάρος:
$$x'_{ij} = \frac{f^{-1}(\theta, x_{il})}{1 - p(x_{ij}) ln\,(p(x_{ij}))}, l \neq j$$
- Μέσω αναλογίας της αναθεωρημένης συνάρτησης εντροπίας με κοινή πληροφορία (*Ratio of Single Entropy with Mutual statistic*) (SREM): θεωρώντας σαν αναλογία της εντροπίας την πιθανότητα της κατηγορικής μεταβλητής, η αντίστροφη συνάρτηση της πιθανότητας της συσχετισμένης μεταβλητής χρησιμοποιείται ως βάρος:
$$x'_{ij} = \frac{p(x_{ij}) ln\,(p(x_{ij})) f^{-1}(\theta, x_{il})}{\sum_{k=1}^{n} p(x_{ij}) ln\,(p(x_{ij}))}, l \neq j$$
- Μέσω της αναθεωρημένης συνάρτησης εντροπίας χωρίς κοινή πληροφορία (*Single Entropy with Self statistic*) (SET): η αντίστροφη συνάρτηση της πιθανότητας των κατηγορικών μεταβλητών χρησιμοποιείται ως βάρος:
$$x'_{ij} = \frac{f^{-1}(\theta, p(x_{ij}))}{1 - p(x_{ij}) ln\,(p(x_{ij}))}$$
- Μέσω αναλογίας της αναθεωρημένης συνάρτησης εντροπίας χωρίς κοινή πληροφορία (*Single Entropy Ratio with Self statistic*) (SRT): θεωρώντας σαν αναλογία της εντροπίας την

πιθανότητα της κατηγορικής μεταβλητής, η αντίστροφη συνάρτηση της πιθανότητας των κατηγορικών μεταβλητών χρησιμοποιείται ως βάρος:

$$x'_{ij} = \frac{p(x_{ij}) ln\,(p(x_{ij})) f^{-1}(\theta, p(x_{ij}))}{\sum_{k=1}^{n} p(x_{ij}) ln\,(p(x_{ij}))}$$

Στη βιβλιογραφία, συγκρίθηκαν οι παραπάνω τύποι, όσον αφορά την συσχέτιση και την διακύμανση των μεταβλητών. Παρατηρήθηκε πως για στατικά δεδομένα (δηλαδή δεδομένα που δεν αναμένεται να αλλάξουν στον χρόνο), η εφαρμογή διαφορετικών τύπων για την μετατροπή των δεδομένων, ενδέχεται να βελτιστοποιήσει την ομαδοποίηση. Επίσης, όταν ο ερευνητής διαθέτει δεδομένα όπου ο αριθμός των μεταβλητών παραμένει ο ίδιος, αλλά οι τιμές τους αλλάζουν, οι SREM και SRT αποδείχθηκαν πως είναι καλύτεροι. Από την άλλη μεριά, όταν ο ερευνητής διαθέτει δεδομένα όπου τόσο ο αριθμός των μεταβλητών, όσο και οι τιμές τους αλλάζουν, οι SEM και SET αποδείχθηκαν πως είναι καλύτεροι. Τέλος, με τους προτεινόμενους τύπους για τη μετατροπή των δεδομένων αυξάνεται η ταχύτητα του αλγόριθμου και μειώνεται η υπολογιστική του πολυπλοκότητα.

Είναι σημαντικό πως μετά τη μετατροπή των τύπων δεδομένων, για να διερευνηθεί η πληροφορία που διατηρείται, θα πρέπει να λάβουμε υπόψιν πως οι τιμές που έχουν μεγαλύτερη πιθανότητα εμφάνισης στο σύνολο δεδομένων, έχουν μεγαλύτερη σημασία από το μεσαίο τμήμα της κατανομής πιθανότητας. Επίσης, η συνολική εντροπία θα πρέπει να μειωθεί μετά τη μετατροπή των δεδομένων και η διακύμανση της εντροπίας των κατηγορικών μεταβλητών θα πρέπει να αυξηθεί. Τέλος, η συσχέτιση μεταξύ των μεταβλητών μπορεί να μειωθεί ή εξαλειφθεί, μετά από κάποια συγκεκριμένη μετατροπή.

Ακολούθως, θα αναπτύξουμε μερικά μέτρα αξιολόγησης, που μπορούν να εφαρμοσθούν για την ποιότητα των αλγορίθμων. Ας υποθέσουμε ότι διαθέτουμε α άτομα που αντιστοιχούν σε σωστές κατηγορίες από ένα n σύνολο ατόμων. Ας συμβολίσουμε με n' το συνολικό αριθμό των ατόμων που έχουν ταξινομηθεί σε κατηγορίες, χωρίς να περιέχει θόρυβο και μη διαβασμένα άτομα. Ας θεωρήσουμε k το σύνολο των ομάδων. Τα μέτρα αξιολόγησης ορίζονται ως εξής:

- Ορθότητα ομαδοποίησης *Ac*(*clustering accuracy):*

$$Ac = \sum_{i=1}^{k} \frac{\alpha_i}{n'_i}$$

  Όσο μεγαλύτερη είναι η τιμή της ακρίβειας, τόσο καλύτερη ομαδοποίηση έχει γίνει. Ιδανικά, όταν παίρνει την τιμή 1, τότε έχει επιτευχθεί η απόλυτα σωστή ομαδοποίηση.
- Ευαισθησία ομαδοποίησης *Rc*(*clustering recall):*

$$Rc = \sum_{j=1}^{k} \frac{\alpha_j}{n_i}$$

- Μέση καθαρότητα ομαδοποίησης *Pu*(*average clustering purity):*

$$Pu = \frac{\sum_{j=1}^{k} \frac{max_j(a)}{n_i}}{k}$$

  Όσο μεγαλύτερη είναι η τιμή της μέσης καθαρότητας, τόσο καλύτερη ομαδοποίηση έχει γίνει.

- F-measure:

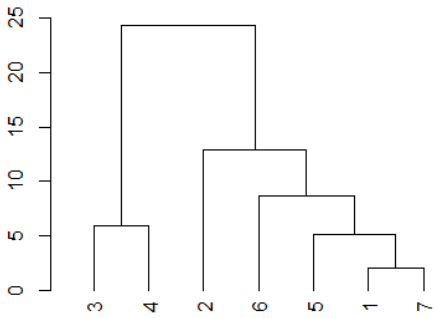
$$F = \frac{2 * Ac * Rc}{Ac + Rc}$$

Το μέτρο F-measure συνδυάζει την ακρίβεια και την ευαισθησία της ομαδοποίησης. Όσο μεγαλύτερη είναι η τιμή του F-measure, τόσο καλύτερη ομαδοποίηση έχει γίνει.

Είναι ενδιαφέρον να σημειωθεί πως για τους Liu, Yang και He (2017) τα αποτελέσματα της ομαδοποίησης με την χρήση του αλγόριθμου EDCPA βελτιώθηκαν σημαντικά, ειδικά στην ανακάλυψη του θορύβου και στον χρόνο που χρειάστηκε για να ολοκληρωθεί η ομαδοποίηση.

3.6 Συσσωρευτική ιεραρχική μέθοδος για μικτά δεδομένα

Στις συσσωρευτικές μεθόδους, όπως έχει αναφερθεί προηγουμένως, κάθε άτομο αντιστοιχεί σε μία ομάδα και με συνεχόμενες συγχωνεύσεις, τοποθετούνται όλα σε μία ομάδα. Ένας συνήθης τρόπος γραφικής αναπαράστασης των διαδοχικών συγχωνεύσεων είναι το δενδρόγραμμα. Σε κάθε βήμα, το δενδρόγραμμα ενώνει με μία γραμμή τα άτομα που συγχωνεύθηκαν. Στο τελικό βήμα, όλες οι παρατηρήσεις είναι ενωμένες με κάποιο μονοπάτι. Στον έναν άξονα, παρουσιάζονται οι παρατηρήσεις και στον άλλον η τιμή της απόστασης της συνένωσης που έγινε. Συνήθως, τα άτομα πριν την πρώτη ένωση ονομάζονται φύλλα, οι γραμμές που ενώνουν τα άτομα ονομάζονται κλαδιά και η τελική ένωση ονομάζεται ρίζα. Τέλος, μία ομάδα στο *i* επίπεδο είναι η ένωση των ομάδων-παιδιών στο *i*-1 επίπεδο. Παρακάτω, απεικονίζεται το δενδρόγραμμα από το παράδειγμα με τις ράτσες των σκύλων:



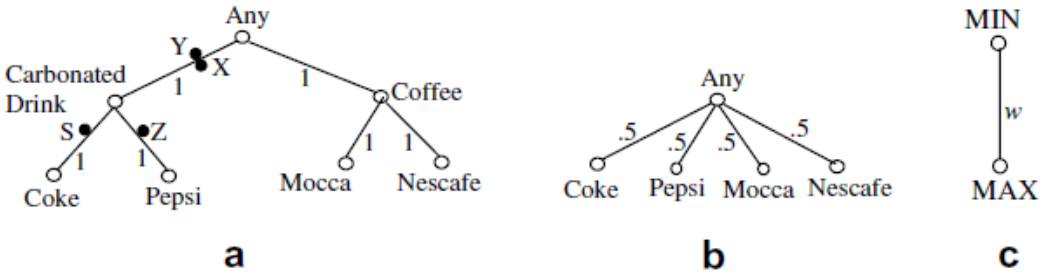
Σχήμα 3: Δενδρόγραμμα των μετρήσεων κάτω γνάθου σε διαφορετικές ράτσες σκύλων

Λόγω της δυσκολίας του υπολογισμού ενός μέτρου απόστασης στις κατηγορικές μεταβλητές, οι Hsu, Chen, και Su (2007) πρότειναν ένα καινούριο σχήμα για την αναπαράσταση των αποστάσεων, το οποίο ονόμασαν ιεραρχία απόστασης (*distance hierarchy*). Διευκολύνει την έκφραση της ομοιότητας ανάμεσα στις κατηγορικές παρατηρήσεις και ενοποιεί το μέτρο απόστασης ανάμεσα στις αριθμητικές και κατηγορικές μεταβλητές. Ως εκ τούτου, η συγκεκριμένη πρόταση μπορεί να διαχειριστεί τόσο αριθμητικά όσο και κατηγορικά δεδομένα. Επίσης, συντελεί στην εξέταση της ομοιότητας των κατηγορικών δεδομένων από τον εκάστοτε ερευνητή. Εφόσον υπολογιστεί η ιεραρχία απόστασης, στη συνέχεια μπορεί να εφαρμοστεί στη συσσωρευτική ιεραρχική μέθοδο.

Η ιεραρχία απόστασης συνδέει το κάθε κλαδί με ένα βάρος, το οποίο αντιπροσωπεύει την απόσταση. Ο λόγος που γίνεται αυτό είναι για να γίνει εφικτός ο υπολογισμός των αποστάσεων μεταξύ των παρατηρήσεων. Υπάρχουν διάφοροι τρόποι για τον ορισμό του βάρους. Ένας τρόπος είναι πρώτα να καθοριστεί η απόσταση μεταξύ δύο ομάδων-παιδιών και ύστερα να κατανεμηθεί σαν βάρος στα κλαδιά της διαδρομής των δύο ομάδων-παιδιών. Ένας

εναλλακτικός τρόπος είναι να δοθεί διαφορετικό βάρος ανάλογα με το επίπεδο του δενδρογράμματος. Για παράδειγμα, μεγαλύτερα βάρη δίνονται στα κλαδιά που είναι κοντά στη ρίζα και μικρότερα στα κλαδιά που είναι κοντά στα φύλλα. Για αποφυγή της πολυπλοκότητας, οι Hsu, Chen, και Su (2007) ορίζουν το βάρος κάθε κλαδιού ίσο με 1, εκτός εάν ορίζεται ρητά κάτι διαφορετικό.

Για διευκόλυνση θα αναπτύξουμε κάποιες έννοιες μέσω ενός εύκολου παραδείγματος, όπως φαίνεται στο κάτωθι σχήμα:



Σχήμα 4: (a) Η ιεραρχία απόστασης με βάρος 1 (b) Η ιεραρχία απόστασης με δύο επίπεδα μέσω της απλής αντιστοίχισης (c) Απομονωμένη (*Degenerated*) ιεραρχία απόστασης με βάρος *w=(max-min)* της αριθμ.μεταβλητής από τους Hsu, Chen, και Su (2007).

Στην ιεραρχία απόστασης, ένα σημείο *X* αποτελείται από δύο μέρη, την άγκυρα (*anchor*) και την αντιστάθμιση (*offset*), όπου η τελευταία είναι θετική πραγματική τιμή. Η άγκυρα είναι ένα φύλλο, ενώ η αντιστάθμιση είναι η απόσταση του *X* από τη ρίζα. Το σημείο *X* ορίζεται ως (N_X,d_X), όπου anchor(*X*)= N_X και offset(*X*)= d_X. Για παράδειγμα, στο Σχήμα 4(a), το σημείο *X*=(Pepsi,0.3), με anchor(*X*)= Pepsi και offset(*X*)= 0.3.

Στην ιεραρχία απόστασης δύο σημεία θεωρούνται *X,Y* ισοδύναμα, όταν βρίσκονται στην ίδια θέση του δενδρογράμματος. Όμως, αυτό δεν σημαίνει πως έχουν απαραίτητα την ίδια τιμή στην άγκυρα. Για παράδειγμα, στο Σχήμα 4(a), παρατηρούμε πως το σημείο *X* είναι ισοδύναμο με το σημείο *Y*, μολονότι το *X*=(Pepsi,0.3) και το *Y*=(Coke,0.3).

Ένα σημείο *X* θεωρείται πρόγονος (*ancestor*) ενός άλλου σημείου *Y*, όταν το *X* βρίσκεται στα κλαδιά που ενώνουν το *Y* με τη ρίζα. Ο ελάχιστος κοινός πρόγονος (*lowest common ancestor*), LCA (*X, Y*) του *X* και του *Y* ορίζεται ως το *X*, αν το *X* είναι πρόγονος του *Y*. Όμως, αν το *X* δεν είναι πρόγονος του *Y*, τότε LCA (*X, Y*) ορίζεται η ένωση που είναι πιο μακριά από τη ρίζα και είναι πρόγονος του *X* και του *Y*. Το ελάχιστο κοινό σημείο (*lowest common point*), LCP (*X, Y*) ορίζεται ως το *X* (ή το *Y*) αν το *X* είναι ισοδύναμο με το *Y*. Σε διαφορετική περίπτωση, ορίζεται ως ο LCA (*X, Y*). Παραδείγματος χάριν στο Σχήμα 4(a), τα *X* και *Y* είναι ισοδύναμα και είναι πρόγονοι των *S* και *Z*. Σύμφωνα με τα παραπάνω, Ο LCA (*S, Z*) είναι η ένωση που ονομάζεται ως «Carbonated Drink». Το LCP (*S, Z*) είναι επίσης η ένωση «Carbonated Drink» αφού τα *S* και *Z* δεν είναι ισοδύναμα. Το LCP (*X, Y*) είναι είτε το *X* είτε το *Y* αφού είναι ισοδύναμα.

Η απόσταση ανάμεσα σε δύο σημεία είναι ουσιαστικά το συνολικό βάρος που υπάρχει μεταξύ τους. Ας συμβολίσουμε με $d_{LCP(X,Y)}$ την απόσταση από την ρίζα έως το ελάχιστο κοινό σημείο των *X* και *Y*. Τότε η απόσταση μεταξύ των *X* και *Y* ορίζεται ως εξής:

$$|X - Y| = d_X + d_Y - 2d_{LCP(X,Y)}.$$

Υποθέτοντας ότι τα σημεία στο Σχήμα 4(a) είναι: *X*=(Pepsi,0.3), *Y*=(Coke,0.3), *Z*=(Pepsi,1.4) και *S*=(Coke,1.4), τότε η απόσταση ανάμεσα στο *X* και *Y* ισούται με μηδέν (αφού είναι ισδούναμα). Η απόσταση ανάμεσα στο *S* και *Y* είναι ίση με |1.4-0.3-2*0.3|=1.1. Αντίστοιχα, η απόσταση ανάμεσα στο *S* και *Z* είναι ίση με |1.4+1.4-2*1|=0.8, αφού το LCP (*S*, *Z*) είναι η ένωση «Carbonated Drink».

Από την άλλη μεριά, η ιεραρχία απόστασης για μία αριθμητική μεταβλητή, είναι μία απομονωμένη (*degenerated*) ιεραρχία. Αποτελείται από ένα κλαδί, την ρίζα *min* και το φύλλο *max*. Το βάρος είναι το εύρος της αριθμητικής μεταβλητής, δηλαδή *w=max-min*. Ένα σημείο *b* σε μια ιεραρχία απόστασης αριθμητικών δεδομένων ορίζεται ως (*max*, d_b) όπου η άγκυρα είναι το *max* και η αντιστάθμιση d_b είναι η απόσταση από το σημείο έως τη ρίζα *min*.

Ακολούθως, θα περιγραφεί ο τρόπος αντιστοίχισης των δεδομένων σε ιεραρχίες απόστασης. Κάθε άτομο x_i από το σύνολο δεδομένων συνδέεται με μία ιεραρχία απόστασης. Αυτή η συσχετισμένη ιεραρχία απόστασης (*associated distance hierarchy*) συμβολίζεται ως $dh=\{dh_1,dh_2,\ldots,dh_n\}$. Για τις κατηγορικές μεταβλητές, η αντιστοίχιση είναι απλή: η μεταβλητή συσχετίζεται σε μία dh_i με τέτοιο τρόπο που το πεδίο τιμών της αντιστοιχεί με το σύνολο των φύλλων της dh_i, δηλαδή για κάθε x_i^c που ανήκει στο σύνολο των πιθανών τιμών της p_c, υπάρχει αντιστοίχιση σε ένα φύλλο μίας dh_i. Για τις αριθμητικές μεταβλητές, η αντιστοίχιση γίνεται ως εξής: η μεταβλητή συσχετίζεται σε μία απομονωμένη dh_j με την διαφορά της μέγιστης τιμής της μεταβλητής από το σημείο, δηλαδή $w_i = max_{x_i^r} - min_i$, όπου το min_i είναι η ελάχιστη τιμή της αριθμητικής μεταβλητής. Αυτό επιτρέπει τη σύγκριση αριθμητικών τιμών με κατηγορικές τιμές στην ίδια ιεραρχία.

Κάθε τιμή ενός ατόμου μπορεί να αντιστοιχιστεί σε ένα σημείο στην συσχετισμένη ιεραρχία απόστασης. Πιο συγκεκριμένα, για μια κατηγορική τιμή x_i^c, η αντιστοίχιση (*mapping*) $h_i(x_i^c)$ αναθέτει το x_i^c σε ένα σημείο *b* στο dh_i και το *b* παίρνει την τιμή $(N_b, d_b) = (x_i^c, d_{x_i^c})$. Για μια αριθμητική τιμή x_i^r, η αντιστοίχιση $h_i(x_i^r)$ αναθέτει το x_i^r σε ένα σημείο *b* με την τιμή (*max*, x_i^r - min_i) στο dh_i, όπου το *max* είναι το μοναδικό φύλλο.

Για να γίνουν περισσότερο κατανοητά τα παραπάνω, παρουσιάζεται παρακάτω ένα παράδειγμα από το Σχήμα 4. Ας υποθέσουμε ότι έχουμε *x*=(Coke,70), με dh_{FD} όπως φαίνεται στο Σχήμα 4(a) και dh_{Amt} όπως φαίνεται στο Σχήμα 4(c) με max_{Amt}=100 και min_{Amt}=0. (όπου FD είναι «Favorite Drink» και Amt είναι η ποσότητα (*Amount*)). Η $h_{FD}(x)$ είναι ένα σημείο στις ενώσεις του επιπέδου Coke της dh_{FD}, δηλαδή $h_{FD}(x) = (\text{Coke}, 2)$. Η αντιστοίχιση $h_{Amt}(x)$ είναι ένα σημείο με την τιμή (*max*,70) στην dh_{Amt}, δηλαδή $h_{Amt}(x) = (\text{MAX}, 70)$.

Εφόσον έχουμε ορίσει όλες τις απαραίτητες έννοιες, θα προχωρίσουμε στον ορισμό της απόστασης ανάμεσα σε δύο άτομα. Έστω ότι διαθέτουμε δύο άτομα x_i, x_j από το σύνολο των δεδομένων μας και τις αντιστοιχίες $h(x_i)$ και $h(x_j)$. Η απόσταση μεταξύ του x_i και του x_j ορίζεται ως εξής:

$$d_L(x_i, x_j) = (\sum_{k=1}^{n} w_k(x_{ik} - x_{jk})^L)^{1/L} = (\sum_{k=1}^{n} w_k(h_k(x_i) - h_k(x_j))^L)^{1/L}.$$

Η απόσταση επιτρέπει στον ερευνητή να χρησιμοποιήσει διαφορετικά βάρη στις μεταβλητές των δεδομένων. Το βάρος w_i ορίζεται από τον εκάστοτε ερευνητή και μπορεί να χρησιμοποιηθεί, σύμφωνα με τις γνώσεις που έχει στο συγκεκριμένο τομέα, για να δώσει μεγαλύτερη ή μικρότερη βαρύτητα στις μεταβλητές. Η παράμετρος *L* είναι μια σταθερά που καθορίζει τον τύπο της απόστασης. Όταν το $L = 1$, η απόσταση είναι παρόμοια με την

σταθμισμένη απόσταση Manhattan, ενώ όταν το $L = 2$, η απόσταση είναι παρόμοια με την σταθμισμένη Ευκλείδεια απόσταση.

Είναι ενδιαφέρον να σημειωθεί ότι το βάρος w_i μπορεί να χρησιμοποιηθεί για να αντιμετωπιστεί την επίδραση μικτής έντασης (*mixed depth effect*). Είναι ένα φαινόμενο που συμβαίνει όταν ένα σύνολο δεδομένων περιέχει μεταβλητές με διάφορα επίπεδα ιεραρχίας απόστασης. Αυτό μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα κατά τον υπολογισμό της συνολικής απόστασης, καθώς οι μεταβλητές με χαμηλή ιεραρχία απόστασης έχουν μεγαλύτερη επιρροή στον υπολογισμό της από τις μεταβλητές με υψηλή ιεραρχία απόστασης. Ένας τρόπος αντιμετώπισης αυτού του φαινομένου είναι να οριστεί το βάρος αναλογικά με το ύψος της ιεραρχίας απόστασης. Για παράδειγμα: $w = c * height(dh)$ ή $w = c * height(dh)/max\ (height(dh_i))$, όπου c είναι μία σταθερά.

Στην προσέγγιση που παρουσιάστηκε, ενοποιούνται κάποιες προσεγγίσεις για τον χειρισμό των κατηγορικών δεδομένων, όπως η απλή αντιστοίχιση και η δυαδική κωδικοποίηση. Η απλή αντιστοίχιση μοντελοποιείται συσχετίζοντας κάθε κατηγορική μεταβλητή με μια ιεραρχία απόστασης δύο επιπέδων, όπου το βάρος κάθε ένωσης είναι ίσο με 0.5. Δύο άτομα που έχουν την ίδια τιμή σε μία κατηγορική μεταβλητή αντιστοιχούν στην ίδια θέση στην ιεραρχία και έχουν διαφορά ίση με το μηδέν, ενώ δύο διαφορετικές τιμές θα αντιστοιχούν σε διαφορετικά φύλλα και θα έχουν διαφορά ίση με τη μονάδα. Η δυαδική κωδικοποίηση εφαρμόζεται μετατρέποντας τις κατηγορικές σε ένα σύνολο δυαδικών μεταβλητών, όπου κάθε δυαδική μεταβλητή αντιπροσωπεύει μια διακριτή τιμή της αρχικής μεταβλητής. Στη συνέχεια, συσχετίζεται κάθε νέα δυαδική μεταβλητή με μια απομονωμένη ιεραρχία αριθμητικής απόστασης που έχει ρίζα *min*, φύλλο *max* και έναν κλαδί με βάρος 1. Οι ελάχιστες και μέγιστες τιμές κάθε δυαδικής μεταβλητής είναι 0 και 1.

Όπως αναφέραμε προηγουμένως, το σχήμα της ιεραρχίας απόστασης ενσωματώνεται στον αλγόριθμο που αφορά τις συσσωρευτικές ιεραρχικές μεθόδους. Με αυτόν τον τρόπο μπορούμε να εκφράσουμε τις ομοιότητες μεταξύ κατηγορικών τιμών και παράγει καλύτερα αποτελέσματα ομαδοποίησης όταν οι κατηγορικές τιμές εμφανίζουν διάφορα επίπεδα ομοιότητας. Η ολοκληρωμένη προσέγγιση περιλαμβάνει τον υπολογισμό του πίνακα απόστασης γειτνίασης (*adjacency distance matrix*) μέσω της προαναφερθείσας προσέγγισης και τη χρήση του ως είσοδο για τον αλγόριθμο ιεραρχικής ομαδοποίησης. Τα αλγοριθμικά βήματα της συσσωρευτικής ιεραρχικής μεθόδου δεν διαφέρουν από αυτά που έχουμε περιγράψει ήδη.

Τέλος, θα αναφέρουμε μερικές μεθόδους αξιολόγησης της ομαδοποίησης όσον αφορά την επίδραση των διαφορετικών τύπων μεταβλητών. Έστω ότι το μέγεθος του συνόλου δεδομένων είναι $|D|$ και το μέγεθος μίας ομάδας είναι $|C|$. Ας συμβολίσουμε με $P(A = V|C_k)$ την δεσμευμένη πιθανότητα του ενδεχομένου η μεταβλητή A να πάρει την τιμή V δοθέντος ότι βρίσκεται στην ομάδα C και με $P(A = V)$ την συνολική πιθανότητα η μεταβλητή A να παίρνει την τιμή V.

- Κατηγορική χρησιμότητα (*categorical utility)*:

  $$CU = \sum_k (\frac{|C_k|}{|D|} \sum_i \sum_j [P(A_i = V_{ij}|C_k)^2 - P(A_i = V_{ij})^2]),$$

  όπου V_{ij} είναι πιθανή τιμή της μεταβλητής A_i. Η κατηγορική χρησιμότητα αποσκοπεί στο να μετρήσει εάν η ομαδοποίηση βελτιώνει την πιθανότητα παρόμοιων ατόμων να εμπίπτουν στην ίδια ομάδα. Όσο μεγαλύτερη τιμή παίρνει η CU, τόσο καλύτερη είναι η επίδοση της ομαδοποίησης.

- Τετραγωνικό σφάλμα (*squared error)*:

$$SE_A = \sum_{m=1}^{k} \sum_{i \in C_m} (i_A - \mu_{m,A})^2$$

όπου i_A είναι η τιμή της μεταβλητής A για την παρατήρηση i και $\mu_{m,A}$ είναι ο μέσος όρος της μεταβλητής A της ομάδας C_m. Όσο μικρότερη είναι η τιμή του τετραγωνικού σφάλματος, τόσο καλύτερη είναι η ποιότητα της ομαδοποίησης. Επιπροσθέτως, το τετραγωνικό σφάλμα μπορεί να χρησιμοποιηθεί για να αναλυθεί κατά πόσο οι αριθμητικές τιμές στις ομάδες διαφέρουν από τις μέσες τιμές τους.

- Τυπική υπόκλιση (*standard deviation*):

$$SD_A = (\frac{1}{k-1} \sum_{i=1}^{k} (\mu_{i,A} - \frac{1}{k} \sum_{j=1}^{k} \mu_{j,A})^2)^{1/2}$$

όπου $\mu_{i,A}$ είναι ο μέσος όρος της μεταβλητής A που αφορά την ομάδα i και $\mu_{j,A}$ είναι ο μέσος όρος της μεταβλητής A που αφορά την ομάδα j. Όσο μεγαλύτερη είναι η τιμή της τυπικής απόκλισης, τόσο πιο χωρισμένες είναι οι μέσες τιμές.

ΚΕΦΑΛΑΙΟ 4

Σύγκριση μεθόδων μέσω προσομοιωμένων δεδομένων

4.1 Δημιουργία του προσομοιωμένου συνόλου δεδομένων

Η ξενοδοχειακή βιομηχανία παίζει καθοριστικό ρόλο στην παγκόσμια οικονομία, με τα ξενοδοχεία να εξυπηρετούν ταξιδιώτες ως βασικοί πάροχοι φιλοξενίας παγκοσμίως. Η κατανόηση των ποικίλων χαρακτηριστικών και προτιμήσεων των επισκεπτών του ξενοδοχείου είναι πρωταρχικής σημασίας στη διαχείριση του ξενοδοχείου και τις στρατηγικές μάρκετινγκ. Καθώς η τεχνολογία εξελίσσεται, οι προσεγγίσεις που βασίζονται σε δεδομένα έχουν καταστεί ζωτικής σημασίας για την ανάλυση των δεδομένων των επισκεπτών, ώστε οι επιχειρήσεις να αποκτήσουν γνώσεις που μπορούν να οδηγήσουν σε βελτιωμένες εμπειρίες των επισκεπτών και ωφέλιμα επιχειρησιακά αποτελέσματα.

Επιλέχθηκε για τις ανάγκες της παρουσίασης, να αναλυθεί ένα σενάριο που αφορά αυτόν τον τομέα, για μία υποθετική τουριστική επιχείρηση που θέλει να προβλέψει το καταλληλότερο πακέτο παροχών για έναν νεοεισερχόμενο πελάτη. Η προσομοίωση δεδομένων μας επιτρέπει να μελετήσουμε αυτό το υποθετικό σενάριο και να αξιολογήσουμε τις μεθόδους ομαδοποίησης χωρίς να βασιστούμε σε πραγματικά δεδομένα, τα οποία μπορεί να είναι περιορισμένα ή μη διαθέσιμα για συγκεκριμένους στόχους.

Το σύνολο δεδομένων που δημιουργήθηκε, αποτελείται από τρεις ομάδες («A», «B», «C»), που η καθεμία αντιπροσωπεύει ένα διαφορετικό τμήμα των επισκεπτών με μοναδικά χαρακτηριστικά. Οι τιμές για τις μεταβλητές επιλέχθηκαν αυθαίρετα, δίνοντας μεγαλύτερη έμφαση άτομα να ανήκουν στην ομάδα Α, άτομα μικρότερης ηλικίας με μικρότερο εισόδημα και κόστος διαμονής, που προέρχονται κυρίως από Μεσογειακές χώρες. Στην ομάδα Β ανήκουν άτομα μεγαλύτερης ηλικίας με μεγαλύτερο εισόδημα και κόστος διαμονής, που προέρχονται κυρίως από Βόρεια Ευρώπη. Στην ομάδα C ανήκουν άτομα μέσης ηλικίας με μέσο εισόδημα και κόστος διαμονής, που προέρχονται κυρίως από Δυτική Ευρώπη.

Τα προσομοιωμένα δεδομένα δημιουργήθηκαν με βάση διάφορες ιδιότητες των επισκεπτών, όπως η εθνικότητα, η ηλικία, το φύλο, τον τύπο του δωματίου προτίμησης, την επιλογή πρωινού, την τιμή της διαμονής, την αξιολόγηση των επισκεπτών και το εισόδημα. Η ομαδοποίηση των επισκεπτών σε διαφορετικές ομάδες πραγματοποιείται για να εντοπίσει τα πρότυπα, ομοιότητες και διαφορές μεταξύ των επισκεπτών. Επιπλέον, συνδέοντας κάθε επισκέπτη με μία πραγματική ομάδα, μπορεί να επιτευχθεί καλύτερη ακρίβεια και αποτελεσματικότητα των μεθόδων ομαδοποίησης που εφαρμόζονται. Για να ελέγξουμε την ευαισθησία των αλγορίθμων στο είδος των μεταβλητών, θα παρουσιάσουμε δύο εφαρμογές σε κάθε αλγόριθμο. Στην πρώτη εφαρμογή, θα υπάρχει ισχυρός διαχωρισμός σε αριθμητικές μεταβλητές και στη δεύτερη θα υπάρχει ισχυρός διαχωρισμός σε κατηγορικές μεταβλητές.

Η ανάλυση των δεδομένων θα πραγματοποιηθεί με την R, μια δημοφιλή γλώσσα προγραμματισμού για την υπολογιστική στατιστική και την ανάλυση δεδομένων. Η εφαρμογή των μεθόδων ομαδοποίησης επωφελείται από την ευελιξία και την αποτελεσματικότητα της γλώσσας R, η οποία είναι μια ισχυρή γλώσσα προγραμματισμού για χειρισμό δεδομένων,

οπτικοποίηση και μοντελοποίηση. Ο ολοκληρωμένος κώδικας των υπολογισμών βρίσκεται στο παράρτημα της εργασίας.

Το προσομοιωμένο σύνολο δεδομένων αποτελείται από 1.000 επισκέπτες και 8 μεταβλητές. Οι τρεις ομάδες δημιουργήθηκαν ως εξής: η πρώτη και η δεύτερη ομάδα αποτελείται από 300 επισκέπτες και η τρίτη αποτελείται από 400 επισκέπτες. Όσον αφορά τις μεταβλητές, το προσομοιωμένο σύνολο δεδομένων απαρτίζεται από τρεις αριθμητικές και πέντε κατηγορικές. Οι αριθμητικές μεταβλητές είναι η ηλικία των επισκεπτών, η τιμή της διαμονής και το εισόδημα των επισκεπτών. Οι κατηγορικές μεταβλητές είναι η εθνικότητα, το φύλο, ο τύπος του δωματίου προτίμησης και η επιλογή πρωινού των επισκεπτών και η διατάξιμη μεταβλητή που είναι η αξιολόγηση. Οι κατηγορίες για την εθνικότητα των επισκεπτών είναι: Έλληνας, Γάλλος, Ιταλός, Γερμανός, Βρετανός, Ισπανός, Σουηδός και Δανός, ενώ για τον τύπο του δωματίου προτίμησης των επισκεπτών είναι: Deluxe Double, Standard Twin, Executive Suite και Superior King. Τέλος, οι κατηγορίες για την αξιολόγηση είναι 1 Star, 2 Stars, 3 Stars, 4 Stars και 5 Stars και για την επιλογή πρωινού των επισκεπτών είναι: Ναι και Όχι. Θα παρουσιάσουμε μία σύνοψη των μεταβλητών των τριών ομάδων που αφορά την πρώτη εφαρμογή:

- **Ομάδα Α**
  - Εθνικότητα: οι κατηγορίες της εθνικότητας έχουν αντίστοιχες πιθανότητες: 0.20, 0.10, 0.20, 0.10, 0.10, 0.20, 0.05, 0.05.
  - Ηλικία: κυμαίνεται από 18 έως 40 χρόνια.
  - Φύλο: οι κατηγορίες του φύλου έχουν τις αντίστοιχες πιθανότητες: 0.55, 0.45.
  - Τύπος του δωματίου προτίμησης: οι κατηγορίες του τύπου δωματίου προτίμησης έχουν τις αντίστοιχες πιθανότητες: 0.30, 0.40, 0.15, 0.15.
  - Επιλογή πρωινού: οι κατηγορίες της επιλογής του πρωινού έχουν τις αντίστοιχες πιθανότητες: 0.45, 0.55.
  - Τιμή της διαμονής: παράγεται από μια κανονική κατανομή με μέσο όρο 300 και τυπική απόκλιση 75.
  - Αξιολόγηση: κυμαίνεται από 1 έως 5 αστέρια με πιθανότητες: 0.10, 0.10, 0.30, 0.30, 0.20.
  - Εισόδημα: παράγεται από μια κανονική κατανομή με μέσο όρο 8,000 και τυπική απόκλιση 1,000.

- **Ομάδα Β**
  - Εθνικότητα: οι κατηγορίες της εθνικότητας έχουν τις αντίστοιχες πιθανότητες: 0.05, 0.10, 0.05, 0.20, 0.10, 0.10, 0.20, 0.20.
  - Ηλικία: κυμαίνεται από 40 έως 70 χρόνια.
  - Φύλο: οι κατηγορίες του φύλου έχουν τις αντίστοιχες πιθανότητες: 0.45, 0.55.
  - Τύπος του δωματίου προτίμησης: οι κατηγορίες του τύπου δωματίου προτίμησης έχουν τις αντίστοιχες πιθανότητες: 0.25, 0.20, 0.25, 0.30.
  - Επιλογή πρωινού: οι κατηγορίες της επιλογής του πρωινού έχουν τις αντίστοιχες πιθανότητες 0.65, 0.35.
  - Τιμή της διαμονής: παράγεται από μια κανονική κατανομή με μέσο όρο 800 και τυπική απόκλιση 100.
  - Αξιολόγηση: κυμαίνεται από 1 έως 5 αστέρια με πιθανότητες: 0.10, 0.30, 0.30, 0.20, 0.10.
  - Εισόδημα: παράγεται από μια κανονική κατανομή με μέσο όρο 20,000 και τυπική απόκλιση 1,000.

- **Ομάδα C**
  - Εθνικότητα: οι κατηγορίες της εθνικότητας έχουν τις αντίστοιχες πιθανότητες: 0.05, 0.20, 0.05, 0.20, 0.20, 0.20, 0.05, 0.05.
  - Ηλικία: κυμαίνεται από 30 έως 50 χρόνια.
  - Φύλο: οι κατηγορίες του φύλου έχουν τις αντίστοιχες πιθανότητες: 0.50, 0.50.
  - Τύπος του δωματίου προτίμησης: οι κατηγορίες του τύπου δωματίου προτίμησης έχουν τις αντίστοιχες πιθανότητες: 0.30, 0.30, 0.20, 0.20.
  - Επιλογή πρωινού: οι κατηγορίες της επιλογής του πρωινού έχουν τις αντίστοιχες πιθανότητες: 0.50, 0.50.
  - Τιμή της διαμονής: παράγεται από μια κανονική κατανομή με μέσο όρο 450 και τυπική απόκλιση 100.
  - Αξιολόγηση: κυμαίνεται από 1 έως 5 αστέρια με πιθανότητες: 0.20, 0.20, 0.20, 0.20, 0.20.
  - Εισόδημα: παράγεται από μια κανονική κατανομή με μέσο όρο 15,000 και τυπική απόκλιση 1,000.

Ακολούθως, φαίνονται κάποια βασικά περιγραφικά χαρακτηριστικά των αριθμητικών μεταβλητών και η αναλογία των κατηγοριών των κατηγορικών μεταβλητών του τελικού προσομοιωμένου συνόλου δεδομένων:

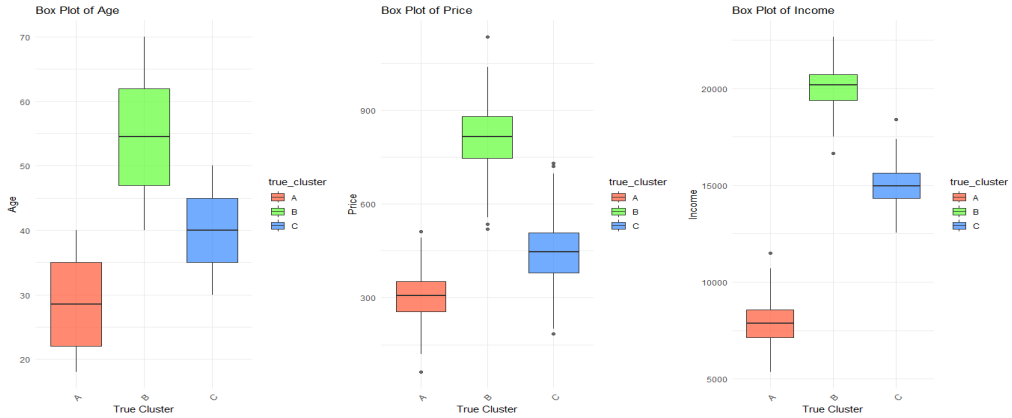
```
 Nationality       Age          Gender             Room_Type    Breakfast     Price            Rating        Income
British:141   Min.   :18.0   Female:484   Deluxe Double  :279   No :475   Min.   :  62.6   1 Star :139   Min.   : 5353
Danish : 91   1st Qu.:33.0   Male  :516   Executive Suite:208   Yes:525   1st Qu.: 332.9   2 Stars:203   1st Qu.: 8881
French :152   Median :40.0                Standard Twin  :289             Median : 451.8   3 Stars:267   Median :14948
German :153   Mean   :41.1                Superior King  :224             Mean   : 513.6   4 Stars:202   Mean   :14392
Greek  :104   3rd Qu.:48.0                                                3rd Qu.: 724.0   5 Stars:189   3rd Qu.:19156
Italian: 91   Max.   :70.0                                                Max.   :1133.7                 Max.   :22657
Spanish:178
Swedish: 90
```

Όσον αφορά τις κατηγορικές μεταβλητές, παρατηρούμε ότι 525 επισκέπτες επέλεξαν πρωινό, ενώ 475 δεν επέλεξαν πρωινό. Επίσης, 484 επισκέπτες ήταν γυναίκες και 516 άντρες. Όσον αφορά την εθνικότητα, παρατηρούμε πως μεγαλύτερο πλήθος προέρχονται από την Ισπανία, Γερμανία και Γαλλία και λιγότεροι επισκέπτες προέρχονται από την Σουηδία. Για την επιλογή του τύπου του δωματίου προτίμησης, παρατηρούμε πως περισσότεροι επισκέπτες επέλεξαν Standard Twin και Deluxe Double. Τέλος, περισσότεροι επισκέπτες αξιολόγησαν το ξενοδοχείο με τρία αστέρια και λιγότεροι με ένα αστέρι.

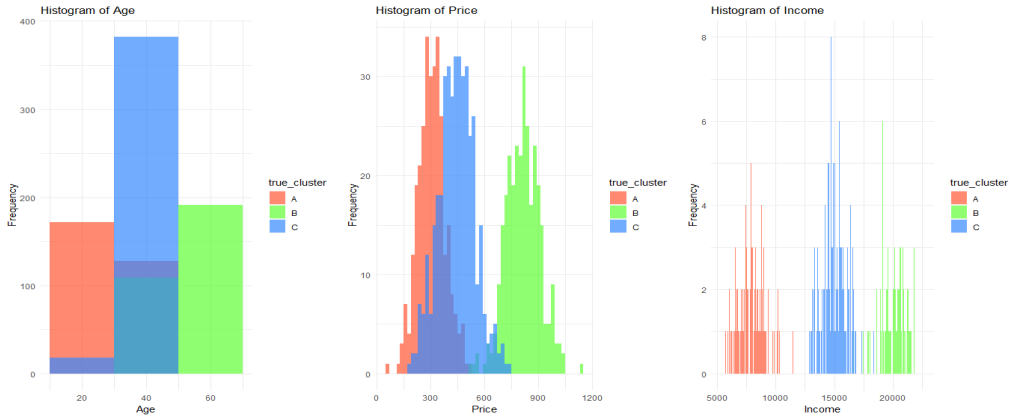
Για τις αριθμητικές μεταβλητές παρατηρούμε πως το ελάχιστο εισόδημα ήταν 5,353 ευρώ, ενώ το μέγιστο ήταν 22,657 ευρώ. Ο μέσος όρος εισοδήματος των επισκεπτών ανέρχεται στα 14,392 ευρώ και η διάμεσος είναι 14,948, το οποίο σημαίνει πως οι μισοί επισκέπτες είχαν εισόδημα πάνω από 14,948 και οι άλλοι μισοί λιγότερο από αυτό το ποσό. Όσον αφορά την ηλικία, οι πιο μικροί ηλικιακά ήταν 18 ετών ενώ οι μεγαλύτεροι 70 ετών. Ο μέσος όρος ηλικίας ήταν 41.10 έτη και η διάμεσος 40, δηλαδή οι μισοί επισκέπτες ήταν πάνω από 40 ετών και οι άλλοι μισοί κάτω από 40 ετών. Τέλος, η ελάχιστη τιμή του δωματίου ήταν 62.60 και η μέγιστη 1,133.70. Ο μέσος όρος της τιμής του δωματίου ήταν 513.60 και η διάμεσος ήταν 451.80, το οποίο σημαίνει πως τα μισά δωμάτια κόστιζαν πάνω από 451.80 και τα άλλα μισά κόστιζαν λιγότερο από 451.80.

Εν συνεχεία, παρουσιάζονται κάποιες γραφικές παραστάσεις προκειμένου να αποκτηθεί μία αρχική εικόνα για το προσομοιωμένο σύνολο δεδομένων. Παρουσιάζονται τα θηκογράμματα και ιστογράμματα για τις αριθμητικές μεταβλητές και τα ραβδογράμματα για τις κατηγορικές μεταβλητές, συγκριτικά με τις τρείς ομάδες.

Κατασκευάζοντας τα θηκογράμματα και τα ιστογράμματα για τις αριθμητικές μεταβλητές, παρατηρούμε ότι οι ομάδες είναι ισχυρά διαχωρισμένες όσον αφορά το εισόδημα, καλά διαχωρισμένες για τιμή της διαμονής και σχετικά καλά για την ηλικία.

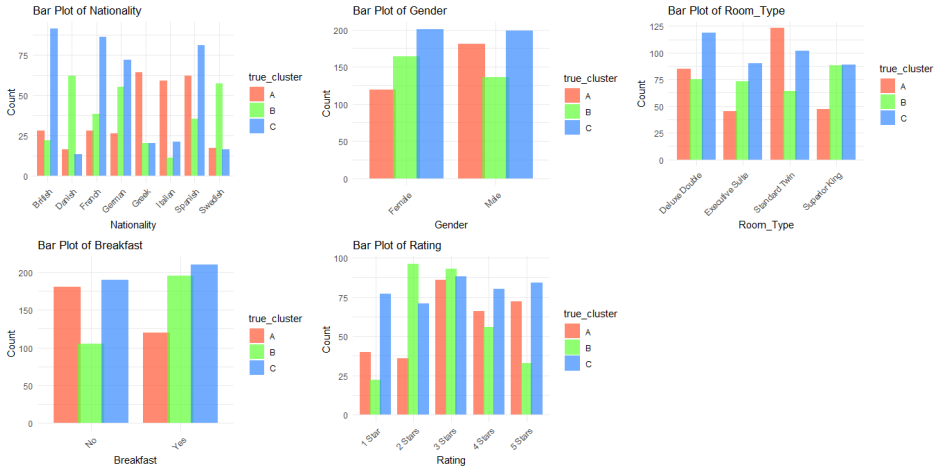


Σχήμα 5: Θηκογράμματα για τις αριθμητικές μεταβλητές



Σχήμα 6: Ιστογράμματα για τις αριθμητικές μεταβλητές

Τέλος, με τα κάτωθι ραβδογράμματα των κατηγορικών μεταβλητών, εύκολα μπορεί να παρατηρήσει κανείς πως δεν υπάρχει ισχυρός διαχωρισμός στις ομάδες σχεδόν σε καμία μεταβλητή.



Σχήμα 7: Ραβδογράμματα για τις κατηγορικές μεταβλητές

Στη συνέχεια, θα παρουσιάσουμε μία σύνοψη των μεταβλητών των τριών ομάδων που αφορά την δεύτερη εφαρμογή:

- **Ομάδα Α**
  - Εθνικότητα: οι κατηγορίες της εθνικότητας έχουν αντίστοιχες πιθανότητες: 0.30, 0.02, 0.30, 0.02, 0.02, 0.30, 0.02, 0.02.
  - Ηλικία: κυμαίνεται από 18 έως 40 χρόνια.
  - Φύλο: οι κατηγορίες του φύλου έχουν τις αντίστοιχες πιθανότητες: 0.60, 0.40.
  - Τύπος του δωματίου προτίμησης: οι κατηγορίες του τύπου δωματίου προτίμησης έχουν τις αντίστοιχες πιθανότητες: 0.20, 0.60, 0.10, 0.10.
  - Επιλογή πρωινού: οι κατηγορίες της επιλογής του πρωινού έχουν τις αντίστοιχες πιθανότητες: 0.05, 0.95.
  - Τιμή της διαμονής: παράγεται από μια κανονική κατανομή με μέσο όρο 300 και τυπική απόκλιση 75.
  - Αξιολόγηση: κυμαίνεται από 1 έως 5 αστέρια με πιθανότητες: 0.10, 0.10, 0.10, 0.10, 0.60.
  - Εισόδημα: παράγεται από μια κανονική κατανομή με μέσο όρο 13,000 και τυπική απόκλιση 1,500.

- **Ομάδα Β**
  - Εθνικότητα: οι κατηγορίες της εθνικότητας έχουν τις αντίστοιχες πιθανότητες 0.02, 0.02, 0.02, 0.30, 0.02, 0.02, 0.30, 0.30.
  - Ηλικία: κυμαίνεται από 35 έως 70 χρόνια.
  - Φύλο: οι κατηγορίες του φύλου έχουν τις αντίστοιχες πιθανότητες: 0.40, 0.60.
  - Τύπος του δωματίου προτίμησης: οι κατηγορίες του τύπου δωματίου προτίμησης έχουν τις αντίστοιχες πιθανότητες: 0.10, 0.10, 0.20, 0.60.
  - Επιλογή πρωινού: οι κατηγορίες της επιλογής του πρωινού έχουν τις αντίστοιχες πιθανότητες 0.95, 0.05.
  - Τιμή της διαμονής: παράγεται από μια κανονική κατανομή με μέσο όρο 500 και τυπική απόκλιση 100.
  - Αξιολόγηση: κυμαίνεται από 1 έως 5 αστέρια με πιθανότητες: 0.10, 0.60, 0.10, 0.10, 0.10.
  - Εισόδημα: παράγεται από μια κανονική κατανομή με μέσο όρο 17,000 και τυπική απόκλιση 1,500.

- **Ομάδα C**
  - Εθνικότητα: οι κατηγορίες της εθνικότητας έχουν τις αντίστοιχες πιθανότητες: 0.05, 0.20, 0.05, 0.20, 0.20, 0.20, 0.05, 0.05.
  - Ηλικία: κυμαίνεται από 25 έως 60 χρόνια.
  - Φύλο: οι κατηγορίες του φύλου έχουν τις αντίστοιχες πιθανότητες: 0.50, 0.50.
  - Τύπος του δωματίου προτίμησης: οι κατηγορίες του τύπου δωματίου προτίμησης έχουν τις αντίστοιχες πιθανότητες: 0.60, 0.20, 0.10, 0.10.
  - Επιλογή πρωινού: οι κατηγορίες της επιλογής του πρωινού έχουν τις αντίστοιχες πιθανότητες: 0.70, 0.30.
  - Τιμή της διαμονής: παράγεται από μια κανονική κατανομή με μέσο όρο 400 και τυπική απόκλιση 100.
  - Αξιολόγηση: κυμαίνεται από 1 έως 5 αστέρια με πιθανότητες: 0.10, 0.10, 0.10, 0.60, 0.10.

- Εισόδημα: παράγεται από μια κανονική κατανομή με μέσο όρο 15,000 και τυπική απόκλιση 1,500.

Ακολούθως, φαίνονται κάποια βασικά περιγραφικά χαρακτηριστικά των αριθμητικών μεταβλητών και η αναλογία των κατηγοριών των κατηγορικών μεταβλητών του τελικού προσομοιωμένου συνόλου δεδομένων:

```
 Nationality       Age             Gender              Room_Type     Breakfast      Price             Rating          Income     
British:100   Min.   :18.00   Female:470   Deluxe Double    :313   No :414   Min.   : 62.6   1 Star :109   Min.   : 9030  
Danish :105   1st Qu.:31.00   Male  :530   Executive Suite:137   Yes:586   1st Qu.:314.5   2 Stars:252   1st Qu.:13407  
French : 98   Median :39.50                Standard Twin   :304             Median :396.8   3 Stars: 97   Median :14935  
German :185   Mean   :41.16                Superior King   :246             Mean   :399.8   4 Stars:324   Mean   :14974  
Greek  :113   3rd Qu.:51.00                                                 3rd Qu.:481.0   5 Stars:218   3rd Qu.:16492  
Italian:115   Max.   :70.00                                                 Max.   :792.5                 Max.   :21271  
Spanish:169                                                                                                              
Swedish:115                                                                                                              
```

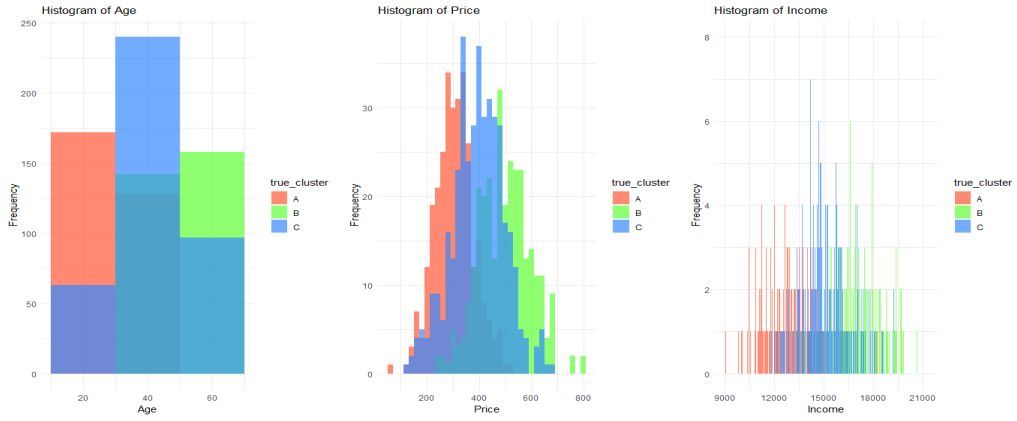
Όσον αφορά τις κατηγορικές μεταβλητές, παρατηρούμε ότι 586 επισκέπτες επέλεξαν πρωινό, ενώ 414 δεν επέλεξαν πρωινό. Επίσης, 470 επισκέπτες ήταν γυναίκες και 530 άντρες. Όσον αφορά την εθνικότητα, παρατηρούμε πως μεγαλύτερο πλήθος προέρχονται από την Γερμανία και Ισπανία και λιγότεροι επισκέπτες προέρχονται από την Γαλλία. Για την επιλογή του τύπου του δωματίου προτίμησης, παρατηρούμε πως περισσότεροι επισκέπτες επέλεξαν Deluxe Double και Standard Twin. Τέλος, περισσότεροι επισκέπτες αξιολόγησαν το ξενοδοχείο με τέσσερα αστέρια και λιγότεροι με τρία αστέρια.

Για τις αριθμητικές μεταβλητές παρατηρούμε πως το ελάχιστο εισόδημα ήταν 9,030 ευρώ, ενώ το μέγιστο ήταν 21,271 ευρώ. Ο μέσος όρος εισοδήματος των επισκεπτών ανέρχεται στα 14,974 ευρώ και η διάμεσος είναι 14,935, το οποίο σημαίνει πως οι μισοί επισκέπτες είχαν εισόδημα πάνω από 14,935 και οι άλλοι μισοί λιγότερο από αυτό το ποσό. Όσον αφορά την ηλικία, οι πιο μικροί ηλικιακά ήταν 18 ετών ενώ οι μεγαλύτεροι 70 ετών. Ο μέσος όρος ηλικίας ήταν 41.16 έτη και η διάμεσος 39.50, δηλαδή οι μισοί επισκέπτες ήταν πάνω από 39.50 ετών και οι άλλοι μισοί κάτω από 39.50 ετών. Τέλος, η ελάχιστη τιμή του δωματίου ήταν 62.60 και η μέγιστη 792.50. Ο μέσος όρος της τιμής του δωματίου ήταν 399.80 και η διάμεσος ήταν 396.80, το οποίο σημαίνει πως τα μισά δωμάτια κόστιζαν πάνω από 396.80 και τα άλλα μισά κόστιζαν λιγότερο από 396.80.

Ακολούθως, παρουσιάζονται οι ίδιες γραφικές παραστάσεις που έγιναν και για την πρώτη εφαρμογή, προκειμένου να αποκτηθεί μία αρχική εικόνα για το προσομοιωμένο σύνολο δεδομένων.

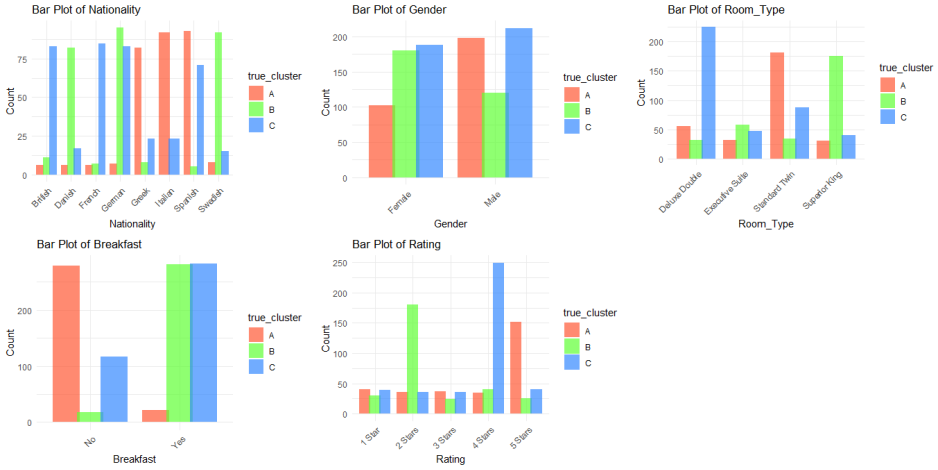
Κατασκευάζοντας τα θηκογράμματα και τα ιστογράμματα για τις αριθμητικές μεταβλητές, παρατηρούμε ότι οι ομάδες είναι όμοιες για όλες τις αριθμητικές μεταβλητές.

Σχήμα 8: Θηκογράμματα για τις αριθμητικές μεταβλητές



Σχήμα 9: Ιστογράμματα για τις αριθμητικές μεταβλητές

Τέλος, με τα κάτωθι ραβδογράμματα των κατηγορικών μεταβλητών, εύκολα μπορεί να παρατηρήσει κανείς πως υπάρχει ισχυρός διαχωρισμός στις ομάδες για τις μεταβλητές αξιολόγηση, τύπος του δωματίου προτίμησης και εθνικότητα.



Σχήμα 10: Ραβδογράμματα για τις κατηγορικές μεταβλητές

4.2 Εφαρμογή της μεθόδου k-means

Όπως έχει αναφερθεί προηγουμένως, εφαρμόσθηκε κανονικοποίηση για τις αριθμητικές μεταβλητές. Για να υπολογιστεί η σημαντικότητα των αριθμητικών μεταβλητών, θα πρέπει πρώτα να γίνει η διακριτοποίησή τους. Ο αριθμός των διαστημάτων υπολογίστηκε με βάση το μέσο όρο των τιμών των διακριτών μεταβλητών. Στη συγκεκριμένη περίπτωση είναι ίσος με 5. Στη συνέχεια, παρουσιάζονται οι πρώτες έξι γραμμές του προσομοιωμένου συνόλου δεδομένων με την πρώτη εφαρμογή, μετά την διακριτοποίηση, και οι κατηγορίες των αριθμητικών μεταβλητών κατά τη διακριτοποίηση.

Nationality	Gender	Room_Type	Breakfast	Age_Category	Price_Category	Income_Category
Spanish	Female	Executive Suite	Yes	u2	u2	u1
French	Female	Superior King	No	u2	u1	u1
British	Male	Deluxe Double	No	u2	u3	u1
Greek	Male	Standard Twin	Yes	u3	u2	u1
Greek	Female	Standard Twin	No	u2	u2	u2
French	Male	Standard Twin	No	u1	u1	u1

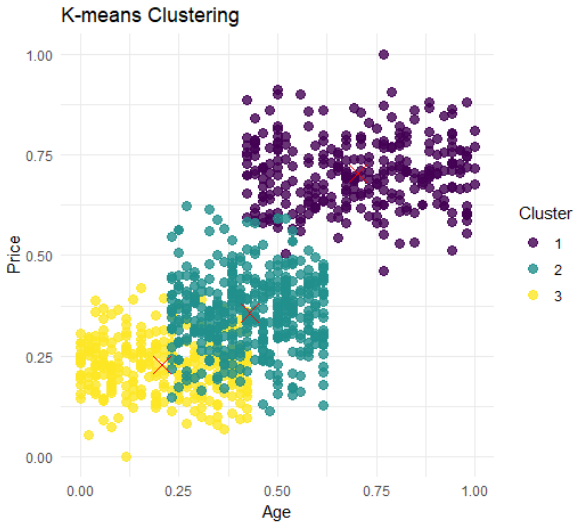
Attribute	Category	attr_min_col	attr_max_col
Age	u1	0.0000000	0.1923077
Age	u2	0.2115385	0.3846154
Age	u3	0.4038462	0.5961538
Age	u4	0.6153846	0.7884615
Age	u5	0.8076923	1.0000000
Income	u1	0.0000000	0.1997253
Income	u2	0.2014052	0.3540461
Income	u3	0.4147355	0.5998779
Income	u4	0.6009556	0.7998893
Income	u5	0.8003177	1.0000000
Price	u1	0.0000000	0.1999573
Price	u2	0.2007263	0.3989563
Price	u3	0.4008798	0.5989450
Price	u4	0.6015090	0.7992194
Price	u5	0.8014260	1.0000000

Ακολούθως, υπολογίσθηκε η σημαντικότητα των αριθμητικών μεταβλητών και εφαρμόσθηκε ο αλγόριθμος *k*-means κάνοντας χρήση των αποστάσεων και του υπολογισμού των κέντρων των ομάδων, όπως έχει αναλυθεί στην Ενότητα 3.3. Εφόσον έχουμε ορίσει εξαρχής ότι το προσομοιωμένο σύνολο δεδομένων αποτελείται από τρεις ομάδες, τόσες θα οριστούν και για την εφαρμογή του αλγόριθμου *k*-means. Επιπροσθέτως, ο αλγόριθμος *k*-means πραγματοποιήθηκε για εκατό επαναλήψεις. Τα αποτελέσματα της ομαδοποίησης του αλγόριθμου *k*-means είναι τα εξής:

	kmeans_results$cluster_assignments		
true_cluster	1	2	3
A	0	0	300
B	297	3	0
C	0	400	0

Παρατηρούμε πως για την ομάδα B, ομαδοποιήθηκαν σωστά 297 επισκέπτες στην πρώτη ομάδα και 3 λανθασμένα στη δεύτερη. Ο αλγόριθμος *k*-means, ομαδοποίησε σωστά τις ομάδες A και C στην τρίτη και δεύτερη ομάδα αντίστοιχα. Συνεπώς, ο αλγόριθμος *k*-means έχει μία πολύ καλή ομαδοποίηση για την πρώτη εφαρμογή.

Ενδεικτικά, παρουσιάζεται το διάγραμμα διασποράς για την ηλικία και την τιμή της διαμονής.

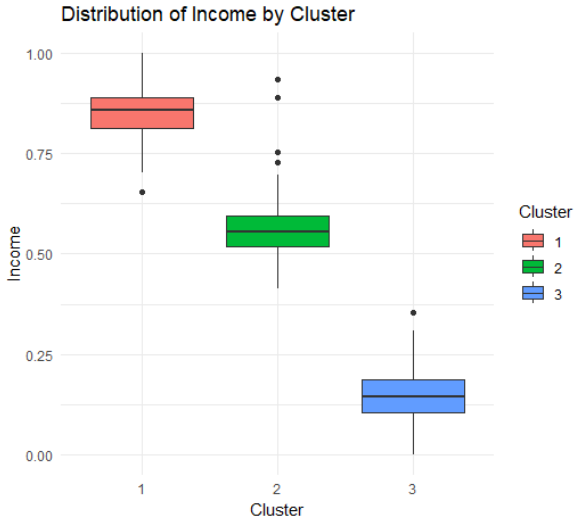


Σχήμα 11: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονής μετά την εφαρμογή του αλγόριθμου *k*-means

Σε αυτό το διάγραμμα διασποράς κάθε παρατήρηση χρωματίζεται σύμφωνα με την ομάδα που έχει τοποθετηθεί από τον αλγόριθμο *k*-means και τα κόκκινα σημάδια «X»

αντιπροσωπεύουν τα κέντρα των ομάδων. Είναι εύκολο να παρατηρήσει κανείς πως οι παρατηρήσεις είναι πολύ καλά διαχωρισμένες. Τέλος, η δεύτερη και τρίτη ομάδα είναι πιο κοντά μεταξύ τους συγκριτικά με την πρώτη ομάδα.

Αντίστοιχα, από το κάτωθι θηκόγραμμα που αφορά το εισόδημα παρατηρούμε πως όλες οι ομάδες είναι ισχυρά διαχωρισμένες μετά την εφαρμογή του αλγορίθμου *k*-means.



Σχήμα 12: Θηκόγραμμα για το εισόδημα μετά την εφαρμογή του αλγόριθμου *k*-means

Προχωρώντας με τη δεύτερη εφαρμογή, παρατηρούμε πως η διακριτοποίηση των αριθμητικών μεταβλητών είναι κάπως διαφορετική από την πρώτη εφαρμογή, όπως ήταν αναμενόμενο. Ακολούθως, παρουσιάζονται οι πρώτες έξι γραμμές του προσομοιωμένου συνόλου δεδομένων με την δεύτερη εφαρμογή, μετά την διακριτοποίηση, και οι κατηγορίες των αριθμητικών μεταβλητών κατά τη διακριτοποίηση.

Nationality	Gender	Room_Type	Breakfast	Age_Category	Price_Category	Income_Category
Spanish	Female	Deluxe Double	No	u2	u3	u2
Greek	Female	Superior King	No	u2	u2	u3
Greek	Male	Standard Twin	No	u2	u3	u3
Spanish	Male	Standard Twin	No	u3	u2	u3
Italian	Female	Standard Twin	No	u2	u3	u3
Greek	Male	Standard Twin	No	u1	u2	u1

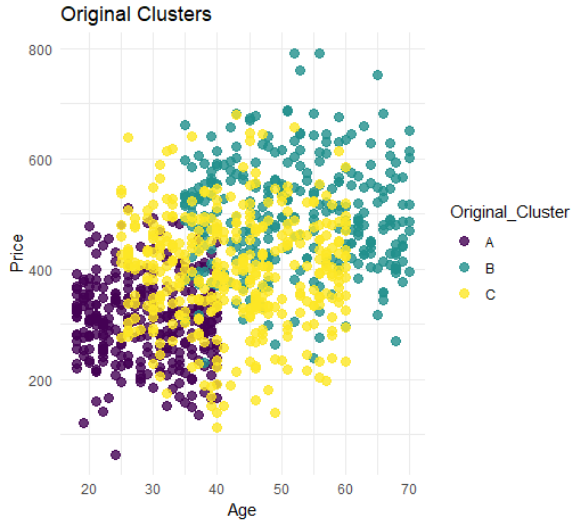
Attribute	Category	attr_min_col	attr_max_col
Age	u1	0.0000000	0.1923077
Age	u2	0.2115385	0.3846154
Age	u3	0.4038462	0.5961538
Age	u4	0.6153846	0.7884615
Age	u5	0.8076923	1.0000000
Income	u1	0.0000000	0.1994951
Income	u2	0.2011285	0.3983591
Income	u3	0.4001696	0.5998814
Income	u4	0.6005532	0.7994149
Income	u5	0.8016636	1.0000000
Price	u1	0.0000000	0.1993200
Price	u2	0.2012825	0.3999285
Price	u3	0.4004657	0.5961791
Price	u4	0.6006612	0.7998908
Price	u5	0.8003773	1.0000000

Όμως, η εφαρμογή του αλγόριθμου *k*-means, στο σύνολο δεδομένων που υπάρχει ισχυρός διαχωρισμός σε κατηγορικές μεταβλητές, δεν είναι τόσο καλή όσο σύνολο δεδομένων που υπάρχει ισχυρός διαχωρισμός σε αριθμητικές μεταβλητές. Από τα παρακάτω αποτελέσματα, παρατηρούμε πως για την ομάδα Α, ομαδοποιήθηκαν σωστά 291 επισκέπτες στη δεύτερη

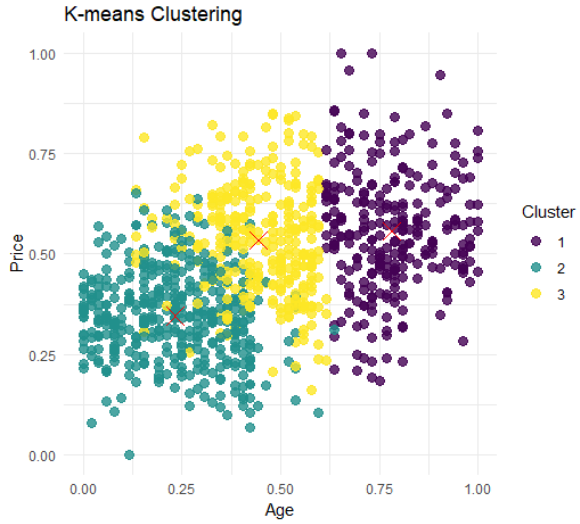
ομάδα και 9 λανθασμένα στην τρίτη. Όμως, τις ομάδες B και C δεν μπορούμε να τις ξεχωρίσουμε πολύ εύκολα. Θα μπορούσαμε να πούμε πως για την ομάδα B, ομαδοποίησε σωστά 165 επισκέπτες στην πρώτη ομάδα, 2 λανθασμένα στη δεύτερη ομάδα και 133 λανθασμένα στην τρίτη. Αντίστοιχα, για την ομάδα C, ομαδοποίησε σωστά 180 επισκέπτες στην τρίτη ομάδα, 100 λανθασμένα στην πρώτη ομάδα και 120 λανθασμένα στη δεύτερη. Συνεπώς, ο αλγόριθμος *k*-means έχει μία αρκετή καλή ομαδοποίηση μόνο για την ομάδα A και μία μέτρια ομαδοποίηση για τις ομάδες B και C.

	kmeans_results$cluster_assignments		
true_cluster	1	2	3
A	0	291	9
B	165	2	133
C	100	120	180

Ενδεικτικά, παρουσιάζεται το διάγραμμα διασποράς για την ηλικία και την τιμή της διαμονής, πριν και μετά την εφαρμογή του αλγορίθμου *k*-means. Είναι εύκολο να παρατηρήσει κανείς πως οι παρατηρήσεις δεν είναι πολύ καλά διαχωρισμένες. Τέλος, η δεύτερη και τρίτη ομάδα, μετά την εφαρμογή του αλγορίθμου *k*-means, είναι πιο κοντά μεταξύ τους συγκριτικά με την πρώτη ομάδα.

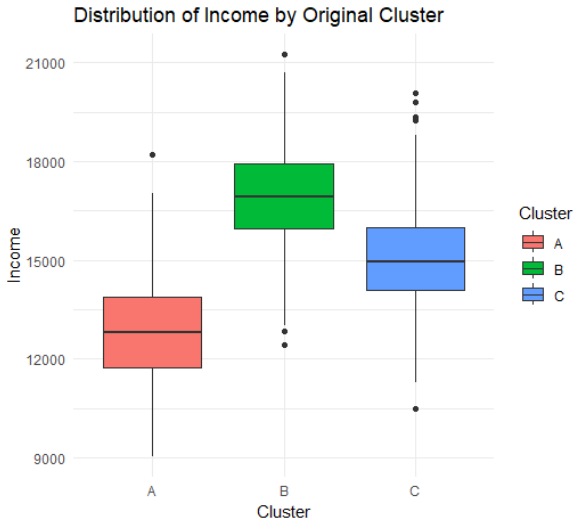


Σχήμα 13: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονής πριν την εφαρμογή του αλγόριθμου *k*-means

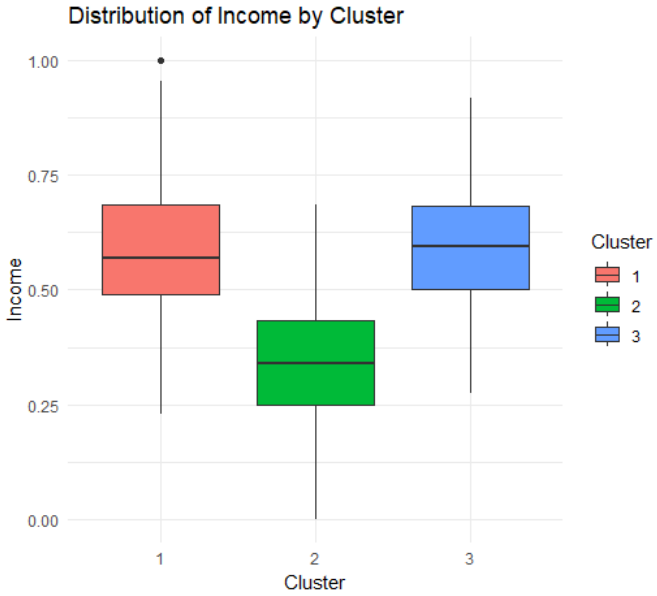


Σχήμα 14: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονής μετά την εφαρμογή του αλγόριθμου *k*-means

Όλα τα παραπάνω, μπορούν να επιβεβαιωθούν και από τα κάτωθι θηκογράμματα που αφορούν το εισόδημα, πριν και μετά την εφαρμογή του αλγορίθμου *k*-means. Η δεύτερη ομάδα μοιάζει αρκετά με την αρχική ομάδα Α, ενώ δεν μπορούμε να διακρίνουμε πολύ εύκολα την πρώτη και τρίτη ομάδα, σε ποιες αρχικές ομάδες αντιστοιχούν.



Σχήμα 15: Θηκόγραμμα για το εισόδημα πριν την εφαρμογή του αλγόριθμου *k*-means



Σχήμα 16: Θηκόγραμμα για το εισόδημα μετά την εφαρμογή του αλγόριθμου *k*-means

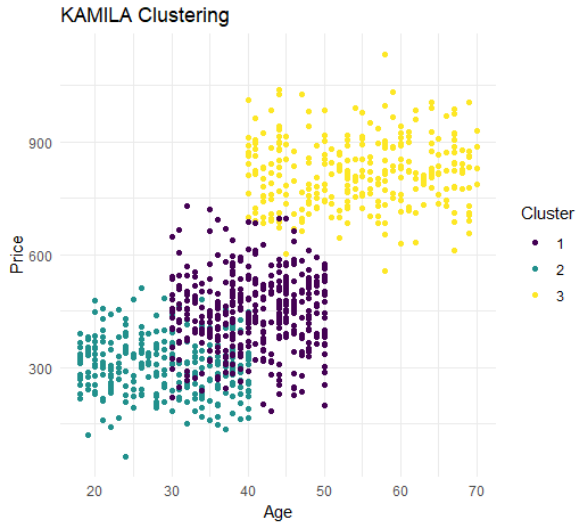
4.3 Εφαρμογή της μεθόδου KAMILA

Η εφαρμογή του αλγόριθμου KAMILA έγινε με την χρήση της εντολής «kamila» (Foss και Markatou (2018)). Εφόσον έχουμε ορίσει εξαρχής ότι το προσομοιωμένο σύνολο δεδομένων αποτελείται από τρεις ομάδες, τόσες θα οριστούν και για την εφαρμογή του αλγόριθμου KAMILA. Επίσης, θα γίνουν 100 αρχικοποιήσεις, λόγω του μεγέθους του προσομοιωμένου συνόλου δεδομένων. Τα αποτελέσματα της ομαδοποίησης του αλγόριθμου KAMILA, για την πρώτη εφαρμογή, είναι τα εξής:

	kamila_results$finalMemb		
true_cluster	1	2	3
A	0	300	0
B	4	0	296
C	399	1	0

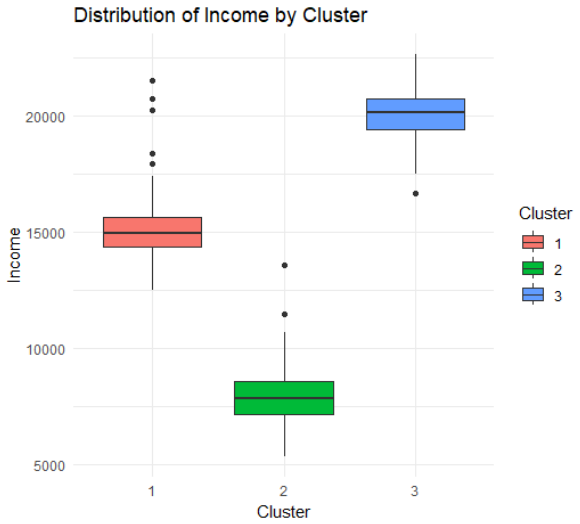
Παρατηρούμε πως για την ομάδα Α, ταξινομήθηκαν σωστά όλοι οι επισκέπτες στη δεύτερη ομάδα. Αντίστοιχα, για την ομάδα Β ομαδοποιήθηκαν σωστά 296 επισκέπτες στην τρίτη ομάδα και 4 λανθασμένα στην πρώτη. Τέλος, για την ομάδα C ομαδοποιήθηκαν σωστά 399 επισκέπτες στην πρώτη ομάδα και 1 λανθασμένα στη δεύτερη. Επομένως, ο αλγόριθμος KAMILA έχει μία πολύ καλή ομαδοποίηση για όλες τις ομάδες.

Ενδεικτικά, παρουσιάζεται το διάγραμμα διασποράς για την ηλικία και την τιμή της διαμονής, μετά την εφαρμογή του αλγόριθμου KAMILA. Το συγκεκριμένο διάγραμμα διασποράς, χρωματίζει την κάθε παρατήρηση σύμφωνα με την ομάδα που έχει τοποθετηθεί από τον αλγόριθμο KAMILA. Ενώ η πρώτη και η δεύτερη ομάδα είναι πολύ όμοιες μεταξύ τους, ο αλγόριθμος KAMILA ταξινόμησε μόνο έναν επισκέπτη λανθασμένα. Αντίστοιχα, από την τρίτη ομάδα ταξινόμησε λανθασμένα τέσσερις επισκέπτες στην πρώτη, που είναι στα όρια των δύο ομάδων, όπως φαίνεται από το διάγραμμα διασποράς.



Σχήμα 17: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονή μετά την εφαρμογή του αλγόριθμου KAMILA

Τέλος, από το κάτωθι θηκόγραμμα για το εισόδημα, μετά την εφαρμογή του αλγόριθμου KAMILA, φαίνεται πόσο καλά λειτούργησε ο αλγόριθμος KAMILA για τη συγκεκριμένη μεταβλητή.

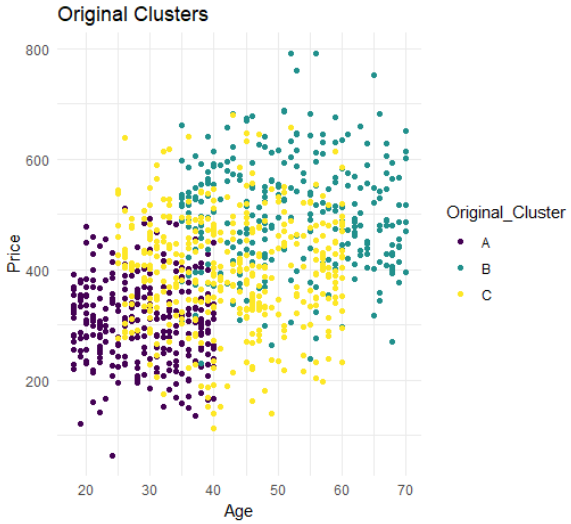


Σχήμα 18: Θηκόγραμμα για το εισόδημα μετά την εφαρμογή του αλγόριθμου KAMILA

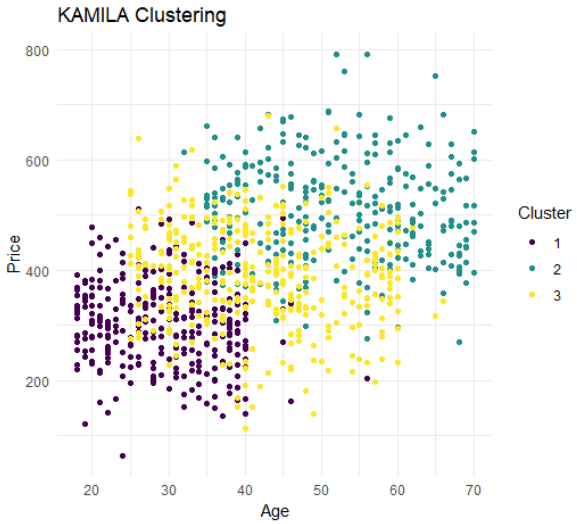
Όμως, η εφαρμογή του αλγόριθμου KAMILA, στο σύνολο δεδομένων που υπάρχει ισχυρός διαχωρισμός σε κατηγορικές μεταβλητές, δεν είναι τόσο καλή όσο σύνολο δεδομένων που υπάρχει ισχυρός διαχωρισμός σε αριθμητικές μεταβλητές. Από τα παρακάτω αποτελέσματα, παρατηρούμε πως για την ομάδα A, ομαδοποιήθηκαν σωστά 294 επισκέπτες στην πρώτη ομάδα και 6 λανθασμένα στην τρίτη. Αντίστοιχα, για την ομάδα B ομαδοποιήθηκαν σωστά 283 επισκέπτες στη δεύτερη ομάδα και 17 λανθασμένα στην τρίτη. Τέλος, για την ομάδα C ομαδοποιήθηκαν σωστά 348 επισκέπτες στην τρίτη ομάδα, 23 λανθασμένα στην πρώτη και 29 λανθασμένα στη δεύτερη. Συνεπώς, ο αλγόριθμος KAMILA έχει μία αρκετή καλή ομαδοποίηση για την ομάδα A και μία σχετικά μέτρια ομαδοποίηση για τις ομάδες B και C.

```
             kamila_results$finalMemb
true_cluster   1   2   3
           A 294   0   6
           B   0 283  17
           C  23  29 348
```

Ενδεικτικά, παρουσιάζεται το διάγραμμα διασποράς για την ηλικία και την τιμή της διαμονής, πριν και μετά την εφαρμογή του αλγορίθμου KAMILA. Είναι εύκολο να παρατηρήσει κανείς πως ενώ οι παρατηρήσεις δεν είναι καλά διαχωρισμένες, ο αλγόριθμος KAMILA έχει κάνει σε γενικές γραμμές, μία καλή ομαδοποίηση.



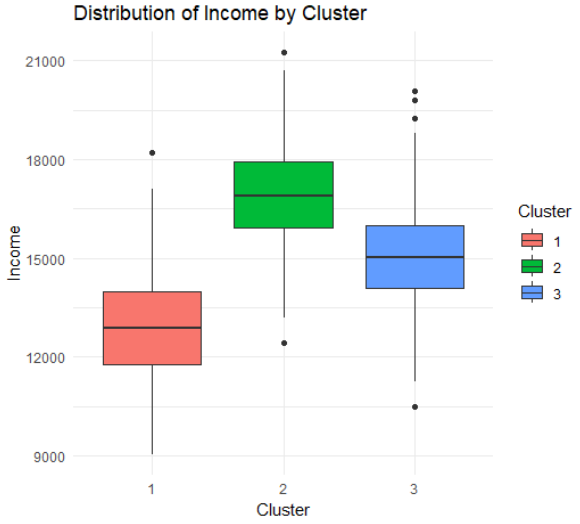
Σχήμα 19: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονής πριν την εφαρμογή του αλγόριθμου KAMILA



Σχήμα 20: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονής μετά την εφαρμογή του αλγόριθμου KAMILA

Όλα τα παραπάνω, μπορούν να επιβεβαιωθούν και από τα κάτωθι θηκογράμματα που αφορούν το εισόδημα, πριν και μετά την εφαρμογή του αλγορίθμου KAMILA. Και οι τρεις ομάδες μοιάζουν αρκετά με τις αντίστοιχες αρχικές ομάδες, όσον αφορά το εισόδημα.

Σχήμα 21: Θηκόγραμμα για το εισόδημα πριν την εφαρμογή του αλγόριθμου KAMILA



Σχήμα 22: Θηκόγραμμα για το εισόδημα μετά την εφαρμογή του αλγόριθμου KAMILA

4.4 Εφαρμογή της μεθόδου DBSCAN

Όπως έχουμε αναφέρει και στη θεωρητική παρουσίαση του αλγόριθμου DBSCAN, θα πρέπει πριν την εφαρμογή του να γίνει μετατροπή των κατηγορικών μεταβλητών. Για τις ανάγκες της παρουσίασης, χρησιμοποιήθηκαν οι τύποι BE και SE. Στη συνέχεια παρουσιάζονται οι αντίστοιχες μετατροπές, για την πρώτη εφαρμογή:

- BE

```
Column: Nationality
column
  British    Danish    French    German     Greek   Italian   Spanish   Swedish
0.2762183 0.2181175 0.2863490 0.2872296 0.2353899 0.2181175 0.3072230 0.2167151

Column: Gender
column
   Female      Male
0.3512245 0.3414106

Column: Room_Type
column
  Deluxe Double Executive Suite   Standard Twin   Superior King
       0.3561556       0.3266052       0.3587440       0.3351285

Column: Breakfast
column
       No       Yes
0.3536092 0.3382874

Column: Rating
column
   1 Star   2 Stars   3 Stars   4 Stars   5 Stars
0.2742861 0.3236935 0.3525753 0.3230965 0.3148756
```

- SE

```
Column: Nationality
column
   British     Danish     French     German      Greek    Italian    Spanish    Swedish
0.11048266 0.07470544 0.11816389 0.11885992 0.08418395 0.07470544 0.13616652 0.07396966

Column: Gender
column
   Female      Male
0.3581936 0.3846697

Column: Room_Type
column
  Deluxe Double Executive Suite   Standard Twin   Superior King
       0.2057286       0.1567912       0.2126964       0.1677741

Column: Breakfast
column
       No       Yes
0.3509137 0.3922924

Column: Rating
column
   1 Star   2 Stars   3 Stars   4 Stars   5 Stars
0.1090807 0.1533588 0.1974012 0.1526722 0.1437398
```

Στη συνέχεια, εφαρμόσθηκε η εναλλακτική προσέγγιση για τον προσδιορισμό της τιμής του Eps. Παρατηρούμε πως και για τους δύο τύπους τα αποτελέσματα είναι τα ίδια:

- BE

```
Eps value: 3011.857
Best MinPts: 279
```

- SE

```
Eps value: 3011.857
Best MinPts: 279
```

Τα αποτελέσματα της ομαδοποίησης του αλγόριθμου DBSCAN για τους δύο τύπους είναι τα εξής:

- BE

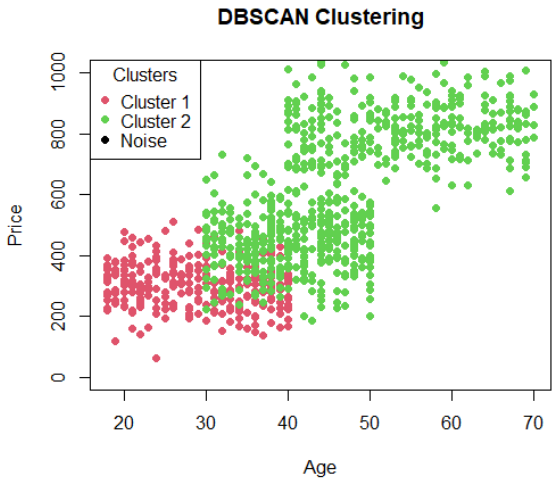
```
             dbscan_BE_results$cluster
true_cluster   1   2
           A 300   0
           B   0 300
           C   0 400
```

- SE

```
             dbscan_SE_results$cluster
true_cluster   1   2
           A 300   0
           B   0 300
           C   0 400
```

Παρατηρούμε πως και με τους δύο τύπους τα αποτελέσματα είναι ακριβώς τα ίδια. Για την ομάδα Α, ομαδοποιήθηκαν σωστά 300 επισκέπτες. Όμως, οι ομάδες Β και C ομαδοποιήθηκαν μαζί.

Επιπροσθέτως, παρουσιάζεται το διάγραμμα διασποράς για την ηλικία και την τιμή της διαμονή. Είναι εύκολο να παρατηρήσει κανείς ότι ο αλγόριθμος DBSCAN κατάφερε να ταξινομήσει τους επισκέπτες σε δύο ομάδες, βάσει των πυκνοτήτων των δεδομένων.



Σχήμα 23: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονή μετά την εφαρμογή του αλγόριθμου DBSCAN

Αντίστοιχα, για τη δεύτερη εφαρμογή έχουμε παρόμοια αποτελέσματα. Οι αντίστοιχες μετατροπές παρουσιάζονται παρακάτω:

- BE

```
Column: Nationality
column
  British    Danish    French    German     Greek   Italian   Spanish   Swedish
0.2302585 0.2366485 0.2276332 0.3121689 0.2463815 0.2487247 0.3004578 0.2487247

Column: Gender
column
   Female      Male
0.3548606 0.3364855

Column: Room_Type
column
  Deluxe Double Executive Suite   Standard Twin   Superior King
      0.3635658       0.2723251       0.3619812       0.3449962

Column: Breakfast
column
       No       Yes
0.3651022 0.3131792

Column: Rating
column
   1 Star   2 Stars   3 Stars   4 Stars   5 Stars
0.2415884 0.3473382 0.2263053 0.3651518 0.3320707
```

- SE

```
Column: Nationality
column
   British     Danish     French     German      Greek    Italian    Spanish    Swedish
0.08128373 0.08490691 0.07982840 0.14098795 0.09066245 0.09209396 0.12995424 0.09209396

Column: Gender
column
   Female      Male
0.3468992 0.3965625

Column: Room_Type
column
  Deluxe Double Executive Suite   Standard Twin   Superior King
       0.2295452       0.1076769       0.2232043       0.1829001

Column: Breakfast
column
       No       Yes
0.3032740 0.4462453

Column: Rating
column
    1 Star    2 Stars    3 Stars    4 Stars    5 Stars
0.08779077 0.18703545 0.07909939 0.23733624 0.16365497
```

Οι τιμές για Eps και MinPts παρουσιάζονται στη συνέχεια. Παρατηρούμε πως και για τους δύο τύπους τα αποτελέσματα είναι τα ίδια:

- BE

```
Eps value: 5239.261
Best MinPts: 977
```

- SE

```
Eps value: 5239.261
Best MinPts: 977
```

Τα αποτελέσματα της ομαδοποίησης του αλγόριθμου DBSCAN για τους δύο τύπους είναι τα εξής:

- BE

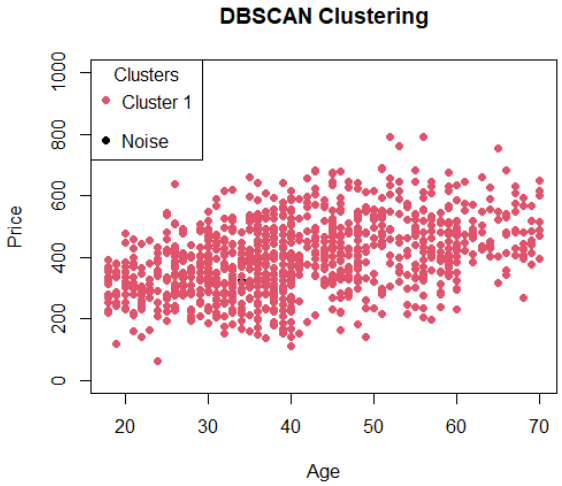
```
             dbscan_BE_results$cluster
true_cluster   0   1
           A   1 299
           B   0 300
           C   0 400
```

- SE

```
             dbscan_SE_results$cluster
true_cluster   0   1
           A   1 299
           B   0 300
           C   0 400
```

Παρατηρούμε πως και με τους δύο τύπους τα αποτελέσματα είναι ακριβώς τα ίδια. Για τη δεύτερη εφαρμογή ο αλγόριθμος DBSCAN είχε χειρότερα αποτελέσματα, καθώς δεν μπόρεσε τελικά να ξεχωρίσει τις ισχυρά διαχωρισμένες κατηγορικές μεταβλητές και ταξινόμησε 999 επισκέπτες σε μία ομάδα και έναν επισκέπτη τον θεώρησε σαν θόρυβο.

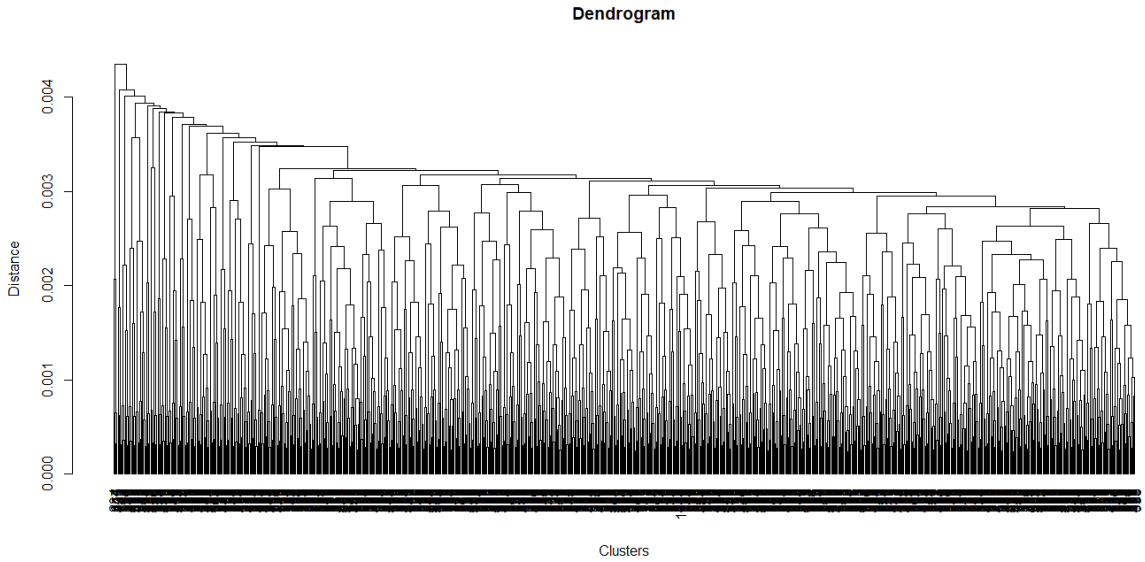
Τέλος, παρουσιάζεται το διάγραμμα διασποράς για την ηλικία και την τιμή της διαμονή. Ο αλγόριθμος DBSCAN, για τη δεύτερη εφαρμογή, δεν είχε καθόλου καλή ομαδοποίηση.



Σχήμα 24: Γράφημα διασποράς για την ηλικία και την τιμή της διαμονή μετά την εφαρμογή του αλγόριθμου DBSCAN

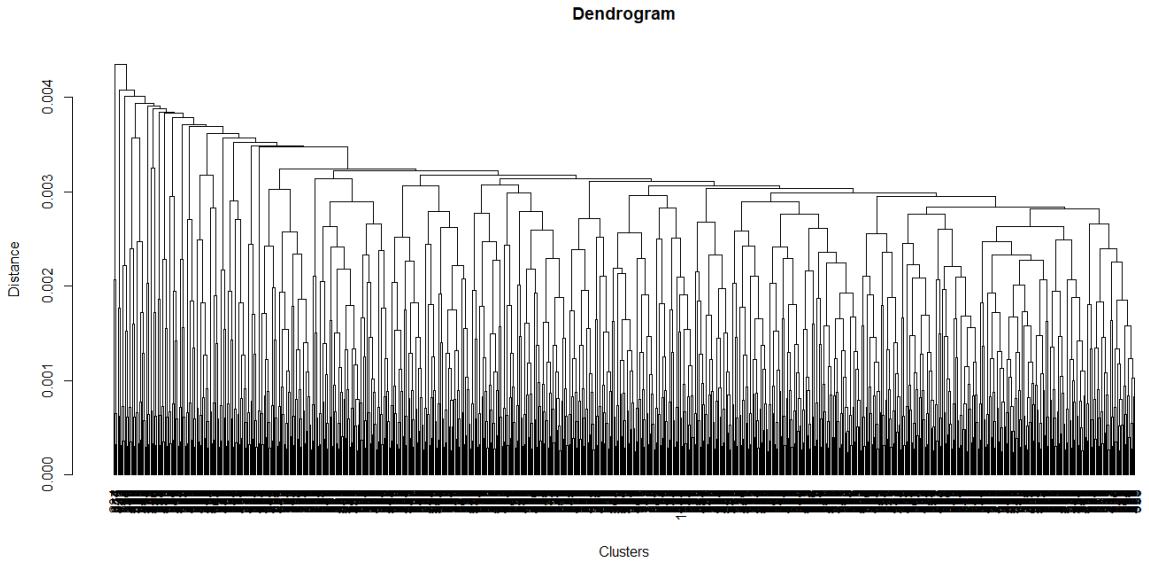
4.5 Εφαρμογή της συσσωρευτικής ιεραρχικής μεθόδου

Για να εφαρμοσθεί η συσσωρευτική ιεραρχική μέθοδος, πρώτα υπολογίστηκαν οι απαραίτητες ιεραρχίες απόστασης για κάθε κατηγορική μεταβλητή. Για τις ανάγκες της παρουσίασης, δόθηκε βάρος ίσο με 0.5 και η παράμετρος L ορίσθηκε ίση με 2. Επίσης, χρησιμοποιήθηκαν οι μέθοδοι της πλήρους συνένωσης και της μέσης συνένωσης. Τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο της πλήρους συνένωσης, για την πρώτη εφαρμογή, εμφανίζονται παρακάτω:



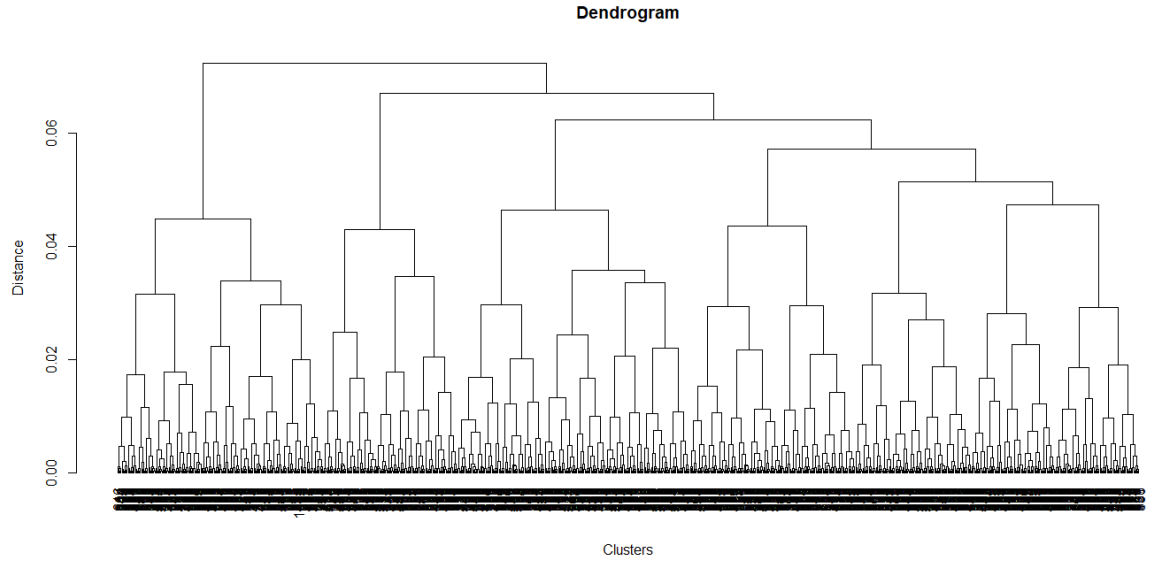
Σχήμα 25: Δενδρόγραμμα του συσσωρευτικού αλγόριθμου με τη μέθοδο της πλήρους συνένωσης

Τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο της μέσης συνένωσης παρουσιάζονται παρακάτω:



Σχήμα 26: Δενδρόγραμμα του συσσωρευτικού αλγόριθμου με τη μέθοδο της μέσης συνένωσης

Τέλος, τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο του Ward παρουσιάζονται παρακάτω:

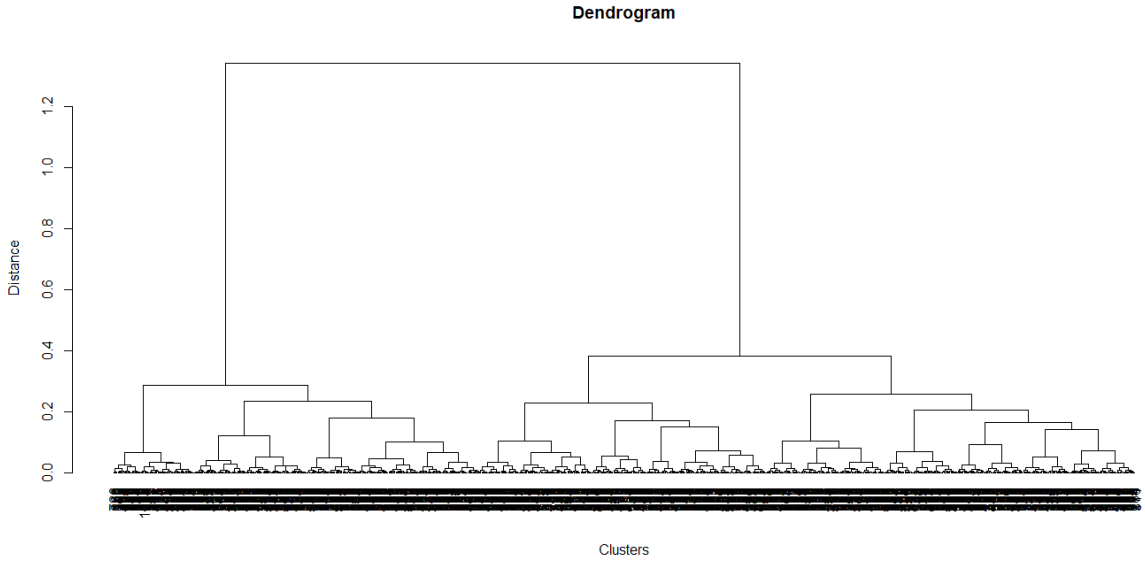


Σχήμα 27: Δενδρόγραμμα του συσσωρευτικού αλγόριθμου με τη μέθοδο του Ward

Παρατηρούμε πως τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο του Ward τείνει να σχηματίζει πιο συμπαγείς ομάδες. Εν αντιθέσει, τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο της πλήρους συνένωσης και της μέσης συνένωσης δημιουργεί πιο επιμήκεις ομάδες. Τα αποτελέσματα της ομαδοποίησης του ιεραρχικού αλγόριθμου με τη μέθοδο του Ward είναι τα εξής:

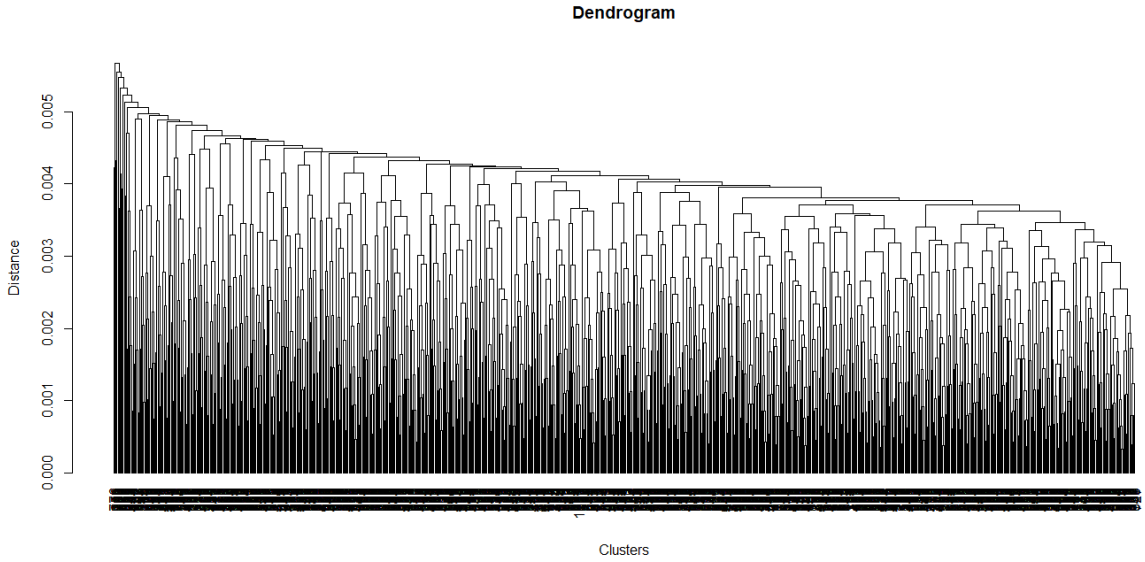
	clusters		
true_cluster	1	2	3
A	200	40	60
B	200	40	60
C	267	51	82

Αντίστοιχα, παρουσιάζονται τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο της πλήρους συνένωσης, για τη δεύτερη εφαρμογή:



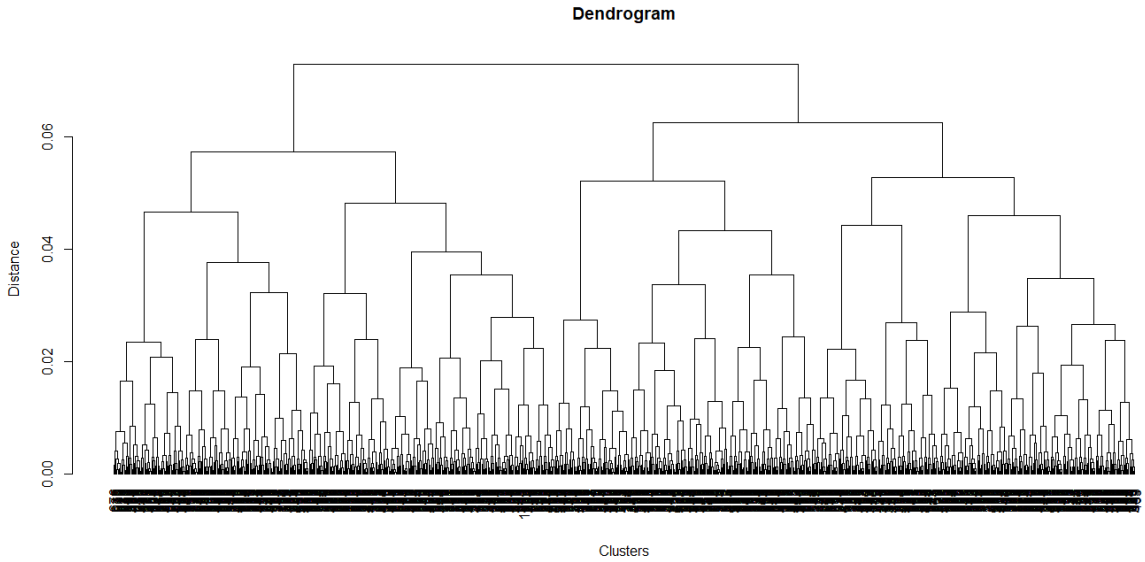
Σχήμα 28: Δενδρόγραμμα του συσσωρευτικού αλγόριθμου με τη μέθοδο της πλήρους συνένωσης

Τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο της μέσης συνένωσης παρουσιάζονται παρακάτω:



Σχήμα 29: Δενδρόγραμμα του συσσωρευτικού αλγόριθμου με τη μέθοδο της μέσης συνένωσης

Τέλος, τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο του Ward παρουσιάζονται παρακάτω:



Σχήμα 30: Δενδρόγραμμα του συσσωρευτικού αλγόριθμου με τη μέθοδο του Ward

Παρατηρούμε πως τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο του Ward τείνει να σχηματίζει πιο συμπαγείς ομάδες. Με τη μέθοδο της πλήρους συνένωσης παρατηρούμε πως ενώνει σε αρχικά στάδια πολλούς επισκέπτες, με αποτέλεσμα η τελευταία ένωση να είναι πολύ μακριά. Εν αντιθέσει, τα αποτελέσματα του συσσωρευτικού αλγόριθμου με τη μέθοδο της μέσης συνένωσης δημιουργεί πιο επιμήκεις ομάδες. Τα αποτελέσματα της ομαδοποίησης του ιεραρχικού αλγόριθμου με τη μέθοδο του Ward είναι τα εξής:

	clusters		
true_cluster	1	2	3
A	132	95	73
B	130	96	74
C	170	121	109

4.6 Σύγκριση των αποτελεσμάτων

Στις προηγούμενες ενότητες, παρουσιάστηκαν τα αποτελέσματα από την εφαρμογή των τεσσάρων αλγορίθμων στο προσομοιωμένο σύνολο δεδομένων. Έγιναν δύο εφαρμογές σε κάθε αλγόριθμο: στην πρώτη εφαρμογή, υπήρχε ισχυρός διαχωρισμός σε αριθμητικές μεταβλητές και στη δεύτερη ισχυρός διαχωρισμός σε κατηγορικές μεταβλητές. Από τα αποτελέσματα, για την πρώτη εφαρμογή, καταλήγουμε στο ότι ο αλγόριθμος KAMILA και *k*-means είχε την καλύτερη ομαδοποίηση από όλους. Την χειρότερη ομαδοποίηση είχε ο αλγόριθμος DBSCAN. Ενώ κατάφερε να ταξινομήσει σωστά την ομάδα Α, ομαδοποίησε μαζί τις ομάδες Β και C. Τέλος, ο συσσωρευτικός αλγόριθμος με τη μέθοδο της του Ward είχε καλύτερα αποτελέσματα από τον συσσωρευτικό αλγόριθμο με τη μέθοδο της πλήρους συνένωσης και της μέσης συνένωσης. Παρουσιάζονται τα συνολικά ποσοστά ορθών ταξινομήσεων:

	k-means				KAMILA				DBSCAN				Agglomerative			
	1	2	3	Συνολικό %Ορθών Ταξινομήσεων	1	2	3	Συνολικό %Ορθών Ταξινομήσεων	1	2	3	Συνολικό %Ορθών Ταξινομήσεων	1	2	3	Συνολικό %Ορθών Ταξινομήσεων
A	0	0	300	99,7%	0	300	0	99,5%	300	0	0	60,0%	200	40	60	32,2%
B	297	3	0		4	0	296		0	300	0		200	40	60	
C	0	400	0		399	1	0		0	400	0		267	51	82	

Πίνακας 3: Συνολικά ποσοστά ορθών ταξινομήσεων ανά αλγόριθμο για την πρώτη εφαρμογή

Αντίστοιχα, για τη δεύτερη εφαρμογή, από τα αποτελέσματα καταλήγουμε στο ότι ο αλγόριθμος KAMILA είχε την καλύτερη ομαδοποίηση από όλους. Ο αλγόριθμος *k*-means έχει μία αρκετή καλή ομαδοποίηση μόνο για την ομάδα A και μία μέτρια ομαδοποίηση για τις ομάδες B και C. Την χειρότερη ομαδοποίηση είχε ο αλγόριθμος DBSCAN, που ομαδοποίησε όλους τους επισκέπτες σε μία ομάδα, πέρα από έναν που τον θεώρησε σαν θόρυβο. Τέλος, ο συσσωρευτικός αλγόριθμος με τη μέθοδο της του Ward είχε καλύτερα αποτελέσματα από τον συσσωρευτικό αλγόριθμο με τη μέθοδο της πλήρους συνένωσης και της μέσης συνένωσης. Παρουσιάζονται τα συνολικά ποσοστά ορθών ταξινομήσεων:

	***k*-means**				**KAMILA**				**DBSCAN**					**Agglomerative**			
	1	2	3	Συνολικό %Ορθών Ταξινομήσεων	1	2	3	Συνολικό %Ορθών Ταξινομήσεων	0	1	2	3	Συνολικό %Ορθών Ταξινομήσεων	1	2	3	Συνολικό %Ορθών Ταξινομήσεων
A	0	291	9	63,6%	294	0	6	92,5%	1	299	0	0	29,9%	132	95	73	34%
B	165	2	133		0	283	17		0	300	0	0		130	96	74	
C	100	120	180		23	29	348		0	400	0	0		170	121	109	

Πίνακας 4: Συνολικά ποσοστά ορθών ταξινομήσεων ανά αλγόριθμο για τη δεύτερη εφαρμογή

ΚΕΦΑΛΑΙΟ 5

Εφαρμογές

Το συγκεκριμένο κεφάλαιο, έχει ως σκοπό την αναφορά εφαρμογών που έγιναν για τις τέσσερις μεθόδους που έχουμε παρουσιάσει. Αυτές οι εφαρμογές έχουν ανακτηθεί από την βιβλιογραφία, όπου εφαρμόσθηκαν σε πραγματικά δεδομένα, και θα γίνει μία συνοπτική περιγραφή των αποτελεσμάτων.

5.1 Εφαρμογές της μεθόδου k-means για μικτά δεδομένα

Παρακάτω θα περιγράφουν τέσσερις εφαρμογές του αλγόριθμου *k*-means όπως έγιναν από τους Ahmad και Dey (2007). Τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι: Iris, Vote, Heart disease και Australian credit data και ελήφθησαν από το αποθετήριο UCI (https://archive.ics.uci.edu). Το σύνολο δεδομένων Iris απαρτίζεται μόνο από αριθμητικές μεταβλητές. Το σύνολο δεδομένων Vote απαρτίζεται μόνο από κατηγορικές μεταβλητές. Τα σύνολα δεδομένων Heart disease και Australian credit data είναι μικτά σύνολα δεδομένων, δηλαδή δεδομένα που περιλαμβάνουν τόσο αριθμητικές όσο και κατηγορικές μεταβλητές.

Για να αξιολογηθεί η ποιότητα της ομαδοποίησης, έγινε η υπόθεση ότι τα σύνολα δεδομένων είναι προ-ταξινομημένα (*pre-classified*), δηλαδή κάθε άτομο έχει ήδη κάποιου είδους ετικέτας/ταμπέλας. Είναι σημαντικό να σημειωθεί πως οι ετικέτες/ταμπέλες που έχουν αντιστοιχηθεί σε κάθε άτομο δεν χρησιμοποιούνται κατά τη διαδικασία της δημιουργίας ομάδων, καθώς πρέπει να εξασφαλιστεί μια αμερόληπτη αξιολόγηση. Επιπροσθέτως, μετράται ο βαθμός επικάλυψης (*overlap*) μεταξύ της ομαδοποίησης που εφαρμόσθηκε και της πραγματικής ομαδοποίησης. Ο αριθμός των διαστημάτων για την διακριτοποίηση των αριθμητικών μεταβλητών, όπως έχει αναλυθεί στην Ενότητα 3.3, τίθεται ίσος με πέντε, εκτός εάν ορίζεται διαφορετικά. Τέλος, παρουσιάζονται τα αποτελέσματα της ομαδοποίησης όπως προέκυψαν από τον αλγόριθμο *k*-means, όπου έχει χρησιμοποιηθεί μέθοδος τυχαίας αρχικοποίησης.

- **Iris**

Το συγκεκριμένο σύνολο δεδομένων αποτελείται από τέσσερις αριθμητικές μεταβλητές και 150 λουλούδια, τα οποία είναι ισόνομα κατανεμημένα σε τρεις διαφορετικές κατηγορίες: Iris Setosa, Iris Versicolour και Iris Virginica. Εφόσον όλες οι μεταβλητές είναι αριθμητικές, η συνάρτηση κόστους θα υπολογιστεί μόνο για αυτές, με βάρος την σημαντικότητα αυτών, όπως έχει οριστεί προηγουμένως. Στον Πίνακα 5 παρουσιάζονται οι τιμές της σημαντικότητας κάθε μεταβλητής:

Significance of attributes for Iris data

Attribute of Iris data	A1	A2	A3	A4
Significance (w_l)	0.70	0.67	0.78	0.77

Πίνακας 5: Σημαντικότητα κάθε αριθμητικής μεταβλητής για το σύνολο δεδομένων Iris από τους Ahmad και Dey (2007)

Είναι εύκολο να παρατηρήσει κανείς πως η τρίτη μεταβλητή είναι η πιο σημαντική, με πολύ μικρή διαφορά από την τέταρτη. Ακολουθούν η πρώτη και η δεύτερη μεταβλητή. Η σειρά των μεταβλητών ως προς την σημαντικότητά τους είναι συνεπής με τα ήδη υπάρχοντα

αποτελέσματα. Τέλος, παρουσιάζονται τα αποτελέσματα της ομαδοποίησης των δεδομένων με τον αλγόριθμο *k*-means:

Cluster recovery result for iris data set with proposed algorithm

Cluster no.	Iris Setosa	Iris Versicolour	Iris Virginica
1	50	0	0
2	0	47	5
3	0	3	45

Πίνακας 6: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Iris από τους Ahmad και Dey (2007)

Ο αλγόριθμος *k*-means εκτελέσθηκε πάνω από 100 φορές και τα αποτελέσματα είναι ο μέσος όρος αυτών. Όπως φαίνεται από τον Πίνακα 6, συνολικά οχτώ λουλούδια δεν ταξινομήθηκαν στη σωστή ομάδα. Τρία από αυτά αφορούν την κατηγορία Iris Versicolour (0.06% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην $3^{η}$ ομάδα αντί για την $2^{η}$. Αντίστοιχα, πέντε από αυτά αφορούν την κατηγορία Iris Virginica (0.1% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην $2^{η}$ ομάδα αντί για την $3^{η}$. Τέλος, η τιμή της μέσης μικροακρίβειας είναι ίση με 0.95, η οποία είναι πολύ κοντά στη μονάδα. Συνεπώς, ο αλγόριθμος *k*-means είχε αποτελέσματα πολύ κοντά στα ήδη υπάρχοντα αποτελέσματα.

- **Vote**

Το σύνολο δεδομένων Vote είναι ένα σύνολο από 435 άτομα, για τα οποία καταγράφηκαν 16 κατηγορικές μεταβλητές. Τα δεδομένα ανήκουν σε δύο κατηγορίες: Δημοκρατικοί 168 άτομα και Ρεπουμπλικάνοι 267 άτομα. Τα αποτελέσματα προήλθαν από τον μέσο όρο των πάνω από 100 εκτελέσεων του αλγόριθμου. Εφαρμόσθηκε ο αλγόριθμος *k*-means για να ανακαλυφθούν κοινά μοτίβα στις κατανομές ψηφοφορίας. Το σύνολο των ατόμων που δεν βρίσκονται στις σωστές ομάδες τους είναι 58. Από αυτά 13 αφορούν τους Ρεπουμπλικάνους (0.08% επί της πραγματικής ομάδας), οι οποίοι ταξινομήθηκαν στην $2^{η}$ ομάδα αντί για την $1^{η}$. Αντίστοιχα, 45 από αυτά αφορούν τους Δημοκρατικούς (0.17% επί της πραγματικής ομάδας), που ταξινομήθηκαν στην $1^{η}$ ομάδα αντί για την $2^{η}$. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 7:

Cluster recovery result for Vote data set with proposed algorithm

Cluster no.	No. of republican	No. of democrat
1	155	45
2	13	222

Πίνακας 7: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Vote από τους Ahmad και Dey (2007)

Όμως, παρατηρήθηκε πως υπήρχαν ακραίες τιμές στο σύνολο δεδομένων, κάτι το οποίο φαίνεται και από την τιμή της τυπικής απόκλισης. Για αυτόν τον λόγο, χρησιμοποιήθηκε μία συγκεκριμένη μέθοδος για την αφαίρεση αυτών. Εισάχθηκε μια παράμετρος που ονομάζεται *r*, η οποία είναι ένας πραγματικός αριθμός που χρησιμεύει ως μια απόσταση κατωφλίου (*threshold distance*). Αν η απόσταση μεταξύ ενός ατόμου και του πλησιέστερου κέντρου ομάδας είναι μεγαλύτερη από *r*, το άτομο θεωρείται ακραία τιμή και αφαιρείται από το σύνολο. Η τιμή του *r* ορίζεται ως η διπλάσια μέση απόσταση όλων των παρατηρήσεων από τα αντίστοιχα κέντρα ομάδων. Τελικά, εφαρμόζοντας ξανά τον αλγόριθμο, όπως και προηγουμένως, 31 άτομα δεν τοποθετήθηκαν στις σωστές ομάδες. Από αυτά 6 αφορούν τους Ρεπουμπλικάνους (0.04% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην $2^{η}$ ομάδα αντί για την $1^{η}$. Αντίστοιχα, 25 από αυτά αφορούν τους Δημοκρατικούς (0.11% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην $1^{η}$ ομάδα αντί για την $2^{η}$. Τέλος, η τιμή της μέσης μικροακρίβειας είναι ίση με 0.92, η οποία είναι πολύ κοντά στη μονάδα. Κατά συνέπεια,

ο αλγόριθμος *k*-means είχε αποτελέσματα πολύ κοντά στα ήδη υπάρχοντα αποτελέσματα. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 8:

Cluster recovery result for Vote data set with our method after outlier removal

Cluster no.	No. of republican	No. of democrat
1	141	25
2	6	200

Πίνακας 8: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Vote μετά την αφαίρεση ακραίων τιμών από τους Ahmad και Dey (2007)

- **Heart disease data**

Αυτά τα δεδομένα προέρχονται από το Κλινικό Κέντρο του Κλίβελαντ και ένα μικτό σύνολο δεδομένων, το οποίο απαρτίζεται από οκτώ κατηγορικές και πέντε αριθμητικές μεταβλητές. Από τα 303 άτομα που αναφέρονται, η μια κατηγορία περιλαμβάνει 164 φυσιολογικούς ασθενείς και η άλλη 139 ασθενείς με καρδιακή νόσο. Ο αριθμός των διαστημάτων, θεωρείται ότι είναι ο μέσος όρος των τιμών των κατηγορικών μεταβλητών, ο οποίος είναι ίσος με 3. Ο αλγόριθμος *k*-means εκτελέσθηκε πάνω από 100 φορές και τα αποτελέσματα είναι ο μέσος όρος αυτών. Ακολουθούν τα αποτελέσματα της ανάλυσης στον Πίνακα 9:

Cluster recovery for Heart disease data set with our proposed algorithm

Cluster no.	Normal	Heart patient
1	139	21
2	25	118

Πίνακας 9: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Heart disease από τους Ahmad και Dey (2007)

Όπως φαίνεται από τον Πίνακα 9, συνολικά 46 άτομα δεν ταξινομήθηκαν στη σωστή ομάδα. Από αυτά 25 αφορούν τους φυσιολογικούς ασθενείς (0.15% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην $2^{η}$ ομάδα αντί για την $1^{η}$. Αντίστοιχα, 21 από αυτά αφορούν τους ασθενείς με καρδιακή νόσο (0.15% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην $1^{η}$ ομάδα αντί για την $2^{η}$. Τέλος, η τιμή της μέσης μικροακρίβειας είναι ίση με 0.85, η οποία είναι πολύ κοντά στη μονάδα. Συνεπώς, ο αλγόριθμος *k*-means είχε αρκετά αποτελεσματική ομαδοποίηση συγκριτικά με τα υπάρχοντα αποτελέσματα.

- **Australian credit data**

Τα δεδομένα λήφθηκαν από οργανισμό πιστωτικών καρτών της Αυστραλίας και αφορά τις αιτήσεις για πιστωτικές κάρτες που υποβάλλονται σε μία μεγάλη τράπεζα. Είναι ένα μικτό σύνολο δεδομένων με οκτώ κατηγορικές και έξι αριθμητικές μεταβλητές. Τα δεδομένα μας περιέχουν σύνολο 690 πελατών. Οι πελάτες τους χωρίζονται σε δύο κατηγορίες: αυτοί που έχουν θετική πιστοληπτική ικανότητα (307) και εκείνοι που έχουν αρνητική πιστοληπτική ικανότητα (383). Ο αριθμός των διαστημάτων, θεωρείται ότι είναι ο μέσος όρος των τιμών των κατηγορικών μεταβλητών, ο οποίος είναι ίσος με 5. Τα αποτελέσματα προήλθαν από τον μέσο όρο των πάνω από 100 εκτελέσεων. Παραθέτουμε τον Πίνακα 10 με τα αποτελέσματα του αλγόριθμου *k*-means:

Cluster recovery for Australian credit data set with our proposed algorithm

Cluster	Credit (negative)	Credit (positive)
1	321	19
2	62	288

Πίνακας 10: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Australian credit από τους Ahmad και Dey (2007)

Όπως φαίνεται από τον Πίνακα 10, συνολικά 81 πελάτες δεν ταξινομήθηκαν στη σωστή ομάδα. Από αυτά 62 αφορούν τους πελάτες που έχουν αρνητική πιστοληπτική ικανότητα

(0.16% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην 2η ομάδα αντί για την 1η. Αντίστοιχα, 19 από αυτούς αφορούν τους πελάτες που έχουν θετική πιστοληπτική ικανότητα (0.06% επί της πραγματικής ομάδας), όπου ταξινομήθηκαν στην 1η ομάδα αντί για την 2η. Τέλος, η τιμή της μέσης μικροακρίβειας είναι ίση με 0.88. Επομένως, ο αλγόριθμος *k*-means είχε σχετικά καλή ομαδοποίηση συγκριτικά με τα ήδη υπάρχοντα αποτελέσματα.

5.2 Εφαρμογές της μεθόδου KAMILA για μικτά δεδομένα

Στη συνέχεια, θα περιγράφουν τρεις εφαρμογές του αλγόριθμου KAMILA όπως έγιναν από τους Foss, Markatou, Ray και Heching (2016). Τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι: Australian credit data, Cylinder bands και The insurance company benchmark (COIL 2000), τα οποία ελήφθησαν από το αποθετήριο UCI (https://archive.ics.uci.edu). Και τα τρία σύνολα δεδομένων απαρτίζονται τόσο από αριθμητικές όσο και κατηγορικές μεταβλητές. Κατά την αξιολόγηση της απόδοσης του αλγόριθμου KAMILA, ο αριθμός των ομάδων ορίστηκε να είναι ίσος με τον πραγματικό αριθμό των κατηγοριών στο σύνολο δεδομένων. Αυτό γίνεται για να εκτιμηθεί πόσο καλά ο αλγόριθμος KAMILA μπορεί να προσδιορίσει την πραγματική δομή των δεδομένων. Με άλλα λόγια, η εκτίμηση γίνεται με την υπόθεση ότι ο πραγματικός αριθμός των ομάδων είναι γνωστός και ο αλγόριθμος δοκιμάζεται για την ικανότητά του να τις αναγνωρίσει σωστά. Στη συνέχεια, παρουσιάζουμε έναν πίνακα που περιέχει μία λεπτομερή περιγραφή των συγκεκριμένων συνόλων δεδομένων, καθώς και τη μεταβλητή με τα πραγματικά αποτελέσματα:

Key characteristics of the data sets analyzed

Data set name	# Obs	# Continuous Variables	# Categorical Variables	Outcome Variable
Australian Credit	690	6	8	Acc (44%) Rej (56%)
Bands	516	7	13	Band (41%) No Band (59%)
The Insurance Company Benchmark (COIL 2000)	9822	3	38	Successful hedonists (9.8%) Career loners (0.8%) Retired and religious (9.0%) Farmers (5.0%) Driven growers (8.4%) Living well (9.6%) Family with grown ups (27.4%) Average family (15.4%) Cruising seniors (3.3%) Conservative families (11.3%)

Πίνακας 11: Περιγραφή των συνόλων δεδομένων για την εφαρμογή του KAMILA από τους Foss, Markatou, Ray και Heching (2016)

- **Australian credit data**

Αυτό το σύνολο δεδομένων είναι το ίδιο που εφαρμόσθηκε και στον αλγόριθμο *k*-means. Αξίζει να σημειωθεί πως αυτό το σύνολο δεδομένων λειτουργεί ως ένα σημείο αναφοράς για την αξιολόγηση αλγορίθμων επιβλεπόμενης μάθησης. Όμως, για τους αλγορίθμους μη

επιβλεπόμενης μάθησης, η πρόκληση έγκειται στη διάκριση των υποκείμενων κατηγοριών αποτελεσμάτων και στην κατανομή των παρατηρήσεων χωρίς τη βοήθεια ενός συνόλου εκπαίδευσης (*training set*). Ακολουθούν τα αποτελέσματα της ανάλυσης στον Πίνακα 12:

Results of Australian Credit data set analysis, purity and macro precision/recall

	Purity	Purity % over M–S	Macro precision/recall	Macro P/R % over M–S
KAMILA	0.775	93.5%	0.808/0.755	96.5%/ 92.2%

Πίνακας 12: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Australian credit KAMILA από τους Foss, Markatou, Ray και Heching (2016)

Παρατηρούμε πως ο αλγόριθμος KAMILA έχει μία αρκετά καλή ομαδοποίηση. Η τιμή της καθαρότητας είναι ίση με 0.775, η τιμή της μάκρο-ακρίβεια είναι ίση με 0.808 και η τιμή της μάκρο-ευαισθησίας είναι ίση με 0.755.

- **Cylinder bands**

Αυτό το σύνολο εξετάζει την εμφάνιση καθυστερήσεων διεργασιών που αναφέρονται ως κυλινδρικές λωρίδες (*Cylinder bands*) στην εκτύπωση βαθυτυπίας. Περιλαμβάνει μεταβλητές που χαρακτηρίζουν διάφορες πτυχές της διαδικασίας εκτύπωσης, όπως το μέγεθος του κυλίνδρου και ο τύπος του χαρτιού. Τα δεδομένα περιέχουν 516 αντικείμενα, τα οποία χωρίζονται σε δύο κατηγορίες: αυτά που έχουν λωρίδα (212) και αυτά που δεν έχουν λωρίδα (304). Τα αρχικά δεδομένα προ-επεξεργάστηκαν με βάση τα κριτήρια που ακολουθούν: οι γραμμές που περιείχαν από μία έως οκτώ ελλειπούσες τιμές για μία μεταβλητή, εξαιρέθηκαν από ολόκληρο το σύνολο δεδομένων. Επιπλέον, οι μεταβλητές με οκτώ ή περισσότερες ελλειπούσες τιμές καταργήθηκαν εντελώς. Τέλος, από το σύνολο δεδομένων εξαιρέθηκαν και οι μεταβλητές με δύο τιμές, όπου η μία από αυτές σημειώθηκε μόνο μία φορά. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 13:

Results of cylinder bands data set analysis, purity metric

	Purity	Purity % over M–S	Macro Precision/Recall	Macro P/R % over M–S
KAMILA	0.671	113.8%	0.662/0.665	112.4%/132.9%

Πίνακας 13: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Cylinder bands KAMILA από τους Foss, Markatou, Ray και Heching (2016)

Παρατηρούμε πως ο αλγόριθμος KAMILA έχει μία σχετικά καλή ομαδοποίηση. Η τιμή της καθαρότητας είναι ίση με 0.671, η τιμή της μάκρο-ακρίβεια είναι ίση με 0.662 και η τιμή της μάκρο-ευαισθησίας είναι ίση με 0.665.

- **The insurance company benchmark (COIL 2000)**

Το συγκεκριμένο σύνολο δεδομένων χρησιμοποιήθηκε στον διαγωνισμό του 2000 CoIL (Computational Intelligence and Learning) και αφορά πελάτες που σχετίζονται με μια ασφαλιστική εταιρεία. Περιλαμβάνει μεταβλητές που σκιαγραφούν τα ποικίλα χαρακτηριστικά των υφιστάμενων και των μελλοντικών πελατών, όπως η οικογενειακή κατάσταση, η ιδιοκτησία στο σπίτι και το μορφωτικό επίπεδο. Η εφαρμογή του αλγόριθμου KAMILA είχε

ως στόχο να αξιολογήσει την αποτελεσματικότητά του στην ακριβή αναγνώριση των τύπων των πελατών, οι οποίοι ταξινομούνται σε δέκα διαφορετικές κατηγορίες. Τα δεδομένα περιέχουν 9,822 πελάτες, οι οποίοι χωρίζονται σε δέκα κατηγορίες: Successful hedonists (963), Career loners (79), Retired and religious (884), Farmers (490), Driven growers (825), Living well (943), Family with grown ups (2,691), Average family (1,513), Cruising seniors (324) και Conservative families (1,110).

Είναι εύκολο να παρατηρήσει κανείς, πως το συγκεκριμένο σύνολο δεδομένων είναι πολύ μεγαλύτερο από τα προηγούμενα, ως προς το πλήθος των ατόμων και το πλήθος των μεταβλητών. Δεδομένου ότι ο διαχωρισμός των πελατών είναι μια διαδεδομένη εφαρμογή για τις τεχνικές ομαδοποίησης, αυτή η εφαρμογή είναι πιο σχετική συγκριτικά με τις άλλες δύο. Τέλος, παρουσιάζονται τα αποτελέσματα της ομαδοποίησης των δεδομένων με τον αλγόριθμο KAMILA:

Results of insurance data set analysis, purity metric.

	Purity	Purity % over M–S	Macro Precision/Recall	Macro P/R % over M–S
KAMILA	0.354	106.1%	0.461/0.225	118.8%/104.6%

Πίνακας 14: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων The insurance company benchmark KAMILA από τους Foss, Markatou, Ray και Heching (2016)

Παρατηρούμε πως ο αλγόριθμος KAMILA έχει μία σχετικά μέτρια ομαδοποίηση. Η τιμή της καθαρότητας είναι ίση με 0.354, η τιμή της μάκρο-ακρίβειας είναι ίση με 0.461 και η τιμή της μάκρο-ευαισθησίας είναι ίση με 0.225.

5.3 Εφαρμογές της μεθόδου DBSCAN για μικτά δεδομένα

Παρακάτω θα περιγράφουν τρεις εφαρμογές του αλγόριθμου EPDCA όπως έγιναν από τους Liu, Yang και He (2017). Τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι: Adult- Census Income, Statlog (Heart) και Australian credit data και ελήφθησαν από το αποθετήριο UCI (https://archive.ics.uci.edu). Και τα τρία σύνολα δεδομένων απαρτίζονται τόσο από αριθμητικές όσο και κατηγορικές μεταβλητές. Όπως έχουμε αναφέρει σε προηγούμενο κεφάλαιο, ο αλγόριθμος EPDCA μπορεί να διαχειριστεί μικτά δεδομένα μέσω κατάλληλης μετατροπής των κατηγορικών δεδομένων. Η σειρά με την οποία δημιουργούνται οι ομάδες δεν επηρεάζει τα τελικά αποτελέσματα της ομαδοποίησης που λαμβάνονται από τον αλγόριθμο EPDCA. Αυτό σημαίνει ότι ο αλγόριθμος δεν είναι ευαίσθητος στην αρχική σειρά των σημείων δεδομένων και μπορεί να παράγει συνεπή αποτελέσματα ομαδοποίησης ανεξάρτητα από τη σειρά με την οποία ανακαλύπτονται οι ομάδες. Σε όλα τα σύνολα δεδομένων, για την μετατροπή των κατηγορικών δεδομένων, εφαρμόσθηκαν οι SE, SET και SRT τύποι, όπως έχουν αναλυθεί προηγουμένως. Τέλος, τα άτομα στο σύνολο δεδομένων αναδιατάχθηκαν τυχαία πριν την εκτέλεση του αλγόριθμου EPDCA και τα αποτελέσματα προήλθαν από τον μέσο όρο δέκα εκτελέσεων. Ακολούθως, παρουσιάζουμε έναν πίνακα που περιέχει μία λεπτομερή περιγραφή των συγκεκριμένων συνόλων δεδομένων:

Dataset	No. of attr.	No. of numeric attr.	No. of categorical attr.	No. of classes	No. of instances
Adult	14	6	8	2	32561
Australia credit	14	6	8	2	690
Heart	13	5	8	2	270

Πίνακας 15: Περιγραφή των συνόλων δεδομένων για την εφαρμογή του DBSCAN από τους Liu, Yang και He (2017)

- **Adult - Census Income**

Το σύνολο δεδομένων Adult, γνωστό και ως Census Income, είναι ένα σύνολο από 32,561 άτομα, που περιγράφονται από έξι αριθμητικές και οχτώ κατηγορικές μεταβλητές. Με την εφαρμογή του αλγόριθμου EPDCA στο συγκεκριμένο σύνολο δεδομένων, γίνεται προσπάθεια να ομαδοποιηθούν τα άτομα με εισόδημα μεγαλύτερο των 50,000 ετησίως. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 16:

Clustering results

Dataset	Conversion	Parameter		*Pu*	*F*
Adult	SE	0.5	4	0.5276	0.7890
	SET	1.6	4	0.6433	0.8458
	SRT	1.7	4	0.6474	0.8743

Πίνακας 16: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Adult - Census Income από τους Liu, Yang και He (2017)

Οι στήλες “Parameter” αναφέρονται στις τιμές Eps και Minpts, αντίστοιχα. Παρατηρούμε πως ο αλγόριθμος EPDCA έχει μία σχετικά καλή ομαδοποίηση. Εφαρμόζοντας τον τύπο SE, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.5276 και η τιμή του F-measure είναι ίση με 0.7890. Αντίστοιχα, με την εφαρμογή του τύπου SET, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.6433 και η τιμή του F-measure είναι ίση με 0.8458. Τέλος, εφαρμόζοντας τον τύπο SRT, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.6474 και η τιμή του F-measure είναι ίση με 0.8743. Είναι εύκολα κατανοητό πως ο τύπος SRT έχει την καλύτερη ομαδοποίηση από τους τρεις.

- **Australian credit data**

Αυτό το σύνολο δεδομένων είναι το ίδιο που εφαρμόσθηκε στους αλγορίθμους *k*-means και KAMILA. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 17:

Clustering results

Dataset	Conversion	Parameter		*Pu*	*F*
Australia credit	SE	0.59	8	0.8333	0.9829
	SET	1.6	8	0.7023	0.7858
	SRT	3.15	10	0.5035	0.7327

Πίνακας 17: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Australian credit από τους Liu, Yang και He (2017)

Οι στήλες “Parameter” αναφέρονται στις τιμές Eps και Minpts, αντίστοιχα. Παρατηρούμε πως ο αλγόριθμος EPDCA έχει μία αρκετά καλή ομαδοποίηση. Εφαρμόζοντας τον τύπο SE, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.8333 και η τιμή του F-measure είναι ίση με 0.9829. Αντίστοιχα, με την εφαρμογή του τύπου SET, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.7023 και η τιμή του F-measure είναι ίση με 0.7858. Τέλος, εφαρμόζοντας τον τύπο SRT, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.5035 και η τιμή του F-measure είναι ίση με 0.7327. Είναι εύκολα κατανοητό πως ο τύπος SE έχει την καλύτερη ομαδοποίηση από τους τρεις.

- **Statlog (Heart)**

Το σύνολο δεδομένων Statlog (Heart), είναι ένα σύνολο από 270 άτομα, που περιγράφονται από πέντε αριθμητικές και οκτώ κατηγορικές μεταβλητές. Με την εφαρμογή του αλγόριθμου EPDCA στο συγκεκριμένο σύνολο δεδομένων, γίνεται προσπάθεια να ομαδοποιήσει τα άτομα που έχουν καρδιακή νόσο ή όχι. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 18:

Clustering results

Dataset	Conversion	Parameter		*Pu*	*F*
Heart	SE	0.42	4	0.7178	0.5811
	SET	1.5	4	0.5912	0.7574
	SRT	0.9	4	0.5912	0.7574

Πίνακας 18: Αποτελέσματα ομαδοποίησης για το σύνολο δεδομένων Statlog από τους Liu, Yang και He (2017)

Οι στήλες “Parameter” αναφέρονται στις τιμές Eps και Minpts, αντίστοιχα. Παρατηρούμε πως ο αλγόριθμος EPDCA έχει μία σχετικά καλή προς μέτρια ομαδοποίηση. Εφαρμόζοντας τον τύπο SE, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.7178 και η τιμή του F-measure είναι ίση με 0.5811. Αντίστοιχα, με την εφαρμογή των τύπων SET και SRT, η τιμή της μέσης καθαρότητας ομαδοποίησης είναι ίση με 0.5912 και η τιμή του F-measure είναι ίση με 0.7574. Αυτό μπορεί να συμβαίνει λόγω των τιμών των παραμέτρων που χρησιμοποιήθηκαν, δηλαδή η τιμή του Eps συνέβαλε στα ίδια αποτελέσματα. Αφού το μέτρο F-measure συνδυάζει την ακρίβεια και την ευαισθησία της ομαδοποίησης, η επιλογή του τύπου SET ή SRT θα ήταν καλύτερη από την επιλογή του τύπου SE.

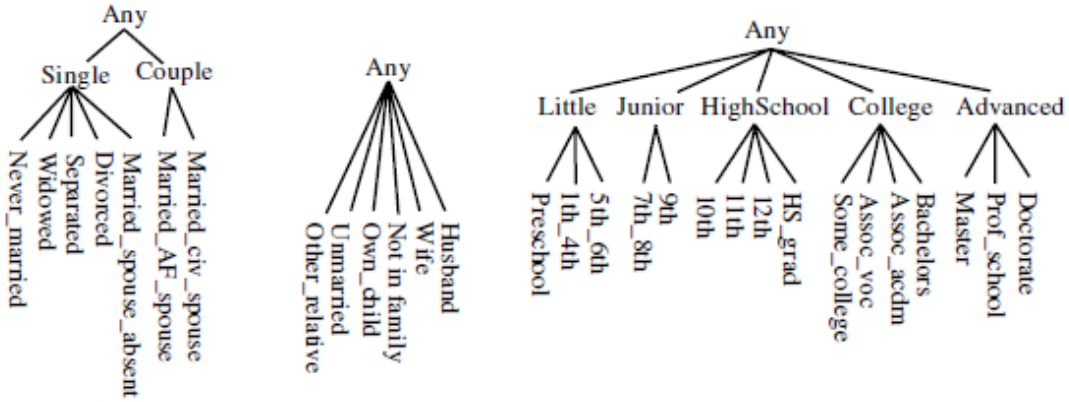
5.4 Εφαρμογή της συσσωρευτικής ιεραρχικής μεθόδου για μικτά δεδομένα

Όσον αφορά τον αλγόριθμο της συσσωρευτικής ιεραρχικής μεθόδου, θα παρουσιάσουμε μία εφαρμογή σε πραγματικά δεδομένα που έγινε από τους Hsu, Chen, και Su (2007). Το σύνολο δεδομένων που εφαρμόσθηκε ο αλγόριθμος είναι το Adult data και ελήφθη από το αποθετήριο UCI (https://archive.ics.uci.edu). Είναι ένα μικτό σύνολο δεδομένων που αφορά 48,842 άτομα και απαρτίζεται από οχτώ κατηγορικές και έξι αριθμητικές μεταβλητές. Επίσης, στο σύνολο δεδομένων υπάρχει και μία μεταβλητή που δείχνει εάν κάποιο άτομο έχει εισόδημα μεγαλύτερο των 50,000 ετησίως. Η συγκεκριμένη μεταβλητή δεν χρησιμοποιείται στην εφαρμογή του αλγόριθμου. Η εφαρμογή του αλγόριθμου στο συγκεκριμένο σύνολο δεδομένων, γίνεται για να ομαδοποιήσει τα άτομα με εισόδημα μεγαλύτερο των 50,000 ετησίως.

Λόγω της υπολογιστικής πολυπλοκότητας των ιεραρχικών αλγορίθμων, επιλέχθηκαν τυχαία 1,000 άτομα από το αρχικό σύνολο δεδομένων για την εφαρμογή του αλγόριθμου. Αυτά τα άτομα είναι ένα αντιπροσωπευτικό δείγμα του αρχικού συνόλου δεδομένων και επιλέχθηκαν

για να μειώσουν την υπολογιστική επιβάρυνση του αλγόριθμου. Επίσης, από αυτά τα 1,000 άτομα, το 76% είχε τιμή μισθού πάνω από πενήντα χιλιάδες, η οποία είναι η ίδια κατανομή με το αρχικό σύνολο δεδομένων, δηλαδή το δείγμα είναι αντιπροσωπευτικό του αρχικού συνόλου. Αυτό είναι ιδιαίτερα σημαντικό επειδή εξασφαλίζει ότι τα αποτελέσματα της ομαδοποίησης που λαμβάνονται από το δείγμα θα είναι παρόμοια με εκείνα που λαμβάνονται από το αρχικό σύνολο δεδομένων. Τέλος, επιλέχθηκαν τρεις κατηγορικές και τέσσερις αριθμητικές μεταβλητές για την εφαρμογή του αλγόριθμου.

Ο αλγόριθμος εφαρμόσθηκε δύο φορές. Την πρώτη φορά, ορίστηκε βάρος 1 σε κάθε κλαδί, υποδεικνύοντας ότι όλα τα κλαδιά θεωρήθηκαν εξίσου σημαντικά. Στην δεύτερη φορά, διπλασιάστηκαν τα βάρη που αφορούν τις κατηγορικές μεταβλητές, για να εξεταστεί η επιρροή τους. Η απόσταση μεταξύ δύο ομάδων στον αλγόριθμο υπολογίστηκε χρησιμοποιώντας τη μέθοδο της μέσης συνένωσης, όπως έχει εξηγηθεί προηγουμένως. Οι ιεραρχίες απόστασης που δόθηκαν στις τρεις κατηγορικές μεταβλητές, για την πρώτη εφαρμογή, παρουσιάζονται στον Πίνακα 19.



Πίνακας 19: Ιεραρχίες απόστασης κάθε κατηγορικής μεταβλητής από τους Hsu, Chen, και Su (2007)

Στον Πίνακα 20, παρουσιάζονται οι τιμές των τυπικών αποκλίσεων και των τετραγωνικών σφαλμάτων των αριθμητικών μεταβλητών:

Comparisons on inter_cluster standard deviation (SD) of cluster means, and squared error (SE), on numerical attributes from the clusters formed

	Age		Capital_gain		Capital_loss		Hours_per_week	
	SD	SE	SD	SE (M)	SD	SE (K)	SD	SE
DH	12.47	729	48,480	33	63.09	683	9.70	986
DH_2C	6.71	957	749	169	58.97	820	4.49	846

Πίνακας 20: Αποτελέσματα ομαδοποίησης για τις αριθμητικές μεταβλητές από τους Hsu, Chen, και Su (2007)

Είναι εύκολο να παρατηρήσει κανείς πως η δεύτερη εφαρμογή του αλγόριθμου, οδήγησε σε υψηλότερα τετραγωνικά σφάλματα και μειωμένες τυπικές αποκλίσεις, σε σύγκριση με την πρώτη εφαρμογή. Συνεπώς, όσον αφορά τις αριθμητικές μεταβλητές, η πρώτη εφαρμογή είχε καλύτερα αποτελέσματα.

Στον Πίνακα 21, παρουσιάζονται τα αποτελέσματα της κατηγορικής χρησιμότητας. Οι τιμές της κατηγορικής χρησιμότητας υπολογίστηκαν σε δύο επίπεδα: Leaf και Level_1. Το επίπεδο Leaf υπολογίστηκε σε επίπεδο φύλλου, το οποίο σημαίνει ότι η κατηγορική χρησιμότητα υπολογίστηκε με βάση την ομοιότητα μεταξύ των παρατηρήσεων μέσα σε κάθε ομάδα χρησιμοποιώντας τις αρχικές κατηγορικές τιμές. Από την άλλη πλευρά, το επίπεδο Level_1 υπολογίστηκε μετά τη ένωση των τιμών σε ένα επίπεδο. Αυτό σημαίνει ότι η κατηγορική

χρησιμότητα υπολογίστηκε με βάση την ομοιότητα μεταξύ των παρατηρήσεων σε κάθε ομάδα χρησιμοποιώντας τις γενικευμένες κατηγορηματικές τιμές στο επίπεδο της συνένωσης της ιεραρχίας απόστασης.

Comparisons on category utility of categorical attributes in leaf level and level one of distance hierarchies

	Leaf CU	Level_1 CU	Increased (%)
DH	1.0809	1.1757	8.77
DH_2C	1.1343	1.2291	8.35

Πίνακας 21: Αποτελέσματα ομαδοποίησης για τις κατηγορικές μεταβλητές από τους Hsu, Chen, και Su (2007)

Παρατηρούμε πως και με τις δύο εφαρμογές, ο αλγόριθμος έχει καταλήξει σε μία σχετικά καλή ομαδοποίηση. Επιπλέον, στο επίπεδο Level_1 υπάρχει μεγαλύτερη τιμή της κατηγορικής χρησιμότητας από το επίπεδο Leaf. Ο ρυθμός αύξησης της κατηγορικής χρησιμότητας στην πρώτη εφαρμογή είναι μεγαλύτερος από την δεύτερη. Επομένως, όσον αφορά τις κατηγορικές μεταβλητές, η πρώτη εφαρμογή είχε καλύτερα αποτελέσματα.

5.5 Συγκρίσεις

Σε αυτή την ενότητα, θα παρουσίασουμε τις τιμές των δεικτών ποιότητας, που αφορούν τα σύνολα δεδομένων, για τα οποία εφαρμόσθηκαν περισσότεροι του ενός αλγορίθμου, τα οποία είναι τα: Heart Disease, Adult και Australian credit. Στο Heart Disease εφαρμόσθηκαν οι *k*-means και DBSCAN, στο Adult εφαρμόσθηκαν η συσσωρευτική ιεραρχική και ο DBSCAN και στο Australian credit εφαρμόσθηκαν οι *k*-means, KAMILA και DBSCAN.

Όπως φαίνεται από τον Πίνακα 22, για το Heart Disease σύνολο δεδομένων, η τιμή της μέσης μικροακρίβειας για τον αλγόριθμο *k*-means είναι ένδειξη της σχετικά υψηλής ακρίβειάς του στις σωστές ταξινομήσεις των ατόμων. Όσον αφορά τον αλγόριθμο DBSCAN, οι τιμές της μέσης καθαρότητας ομαδοποίησης και του F-measure, και για τους τρεις τρόπους μετατροπής των κατηγορικών δεδομένων είναι αρκετά μικρότερες από την τιμή της μέσης μικροακρίβειας για τον αλγόριθμο *k*-means. Τελικά, ο αλγόριθμος *k*-means συμπεριφέρεται καλύτερα από τον DBSCAN στο συγκεκριμένο σύνολο δεδομένων.

k-means	DBSCAN		
Micro-p	Conversion	Pu	F
0.85	SE	0.7178	0.5811
	SET	0.5912	0.7574
	SRT	0.5912	0.7574

Πίνακας 22: Σύγκριση των *k*-means και DBSCAN του Heart Disease dataset με βάση τις τιμές των δεικτών ποιότητας

Στη συνέχεια, παρουσιάζονται οι δείκτες ποιότητας που αφορούν το Adult σύνολο δεδομένων. Όσον αφορά τη συσσωρευτική ιεραρχική μέθοδο, παρατηρούμε πως η πρώτη εφαρμογή έχει καλύτερα αποτελέσματα από την δεύτερη. Από την άλλη μεριά, η καλύτερη ομαδοίηση για τον αλγόριθμο DBSCAN είναι με τη χρήση του τύπου SRT. Τελικά, η επιλογή μεταξύ της συσσωρευτικής ιεραρχικής και του DBSCAN εξαρτάται από τους στόχους του εκάστοτε ερευνητή. Αν η εξερεύνηση της ιεραρχική δομής είναι σημαντική, η συσσωρευτική ιεραρχική μέθοδος με τη δεύτερη εφαρμογή είναι καταλληλότερη. Εάν είναι κρίσιμης

σημασίας ο εντοπισμός καλά καθορισμένων ομάδων, ο αλγόριθμος DBSCAN με τη χρήση του τύπου SRT είναι καλύτερος.

Agglomerative			DBSCAN		
Distance Hierarchy	Leaf CU	Level_1 CU	Conversion	Pu	F
DH	1.0809	1.1757	SE	0.5276	0.7890
DH_2C	1.1343	1.2291	SET	0.6433	0.8458
			SRT	0.6474	0.8743

Πίνακας 23: Σύγκριση της συσσωρευτικής ιεραρχικής μεθόδου και του DBSCAN του Adult dataset με βάση τις τιμές των δεικτών ποιότητας

Τέλος, παρουσιάζονται οι δείκτες ποιότητας που αφορούν το Australian credit σύνολο δεδομένων. Όσον αφορά τον αλγόριθμο *k*-means, η τιμή της μέσης μικροακρίβειας υποδεικνύει την σχετικά υψηλή ακρίβειά του στις σωστές ταξινομήσεις των ατόμων. Για τον αλγόριθμο KAMILA, οι τιμές της καθαρότητας, μάκρο-ακρίβειας και μάκρο-ευαισθησίας δείχνουν ότι έχει γίνει μία αρκετά καλή ομαδοποίηση. Ο αλγόριθμος DBSCAN, με τη χρήση του τύπου SE, έχει την καλύτερη ομαδοποίηση. Η επιλογή του καλύτερου αλγορίθμου εξαρτάται από τους συγκεκριμένους δείκτες μέτρησεις, που ο εκάστοτε ερευνητής θεωρεί πιο σημαντικές.

k-means	KAMILA			DBSCAN		
Micro-p	Purity	Macro precision	Macro recall	Conversion	Pu	F
0.88	0.775	0.808	0.755	SE	0.8333	0.9829
				SET	0.7023	0.7858
				SRT	0.5035	0.7327

Πίνακας 24: Σύγκριση των *k*-means, KAMILA και DBSCAN του Australian credit dataset με βάση τις τιμές των δεικτών ποιότητας

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα

Η πολυμεταβλητή ανάλυση αφορά τις μεθόδους που χρησιμοποιεί ένας ερευνητής για να κατανοήσει, περιγράψει και εξερευνήσει δεδομένα που περιέχουν πολλά χαρακτηριστικά των ίδιων ατόμων. Συχνά, οι αναλύσεις εστιάζουν στην κατανόηση της δομής και των μεταβλητών των δεδομένων συνολικά. Μία σημαντική εφαρμογή είναι η ανάλυση σε συστάδες, όπου οι παρατηρήσεις ταξινομούνται σε ομάδες βάσει κοινών χαρακτηριστικών που περιλαμβάνονται σε κάποιες μεταβλητές. Η χρήση πολυμεταβλητής ανάλυσης αποτελεί σημαντικό εργαλείο για την αποτελεσματική ανάλυση και ερμηνεία δεδομένων.

Σκοπός αυτής της εργασίας ήταν η παρουσίαση μεθόδων ομαδοποίησης μικτών δεδομένων. Πιο συγκεκριμένα, αναπτύχθηκε η έννοια των μικτών δεδομένων, τα οποία μπορούν να εμφανισθούν σε πολλούς τομείς επιστημών και επιχειρήσεων. Υλοποιήθηκαν τέσσερις αλγορίθμους ομαδοποίησης: *k*-means, KAMILA, DBSCAN και συσσωρευτικός ιεραρχικός. Πραγματοποιήθηκε η εφαρμογή τους σε ένα προσομοιωμένο σύνολο δεδομένων και παρουσιάστηκαν εφαρμογές που ανακτήθηκαν από την βιβλιογραφία.

Εξετάζοντας τους συγκεκριμένους αλγορίθμους ομαδοποίησης, επιβεβαιώθηκε η ωφελιμότητα και η αναγκαιότητα της χρήσης αλγορίθμων ομαδοποίησης. Το μεγαλύτερο πλεονέκτημα αυτών των μεθόδων είναι η αναγνώριση φυσικών ομάδων ή μοτίβων δεδομένων που ενδέχεται να μην είναι ορατά. Μπορεί να αποκαλύψει ομοιότητες μεταξύ των ατόμων, επιτρέποντας την καλύτερη κατανόηση των δεδομένων. Από τα αποτελέσματα που παρουσιάστηκαν, οι αλγόριθμοι *k*-means και KAMILA είχαν την καλύτερη απόδοση.

Τελικά, η ανάλυση κατά συστάδες παραμένει ένα πολύτιμο εργαλείο για την ανάλυση δεδομένων και την αναγνώριση προτύπων. Κατά την εφαρμογή της ο ερευνητής θα χρειαστεί να επιλέξει τον βέλτιστο αλγόριθμο ομαδοποίησης και να εφαρμόσει κατάλληλες τεχνικές προπαρασκευής δεδομένων, με βάση τα χαρακτηριστικά των δεδομένων και το συγκεκριμένο πρόβλημα που αντιμετωπίζει.

7.Παράρτημα - Κώδικας R

2ο Κεφάλαιο - mixed datasets

```
# Load required library
library(data.table)

#Read the Bank Marketing data
Bank_Marketing_DataSet <-fread("bank-full.csv")
Bank_Marketing_DataSet <- Bank_Marketing_DataSet[,-17]

#Print the first rows
head(Bank_Marketing_DataSet)

#Read the Food Delivery data
Food_Delivery_Dataset <- fread("train.csv")

#Print the first rows
head(Food_Delivery_Dataset)
```

3ο Κεφάλαιο - andrews_dog

```
# Load required libraries
library(data.table)
library(andrews)
library(zoom)

#Read the data
data <- fread("data.csv")

# Create Andrews curves plot
andrews(data,palcol= object_names)
zm()
```

3ο Κεφάλαιο - dendrogram

```
# Load required libraries
library(data.table)

#Read the dataset
data <- fread("data.csv")

# Compute hierarchical clustering
```

```
hc <- hclust(dist(data))

# Create dendrogram
dend <- as.dendrogram(hc)

# Plot dendrogram
plot(dend)
```

4ο Κεφάλαιο - δημιουργία του προσομοιωμένου συνόλου δεδομένων – πρώτη εφαρμογή

```
# Set seed for reproducibility
set.seed(20015)

# Sample size for each cluster
size1 <- 300
size2 <- 300
size3 <- 400

#Generate the true cluster information
true_cluster <- c(rep("A",size1),rep("B",size2),rep("C",size3))

#Defining the categorical levels
nat <- factor(c("Greek", "French", "Italian",
"German","British","Spanish","Swedish","Danish"))
gen <- factor(c("Male", "Female"))
rt <- factor(c("Deluxe Double", "Standard Twin", "Executive Suite",
"Superior King"))
bf <- factor(c("Yes","No"))
rg <- factor(c("1 Star","2 Stars","3 Stars","4 Stars","5 Stars"))

# Simulate hotel data with three different clusters
hotel_data_c1 <- data.frame(
  Nationality = sample(nat , size1, replace = TRUE, prob = c(0.20, 0.10,
0.20, 0.10, 0.10, 0.20, 0.05, 0.05)),
  Age = sample(18:40, size1, replace = TRUE),
  Gender = sample(gen , size1, replace = TRUE, prob = c(0.55, 0.45)),
  Room_Type = sample(rt, size1, replace = TRUE, prob = c(0.3, 0.4, 0.15,
0.15)),
  Breakfast = sample(bf , size1, replace = TRUE, prob = c(0.45, 0.55)),
  Price = rnorm(size1, 300, 75),
  Rating = sample(rg, size1, replace = TRUE, prob = c(0.1, 0.1, 0.3, 0.3,
0.2)),
  Income = rnorm(size1, 8000, 1000)
)
```

```
hotel_data_c2 <- data.frame(
  Nationality = sample(nat , size2, replace = TRUE, prob = c(0.10, 0.05, 0.05, 0.20, 0.10, 0.05, 0.20, 0.20)),
  Age = sample(40:70, size2, replace = TRUE),
  Gender = sample(gen , size2, replace = TRUE, prob = c(0.45, 0.55)),
  Room_Type = sample(rt, size2, replace = TRUE, prob = c(0.25, 0.20, 0.25, 0.30)),
  Breakfast = sample(bf , size2, replace = TRUE, prob = c(0.65, 0.35)),
  Price = rnorm(size2, 800, 100),
  Rating = sample(rg, size2, replace = TRUE, prob = c(0.1, 0.3, 0.3, 0.2, 0.1)),
  Income = rnorm(size2, 20000, 1000)
)


hotel_data_c3 <- data.frame(
  Nationality = sample(nat , size3, replace = TRUE, prob = c(0.05, 0.20, 0.05, 0.20, 0.20, 0.20, 0.05, 0.05)),
  Age = sample(30:50, size3, replace = TRUE),
  Gender = sample(gen , size3, replace = TRUE, prob = c(0.5, 0.5)),
  Room_Type = sample(rt, size3, replace = TRUE, prob = c(0.3, 0.3, 0.20, 0.20)),
  Breakfast = sample(bf , size3, replace = TRUE, prob = c(0.5, 0.5)),
  Price = rnorm(size3, 450, 100),
  Rating = sample(rg, size3, replace = TRUE, prob = c(0.2, 0.2, 0.2, 0.2, 0.2)),
  Income = rnorm(size3, 15000, 1000)
)


hotel_data <-  rbind(hotel_data_c1, hotel_data_c2, hotel_data_c3)


# Summary of the hotel data
summary(hotel_data,maxsum=8)


# Barplot for Room_Type
barplot(table(hotel_data$Room_Type), main = "Room Type Distribution", xlab = "Room Type", ylab = "Count", col = "steelblue", ylim = c(0, 500))


# Boxplot for Price
boxplot(hotel_data$Price, main = "Boxplot of Room Prices", ylab = "Price", col = "lightblue", border = "blue",ylim = c(0, 1000))


# Convert true_cluster to a factor with levels A, B, and C
true_cluster <- factor(true_cluster, levels = c("A", "B", "C"))
```

```
# Scatterplot for Age vs. Income with 3 different colors based on
true_cluster
plot(hotel_data$Age, hotel_data$Income,
     main = "Scatterplot of Age vs. Income",
     xlab = "Age", ylab = "Income",
     col = true_cluster,
     pch = 16, # Sets the shape of the data points to circles
     ylim = c(5000, 25000))
```

4ο Κεφάλαιο - δημιουργία του προσομοιωμένου συνόλου δεδομένων – δεύτερη εφαρμογή

```
# Set seed for reproducibility
set.seed(20015)

# Sample size for each cluster
size1 <- 300
size2 <- 300
size3 <- 400

#Generate the true cluster information
true_cluster <- c(rep("A",size1),rep("B",size2),rep("C",size3))

#Defining the categorical levels
nat <- factor(c("Greek", "French", "Italian",
"German","British","Spanish","Swedish","Danish"))
gen <- factor(c("Male", "Female"))
rt <- factor(c("Deluxe Double", "Standard Twin", "Executive Suite",
"Superior King"))
bf <- factor(c("Yes","No"))
rg <- factor(c("1 Star","2 Stars","3 Stars","4 Stars","5 Stars"))

# Simulate hotel data with three different clusters
hotel_data_c1 <- data.frame(
  Nationality = sample(nat , size1, replace = TRUE, prob = c(0.30, 0.02,
0.30, 0.02, 0.02, 0.30, 0.02, 0.02)),
  Age = sample(18:40, size1, replace = TRUE),
  Gender = sample(gen , size1, replace = TRUE, prob = c(0.60, 0.40)),
  Room_Type = sample(rt, size1, replace = TRUE, prob = c(0.2, 0.6, 0.10,
0.10)),
  Breakfast = sample(bf , size1, replace = TRUE, prob = c(0.05, 0.95)),
  Price = rnorm(size1, 300, 75),
  Rating = sample(rg, size1, replace = TRUE, prob = c(0.1, 0.1, 0.3, 0.3,
0.2)),
  Income = rnorm(size1, 12000, 1500)
```

```
)


hotel_data_c2 <- data.frame(
  Nationality = sample(nat , size2, replace = TRUE, prob = c(0.02, 0.02, 0.02, 0.30, 0.02, 0.02, 0.30, 0.30)),
  Age = sample(35:70, size2, replace = TRUE),
  Gender = sample(gen , size2, replace = TRUE, prob = c(0.40, 0.6)),
  Room_Type = sample(rt, size2, replace = TRUE, prob = c(0.10, 0.10, 0.20, 0.6)),
  Breakfast = sample(bf , size2, replace = TRUE, prob = c(0.95, 0.05)),
  Price = rnorm(size2, 500, 100),
  Rating = sample(rg, size2, replace = TRUE, prob = c(0.1, 0.3, 0.3, 0.2, 0.1)),
  Income = rnorm(size2, 18000, 1500)
)


hotel_data_c3 <- data.frame(
  Nationality = sample(nat , size3, replace = TRUE, prob = c(0.05, 0.20, 0.05, 0.20, 0.20, 0.20, 0.05, 0.05)),
  Age = sample(25:60, size3, replace = TRUE),
  Gender = sample(gen , size3, replace = TRUE, prob = c(0.5, 0.5)),
  Room_Type = sample(rt, size3, replace = TRUE, prob = c(0.6, 0.2, 0.1, 0.1)),
  Breakfast = sample(bf , size3, replace = TRUE, prob = c(0.7, 0.3)),
  Price = rnorm(size3, 400, 100),
  Rating = sample(rg, size3, replace = TRUE, prob = c(0.2, 0.2, 0.2, 0.2, 0.2)),
  Income = rnorm(size3, 15000, 1500)
)


hotel_data <-  rbind(hotel_data_c1, hotel_data_c2, hotel_data_c3)


# Summary of the hotel data
summary(hotel_data,maxsum=8)


# Barplot for Room_Type
barplot(table(hotel_data$Room_Type), main = "Room Type Distribution", xlab = "Room Type", ylab = "Count", col = "steelblue", ylim = c(0, 500))


# Boxplot for Price
boxplot(hotel_data$Price, main = "Boxplot of Room Prices", ylab = "Price", col = "lightblue", border = "blue",ylim = c(0, 1000))


# Convert true_cluster to a factor with levels A, B, and C
true_cluster <- factor(true_cluster, levels = c("A", "B", "C"))
```

```
# Scatterplot for Age vs. Income with 3 different colors based on
true_cluster
plot(hotel_data$Age, hotel_data$Income,
     main = "Scatterplot of Age vs. Income",
     xlab = "Age", ylab = "Income",
     col = true_cluster,
     pch = 16, # Sets the shape of the data points to circles
     ylim = c(5000, 25000))
```

4ο Κεφάλαιο - εφαρμογή της μεθόδου k-means

```
# Load required libraries
library(dplyr)
library(ggplot2)
library(viridis)

# Min-Max Normalization function
min_max_normalization <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

# Columns to normalize
columns_to_normalize <- c("Age", "Price", "Income")

# Apply Min-Max normalization to numerical attributes
hotel_data[columns_to_normalize] <-
lapply(hotel_data[columns_to_normalize], min_max_normalization)

# List of categorical attributes
categorical_attributes <- c("Nationality", "Gender", "Room_Type",
"Breakfast", "Rating")

# Calculate the average number of distinct values for categorical
attributes
average_distinct_values <- ceiling(mean(sapply(categorical_attributes,
function(attr) {
  length(unique(hotel_data[[attr]]))
})))

# Number of intervals for discretization
S <- 5

discretize_numeric <- function(x, S) {
  interval_width <- (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)) / S
```

```
  breaks <- seq(min(x, na.rm = TRUE), max(x, na.rm = TRUE) +
interval_width, interval_width)
  labels <- paste0("u", 1:S)
  categorical_codes <- cut(x, breaks = breaks, labels = FALSE,
include.lowest = TRUE)
  factor(labels[replace(categorical_codes, categorical_codes == (S + 1),
S)])
}

# Discretize numerical attributes
hotel_data_discretized <- mutate(
  hotel_data,
  Age_Category = discretize_numeric(Age, S),
  Price_Category = discretize_numeric(Price, S),
  Income_Category = discretize_numeric(Income, S)
)

# Initialize an empty data frame to store the discretized categories
dcateg_data <- data.frame()

# Merge minimum and maximum values for each class of discretization
for (attr in c("Age", "Price", "Income")) {
  attr_min_col <- paste0("Min_", attr)
  attr_max_col <- paste0("Max_", attr)
  category_col <- paste0(attr, "_Category")

  unique_categories <- unique(hotel_data_discretized[[category_col]])
  min_values <- numeric(length(unique_categories))
  max_values <- numeric(length(unique_categories))

  for (i in seq_along(unique_categories)) {
    category <- unique_categories[i]
    category_data <-
hotel_data_discretized[hotel_data_discretized[[category_col]] == category,
attr]

    min_values[i] <- min(category_data)
    max_values[i] <- max(category_data)
  }

   # Create a data frame with the calculated values and attribute name
  attr_data <- data.frame(
    Attribute = attr,
    Category = unique_categories,
```

```
    attr_min_col = min_values,
    attr_max_col = max_values
  )

  # Bind the current attribute data to the dcateg_data data frame
  dcateg_data <- rbind(dcateg_data, attr_data)

  # Sort dcateg_data by Attribute and Category
  dcateg_data <- dcateg_data[order(dcateg_data$Attribute,
dcateg_data$Category), ]

}

#Print dcateg_data
print(dcateg_data)

# Remove specific columns from the discretized data frame
hotel_data_discretized <- hotel_data_discretized[,
!(names(hotel_data_discretized) %in% c("Age", "Price", "Rating",
"Income"))]

# Print the first few rows of the updated hotel_data
print(head(hotel_data_discretized))

# Define the find_max function
find_max <- function(data, attr1, attr2, val1, val2) {
  # Compute the probability of each value for the first attribute
  prob1 <- table(data[[attr1]]) / nrow(data)

  # Compute the probability of each value for the second attribute
  prob2 <- table(data[[attr2]]) / nrow(data)

  # Compute the probability of co-occurrence for each pair of values
  prob12 <- table(data[[attr1]], data[[attr2]]) / nrow(data)

  # Get the unique values of the second attribute
  vals2 <- unique(data[[attr2]])

  # Initialize the maximum value to negative infinity
  max_val <- -Inf

  # Loop over all unique values of the second attribute
  for (i in seq_along(vals2)) {
```

```
    curr_val <- vals2[i]

    # Check if the combination exists in the prob12 table
    if (!is.null(dim(prob12)) && curr_val %in% rownames(prob12) && val1
%in% rownames(prob12) && val2 %in% rownames(prob12)) {
      # Compute the probability of the first value given the current value
of the second attribute
      prob_val1_given_val2 <- prob12[val1, curr_val] / prob2[curr_val]

      # Compute the probability of the second value given the current value
of the second attribute
      prob_val2_given_val2 <- prob12[val2, curr_val] / prob2[curr_val]
    } else {
      # If the combination does not exist, set probabilities to zero
      prob_val1_given_val2 <- 0
      prob_val2_given_val2 <- 0
    }

    # Update the distance
    dij_xy <- prob_val1_given_val2 + prob_val2_given_val2

    # Update the maximum value if necessary
    if (dij_xy > max_val) {
      max_val <- dij_xy
    }
  }

  # Return the maximum value
  return(1 - max_val)
}

# Define the ALGO_DISTANCE() function
ALGO_DISTANCE <- function(data) {
  cat_attrs <- names(data)[sapply(data, is.factor)]
  num_attrs <- names(data)[sapply(data, is.numeric)]

  dist_mat <- matrix(0, nrow = nrow(data), ncol = nrow(data))

  for (i in 1:(nrow(data) - 1)) {
    for (j in (i + 1):nrow(data)) {
      dist <- 0

      for (attr1 in cat_attrs) {
```

```
        for (attr2 in cat_attrs) {
          max_val <- find_max(data, attr1, attr2, data[i, attr1], data[j,
attr2])
          dist <- dist + (1 - max_val)
        }
      }

      for (attr1 in num_attrs) {
        for (attr2 in num_attrs) {
          if (attr1 != attr2) {
            dist <- dist + (data[i, attr1] - data[j, attr2])^2
          }
        }
      }

      dist_mat[i, j] <- sqrt(dist)
      dist_mat[j, i] <- sqrt(dist)  # Set the corresponding symmetric
element
    }
  }

  return(dist_mat)
}

#Creating a data frame only with the discretized numerical attributes
hotel_data_discretized_num <-
data.frame(hotel_data_discretized$Age_Category,hotel_data_discretized$Price
_Category,hotel_data_discretized$Income_Category)

#Compute the distance matrix for the discretized numerical attributes
distance_matrix_num <- ALGO_DISTANCE(hotel_data_discretized_num)

# Function to compute the significance of a numeric attribute
compute_significance <- function(data, attr, S, dist_mat) {
  # Extract the discretized numeric attribute column from the data
  numeric_column <- data[, attr]

  # Get the unique categorical values from the discretized numeric
attribute
  unique_categories <- unique(numeric_column)

  # Initialize the sum of distances
  sum_distances <- 0
```

```
  # Get the number of unique pairs of categorical values
  num_unique_pairs <- S * (S - 1) / 2

  # Loop over all unique pairs of categorical values
  for (i in 1:(S - 1)) {
    for (j in (i + 1):S) {
      # Get the current pair of categorical values
      cat_val_1 <- unique_categories[i]
      cat_val_2 <- unique_categories[j]

      # Find row indices for the current pair of categorical values in the
distance matrix
      row_indices <- which(numeric_column == cat_val_1)
      col_indices <- which(numeric_column == cat_val_2)

      # Calculate the distance between the categorical values
      distance <- dist_mat[row_indices, col_indices]

      # Update the sum of distances
      sum_distances <- sum_distances + sum(distance)
    }
  }

  # Calculate the significance of the numeric attribute
  significance <- sum_distances / num_unique_pairs

  return(significance)
}

# Create an empty data frame to store the significance values
w_t <- data.frame(Attribute = character(0), Significance = numeric(0))

# Loop through each numeric attribute and calculate its significance
for (attr in names(hotel_data_discretized_num)) {
  # Calculate the significance for the current attribute
  w <- compute_significance(hotel_data_discretized_num, attr, S,
distance_matrix_num)

  # Append the attribute and its significance to the data frame
  row <- data.frame(Attribute = attr, Significance = w)
  w_t <- rbind(w_t, row)
}
```

```
# Transpose the data frame and convert it back to a data frame
w_t <- as.data.frame(t(w_t))

# Set the first row as column names
colnames(w_t) <- w_t[1,]

# Remove the first row
w_t <- w_t[-1,]

# Convert the data frame to numeric
w_t <-as.data.frame(lapply(w_t, as.numeric))

# Renaming all columns
names(w_t) <- c("Age", "Price", "Income")

# Function to compute the center of clusters
compute_cluster_centers <- function(data, clusters, k) {
  num_attributes <- ncol(data)  # Number of attributes

  # Convert non-numeric attributes to factors with all levels
  for (i in 1:num_attributes) {
    if (!is.numeric(data[[i]])) {
      attribute_values <- unique(as.character(data[[i]]))  # Convert to
character and get unique attribute values
      data[[i]] <- factor(as.character(data[[i]]), levels =
attribute_values)
    }
  }

  # Compute the cluster center for each attribute
  cluster_centers <- vector("list", length = num_attributes)

  for (i in 1:num_attributes) {
    if (is.factor(data[[i]])) {
      # For categorical attributes
      attribute_counts <- data.frame(table(data[[i]], clusters))  # Count
occurrences in each cluster
      normalized_counts <- as.data.frame.matrix(xtabs(Freq ~ .,
attribute_counts))  # Convert counts to a matrix
      normalized_counts <- normalized_counts / rowSums(normalized_counts)
# Normalize the counts
      normalized_counts <- t(normalized_counts)
      normalized_counts <- as.data.frame(normalized_counts)
      # Function to join cell value with header and concatenate rows
```

```
      join_cell_with_header_and_row <- function(df) {
        df_strings <- apply(df, 1, function(x) paste(round(x, 8),
names(df), sep = "", collapse = " "))
        df <- data.frame(Concatenated_Row = df_strings)
        return(df)
      }

      # Apply the function to hotel data
      normalized_counts <- join_cell_with_header_and_row(normalized_counts)

      colnames(normalized_counts) <- (colnames(data)[i])

      cluster_centers[[i]] <- normalized_counts
    } else if (is.numeric(data[[i]])) {
      # For numeric attributes, compute the mean
      attribute_mean <- tapply(data[[i]], clusters, mean)
      attribute_mean <- as.data.frame(attribute_mean)
      colnames(attribute_mean) <- colnames(data)[i]
      cluster_centers[[i]] <- attribute_mean
    } else {
      # For other non-numeric attributes, set the cluster center to NA
      cluster_centers[[i]] <- NA
    }
  }

  # Combine cluster_centers into a single data frame
  cluster_centers_df <- as.data.frame(do.call(cbind, cluster_centers))

  # Set the row names to "Cluster 1", "Cluster 2", ..., "Cluster k"
  rownames(cluster_centers_df) <- paste("Cluster", 1:k)

  return(cluster_centers_df)
}

# Function to calculate the cost function
cost_function <- function(x_i, w_t, C_j, C_Jt_C,Nc) {
  # Extract numerical and categorical attributes from the data point x_i
  categorical_data <- x_i[, sapply(x_i, is.factor)]
  numerical_data <- x_i[, sapply(x_i, is.numeric)]

  # Calculate the sum of squared differences for numerical attributes using
a for loop
  sum_squared_numerical <- 0
```

```
  for (t in 1:length(numerical_data)) {
    sum_squared_numerical <- sum_squared_numerical + (w_t[,t] *
(numerical_data[, t] - C_j[t])^2)
  }

  # Create a function to calculate the Ω for categorical attributes

  # Function to calculate the distance matrix using ALGO_DISTANCE function
  compute_distance_matrix <- function(data) {
    cat_attrs <- names(data)[sapply(data, is.factor)]
    num_attrs <- names(data)[sapply(data, is.numeric)]

    dist_mat <- matrix(0, nrow = nrow(data), ncol = nrow(data))

    for (i in 1:(nrow(data) - 1)) {
      for (j in (i + 1):nrow(data)) {
        dist <- 0

        for (attr1 in cat_attrs) {
          for (attr2 in cat_attrs) {
            max_val <- find_max(data, attr1, attr2, data[i, attr1], data[j,
attr2])
            dist <- dist + (1 - max_val)
          }
        }

        for (attr1 in num_attrs) {
          for (attr2 in num_attrs) {
            if (attr1 != attr2) {
              dist <- dist + (data[i, attr1] - data[j, attr2])^2
            }
          }
        }

        dist_mat[i, j] <- sqrt(dist)
        dist_mat[j, i] <- sqrt(dist)  # Set the corresponding symmetric
element
      }
    }

    return(dist_mat)
  }
```

```
  # Function to compute Ω(X,C) given categorical data, cluster's center,
and Nc
  compute_omega <- function(categorical_data, C_Jt_C, Nc) {
    if (!is.data.frame(categorical_data) || !is.data.frame(C_Jt_C)) {
      stop("Inputs 'categorical_data' and 'center_data' must be data
frames.")
    }

    if (Nc <= 0) {
      stop("Nc must be a positive value.")
    }

    # Calculate the distance matrix between categorical_data and
center_data using ALGO_DISTANCE function
    distance_matrix_cat <- compute_distance_matrix(rbind(categorical_data,
C_Jt_C))

    # Extract distances for each categorical attribute from the distance
matrix
    distances_cat <-
colMeans(distance_matrix_cat)[1:nrow(categorical_data)]

    omega_xc <- 0

    for (i in 1:length(distances_cat)) {
      Ai <- unique(categorical_data[[i]])
      pi <- length(Ai)
      weight <- pi / Nc
      omega_xc <- omega_xc + (weight * distances_cat[i])
    }

    return(omega_xc)
  }

  # Calculate the sum of Ω for categorical attributes
  sum_omega_categorical <- (compute_omega(categorical_data, C_Jt_C, Nc))^2

  # Calculate the final distance
  distance <- sum_squared_numerical + sum_omega_categorical

  return(distance)
}

# Function to assign data objects to the closest cluster center
```

```
assign_to_clusters <- function(data, cluster_centers) {
  # Extract numerical and categorical attributes from the data
  categorical_data <- data[, sapply(data, is.factor)]
  numerical_data <- data[, sapply(data, is.numeric)]

  num_objects <- nrow(data)
  num_clusters <- k
  cluster_assignments <- numeric(num_objects)

  for (i in 1:num_objects) {
    distances <- numeric(num_clusters)
    for (j in 1:num_clusters) {
      C_j <- as.numeric(cluster_centers[j, sapply(names(cluster_centers),
function(x) x %in% names(numerical_data))])
      # Function to extract categorical cluster center with attribute names
      extract_categorical_center <- function(cluster_center,
categorical_data) {
        col_names <- names(categorical_data)
        categorical_center <-
as.character(cluster_center[sapply(names(cluster_center), function(x) x
%in% col_names)])
        return(as.data.frame(t(categorical_center), stringsAsFactors =
FALSE))
      }
      # Get the cluster center for cluster 'j'
      cluster_center <- cluster_centers[j, ]
      # Calculate the categorical cluster center for cluster 'j'
      C_Jt_C <- extract_categorical_center(cluster_center,
categorical_data)
      # Add headers to the resulting data frame
      colnames(C_Jt_C) <- names(categorical_data)

      # Calculate the cost for data object i assigned to cluster center j
      distances[j] <- cost_function(data[i, ], w_t, C_j, C_Jt_C,Nc[j,2])
    }

    # Assign the data object to the closest cluster center (min cost)
    closest_cluster <- which.min(distances)
    cluster_assignments[i] <- closest_cluster
  }

  return(cluster_assignments)
}
```

```
#Run the modified k-mean clustering algorithm
k <- 3
max_iterations <-  100
num_objects <- nrow(hotel_data)
num_attributes <- ncol(hotel_data)
clusters <- sample(1:k, num_objects, replace = TRUE)

# Create a list to store objects in each cluster
Nc <- vector("list", k)

# Initialize the cluster sums to zero
for (cluster_id in 1:k) {
  Nc[[cluster_id]] <- 0
}

# Calculate the sum for each cluster
for (i in 1:num_objects) {
  cluster_id <- clusters[i]
  Nc[[cluster_id]] <- Nc[[cluster_id]] + 1
}

# Convert the cluster_sums list to a data frame
Nc <- data.frame(cluster_id = 1:k, sum_of_objects = sapply(Nc, sum))

for (iteration in 1:max_iterations) {
  old_clusters <- clusters

  # Step 1: Compute cluster centers
  cluster_centers <- compute_cluster_centers(hotel_data, clusters,k)

  # Step 2: Assign data objects to the closest cluster center
  clusters <- assign_to_clusters(hotel_data, cluster_centers)

  # Check for convergence
  if (all(clusters == old_clusters)) {
    break
  }
}

#Results of k-means
kmeans_results <- list(cluster_centers = cluster_centers,
cluster_assignments = clusters)
```

```
#Create the cross-tabulation table
xtabs(~true_cluster+kmeans_results$cluster_assignments)


# Combine the cluster assignments with the original data
hotel_data_clustered <- cbind(hotel_data, Cluster =
as.factor(kmeans_results$cluster_assignments))


# Create a scatter plot of the data points, color-coded by cluster labels
ggplot(hotel_data_clustered, aes(x = Age, y = Price, color = Cluster)) +
  geom_point(size = 3, alpha = 0.8) +
  geom_point(data = kmeans_results$cluster_centers, aes(x = Age, y =
Price), color = "red", shape = 4, size = 5) +
  labs(x = "Age", y = "Price", title = "K-means Clustering") +
  scale_color_viridis(discrete = TRUE) +
  theme_minimal()


# Compare distributions of Income across clusters using box plots
ggplot(hotel_data, aes(x = factor(kmeans_results$cluster_assignments), y =
Income, fill = factor(kmeans_results$cluster_assignments))) +
  geom_boxplot() +
  labs(x = "Cluster", y = "Income", title = "Distribution of Income by
Cluster") +
  scale_fill_discrete(name = "Cluster") +
  theme_minimal()


# Create a scatter plot of the data points, color-coded by original
clusters
hotel_data_original <- cbind(hotel_data, Original_Cluster = true_cluster)
ggplot(hotel_data_original, aes(x = Age, y = Price, color =
Original_Cluster)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "Age", y = "Price", title = "Original Clusters") +
  scale_color_viridis(discrete = TRUE) +
  theme_minimal()


# Compare distributions of Income across original clusters using box plots
hotel_data_original <- cbind(hotel_data, Original_Cluster = true_cluster)
ggplot(hotel_data_original, aes(x =
factor(hotel_data_original$Original_Cluster), y = Income, fill =
factor(hotel_data_original$Original_Cluster))) +
  geom_boxplot() +
  labs(x = "Cluster", y = "Income", title = "Distribution of Income by
Original Cluster") +
  scale_fill_discrete(name = "Cluster") +
```

```
  theme_minimal()
```

4ο Κεφάλαιο - εφαρμογή της μεθόδου KAMILA

```
# Load the required libraries
library(kamila)
library(ggplot2)


# Separate numerical and categorical variables
contvar <- hotel_data[, c("Age", "Price", "Income")]
catvar <- hotel_data[, c("Nationality", "Gender", "Room_Type", "Breakfast",
"Rating")]
catDf <- data.frame(apply(catvar, 2, factor), stringsAsFactors = TRUE)
conDf <- data.frame(scale(contvar), stringsAsFactors = TRUE)


# Run KAMILA algorithm
kamila_results <- kamila(conVar = conDf, catFactor = catDf, numClust = 3,
numInit = 100)


#Create the cross-tabulation table
xtabs(~true_cluster+kamila_results$finalMemb)


# Extract the cluster assignments from kamila_results
final_clusters <- kamila_results$finalMemb


# Combine the cluster assignments with the original data
hotel_data_clustered <- cbind(hotel_data, Cluster =
as.factor(kamila_results$finalMemb))


# Create the scatter plot for Age vs. Price with points colored based on
clusters
ggplot(hotel_data_clustered, aes(x = Age, y = Price, color =
factor(Cluster))) +
  geom_point() +
  labs(title = "KAMILA Clustering", x = "Age", y = "Price", color =
"Cluster") +
  scale_color_viridis_d()+
  theme_minimal()


# Compare distributions of Income across clusters using box plots
ggplot(hotel_data_clustered, aes(x = factor(Cluster), y = Income, fill =
factor(Cluster))) +
  geom_boxplot() +
  labs(x = "Cluster", y = "Income", title = "Distribution of Income by
Cluster") +
  scale_fill_discrete(name = "Cluster") +
  theme_minimal()
```

```
# Create a scatter plot of the data points, color-coded by original clusters
hotel_data_original <- cbind(hotel_data, Original_Cluster = true_cluster)
ggplot(hotel_data_original, aes(x = Age, y = Price, color = Original_Cluster)) +
  geom_point() +
  labs(title = "Original Clusters", x = "Age", y = "Price", color = "Original_Cluster") +
  scale_color_viridis_d()+
  theme_minimal()


# Compare distributions of Income across original clusters using box plots
hotel_data_original <- cbind(hotel_data, Original_Cluster = true_cluster)
ggplot(hotel_data_original, aes(x = factor(hotel_data_original$Original_Cluster), y = Income, fill = factor(hotel_data_original$Original_Cluster))) +
  geom_boxplot() +
  labs(x = "Cluster", y = "Income", title = "Distribution of Income by Original Cluster") +
  scale_fill_discrete(name = "Cluster") +
  theme_minimal()
```

4ο Κεφάλαιο - εφαρμογή της μεθόδου DBSCAN - BE

```
# Load required libraries
library(dbscan)
library(dplyr)
library(ggplot2)



# Identify categorical columns
categorical_columns <- sapply(hotel_data, is.factor)
categorical_data <- hotel_data[, categorical_columns]

# Calculate frequency for each categorical variable
frequency_results <- lapply(categorical_data, function(column) {
  frequencies <- table(column)
  frequencies_percentage <- frequencies / sum(frequencies)
  data.frame(Frequency = frequencies, Percentage = frequencies_percentage)
})

# Assign names to the frequency results
names(frequency_results) <- names(categorical_data)
```

```
# Function to calculate BE conversion
calculate_xij <- function(column) {
  probabilities <- table(column) / length(column)
  xij <- -probabilities * log(probabilities)
  xij[is.nan(xij)] <- 0  # Handle the case when p(xi j) is 0
  return(xij)
}

# Calculate BE conversion for each categorical variable
xij_results <- lapply(categorical_data, calculate_xij)

# Assign names to the BE conversion results
names(xij_results) <- names(categorical_data)

# Print BE conversion results
for (i in seq_along(xij_results)) {
  cat("Column:", names(xij_results)[i], "\n")
  print(xij_results[[i]])
  cat("\n")
}

# Replace categorical attributes with BE conversion
hotel_data_BE <- hotel_data
for (i in seq_along(xij_results)) {
  column_name <- names(xij_results)[i]
  hotel_data_BE[[column_name]] <-
xij_results[[i]][match(hotel_data[[column_name]], names(xij_results[[i]]))]
}

# Select the columns to determine Eps
distance_data <- hotel_data_BE

# Calculate distances for each point to its kth nearest neighbor for
different values of k
max_k <- 700
k_distance_values <- lapply(1:max_k, function(k) {
  dbscan::kNNdist(distance_data, k = k + 1)
})

# Extract the maximum distance for each k
max_distances <- sapply(k_distance_values, function(distances) {
  max(distances)
})
```

```
# Create a data frame with k and corresponding maximum distance values
k_distance_df <- data.frame(k = 1:max_k, MaxDistance = max_distances)

# Function to calculate the angle between three points
calculate_angle <- function(p1, p2, p3) {
  a <- sqrt(sum((p2 - p3)^2))
  b <- sqrt(sum((p1 - p3)^2))
  c <- sqrt(sum((p1 - p2)^2))
  angle_rad <- acos((a^2 + b^2 - c^2) / (2 * a * b))
  return(angle_rad)
}

# Find the knee point on the k-distance plot
knee_point <- NULL
for (i in 2:(nrow(k_distance_df) - 1)) {
  distances <- dist(rbind(c(k_distance_df[i - 1, "k"], k_distance_df[i,
"MaxDistance"]),
                          c(k_distance_df[i, "k"], k_distance_df[i,
"MaxDistance"]),
                          c(k_distance_df[i + 1, "k"], k_distance_df[i + 1,
"MaxDistance"])))
  angle_i <- acos((distances[1]^2 + distances[3]^2 - distances[2]^2) / (2 *
distances[1] * distances[3]))
  if (!is.nan(angle_i) && (is.null(knee_point) || angle_i >
knee_point$angle) && angle_i > 0.1) {
    knee_point <- list(index = i, angle = angle_i)
  }
}

# Extract the value of k corresponding to the knee point
knee_k <- knee_point$index

# Print the knee point and its corresponding Eps value
cat("Eps value:", k_distance_df$MaxDistance[knee_k], "\n")

# Specify the range of MinPts values to consider
minPts_values <- 2:1000

# Initialize an empty vector to store the number of clusters
num_clusters <- numeric(length(minPts_values))

# Calculate the number of clusters for each MinPts value
for (i in seq_along(minPts_values)) {
```

```
  minPts <- minPts_values[i]
  dbscan_result <- dbscan(hotel_data_BE, eps =
k_distance_df$MaxDistance[knee_k] , minPts = minPts)
  num_clusters[i] <- length(unique(dbscan_result$cluster))
}

# Create a data frame to store the results
results_df <- data.frame(MinPts = minPts_values, NumClusters =
num_clusters)

# Define a function to find the "elbow point" index
elbow_point <- function(data) {
  # Calculate the differences between consecutive data points
  diff_data <- diff(data)

  # Calculate the second differences
  diff2_data <- diff(diff_data)

  # Find the index where the second differences are minimal
  elbow_index <- which.min(diff2_data) + 1  # +1 to account for the diff
operation

  return(elbow_index)
}

# Find the index of the "elbow point"
elbow_index <- elbow_point(results_df$NumClusters)
best_MinPts <- results_df$MinPts[elbow_index]

cat("Best MinPts:", best_MinPts, "\n")

# Run DBSCAN with selected Eps value and MinPts value
dbscan_BE_results <- dbscan::dbscan(hotel_data_BE, eps =
k_distance_df$MaxDistance[knee_k], MinPts = best_MinPts)

#Create the cross-tabulation table
xtabs(~true_cluster+dbscan_BE_results$cluster)

# Scatterplot of Age vs. Price with Clusters Colored
plot(hotel_data$Age, hotel_data$Price,
     main = "DBSCAN Clustering",
     xlab = "Age", ylab = "Price",
     col = dbscan_BE_results$cluster + 1, # Add 1 to avoid coloring points
with cluster -1 (noise) in black
```

```
     pch = 16,
     ylim = c(0, 1000))
legend("topleft", legend = c("Cluster 1", "Cluster 2", "Noise"),
       col = c(2, 3, 1), pch = 16, title = "Clusters")
```

4ο Κεφάλαιο - εφαρμογή της μεθόδου DBSCAN - SE

```
# Load required libraries
library(dbscan)
library(dplyr)
library(ggplot2)

# Identify categorical columns
categorical_columns <- sapply(hotel_data, is.factor)
categorical_data <- hotel_data[, categorical_columns]

# Calculate frequency for each categorical variable
frequency_results <- lapply(categorical_data, function(column) {
  frequencies <- table(column)
  frequencies_percentage <- frequencies / sum(frequencies)
  data.frame(Frequency = frequencies, Percentage = frequencies_percentage)
})

# Assign names to the frequency results
names(frequency_results) <- names(categorical_data)

# Function to calculate SE conversion
calculate_xij <- function(column) {
  probabilities <- table(column) / length(column)
  xij <- 1 / (1 - probabilities * log(probabilities)) * probabilities
  xij[is.nan(xij)] <- 0  # Handle the case when p(xi j) is 0
  return(xij)
}

# Calculate SE conversion for each categorical variable
xij_results <- lapply(categorical_data, calculate_xij)

# Assign names to the SE conversion results
names(xij_results) <- names(categorical_data)

# Print SE conversion results
for (i in seq_along(xij_results)) {
  cat("Column:", names(xij_results)[i], "\n")
```

```
  print(xij_results[[i]])
  cat("\n")
}

# Replace categorical attributes with SE conversion
hotel_data_SE <- hotel_data
for (i in seq_along(xij_results)) {
  column_name <- names(xij_results)[i]
  hotel_data_SE[[column_name]] <-
xij_results[[i]][match(hotel_data[[column_name]], names(xij_results[[i]]))]
}

# Select the columns to determine Eps
distance_data <- hotel_data_SE

# Calculate distances for each point to its kth nearest neighbor for
different values of k
max_k <- 700
k_distance_values <- lapply(1:max_k, function(k) {
  dbscan::kNNdist(distance_data, k = k)
})

# Extract the maximum distance for each k
max_distances <- sapply(k_distance_values, function(distances) {
  max(distances)
})

# Create a data frame with k and corresponding maximum distance values
k_distance_df <- data.frame(k = 1:max_k, MaxDistance = max_distances)

# Function to calculate the angle between three points
calculate_angle <- function(p1, p2, p3) {
  a <- sqrt(sum((p2 - p3)^2))
  b <- sqrt(sum((p1 - p3)^2))
  c <- sqrt(sum((p1 - p2)^2))
  angle_rad <- acos((a^2 + b^2 - c^2) / (2 * a * b))
  return(angle_rad)
}

# Find the knee point on the k-distance plot
knee_point <- NULL
for (i in 2:(nrow(k_distance_df) - 1)) {
```

```
  distances <- dist(rbind(c(k_distance_df[i - 1, "k"], k_distance_df[i,
"MaxDistance"]),
                          c(k_distance_df[i, "k"], k_distance_df[i,
"MaxDistance"]),
                          c(k_distance_df[i + 1, "k"], k_distance_df[i + 1,
"MaxDistance"])))
  angle_i <- acos((distances[1]^2 + distances[3]^2 - distances[2]^2) / (2 *
distances[1] * distances[3]))
  if (!is.nan(angle_i) && (is.null(knee_point) || angle_i >
knee_point$angle) && angle_i > 0.1) {
    knee_point <- list(index = i, angle = angle_i)
  }
}

# Extract the value of k corresponding to the knee point
knee_k <- knee_point$index

# Print the knee point and its corresponding Eps value
cat("Eps value:", k_distance_df$MaxDistance[knee_k], "\n")

# Specify the range of MinPts values to consider
minPts_values <- 2:1000

# Initialize an empty vector to store the number of clusters
num_clusters <- numeric(length(minPts_values))

# Calculate the number of clusters for each MinPts value
for (i in seq_along(minPts_values)) {
  minPts <- minPts_values[i]
  dbscan_result <- dbscan(hotel_data_SE, eps =
k_distance_df$MaxDistance[knee_k] , minPts = minPts)
  num_clusters[i] <- length(unique(dbscan_result$cluster))
}

# Create a data frame to store the results
results_df <- data.frame(MinPts = minPts_values, NumClusters =
num_clusters)

# Define a function to find the "elbow point" index
elbow_point <- function(data) {
  # Calculate the differences between consecutive data points
  diff_data <- diff(data)

  # Calculate the second differences
```

```
  diff2_data <- diff(diff_data)

  # Find the index where the second differences are minimal
  elbow_index <- which.min(diff2_data) + 1  # +1 to account for the diff operation

  return(elbow_index)
}

# Find the index of the "elbow point"elbow_index <- elbow_point(results_df$NumClusters)
best_MinPts <- results_df$MinPts[elbow_index]

cat("Best MinPts:", best_MinPts, "\n")

# Run DBSCAN with selected Eps value and MinPts value
dbscan_SE_results <- dbscan::dbscan(hotel_data_SE, eps = k_distance_df$MaxDistance[knee_k], MinPts = best_MinPts)

#Create the cross-tabulation table
xtabs(~true_cluster+dbscan_SE_results$cluster)

# Scatterplot of Age vs. Price with Clusters Colored
plot(hotel_data$Age, hotel_data$Price,
     main = "DBSCAN Clustering",
     xlab = "Age", ylab = "Price",
     col = dbscan_SE_results$cluster + 1, # Add 1 to avoid coloring points with cluster -1 (noise) in black
     pch = 16,
     ylim = c(0, 1000))
legend("topleft", legend = c("Cluster 1", "Cluster 2", "Noise"),
       col = c(2, 3, 1), pch = 16, title = "Clusters")
```

4ο Κεφάλαιο - εφαρμογή της συσσωρευτικής ιεραρχικής μεθόδου

```
# Load required libraries
library(cluster)
library(zoom)

# Define the procedure for computing the adjacency distance matrix
compute_adjacency_matrix <- function(data, dh, w, L) {
  # Compute the matrix
  n <- nrow(data)
```

```
  A <- matrix(0, n, n)
  for (i in 1:n) {
    for (j in 1:n) {
      if (i != j) {
        for (p in 1:ncol(data)) {
          if (is.numeric(data[[p]])) {
            # Numeric attribute
            A[i, j] <- A[i, j] + w[p] * abs( (data[i, p] - data[j, p]))^L
          } else {
            # Categorical attribute
            if (data[i, p] == data[j, p]) {
              d <- 0
            } else {
              idx_i <- which(sapply(dh[[p]], function(x) x[[1]]) == data[i,
p])
              idx_j <- which(sapply(dh[[p]], function(x) x[[1]]) == data[j,
p])
              d <- abs(idx_i - idx_j)/ length(levels)
            }
            A[i, j] <- A[i, j] + w[p] * d^L
          }
        }
        A[i, j] <- A[i, j]^(1/L)
      }
    }
  }
  # Return the matrix
  return(A)
}

# Define the distance hierarchy for each categorical attribute
dh <- list()
for (i in 1:ncol(hotel_data)) {
  if (is.numeric(hotel_data[[i]])) {
    # Numeric attribute
    dh[[i]] <- list(0)
  } else {
```

```
    # Categorical attribute
    levels <- levels(hotel_data[[i]])
    dh[[i]] <- lapply(levels, function(level) {
      list(level, length(levels))
    })
  }
}

c <- 0.5
max_height <- max(sapply(dh, length))
w <- sapply(dh, function(x) c * length(x) / max_height)
L <- 2

# Compute the adjacency distance matrix
A <- compute_adjacency_matrix(hotel_data, dh, w, L)

# Perform agglomerative hierarchical clustering with complete linkage
hclust_results_complete <- hclust(as.dist(1 / A), method = 'complete')

# Convert agglomerative hierarchical clustering to dendrogram
dend <- as.dendrogram(hclust_results_complete)

# Plot the dendrogram
plot(dend, main = "Dendrogram", xlab = "Clusters", ylab = "Distance")
zm()

# Perform agglomerative hierarchical clustering with average linkage
hclust_results_average <- hclust(as.dist(1 / A), method = 'average')

# Convert agglomerative hierarchical clustering to dendrogram
dend <- as.dendrogram(hclust_results_average)

# Plot the dendrogram
plot(dend, main = "Dendrogram", xlab = "Clusters", ylab = "Distance")
zm()
```

```
# Perform agglomerative hierarchical clustering with ward linkage
hclust_results_ward <- hclust(as.dist(1 / A), method = 'ward.D')

# Convert agglomerative hierarchical clustering to dendrogram
dend <- as.dendrogram(hclust_results_ward)

# Plot the dendrogram
plot(dend, main = "Dendrogram", xlab = "Clusters", ylab = "Distance")
zm()

#Create the cross-tabulation table for ward method
clusters <- cutree(hclust_results_ward, k = 3)
ctab <- table(clusters, true_cluster)
ctab_transposed <- t(ctab)
print(ctab_transposed)
```

8. Βιβλιογραφία

Ελληνική Βιβλιογραφία

1. Κούτρας, Μ. (2020). *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση - Ανάλυση κατά συστάδες*, Πανεπιστήμιο Πειραιώς, Πειραιάς.
2. Κύρκος, Ε. (2015). *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*, Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις, Ζωγράφου.
3. Ουζούνη, Χ., και Νακάκης, Κ. (2011). *Η αξιοπιστία και η εγκυρότητα των εργαλείων μέτρησης σε ποσοτικές μελέτες*, Νοσηλευτική, **50(2)**, 231-239.

Ξένη Βιβλιογραφία

1. Ahmad, A. and Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data, *Data & Knowledge Engineering*, **63**, 503-527.
2. Box, G. E. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**, 211-252.
3. Enders, C. K. (2010). *Applied missing data analysis*, Guilford Press, New York.
4. Ester, M., Kriegel, H. P., Sander, J. and Xu, X. (1996). Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226-231.
5. Everitt, B. S. and Dunn, G. (1991). *Applied Multivariate Data Analysis*, Arnold, New York.
6. Foss, A., and Markatou, M. (2018). KAMILA: Clustering Mixed-Type Data in R and Hadoop, *Journal of Statistical Software*, **83**, 1-44.
7. Foss, A., Markatou, M., Ray, B. et al (2016). A semiparametric method for clustering mixed data, *Machine Learning,* **105**, 419-458.
8. Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification, *Journal of the Royal Statistical Society, Series C*, **62**, 309-369.
9. Hsu, C.C. , Chen, C.L. and Su, Y.W. (2007). Hierarchical clustering of mixed data based on distance hierarchy, *Information Sciences*, **177**, 4474-4492.
10. Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values, *Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference*, 21-34.
11. Johnson, R. and Wichern D. (2007). *Applied Multivariate Statistical Analysis*, *Sixth Edition,* Pearson, New Jersey.

12. Liu, X. , Yang, Q. , and He, L. (2017). A novel DBSCAN with entropy and probability for mixed data, *Cluster Computing*, **20**, 1313-1323.
13. Manning, C.,Raghavan, P. and Schutze, H. (2008). *Introduction to information retrieval*, Cambridge: Cambridge University Press, Cambridge, England.
14. Manly, B. F. J. and Navarro Alberto, J.A. (2017). *Multivariate Statistical Methods: A primer, Fourth Edition*, Chapman and Hall, London.
15. Modha, D. and Spangler, W. (2003). Feature weighting in k-means clustering, *Machine Learning*, **52**, 217-237.
16. Saxena S. (2023). Here's All you Need to Know About Encoding Categorical Data (with Python code), https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/