**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

**ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

## Πτυχιακή Εργασία

| | |
|---|---|
| Τίτλος Πτυχιακής Εργασίας | Τεχνικές ομαδοποίησης και η χρήση τους σε λογισμικά με εξατομικευμένες υπηρεσίες.<br><br>Clustering Algorithms and their implementation in personalized Services. |
| Ονοματεπώνυμο Φοιτητή | Ναπολέων Κουτσουρίδης |
| Πατρώνυμο | Παναγιώτης |
| Αριθμός Μητρώου | Π16058 |
| Επιβλέπων | Κωνσταντίνα Χρυσαφιάδη , Επίκουρη Καθηγήτρια |

Ημερομηνία Παράδοσης Σεπτέμβριος 2024

# Copyright ©

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαιτέρως την Κα Χρυσαφιάδη Κωνσταντίνα για την ανάληψη της επίβλεψης της παρούσας πτυχιακής εργασίας καθώς και για την υποστήριξη και την καθοδήγησή της κατά την διάρκεια της εκπόνησης της εργασίας.

# Περίληψη

Η μελέτη εξετάζει τη χρήση των αλγορίθμων ομαδοποίησης για την ενίσχυση των εξατομικευμένων υπηρεσιών, εστιάζοντας στη σημασία τους για την ανάλυση δεδομένων και την βελτίωση της εμπειρίας των χρηστών. Οι αλγόριθμοι ομαδοποίησης, όπως οι K-means, η ιεραρχική ομαδοποίηση και ο DBSCAN, αναλύονται αναλυτικά για να κατανοηθεί η αποδοτικότητά τους και η εφαρμογή τους σε διάφορους τομείς. Η μελέτη περιλαμβάνει τη συλλογή και ανάλυση δεδομένων, την εφαρμογή των αλγορίθμων και την αξιολόγηση των αποτελεσμάτων σε πραγματικές συνθήκες εξατομικευμένων υπηρεσιών.

Αναλύοντας τα αποτελέσματα, η μελέτη εντοπίζει ποιοι αλγόριθμοι αποδεικνύονται πιο αποτελεσματικοί σε συγκεκριμένα σενάρια και πώς η εφαρμογή τους επηρεάζει την ποιότητα των εξατομικευμένων υπηρεσιών. Οι συγκρίσεις μεταξύ των διαφορετικών αλγορίθμων αναδεικνύουν τα πλεονεκτήματα και τις αδυναμίες τους, παρέχοντας πολύτιμες πληροφορίες για τη βελτίωση της εφαρμογής τους σε διάφορες επιχειρηματικές και τεχνολογικές περιοχές.

Στη συζήτηση, επισημαίνονται τα κύρια ευρήματα της μελέτης και οι επιπτώσεις τους για την θεωρία και την πρακτική. Εξετάζονται οι περιορισμοί της έρευνας και προτείνονται κατευθύνσεις για μελλοντική έρευνα, με στόχο την περαιτέρω κατανόηση και βελτίωση της χρήσης των αλγορίθμων ομαδοποίησης. Οι προτάσεις που προκύπτουν παρέχουν χρήσιμες κατευθύνσεις για επαγγελματίες και ερευνητές που ασχολούνται με την ανάλυση δεδομένων και την ανάπτυξη εξατομικευμένων υπηρεσιών.

Η επιστημονική περιοχή της μελέτης είναι η "Ανάλυση Δεδομένων και Μηχανική Μάθηση", με ειδικότερη εστίαση στη χρήση αλγορίθμων ομαδοποίησης για την ανάπτυξη εξατομικευμένων υπηρεσιών.

*Λέξεις Κλειδιά:* Αλγόριθμοι Ομαδοποίησης, Εξατομικευμένες Υπηρεσίες, Μηχανική Μάθηση Ανάλυση Δεδομένων, Εφαρμογές Ομαδοποίησης

# Abstract

The study examines the use of clustering algorithms to enhance personalized services, focusing on their importance for data analysis and improving user experience. Clustering algorithms, such as K-means, hierarchical clustering and DBSCAN, are analyzed in detail to understand their efficiency and their application in different domains. The study includes data collection and analysis, implementation of the algorithms and evaluation of the results in real-life personalized service settings.

By analyzing the results, the study identifies which algorithms prove to be most effective in specific scenarios and how their implementation affects the quality of personalized services. Comparisons between different algorithms highlight their strengths and weaknesses, providing valuable insights for improving their application in different business and technological areas.

In the discussion, the main findings of the study and their implications for theory and practice are highlighted. The limitations of the research are discussed and directions for future research are suggested to further understand and improve the use of clustering algorithms. The resulting recommendations provide useful guidance for practitioners and researchers involved in data analysis and the development of personalized services.

The scientific area of study is "Data Analysis and Machine Learning", with a special focus on the use of clustering algorithms for the development of personalized services.

*Key Words:* Clustering Algorithms, Personalized Services, Machine Learning Data Analysis, Clustering Applications

# Αφιερώσεις

Αφιερωμένο στον αγαπημένο μου παππού, τον αληθινό ήρωά μου, που πάντα με στήριζε και μου έδινε δύναμη. Η παρουσία σου στη ζωή μου ήταν ανεκτίμητη και η αγάπη σου πάντα οδηγός μου. Η έμπνευση και η υποστήριξή σου με βοήθησαν να φτάσω ως εδώ, και αυτή η πτυχιακή εργασία είναι αφιερωμένη στη μνήμη σου. Σε ευχαριστώ για όλα.

# Πίνακας Περιεχομένων

## Περιεχόμενα

# Κατάλογος Εικόνων

# Κατάλογος Πινάκων

# 1. Εισαγωγή

## 1.1 Πρόβλημα – Σημαντικότητα του θέματος

Η τεράστια ανάπτυξη των ψηφιακών δεδομένων τα τελευταία χρόνια έχει καταστήσει απαραίτητη την ανάλυση και την εξαγωγή χρήσιμων πληροφοριών από μεγάλους όγκους δεδομένων. Σε αυτό το πλαίσιο, οι αλγόριθμοι ομαδοποίησης αποτελούν κρίσιμα εργαλεία για την αναγνώριση προτύπων και την κατηγοριοποίηση δεδομένων με παρόμοια χαρακτηριστικά. Η ομαδοποίηση δεδομένων, γνωστή και ως clustering, επιτρέπει στους αναλυτές να κατανοήσουν καλύτερα τις δομές και τις σχέσεις μέσα σε σύνολα δεδομένων, παρέχοντας πολύτιμες πληροφορίες για τη λήψη αποφάσεων.

Η αυξανόμενη πολυπλοκότητα των δεδομένων και η ποικιλία των εφαρμογών τους έχουν δημιουργήσει την ανάγκη για εξελιγμένους αλγορίθμους ομαδοποίησης που μπορούν να προσφέρουν ακριβή και αξιόπιστα αποτελέσματα σε σύντομο χρονικό διάστημα. Παρά την ύπαρξη πολλών αλγορίθμων, η επιλογή του καταλληλότερου για μια συγκεκριμένη εφαρμογή παραμένει ένα δύσκολο και περίπλοκο πρόβλημα. Επιπλέον, η εφαρμογή αυτών των αλγορίθμων στις εξατομικευμένες υπηρεσίες απαιτεί προσεκτική ανάλυση και αξιολόγηση της απόδοσής τους σε πραγματικά δεδομένα.

Η ανάλυση δεδομένων μέσω αλγορίθμων ομαδοποίησης έχει σημαντικές επιπτώσεις σε πολλούς τομείς, από την υγειονομική περίθαλψη και το εμπόριο μέχρι τις κοινωνικές επιστήμες και την τεχνολογία. Οι εξατομικευμένες υπηρεσίες, που προσαρμόζουν τα προϊόντα και τις υπηρεσίες στις ανάγκες και τις προτιμήσεις των χρηστών, βασίζονται σε μεγάλο βαθμό στους αλγόριθμους ομαδοποίησης για την αναγνώριση μοτίβων και την κατηγοριοποίηση των χρηστών σε ομάδες με παρόμοια χαρακτηριστικά. Η βελτίωση της ακρίβειας και της αποδοτικότητας των αλγορίθμων αυτών μπορεί να οδηγήσει σε καλύτερες και πιο αποδοτικές εξατομικευμένες υπηρεσίες, προσφέροντας αυξημένη ικανοποίηση των χρηστών και βελτιωμένη επιχειρησιακή αποδοτικότητα.

Η μελέτη αυτή στοχεύει στην κατανόηση και τη σύγκριση διαφόρων αλγορίθμων ομαδοποίησης, διερευνώντας την αποτελεσματικότητά τους στην παροχή εξατομικευμένων υπηρεσιών και συμβάλλοντας στην ανάπτυξη πιο υπηρεσιών.

## 1.2 Σκοπός και στόχοι

Ο σκοπός της παρούσας μελέτης είναι να διερευνήσει και να αξιολογήσει τη χρήση αλγορίθμων ομαδοποίησης στην ανάπτυξη εξατομικευμένων υπηρεσιών, με στόχο τη βελτίωση της εμπειρίας των χρηστών και της αποδοτικότητας των επιχειρήσεων. Μέσω της ανάλυσης και της σύγκρισης διαφορετικών αλγορίθμων, η μελέτη στοχεύει να προσδιορίσει τους πιο αποδοτικούς αλγόριθμους για διάφορες εφαρμογές, καθώς και να προτείνει βελτιώσεις και καινοτόμες λύσεις στον τομέα αυτό.

**Στόχοι της Μελέτης**

1. **Ανασκόπηση και Κατανόηση των Αλγορίθμων Ομαδοποίησης**
   - Να παρουσιάσει και να αναλύσει τους διάφορους τύπους αλγορίθμων ομαδοποίησης, όπως οι K-means, η ιεραρχική ομαδοποίηση και ο DBSCAN.
   - Να εξετάσει τις βασικές αρχές λειτουργίας και τις διαφορές μεταξύ αυτών των αλγορίθμων.
2. **Σύγκριση της Απόδοσης των Αλγορίθμων**
   - Να συγκρίνει την απόδοση των διαφορετικών αλγορίθμων σε διάφορα σύνολα δεδομένων.
   - Να εντοπίσει τους παράγοντες που επηρεάζουν την απόδοση και την ακρίβεια των αλγορίθμων.
3. **Εφαρμογή των Αλγορίθμων σε Εξατομικευμένες Υπηρεσίες**
   - Να εξετάσει πώς οι αλγόριθμοι ομαδοποίησης μπορούν να εφαρμοστούν σε πραγματικά σενάρια εξατομικευμένων υπηρεσιών.

- Να αναλύσει την αποτελεσματικότητα των αλγορίθμων στην κατηγοριοποίηση των χρηστών και στην παροχή προσαρμοσμένων προτάσεων και υπηρεσιών.

4. **Αξιολόγηση και Βελτίωση της Απόδοσης των Αλγορίθμων**
   - Να αξιολογήσει την απόδοση των αλγορίθμων μέσω διαφόρων μεθόδων και κριτηρίων αξιολόγησης.
   - Να προτείνει βελτιώσεις και προσαρμογές στους αλγορίθμους για την επίτευξη καλύτερων αποτελεσμάτων.
5. **Προτάσεις για Πρακτική Εφαρμογή και Μελλοντική Έρευνα**
   - Να παρέχει συστάσεις για την πρακτική εφαρμογή των αλγορίθμων σε διάφορους τομείς.
   - Να εντοπίσει περιοχές για μελλοντική έρευνα και ανάπτυξη στον τομέα της ανάλυσης δεδομένων και των εξατομικευμένων υπηρεσιών.

Με την επίτευξη αυτών των στόχων, η μελέτη θα συμβάλει στην κατανόηση και τη βελτίωση των αλγορίθμων ομαδοποίησης, προωθώντας την ανάπτυξη πιο αποτελεσματικών και αποδοτικών εξατομικευμένων υπηρεσιών.

## 1.3 Συνεισφορά

Η παρούσα μελέτη προσφέρει σημαντικές συνεισφορές στον τομέα της ανάλυσης δεδομένων και της μηχανικής μάθησης, ειδικότερα στην εφαρμογή των αλγορίθμων ομαδοποίησης στις εξατομικευμένες υπηρεσίες. Οι συνεισφορές αυτές κατηγοριοποιούνται σε θεωρητικές, πρακτικές και ερευνητικές, καλύπτοντας ευρύ φάσμα εφαρμογών και προοπτικών.

**Θεωρητική Συνεισφορά**

- Συστηματική Ανασκόπηση: Παρέχει μια ολοκληρωμένη ανασκόπηση και ανάλυση των βασικών αλγορίθμων ομαδοποίησης, προσδιορίζοντας τα πλεονεκτήματα και τα μειονεκτήματά τους.
- Νέα Προοπτική: Εισάγει νέες οπτικές και προσεγγίσεις για την εφαρμογή των αλγορίθμων αυτών στις εξατομικευμένες υπηρεσίες, συνεισφέροντας στη θεωρητική κατανόηση της σχέσης μεταξύ της ομαδοποίησης και της εξατομίκευσης.

**Πρακτική Συνεισφορά**

- Οδηγίες Εφαρμογής: Παρέχει πρακτικές κατευθυντήριες γραμμές για την εφαρμογή και την αξιολόγηση των αλγορίθμων ομαδοποίησης σε πραγματικά δεδομένα και σενάρια.
- Βέλτιστες Πρακτικές: Αναδεικνύει τις βέλτιστες πρακτικές για τη βελτίωση της αποδοτικότητας των αλγορίθμων, προσφέροντας λύσεις και στρατηγικές που μπορούν να εφαρμοστούν άμεσα από επαγγελματίες στον τομέα.

**Ερευνητική Συνεισφορά**

- Αναγνώριση Περιορισμών και Προκλήσεων: Εντοπίζει τους περιορισμούς των τρεχουσών μεθόδων και προτείνει περιοχές για μελλοντική έρευνα και ανάπτυξη, συμβάλλοντας στη συνεχή πρόοδο του τομέα.
- Καινοτόμες Προτάσεις: Προσφέρει καινοτόμες προτάσεις για τη βελτίωση των αλγορίθμων ομαδοποίησης και την ανάπτυξη νέων τεχνικών που μπορούν να βελτιώσουν τις εξατομικευμένες υπηρεσίες.

Η συνεισφορά της μελέτης αυτής έχει την προοπτική να επηρεάσει τόσο τη θεωρητική όσο και την πρακτική προσέγγιση των αλγορίθμων ομαδοποίησης, προσφέροντας πολύτιμες γνώσεις και εργαλεία για ερευνητές, επαγγελματίες και οργανισμούς που ασχολούνται με την ανάλυση δεδομένων και την παροχή εξατομικευμένων υπηρεσιών.

## 1.4 Βασική Ορολογία

**Πίνακας 1 Βασική ορολογία μελέτης**

| | |
|---|---|
| **K-means Clustering** | Δημοφιλής αλγόριθμος που χωρίζει τα δεδομένα σε Κ ομάδες |
| **Hierarchical Clustering** | Τεχνική ομαδοποίησης για ιεράρχηση ομάδων |
| **DBSCAN** | Density-Based Spatial Clustering of Applications with Noise |
| **GMM** | Gaussian Mixture Models |
| **Mean Shift Clustering** | Εντοπίζει περιοχές υψηλής πυκνότητας μετακινώντας επαναληπτικά τα δεδομένα |
| **K-medoids** | Αποτελεί μία παραλλαγή του αλγορίθμου K-means, που έχει ως στόχο τη μείωση της ευαισθησίας στις εξωγενείς τιμές και τα ανώμαλα δεδομένα |
| **Linkage Criteria** | Κριτήρια συνοχής καθορίζουν πώς υπολογίζεται η απόσταση μεταξύ συστάδων |

## 1.5 Διάρθρωση μελέτης

Στο κεφάλαιο 1ο παρουσιάζεται το πρόβλημα της παρούσας μελέτης καθώς και η σημαντικότητα της. Παρουσιάζονται αναλυτικά ο σκοπός και οι στόχοι της , η συνεισφορά καθώς και ένας αναλυτικός πίνακας με τις πιο βασικές ορολογίες. Στο κεφάλαιο 2ο γίνεται μια βιβλιογραφική ανασκόπηση στους βασικότερους αλγόριθμους ομαδοποίησης παρουσιάζοντας αναλυτικά τον τρόπο λειτουργίας τους , ορισμένα παραδείγματα εφαρμογής τους καθώς και τα βασικότερα μειονεκτήματα και πλεονεκτήματα που προκύπτουν από την εφαρμογή τους.

Στο κεφάλαιο 3ο γίνεται μια σύγκριση των αλγόριθμων ομαδοποίησης και παρουσιάζονται ορισμένες εφαρμογές τους. Στη συνέχεια του κεφαλαίου παρουσιάζονται οι εξατομικευμένες υπηρεσίες και ο ρόλος τους καθώς και πως οι εξατομικευμένες υπηρεσίες μπορούν να προσδιοριστούν σε σχέση με τους αλγόριθμούς ομαδοποίησης και τη διαχείριση μεγάλου όγκου δεδομένων. Στο κεφάλαιο 4ο για να μπορέσουν να αποδοθούν οι αλγόριθμοι ομαδοποίησης και η σημασία τους σε σχέση με τις εξατομικευμένες υπηρεσίες παρουσιάζονται δύο σύντομες μελέτες στον τομέα της υγείας και του μάρκετινγκ πωλήσεων. Σε κάθε μελέτη παρουσιάζονται αναλυτικά τα δεδομένα, ο υπολογισμός των αλγόριθμών στη γλώσσα python , η εξαγωγή των αποτελεσμάτων καθώς και η ανάλογη εξήγηση τους με διάφορες προτάσεις για εξατομικευμένες υπηρεσίες.

Στο τέλος της μελέτης παρουσιάζονται αναλυτικά συμπεράσματα σχετικά με την έρευνα πάνω στους αλγόριθμους ομαδοποίησης δεδομένων για παροχή εξατομικευμένων υπηρεσιών καθώς και προτάσεις για μελλοντική έρευνα πάνω στο συγκεκριμένο τομέα.

# 2. Βιβλιογραφική Ανασκόπηση

## 2.1 Ανάλυση συστάδων

Η "Ανάλυση Συστάδων" επικεντρώνεται στη διαδικασία ομαδοποίησης δεδομένων με βάση την ομοιότητά τους, μια μέθοδο που είναι ιδιαίτερα σημαντική όταν δεν υπάρχει προ υπάρχουσα γνώση για τις κατηγορίες που μπορεί να υπάρχουν στα δεδομένα. Αυτή η διαδικασία επιτρέπει στους αναλυτές να ανακαλύψουν κρυμμένα πρότυπα και σχέσεις μεταξύ των δεδομένων, χωρίς να απαιτείται προκαθορισμένη κατηγοριοποίηση. Στο επιχειρηματικό πλαίσιο, η ανάλυση συστάδων μπορεί να χρησιμοποιηθεί για να αποκαλύψει καταναλωτικά πρότυπα, τμηματοποίηση πελατών και άλλες κρίσιμες πληροφορίες που ενισχύουν τη λήψη αποφάσεων. Μια βασική έννοια της ανάλυσης συστάδων είναι ότι αποτελεί μια μη επιβλεπόμενη τεχνική μάθησης, κάτι που σημαίνει ότι δεν απαιτεί προκαθορισμένες κατηγορίες για να λειτουργήσει. Ο στόχος είναι να βρεθούν φυσικές ομάδες μέσα στα δεδομένα, οι οποίες χαρακτηρίζονται από την εσωτερική τους ομοιότητα και τη διαφορά τους από άλλες ομάδες. Για να επιτευχθεί αυτό, χρησιμοποιούνται διάφοροι αλγόριθμοι, όπως οι k-means, η ιεραρχική ανάλυση συστάδων και ο αλγόριθμος DBSCAN, καθένας από τους οποίους έχει τα δικά του πλεονεκτήματα και μειονεκτήματα, ανάλογα με τη φύση των δεδομένων και τους στόχους της ανάλυσης. (Κύρκος, 2015)

Η επιλογή των κατάλληλων χαρακτηριστικών των δεδομένων είναι κρίσιμη για την επιτυχία της ανάλυσης συστάδων. Τα χαρακτηριστικά που επιλέγονται καθορίζουν τον τρόπο με τον οποίο θα ομαδοποιηθούν τα δεδομένα, και συνεπώς η επιλογή αυτή μπορεί να επηρεάσει σημαντικά τα αποτελέσματα της ανάλυσης. Η σωστή προεπεξεργασία των δεδομένων, όπως η κανονικοποίηση και η τυποποίηση, είναι επίσης απαραίτητη για να διασφαλιστεί ότι τα αποτελέσματα είναι αξιόπιστα και αντιπροσωπευτικά. Μια από τις προκλήσεις στην ανάλυση συστάδων είναι η επιλογή του σωστού αριθμού ομάδων (clusters). Αυτή η επιλογή δεν είναι πάντα προφανής και μπορεί να επηρεάσει τα τελικά αποτελέσματα της ανάλυσης. Υπάρχουν διάφορες μέθοδοι που χρησιμοποιούνται για να προσδιοριστεί ο βέλτιστος αριθμός ομάδων, όπως το elbow method και το silhouette score, οι οποίες βοηθούν στον εντοπισμό του σημείου όπου η αύξηση του αριθμού των ομάδων δεν βελτιώνει σημαντικά την εσωτερική συνοχή της ομάδας (Κύρκος, 2015)

Η εφαρμογή της ανάλυσης συστάδων μπορεί να γίνει σε διάφορα πεδία, από το μάρκετινγκ και την οικονομία, έως τη βιολογία και την κοινωνιολογία. Για παράδειγμα, στον τομέα του μάρκετινγκ, η ανάλυση αυτή μπορεί να χρησιμοποιηθεί για την τμηματοποίηση της αγοράς, βοηθώντας τις επιχειρήσεις να αναγνωρίσουν ομάδες πελατών με κοινά χαρακτηριστικά και να προσαρμόσουν τις στρατηγικές τους αναλόγως. Επιπλέον, η ανάλυση συστάδων έχει σημαντική εφαρμογή στη βιολογία, όπως στη γονιδιωματική και την ανάλυση δεδομένων από πειράματα γονιδιακής έκφρασης. Η ομαδοποίηση γονιδίων με παρόμοια πρότυπα έκφρασης μπορεί να βοηθήσει στην ανακάλυψη νέων βιολογικών διαδικασιών ή στη διάγνωση ασθενειών. Παρόμοια, σε κοινωνιολογικές μελέτες, η ανάλυση αυτή μπορεί να βοηθήσει στην κατανόηση κοινωνικών δομών και προτύπων συμπεριφοράς. Επιπροσθέτως, υπάρχουν ορισμένα μειονεκτήματα και προκλήσεις κατά την ανάλυση συστάδων. Ένα από τα κύρια προβλήματα είναι η πολυπλοκότητα των δεδομένων και η δυσκολία στην οπτικοποίηση των αποτελεσμάτων, ειδικά όταν υπάρχουν πολλές διαστάσεις. Επίσης, οι αλγόριθμοι συστάδων μπορεί να επηρεαστούν από θόρυβο στα δεδομένα ή από την παρουσία ακραίων τιμών, κάτι που απαιτεί προσεκτική προεπεξεργασία. (Κύρκος, 2015)

Συνολικά, η ανάλυση των συστάδων αποτελεί ένα κρίσιμο εργαλείο για την κατανόηση σύνθετων δεδομένων και τη λήψη αποφάσεων. Παρά τις προκλήσεις που παρουσιάζει, αποτελεί ένα σημαντικό πλεονέκτημα στην εργαλειοθήκη των δεδομένων, προσφέροντας ευκαιρίες για ανακάλυψη κρυμμένων προτύπων και σχέσεων. Τέλος, η ανάλυση αυτή αποτελεί βασικό στοιχείο της επιχειρηματικής ευφυΐας, παρέχοντας έναν τρόπο για την ενίσχυση της κατανόησης των δεδομένων και της δημιουργίας αξίας μέσω

της ανακάλυψης γνώσης, κάτι που είναι ιδιαίτερα σημαντικό σε ένα περιβάλλον όπου οι αποφάσεις βασίζονται ολοένα και περισσότερο στην ανάλυση δεδομένων.

## 2.2 Ορισμός και Τύποι Αλγορίθμων Ομαδοποίησης

Οι αλγόριθμοι ομαδοποίησης είναι μέθοδοι της μηχανικής μάθησης που χρησιμοποιούνται για την ταξινόμηση ενός συνόλου δεδομένων σε ομάδες (clusters) με βάση τις ομοιότητές τους. Σκοπός της ομαδοποίησης είναι να δημιουργηθούν ομάδες όπου τα δεδομένα εντός της ίδιας ομάδας είναι πιο όμοια μεταξύ τους από ό,τι με δεδομένα σε άλλες ομάδες. Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται σε πολλές εφαρμογές, όπως η αναγνώριση προτύπων, η ανάλυση δεδομένων και η εξόρυξη δεδομένων (Jain & Dubes, 2020).

**Τύποι Αλγορίθμων Ομαδοποίησης**

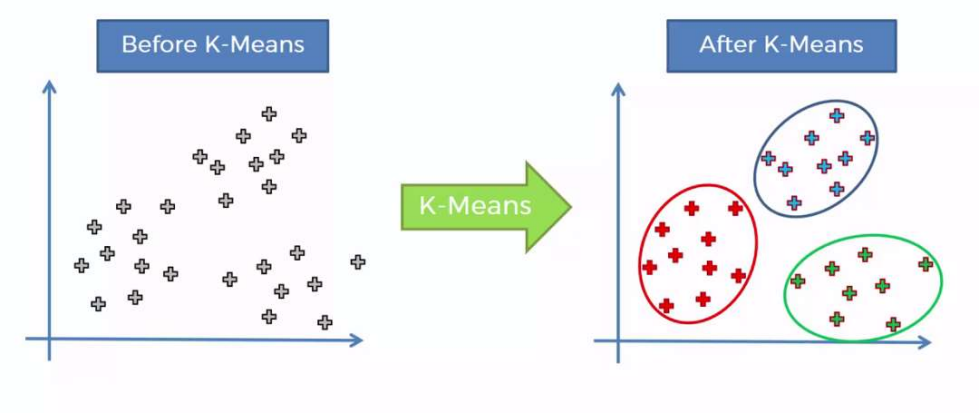
- **K-means Clustering**
  Ένας δημοφιλής αλγόριθμος που χωρίζει τα δεδομένα σε Κ ομάδες με βάση την εγγύτητα των δεδομένων σε κεντροειδή σημεία που ενημερώνονται επαναληπτικά.
- **Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)**
  Δημιουργεί μια ιεραρχική δομή των ομάδων, η οποία μπορεί να αναπαρασταθεί με ένα δενδροειδές διάγραμμα (dendrogram), είτε ξεκινώντας από μεμονωμένα δεδομένα και συγχωνεύοντας τα (αγνωστική μέθοδος), είτε ξεκινώντας από όλα τα δεδομένα σε μία ομάδα και διαιρώντας τα σταδιακά (διαιρετική μέθοδος).
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**
  Ένας αλγόριθμος που εντοπίζει ομάδες με βάση την πυκνότητα των δεδομένων, αναγνωρίζοντας περιοχές υψηλής πυκνότητας και διαχωρίζοντάς τις από περιοχές χαμηλής πυκνότητας.
- **Mean Shift Clustering**
  Εντοπίζει περιοχές υψηλής πυκνότητας μετακινώντας επαναληπτικά τα δεδομένα προς την κατεύθυνση της υψηλότερης πυκνότητας.
- **Gaussian Mixture Models (GMM)**
  Βασίζεται στην υπόθεση ότι τα δεδομένα προέρχονται από ένα μίγμα πολλών κανονικών κατανομών και χρησιμοποιεί τη μέγιστη πιθανότητα για να βρει τις παραμέτρους αυτών των κατανομών.
- **K medoids clustering**
  Ο αλγόριθμος K-medoids, παρόμοιος με τον K-means, αναζητά ομάδες δεδομένων, αλλά αντί να χρησιμοποιεί κεντροειδή (μέσες τιμές) όπως ο K-means, επιλέγει πραγματικά δεδομένα ως κεντροειδή (medoids) για κάθε ομάδα. Το K-medoids είναι πιο ανθεκτικός σε outliers σε σύγκριση με τον K-means, καθώς η επιλογή των medoids από τα δεδομένα μειώνει την ευαισθησία σε ασυνήθιστα σημεία δεδομένων.

### 2.2.1 K-means Clustering

Ο αλγόριθμος K-means Clustering είναι ένα ευρέως χρησιμοποιούμενο εργαλείο στην ανάλυση δεδομένων που στοχεύει στη διαίρεση ενός συνόλου δεδομένων σε έναν προκαθορισμένο αριθμό ομάδων, γνωστές ως κλάστερ. Ο αλγόριθμος λειτουργεί με την εκχώρηση κάθε σημείου δεδομένων σε μία από τις Κ ομάδες με βάση την εγγύτητα του σημείου προς το κέντρο της ομάδας. Ο αλγόριθμος αυτός στην ουσία δημιουργεί τις συστάδες εντοπίζοντας το κέντρο τους με την εφαρμογή της Ευκλείδειας απόστασης μεταξύ του κάθε σημείου του συνόλου των δεδομένων. Το κέντρο της συστάδας υπολογίζεται ως το μέσο όλων των σημείων που ανήκουν στην ομάδα, και ανανεώνεται συνεχώς μέχρι η διαδικασία να καταλήξει σε σταθερά αποτελέσματα. Η διαδικασία αυτή επαναλαμβάνεται έως ότου οι αλλαγές στα κέντρα γίνουν ελάχιστες ή επιτευχθεί ο προκαθορισμένος αριθμός επαναλήψεων (Hartigan,& Wong, 1979).

Η βασική αρχή του K-means είναι η αναζήτηση της βέλτιστης κατανομής των δεδομένων ώστε να ελαχιστοποιηθεί η συνολική απόσταση μεταξύ των σημείων δεδομένων και των αντίστοιχων κέντρων. Αυτός ο αλγόριθμος είναι απλός στην υλοποίηση και γρήγορος στην εκτέλεση, καθιστώντας τον κατάλληλο για ανάλυση μεγάλων συνόλων δεδομένων. Ωστόσο, η αποτελεσματικότητά του εξαρτάται σε μεγάλο βαθμό από την αρχική επιλογή των κεντροειδών, γεγονός που μπορεί να επηρεάσει την ποιότητα των τελικών ομάδων και να οδηγήσει σε τοπικά βέλτιστα λύσεις

Ένας άλλος περιορισμός του K-means είναι η ανάγκη για προκαθορισμένο αριθμό ομάδων, K. Η επιλογή αυτού του αριθμού μπορεί να είναι δύσκολη, ειδικά όταν δεν υπάρχουν σαφή πρότυπα στα δεδομένα. Εάν ο αριθμός των ομάδων δεν επιλεγεί σωστά, μπορεί να προκύψουν εσφαλμένα ή μη ικανοποιητικά αποτελέσματα. Επιπλέον, ο αλγόριθμος τείνει να δημιουργεί ομάδες με σφαιρικό σχήμα, πράγμα που τον καθιστά λιγότερο κατάλληλο για σύνολα δεδομένων που περιλαμβάνουν ομάδες με ασύμμετρα ή πολύπλοκα σχήματα (Jain, 2010).




**Εικόνα 1 Εφαρμογή K-Means**

Πηγή: https://medium.com/@pranav3nov/understanding-k-means-clustering-f5e2e84d2129

**Πώς Λειτουργεί ο K-means**

Ο αλγόριθμος K-means εκτελείται μέσω των παρακάτω βημάτων:

- **Αρχικοποίηση**: Επιλέγονται τυχαία k σημεία δεδομένων ως αρχικά μέσα ή κεντροειδή συστάδων. (centroids).
- **Αντιστοίχιση Σημείων**: Κάθε σημείο δεδομένων αντιστοιχίζεται στο πλησιέστερο κεντροειδές, σχηματίζοντας k συστάδες.
- **Ενημέρωση Κεντροειδών**: Υπολογίζονται τα νέα κεντροειδή ως οι μέσες τιμές (μέσοι όροι) των σημείων δεδομένων που ανήκουν σε κάθε συστάδα.
- **Επανάληψη**: Τα βήματα 2 και 3 επαναλαμβάνονται μέχρι τα κεντροειδή να μην αλλάζουν (ή μέχρι να φτάσουμε σε έναν προκαθορισμένο αριθμό επαναλήψεων) (Mohiuddin, Seraj, Islam, 2022).

Ο αλγόριθμος εκτελεί επαναληπτικά αυτά τα δύο βήματα, μέχρις ότου τα κεντροειδή των συστάδων να μετατοπίζονται ελάχιστα και σε απόσταση μικρότερη από κάποια δοθείσα τιμή κατωφλίου. Ως εναλλακτικό κριτήριο τερματισμού του αλγορίθμου μπορεί να χρησιμοποιηθεί και ο αριθμός επαναλήψεων του αλγορίθμου (Βερύκιος, Καγκλής &Σταυρόπουλος, 2015).

*Αρχικοποίησε τυχαία τα k κεντροειδή των συστάδων μ1,μ2,…,μk.*
*Επανέλαβε{*
*Εξέτασε κάθε δείγμα και ανέθεσε το στη συστάδα με το πλησιέστερο κεντροειδές ( min |x(i) – μk|2 )*
*Επανυπολόγισε τα κεντροειδή υπολογίζοντας το μέσο όρο των δειγμάτων της συστάδας*
*}*

**Παράδειγμα**

Εξετάζεται ένα απλό παράδειγμα με ένα σύνολο δεδομένων που περιλαμβάνει τις τοποθεσίες πελατών μιας επιχείρησης σε δύο διαστάσεις (π.χ., γεωγραφικό πλάτος και μήκος).

1. **Αρχικοποίηση**: Επιλέγεται k = 3 και τυχαία επιλέγονται τρία αρχικά σημεία ως κεντροειδή.
2. **Αντιστοίχιση Σημείων**: Κάθε πελάτης αντιστοιχίζεται στο πλησιέστερο από τα τρία κεντροειδή με βάση την ευκλείδεια απόσταση.
3. **Ενημέρωση Κεντροειδών**: Υπολογίζονται τα νέα κεντροειδή ως οι μέσες τιμές των τοποθεσιών των πελατών σε κάθε συστάδα.
4. **Επανάληψη**: Τα βήματα 2 και 3 επαναλαμβάνονται μέχρι τα κεντροειδή να μην αλλάζουν.

Ας υποθέσουμε ότι οι τοποθεσίες των πελατών είναι οι εξής: (2,3), (3,3), (6,8), (8,8), (3,4), (5,6), (8,9), (3,3). Αρχικά επιλέγουμε τυχαία τα κεντροειδή (2,3), (6,8), και (3,4).

Μετά από αρκετές επαναλήψεις, τα τελικά κεντροειδή ενδέχεται να σταθεροποιηθούν, για παράδειγμα, στα σημεία (2.5, 3.33), (7.33, 8.33), και (3, 3.33). Οι πελάτες θα αντιστοιχηθούν στις συστάδες με βάση την απόσταση από τα τελικά αυτά κεντροειδή.

Ο K-means έχει εφαρμογές σε διάφορους τομείς όπως η ανάλυση αγοράς, η αναγνώριση προτύπων και η επεξεργασία εικόνας. Στην ανάλυση αγοράς, μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση πελατών με βάση τις αγοραστικές τους συνήθειες, επιτρέποντας στις επιχειρήσεις να σχεδιάσουν στρατηγικές marketing που απευθύνονται σε συγκεκριμένες ομάδες. Στην αναγνώριση προτύπων, μπορεί να εντοπίσει κοινά χαρακτηριστικά σε δεδομένα εικόνας ή κειμένου, διευκολύνοντας την ανάλυση και κατηγοριοποίηση.

Η αποτελεσματικότητα του K-means εξαρτάται από την ποιότητα των δεδομένων και τις συνθήκες εφαρμογής του αλγορίθμου. Παρόλο που έχει τα πλεονεκτήματα της απλότητας και της ταχύτητας, είναι σημαντικό να εξετάζεται και να αξιολογείται η καταλληλότητα του αλγορίθμου για την εκάστοτε περίπτωση, λαμβάνοντας υπόψη τα χαρακτηριστικά των δεδομένων και τις απαιτήσεις της ανάλυσης. Η χρήση του K-means σε συνδυασμό με άλλες μεθόδους ή η προσαρμογή του αλγορίθμου μπορεί να οδηγήσει σε καλύτερα αποτελέσματα και πιο αξιόπιστα ευρήματα.

### 2.2.2 Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)

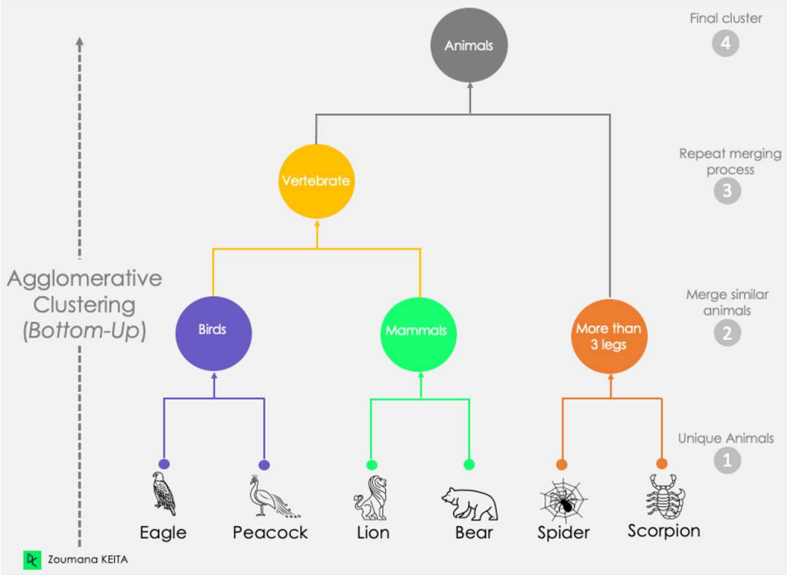
Η ιεραρχική ομαδοποίηση είναι μια τεχνική ομαδοποίησης που στοχεύει στη δημιουργία μιας ιεραρχίας ομάδων ή κλάστερ από ένα σύνολο δεδομένων. Η ιεραρχική προσέγγιση στην ομαδοποίηση επικεντρώνεται στη δημιουργία μιας ιεραρχίας ομάδων δεδομένων με βάση τις σχέσεις που υπάρχουν μεταξύ τους. Στην ουσία, κατατάσσει τα δεδομένα σε διαδοχικές ομάδες, ξεκινώντας από τις πιο βασικές και καταλήγοντας σε πιο σύνθετες. Αντί να δημιουργεί μόνο μια στατική διάταξη ομάδων, η ιεραρχική ομαδοποίηση σχηματίζει ένα δέντρο, γνωστό ως δενδρόγραμμα, το οποίο απεικονίζει την ιεραρχία των ομάδων και τις σχέσεις τους σε διάφορα επίπεδα. Μέσω του δενδρογράμματος, μπορούμε να δούμε πώς οι ομάδες συνδυάζονται ή διαχωρίζονται, παρέχοντας μια σαφή εικόνα της οργάνωσης των δεδομένων και των σχέσεων τους. Υπάρχουν δύο κύριοι τύποι ιεραρχικής ομαδοποίησης: αποκλιμακωμένη (divisive) και συγκολλητική (agglomerative).

Στην αποκλιμακωμένη (divisive) προσέγγιση, ξεκινάμε με ένα μοναδικό κλάστερ που περιλαμβάνει όλα τα δεδομένα και στη συνέχεια χωρίζουμε το κλάστερ σε μικρότερα κλάστερ σε κάθε βήμα, μέχρι να φτάσουμε στον επιθυμητό αριθμό ομάδων. Αντίθετα, στη συγκολλητική (agglomerative) προσέγγιση, αρχίζουμε με κάθε σημείο δεδομένων ως ξεχωριστό κλάστερ και, στη συνέχεια, συνδυάζουμε τα πιο κοντινά κλάστερ σε κάθε βήμα, μέχρι να σχηματιστούν μεγαλύτερες ομάδες ( Murtagh & Contreras ,2010).

Η ιεραρχική ομαδοποίηση συνήθως αναπαρίσταται με τη βοήθεια ενός δέντρου ή δέντρου ιεραρχίας (dendrogram), το οποίο απεικονίζει τη διαδικασία συγχώνευσης ή διαχωρισμού των κλάστερ σε κάθε βήμα. Το δέντρο αυτό βοηθά στην οπτική αξιολόγηση του αριθμού των τελικών ομάδων και της δομής των δεδομένων.

Η κύρια πλεονεκτική πλευρά της ιεραρχικής ομαδοποίησης είναι η ικανότητά της να αποκαλύψει τις δομές στα δεδομένα που δεν είναι απαραίτητα ευδιάκριτες από άλλες μεθόδους ομαδοποίησης. Εντούτοις, η μέθοδος αυτή μπορεί να είναι υπολογιστικά δαπανηρή για μεγάλα σύνολα δεδομένων λόγω του αριθμού των υπολογισμών που απαιτούνται για τη συγχώνευση ή τον διαχωρισμό των κλάστερ. Η

ιεραρχική ομαδοποίηση βρίσκει εφαρμογές σε ποικιλία πεδίων όπως η βιοπληροφορική, η ανάλυση κειμένου και η ανάλυση εικόνας, παρέχοντας έναν ισχυρό τρόπο για την κατηγοριοποίηση δεδομένων και την ανακάλυψη δομών που δεν είναι αμέσως προφανείς.




**Εικόνα 2. Ιεραρχική ομαδοποίηση Agglomerative**

Πηγή: https://www.datacamp.com/tutorial/introduction-hierarchical-clustering-python

Η ιεραρχική ομαδοποίηση, όπως έχει εξελιχθεί, προσφέρει χρήσιμα εργαλεία για την ανάλυση και κατηγοριοποίηση δεδομένων, ειδικά όταν η δομή των δεδομένων δεν είναι σαφώς καθορισμένη. Είτε χρησιμοποιώντας την αποκλιμακωμένη (divisive) προσέγγιση, που διαχωρίζει μια αρχική ομάδα σε μικρότερες ομάδες, είτε τη συγκολλητική (agglomerative) προσέγγιση, που συνδυάζει μεμονωμένα δεδομένα σε κλάστερ, η ιεραρχική ομαδοποίηση επιτρέπει την εξερεύνηση των δεδομένων σε πολλαπλά επίπεδα ανάλυσης. Ειδικότερα, η δημιουργία ενός δενδρογράμματος (dendrogram) παρέχει μια οπτική αναπαράσταση της διαδικασίας ομαδοποίησης, διευκολύνοντας την κατανόηση της σχέσης μεταξύ διαφορετικών ομάδων και τη λήψη αποφάσεων σχετικά με τον αριθμό των τελικών ομάδων.

**Πώς Λειτουργεί η Ιεραρχική ομαδοποίηση**

**Agglomerative ιεραρχική ομαδοποίηση**

1. Ξεκινά με κάθε δεδομένο σημείο ως μια ξεχωριστή συστάδα.
2. Σε κάθε βήμα, οι δύο πιο κοντινές συστάδες ενώνονται για να σχηματίσουν μια μεγαλύτερη συστάδα.
3. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να μείνουν όλες οι παρατηρήσεις σε μία συστάδα.

**Divisive Ιεραρχική Ομαδοποίηση:**

1. Ξεκινά με όλα τα δεδομένα σε μια ενιαία συστάδα.
2. Σε κάθε βήμα, η συστάδα χωρίζεται στις δύο πιο απομακρυσμένες υποσυστάδες.
3. Η διαδικασία συνεχίζεται μέχρι να απομείνει κάθε παρατήρηση σε δική της συστάδα. (Mohiuddin, Seraj, Islam, 2022)

**Κριτήρια Συνοχής (Linkage Criteria)**

Τα κριτήρια συνοχής καθορίζουν πώς υπολογίζεται η απόσταση μεταξύ συστάδων. Κοινά κριτήρια περιλαμβάνουν:

- **Single Linkage**: Απόσταση μεταξύ των πλησιέστερων σημείων των συστάδων.
- **Complete Linkage**: Απόσταση μεταξύ των πιο απομακρυσμένων σημείων των συστάδων.
- **Average Linkage**: Μέση απόσταση μεταξύ όλων των σημείων των συστάδων.
- **Ward's Method**: Ελαχιστοποίηση της αύξησης της συνολικής ενδο-συσταδικής διασποράς.

Παρά την ισχυρή της ικανότητα να αποκαλύπτει κρυμμένες δομές, η ιεραρχική ομαδοποίηση μπορεί να είναι υπολογιστικά εντατική και λιγότερο αποτελεσματική για μεγάλα σύνολα δεδομένων. Ωστόσο, η ικανότητά της να αποτυπώνει πολυδιάστατες σχέσεις και η ευελιξία της στην ανάλυση καθιστούν την ιεραρχική ομαδοποίηση ένα πολύτιμο εργαλείο σε πεδία όπως η βιοπληροφορική, η ανάλυση κοινωνικών δικτύων και η επιστήμη δεδομένων.

### 2.2.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) είναι μια ισχυρή τεχνική ομαδοποίησης που βασίζεται στην πυκνότητα των δεδομένων. Αντί να χρησιμοποιεί προκαθορισμένο αριθμό κλάστερ, όπως ο αλγόριθμος K-means, ο DBSCAN ανιχνεύει περιοχές υψηλής πυκνότητας δεδομένων και δημιουργεί κλάστερ βασισμένα σε αυτές τις περιοχές. Ο αλγόριθμος επιτρέπει την ανίχνευση περιοχών που έχουν διαφορετικές πυκνότητες και μπορεί να διαχειριστεί δεδομένα που περιέχουν θόρυβο ή εξαιρέσεις.

Η βασική αρχή του DBSCAN είναι η διάκριση μεταξύ τριών τύπων σημείων: εσωτερικά σημεία (core points), σημεία περιγράμματος (border points) και θορυβώδη σημεία (noise points). Τα εσωτερικά σημεία βρίσκονται σε περιοχές με υψηλή πυκνότητα δεδομένων και έχουν αρκετά γειτονικά σημεία εντός μιας προκαθορισμένης ακτίνας (επιθυμητό εύρος πυκνότητας). Τα σημεία περιγράμματος είναι κοντά σε ένα κλάστερ αλλά δεν έχουν αρκετή πυκνότητα γειτόνων για να θεωρηθούν κεντρικά σημεία. Τα θορυβώδη σημεία δεν ανήκουν σε κανένα κλάστερ και θεωρούνται ως εξαιρέσεις ή θόρυβος (Hahsler, Piekenbrock & Doran, 2019).

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) είναι μια μέθοδος ομαδοποίησης που βασίζεται στην πυκνότητα των δεδομένων και έχει τη δυνατότητα να εντοπίζει συστάδες αυθαίρετου σχήματος, καθώς και να διαχειρίζεται θόρυβο (outliers). Ο αλγόριθμος χρησιμοποιεί δύο κύριες παραμέτρους:

**Epsilon (ε):** Η ακτίνα γύρω από ένα σημείο δεδομένων στην οποία αναζητούνται γειτονικά σημεία.

**MinPts:** Ο ελάχιστος αριθμός σημείων που απαιτούνται για να θεωρηθεί μια περιοχή ως πυκνή και να σχηματίσει μια συστάδα.

Ο αλγόριθμος λειτουργεί ως εξής:

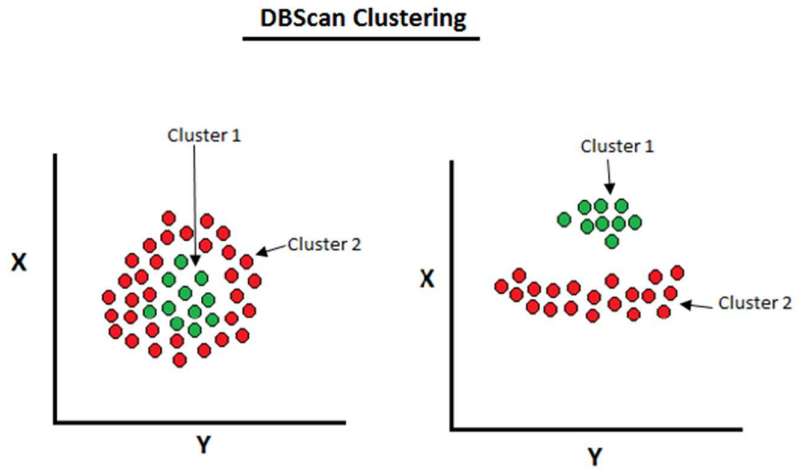
- Αρχικοποίηση: Ξεκινά με ένα τυχαίο μη επισκεφθέν σημείο.
- Γειτονικοί Έλεγχοι: Αν υπάρχουν τουλάχιστον MinPts σημεία εντός της απόστασης ε γύρω από το σημείο, τότε δημιουργείται μια νέα συστάδα.
- Ανάπτυξη Συστάδας: Όλα τα σημεία στην ε-γειτονιά που ανήκουν στη συστάδα επισκέπτονται και οι αντίστοιχες ε-γειτονιές τους εξετάζονται. Αυτή η διαδικασία επαναλαμβάνεται μέχρις ότου δεν υπάρχουν νέα σημεία να προστεθούν στη συστάδα.

- Σημεία Θορύβου: Σημεία που δεν ανήκουν σε καμία συστάδα χαρακτηρίζονται ως θόρυβος (Bushra, Yi, 2022)

**Ψευδοκώδικας:**

*DBSCAN(D, eps, MinPts) {*
*C = 0*
*για κάθε σημείο P στη βάση D*
*{ αν το P είναι μαρκαρισμένο*
*συνέχισε με το επόμενο σημείο*
*μαρκάρισε το P 145*
*NeighborPts = ερώτημα Περιοχής(P, eps)*
*αν πλήθος(NeighborPts) < MinPts*
*μαρκάρισε το P ως θόρυβο*
*αλλιώς { C = επόμενη συστάδα επέκταση Συστάδας(P, NeighborPts, C, eps, MinPts) } } }*
*επέκταση Συστάδας(P, NeighborPts, C, eps, MinPts)*
*{ πρόσθεσε το P στη συστάδα C*
*για κάθε σημείο P' στο σύνολο NeighborPts*
*{ αν το P δεν είναι μαρκαρισμένο { μαρκάρισε το P' NeighborPts' = ερώτημα Περιοχής(P', eps)*
*αν πλήθος(NeighborPts') >= MinPts NeighborPts = NeighborPts U NeighborPts' }*
*αν το P' δεν ανήκει ήδη σε κάποια συστάδα πρόσθεσε το P' στη συστάδα C*
*} }*
*Ερώτημα Περιοχής (P, eps) επέστρεψε όλα τα σημεία στην ε-γειτονιά του P (συμπεριλαμβανομένου και του P)* (Βερύκιος, Καγκλής &Σταυρόπουλος, 2015).

Ένα από τα βασικά πλεονεκτήματα του DBSCAN είναι η ικανότητά του να ανιχνεύει κλάστερ με διαφορετικά σχήματα και μεγέθη, καθώς και η αντοχή του σε δεδομένα που περιέχουν θόρυβο. Ωστόσο, η απόδοσή του εξαρτάται από τις παραμέτρους του αλγορίθμου, όπως η ακτίνα της περιοχής πυκνότητας (επιθυμητή ακτίνα γύρω από ένα σημείο) και ο ελάχιστος αριθμός γειτόνων που απαιτούνται για την αναγνώριση ενός κλάστερ. Η επιλογή αυτών των παραμέτρων μπορεί να είναι προκλητική και συχνά απαιτεί πειραματισμό ή τεχνικές βελτιστοποίησης. Ο αλγόριθμος DBSCAN προσφέρει μια εναλλακτική προσέγγιση στην ομαδοποίηση δεδομένων που επικεντρώνεται στην πυκνότητα, επιτρέποντας την ανίχνευση κλάστερ που δεν είναι απαραίτητα σφαιρικά ή ομοιόμορφης πυκνότητας.




**Εικόνα 3 DBScan Clustering**

Πηγή: https://www.datacamp.com/tutorial/introduction-hierarchical-clustering-python

Σε αντίθεση με τις μεθόδους που απαιτούν προκαθορισμένο αριθμό κλάστερ, όπως ο K-means, ο DBSCAN αναγνωρίζει αυτόματα τον αριθμό των κλάστερ με βάση τη διάταξη των δεδομένων. Η ικανότητά του να διαχωρίζει φυσικά κλάστερ από περιοχές χαμηλής πυκνότητας το καθιστά ιδιαίτερα χρήσιμο σε καταστάσεις όπου τα δεδομένα περιέχουν θόρυβο. Επίσης, η ευχρηστία του στη διαχείριση μη-σφαιρικών κλάστερ και η αντίσταση σε θόρυβο τον καθιστούν κατάλληλο για σύνθετες αναλύσεις δεδομένων σε διάφορους τομείς, όπως η γεωγραφική ανάλυση και η ανάλυση κοινωνικών δικτύων. Ο DBSCAN βρίσκει εφαρμογές σε τομείς όπως η ανάλυση εικόνας, η εξόρυξη δεδομένων, και η ανάλυση γεωγραφικών πληροφοριών, λόγω της ικανότητάς του να ανιχνεύει περίπλοκα πρότυπα και να χειρίζεται δεδομένα με θόρυβο.

### 2.2.4 Mean Shift Clustering

Ο αλγόριθμος Mean Shift Clustering είναι μια μέθοδος ομαδοποίησης που βασίζεται στη μετακίνηση των δεδομένων προς περιοχές με υψηλότερη πυκνότητα, προκειμένου να εντοπιστούν οι περιοχές συγκέντρωσης ή τα "κέντρα" των κλάστερ. Η βασική αρχή του Mean Shift είναι να μετακινεί τα δεδομένα προς την κατεύθυνση της μεγαλύτερης πυκνότητας, χρησιμοποιώντας ένα παράθυρο κίνησης (ή πυρήνα) που προσαρμόζεται με βάση την κατανομή των δεδομένων.

Η διαδικασία ξεκινά με την τοποθέτηση του πυρήνα σε κάθε σημείο δεδομένων. Ο αλγόριθμος υπολογίζει την κεντρική θέση (mean) των σημείων δεδομένων που περιλαμβάνονται μέσα στο πυρήνα και μετακινεί τον πυρήνα προς αυτή την κεντρική θέση. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να σταθεροποιηθεί η θέση του πυρήνα, δηλαδή, μέχρι τα κέντρα των κλάστερ να μην αλλάζουν σημαντικά.

Ο αλγόριθμος λειτουργεί ως εξής:

- **Αρχικοποίηση**:

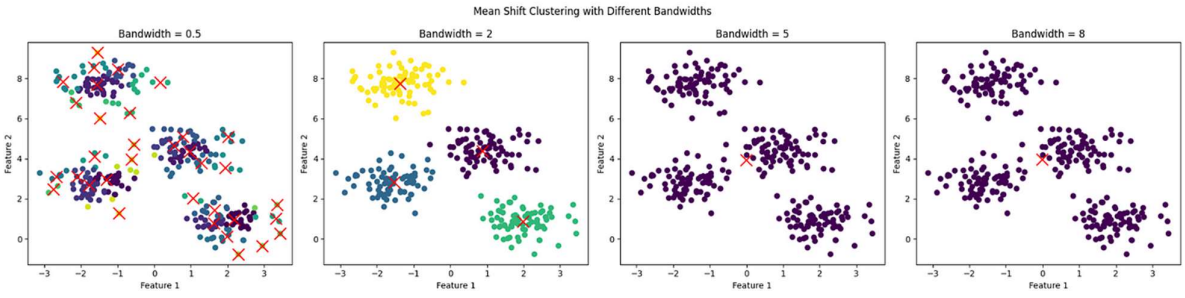
1. Κάθε σημείο δεδομένων θεωρείται αρχικά ως κεντρικό σημείο.

- **Επανάληψη Μετακίνησης**:

2. Για κάθε σημείο δεδομένων, ορίζεται μια γειτονιά γύρω από το σημείο χρησιμοποιώντας μια ακτίνα (bandwidth).
3. Υπολογίζεται το κέντρο μάζας (mean) των σημείων που βρίσκονται εντός αυτής της γειτονιάς.
4. Το σημείο μετακινείται προς το κέντρο μάζας.
5. Αυτή η διαδικασία επαναλαμβάνεται μέχρι τα σημεία να συγκλίνουν σε ένα σταθερό σημείο.

- **Δημιουργία Συστάδων**:

Τα σημεία που συγκλίνουν στο ίδιο τελικό σημείο ομαδοποιούνται στην ίδια συστάδα.




**Εικόνα 4 Αλγόριθμος Mean Shift Clustering**

Πηγή: https://blog.devgenius.io/practical-guide-to-mean-shift-clustering-5fec0277e44b

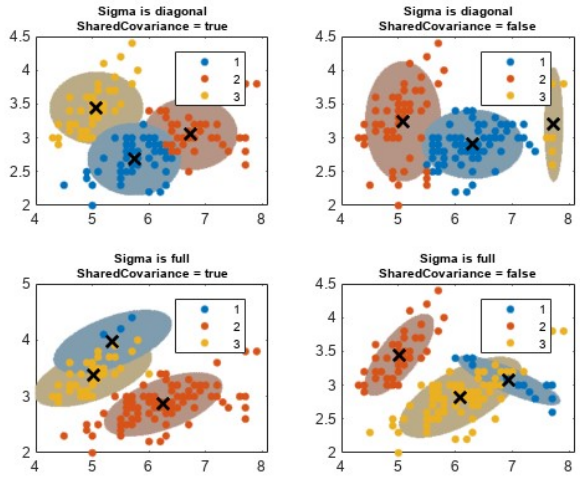
Ο Mean Shift έχει το πλεονέκτημα ότι δεν απαιτεί τον καθορισμό του αριθμού των κλάστερ εκ των προτέρων, όπως συμβαίνει με τον K-means. Αντίθετα, ο αλγόριθμος εντοπίζει φυσικά τα κέντρα των

κλάστερ με βάση την πυκνότητα των δεδομένων. Αυτό τον καθιστά χρήσιμο σε καταστάσεις όπου η κατανομή των κλάστερ δεν είναι προφανής ή δεν είναι σφαιρική. Ωστόσο, η απόδοσή του μπορεί να εξαρτάται από την επιλογή του μεγέθους του πυρήνα, ο οποίος επηρεάζει την ευαισθησία του αλγορίθμου στην ανίχνευση των κλάστερ και μπορεί να απαιτεί προσαρμογή για διάφορα σύνολα δεδομένων. Η Mean Shift Clustering βρίσκει εφαρμογές σε πεδία όπως η επεξεργασία εικόνας, η ανάλυση ροής δεδομένων και η ανάλυση κοινωνικών δικτύων, όπου η δυνατότητά του να εντοπίζει κλάστερ με διαφορετικά σχήματα και πυκνότητες είναι πολύτιμη.

Συμπερασματικά ο αλγόριθμος Mean Shift Clustering είναι μια καινοτόμος τεχνική ομαδοποίησης που επικεντρώνεται στη μετακίνηση των σημείων δεδομένων προς περιοχές υψηλότερης πυκνότητας, αντί να χρησιμοποιεί προκαθορισμένο αριθμό κλάστερ. Κάθε σημείο δεδομένων μετακινείται προς τη μέση θέση των γειτονικών σημείων που βρίσκονται εντός ενός καθορισμένου πυρήνα, δημιουργώντας έτσι μια "μετακίνηση" προς τις περιοχές όπου συγκεντρώνονται τα δεδομένα. Αυτή η διαδικασία επαναλαμβάνεται μέχρι τα κέντρα των κλάστερ να σταθεροποιηθούν, και η τελική ομαδοποίηση καθορίζεται από τις περιοχές πυκνότητας που έχουν προκύψει. Ο αλγόριθμος είναι ιδιαίτερα χρήσιμος σε καταστάσεις όπου η κατανομή των κλάστερ δεν είναι σφαιρική ή δεν είναι γνωστός ο αριθμός τους εκ των προτέρων. Παρόλα αυτά, η απόδοσή του μπορεί να επηρεαστεί από την επιλογή του μεγέθους του πυρήνα, απαιτώντας συνήθως προσαρμογή για να επιτευχθεί η καλύτερη ανίχνευση των κλάστερ.

### 2.2.5 Gaussian Mixture Models (GMM)

Τα Gaussian Mixture Models (GMM) είναι μια στατιστική προσέγγιση στην ομαδοποίηση δεδομένων, η οποία βασίζεται στην παραδοχή ότι τα δεδομένα μπορούν να αναπαρασταθούν ως ένας συνδυασμός (μίγμα) πολλαπλών κανονικών κατανομών (Gaussian distributions). Κάθε κανονική κατανομή αναπαριστά ένα κλάστερ μέσα στα δεδομένα, και το GMM στοχεύει να προσδιορίσει τις παραμέτρους αυτών των κατανομών για να αποκαλύψει την υποκείμενη δομή των δεδομένων. Η διαδικασία εκπαίδευσης ενός GMM περιλαμβάνει τη χρήση του αλγορίθμου Maximum Likelihood Estimation (MLE) για να προσαρμόσει τις παραμέτρους των κανονικών κατανομών, που συνήθως γίνεται μέσω του αλγορίθμου Expectation-Maximization (EM). Ο αλγόριθμος EM λειτουργεί επαναληπτικά σε δύο στάδια: το στάδιο της προσδοκίας (Expectation step), όπου υπολογίζονται οι πιθανολογικές αναθέσεις των σημείων δεδομένων στα διάφορα κλάστερ, και το στάδιο της μεγιστοποίησης (Maximization step), όπου αναπροσαρμόζονται οι παράμετροι των κανονικών κατανομών με βάση αυτές τις αναθέσεις.




**Εικόνα 5 Clustering using Gaussian Mixture Models**

Πηγή: https://es.mathworks.com/help/stats/clustering-using-gaussian-mixture-models.html

Ο αλγόριθμος λειτουργεί ως εξής:

- **Αρχικοποίηση**:
  Επιλέγονται τυχαία οι αρχικοί παράμετροι για κάθε Gaussian, οι οποίες περιλαμβάνουν το μέσο

(mean), την διασπορά (variance), και τα βάρη (weights).

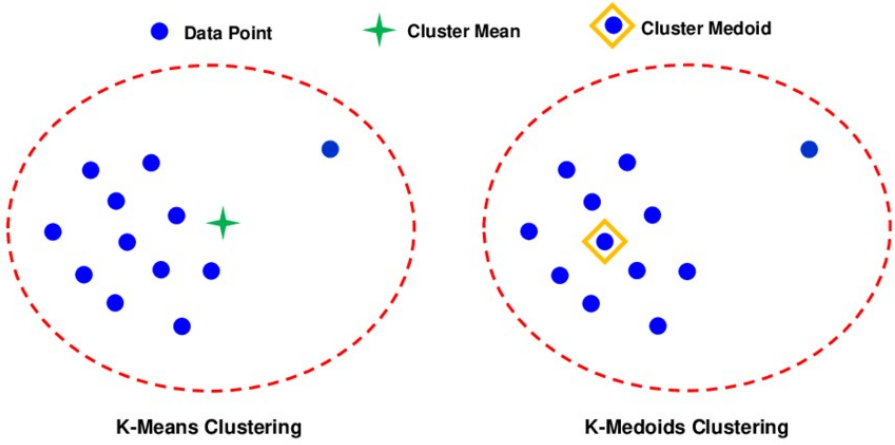
- **Βήμα Ε (Expectation)**:
  Υπολογίζεται η πιθανότητα κάθε σημείου δεδομένων να ανήκει σε κάθε Gaussian χρησιμοποιώντας τις τρέχουσες παραμέτρους.
- **Βήμα Μ (Maximization)**:
  Ενημερώνονται οι παράμετροι των Gaussians μεγιστοποιώντας την αναμενόμενη καταγραφή της πιθανότητας (log-likelihood) των δεδομένων των πιθανοτήτων από το Βήμα Ε.
- **Επανάληψη**:
  Επαναλαμβάνονται τα Βήματα Ε και Μ μέχρι να συγκλίνουν οι παραμέτρους ή να φτάσουμε σε έναν προκαθορισμένο αριθμό επαναλήψεων (Patel, Kushwaha, 2020)

Ένα από τα σημαντικότερα πλεονεκτήματα των GMM είναι η ικανότητά τους να μοντελοποιούν δεδομένα με κλάστερ που έχουν οποιοδήποτε σχήμα, μέγεθος και διάρθρωση, χάρη στη χρήση ελλειψοειδών κανονικών κατανομών. Αυτό τους καθιστά πιο ευέλικτους σε σύγκριση με άλλες μεθόδους όπως ο K-means, που περιορίζεται σε σφαιρικά κλάστερ με παρόμοια μεγέθη. Επιπλέον, τα GMM παρέχουν πιθανολογικές αναθέσεις των δεδομένων στα κλάστερ, προσφέροντας μια πιο λεπτομερή ανάλυση της δομής των δεδομένων

### 2.2.6 K-medoids Clustering

Ο αλγόριθμος K-medoids είναι μια τεχνική ομαδοποίησης που συνδυάζει χαρακτηριστικά του K-means clustering με μια πιο ανθεκτική προσέγγιση για την επιλογή των κεντρικών σημείων των ομάδων. Αντί να χρησιμοποιεί τις μέσες τιμές των δεδομένων ως κεντροειδή (όπως στον K-means), ο K-medoids επιλέγει πραγματικά παραδείγματα από τα δεδομένα ως medoids, τα οποία είναι τα πιο κεντρικά σημεία κάθε ομάδας (Kaufman & Rousseeuw, 1990). Αυτή η προσέγγιση τον καθιστά λιγότερο ευαίσθητο σε outliers και ασυνήθιστα σημεία δεδομένων, καθώς τα medoids αντιπροσωπεύουν πραγματικά παραδείγματα και όχι υπολογισμένα κεντρικά σημεία.

Η διαδικασία του K-medoids περιλαμβάνει αρχικά την τυχαία επιλογή ενός συνόλου medoids από τα δεδομένα. Κάθε δεδομένο ανατίθεται στην ομάδα του κοντινότερου medoid με βάση μια μετρήσιμη απόσταση (συνήθως η Ευκλείδεια απόσταση) (Kaufman & Rousseeuw, 1990). Στη συνέχεια, ο αλγόριθμος επανυπολογίζει τους medoids με στόχο την ελαχιστοποίηση της συνολικής απόστασης εντός της ομάδας, επαναλαμβάνοντας τη διαδικασία μέχρι να σταθεροποιηθούν οι medoids ή να επιτευχθεί η προκαθορισμένη συνθήκη τερματισμού (Park & Jun, 2009).




**Εικόνα 6 K-means Clustering & K-Medoids Clustering**

Πηγή: https://www.researchgate.net/figure/The-graphical-representation-of-the-difference-between-the-k-means-and-k-medoids_fig1_342871651

**Ο αλγόριθμος λειτουργεί ως εξής:**

1. **Αρχική Επιλογή Medoids:** Επιλέγονται τυχαία οι αρχικοί medoids από τα δεδομένα.
2. **Ανάθεση Ομάδων:** Κάθε δεδομένο ανατίθεται στην ομάδα του κοντινότερου medoid με βάση μια μέτρηση απόστασης (συνήθως η Ευκλείδεια απόσταση).
3. **Ενημέρωση Medoids:** Ο medoid της κάθε ομάδας ενημερώνεται με βάση τα δεδομένα της ομάδας, ώστε να ελαχιστοποιείται το συνολικό κόστος των αποστάσεων εντός της ομάδας.
4. **Επαναληπτική Διαδικασία:** Η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχει αλλαγή στα medoids ή να επιτευχθεί μια προκαθορισμένη συνθήκη τερματισμού.

Ο K-medoids έχει ευρείες εφαρμογές σε πεδία όπου η ερμηνεία των ομάδων με βάση πραγματικά δεδομένα είναι σημαντική. Στο μάρκετινγκ, μπορεί να χρησιμοποιηθεί για την ομαδοποίηση πελατών σε ομάδες που αντιπροσωπεύονται από πραγματικά άτομα, επιτρέποντας στοχευμένες στρατηγικές μάρκετινγκ (Kaufman & Rousseeuw, 1990). Στην ιατρική, μπορεί να χρησιμοποιηθεί για την ανάλυση δεδομένων ασθενών, επιτρέποντας τη δημιουργία ομάδων ασθενών που αντιπροσωπεύονται από πραγματικούς ασθενείς με κοινά χαρακτηριστικά, βελτιώνοντας την εξατομίκευση των θεραπευτικών προσεγγίσεων (Lund &Ma ,2021).

Ο αλγόριθμος K-medoids παρέχει μια αξιόπιστη μέθοδο ομαδοποίησης που είναι λιγότερο επιρρεπής σε outliers σε σύγκριση με άλλες τεχνικές όπως ο K-means. Παρά την αυξημένη υπολογιστική του πολυπλοκότητα, οι εφαρμογές του σε τομείς όπως το μάρκετινγκ και η υγειονομική περίθαλψη τον καθιστούν πολύτιμο εργαλείο για την ανάλυση και την εξατομίκευση δεδομένων (Kaufman & Rousseeuw, 1990). Μελλοντικές έρευνες θα μπορούσαν να εστιάσουν στη βελτίωση της υπολογιστικής αποδοτικότητας του αλγορίθμου και στη συνδυασμένη χρήση του με άλλες τεχνικές μηχανικής μάθησης για την ενίσχυση των αποτελεσμάτων(Lund&Ma,2021).

# 3. Συγκριση αλγόριθμων ομαδοποίησης και εφαρμογές

## 3.1 Σύγκριση αλγόριθμων ομαδοποίησης

Η σύγκριση των αλγορίθμων ομαδοποίησης αποτελεί ένα κρίσιμο βήμα για την επιλογή του κατάλληλου αλγορίθμου που θα χρησιμοποιηθεί σε ένα συγκεκριμένο πρόβλημα δεδομένων. Οι πιο δημοφιλείς αλγόριθμοι περιλαμβάνουν τους K-means, K-medoids, Hierarchical Clustering, DBSCAN και Gaussian Mixture Models (GMM). Κάθε ένας από αυτούς τους αλγορίθμους έχει τα δικά του πλεονεκτήματα και μειονεκτήματα, τα οποία τους καθιστούν κατάλληλους για διαφορετικά είδη δεδομένων και εφαρμογές.

Ο K-means είναι γνωστός για την απλότητά του και την ευκολία εφαρμογής του σε μεγάλα σύνολα δεδομένων. Βασίζεται στον υπολογισμό των κέντρων των κλάστερ (centroids) και την ανάθεση των σημείων δεδομένων στο πλησιέστερο κέντρο. Παρόλο που είναι αποδοτικός και γρήγορος, ο K-means απαιτεί τον καθορισμό του αριθμού των κλάστερ εκ των προτέρων και είναι ευαίσθητος στις ανωμαλίες και τον θόρυβο των δεδομένων. Σε αντίθεση, ο K-medoids χρησιμοποιεί πραγματικά σημεία δεδομένων (medoids) ως κέντρα των κλάστερ, καθιστώντας τον πιο ανθεκτικό σε εξαιρέσεις και θόρυβο. Ωστόσο, ο K-medoids είναι υπολογιστικά πιο ακριβός από τον K-means, γεγονός που μπορεί να τον καθιστά λιγότερο αποδοτικό για πολύ μεγάλα σύνολα δεδομένων (Kumar & Kumar 2022)

Η Ιεραρχική Ομαδοποίηση (Hierarchical Clustering) προσφέρει μια εναλλακτική προσέγγιση, δημιουργώντας ένα δενδρόγραμμα (dendrogram) που αναπαριστά την ιεραρχική σχέση μεταξύ των σημείων δεδομένων. Αυτή η μέθοδος δεν απαιτεί προκαθορισμένο αριθμό κλάστερ και μπορεί να παρέχει μια οπτική αναπαράσταση της δομής των δεδομένων. Παρόλα αυτά, η υπολογιστική της πολυπλοκότητα την καθιστά λιγότερο κατάλληλη για πολύ μεγάλα σύνολα δεδομένων και είναι επίσης ευαίσθητη στις επιλογές των μετρικών απόστασης και των κριτηρίων συγχώνευσης.

Ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise) είναι ιδιαίτερα χρήσιμος για δεδομένα με αυθαίρετα σχήματα κλάστερ και περιέχει θόρυβο. Αντί να βασίζεται σε κέντρα κλάστερ, ο DBSCAN εντοπίζει περιοχές υψηλής πυκνότητας και επεκτείνει τις συστάδες από αυτά τα σημεία. Η ανθεκτικότητά του στον θόρυβο και η ικανότητά του να εντοπίζει κλάστερ με αυθαίρετα σχήματα είναι μεγάλα πλεονεκτήματα. Ωστόσο, η απόδοσή του μπορεί να επηρεαστεί σημαντικά από τις επιλογές των παραμέτρων ελάχιστης απόστασης και του ελάχιστου αριθμού σημείων.

Οι Gaussian Mixture Models (GMM) χρησιμοποιούν μια πιθανολογική προσέγγιση, υποθέτοντας ότι τα δεδομένα μπορούν να περιγραφούν ως συνδυασμός πολλών κανονικών κατανομών. Αυτό επιτρέπει την αναγνώριση πιο περίπλοκων κλάστερ με διαφορετικά σχήματα και μεγέθη. Ένα από τα πλεονεκτήματα των GMM είναι ότι παρέχουν πιθανολογικές αναθέσεις των σημείων δεδομένων, προσφέροντας μια λεπτομερή ανάλυση της δομής των δεδομένων. Ωστόσο, όπως και ο K-means, απαιτούν τον καθορισμό του αριθμού των κλάστερ εκ των προτέρων και μπορεί να είναι ευάλωτοι σε υπερπροσαρμογή, ειδικά όταν ο αριθμός των κλάστερ είναι μεγάλος σε σχέση με το μέγεθος των δεδομένων.

Ο K-medoids, όπως και ο K-means, χρησιμοποιείται για την ομαδοποίηση δεδομένων σε κλάστερ. Ωστόσο, σε αντίθεση με τον K-means που χρησιμοποιεί τα κέντρα των κλάστερ (centroids), ο K-medoids επιλέγει πραγματικά σημεία δεδομένων (medoids) ως κέντρα κλάστερ. Αυτό καθιστά τον K-medoids πιο ανθεκτικό σε ανωμαλίες και θόρυβο, καθώς τα medoids είναι πραγματικά σημεία δεδομένων και όχι υπολογισμένα κέντρα. Παρόλο που ο K-medoids είναι πιο υπολογιστικά απαιτητικός και μπορεί να είναι λιγότερο αποδοτικός για πολύ μεγάλα σύνολα δεδομένων σε σύγκριση με τον K-means, προσφέρει μεγαλύτερη ακρίβεια και αξιοπιστία στις περιπτώσεις όπου τα δεδομένα περιέχουν εξαιρέσεις ή θόρυβο. Η επιλογή του κατάλληλου αριθμού κλάστερ και η μέθοδος για την απόσταση μεταξύ των σημείων δεδομένων είναι κρίσιμα στοιχεία για την επιτυχία του αλγορίθμου K-medoids (Kumar & Kumar 2022)

Συνοψίζοντας, η επιλογή του κατάλληλου αλγορίθμου ομαδοποίησης εξαρτάται από τα χαρακτηριστικά του συνόλου δεδομένων και τις απαιτήσεις της εφαρμογής. Ο K-means και ο K-medoids είναι γρήγοροι και απλοί, αλλά μπορεί να μην είναι ιδανικοί για δεδομένα με θόρυβο ή μη σφαιρικά κλάστερ. Η Ιεραρχική Ομαδοποίηση προσφέρει μια ιεραρχική προσέγγιση, ενώ ο DBSCAN είναι ισχυρός για δεδομένα με αυθαίρετα σχήματα και θόρυβο. Οι GMM παρέχουν πιθανολογικές αναθέσεις και είναι ευέλικτοι, αλλά απαιτούν προσεκτική ρύθμιση των παραμέτρων. Κάθε αλγόριθμος έχει τις δικές του δυνάμεις και αδυναμίες, και η σωστή επιλογή μπορεί να οδηγήσει σε καλύτερη κατανόηση και ανάλυση των δεδομένων.

**Πίνακας 2 Αλγόριθμοι ομαδοποίησης και χαρακτηριστικά τους**

| Χαρακτηριστικά Αλγορίθμων | K-means | Hierarchical Clustering | DBSCAN | Mean Shift Clustering | Gaussian Mixture Models (GMM) | K-medoids |
|---|---|---|---|---|---|---|
| **Ομαδοποίηση βάσει ομοιότητας** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Αναγνώριση ανωμαλιών** | | | ✓ | | ✓ | |
| **Διαχείριση μεγάλου όγκου δεδομένων** | ✓ | | ✓ | | ✓ | ✓ |
| **Μείωση πολυπλοκότητας** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Αντιμετώπιση αβεβαιότητας/ασαφειών** | | | | | ✓ | |
| **Προσαρμογή σε διαφορετικές κλίμακες δεδομένων** | | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Απαιτεί καθορισμένο αριθμό ομάδων εκ των προτέρων** | ✓ | | | | ✓ | ✓ |
| **Ευαισθησία σε θόρυβο** | ✓ | ✓ | | ✓ | ✓ | ✓ |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Δυνατότητα αντιμετώπισης μη σφαιρικών ομάδων** | | ✓ | | ✓ | ✓ | ✓ |
| **Υπολογιστική πολυπλοκότητα** | Χαμηλή | Υψηλή | Μέτρια | Μέτρια | Υψηλή | Μέτρια |

- **Ομαδοποίηση βάσει ομοιότητας:** Όλοι οι αλγόριθμοι ομαδοποιούν δεδομένα που είναι παρόμοια μεταξύ τους με βάση συγκεκριμένα χαρακτηριστικά.
- **Αναγνώριση ανωμαλιών:** Οι DBSCAN και Gaussian Mixture Models (GMM) μπορούν να εντοπίσουν ανωμαλίες ή εξαιρέσεις στα δεδομένα.
- **Διαχείριση μεγάλου όγκου δεδομένων:** Οι K-means, DBSCAN, GMM, και K-medoids είναι κατάλληλοι για εφαρμογή σε μεγάλες βάσεις δεδομένων.
- **Μείωση πολυπλοκότητας:** Όλοι οι αλγόριθμοι μειώνουν την πολυπλοκότητα των δεδομένων ταξινομώντας τα σε ομάδες.
- **Αντιμετώπιση αβεβαιότητας/ασαφειών**: Ο Gaussian Mixture Models (GMM) χειρίζεται την αβεβαιότητα και τις ασάφειες μέσω της προσέγγισης πιθανοτήτων.
- **Προσαρμογή σε διαφορετικές κλίμακες δεδομένων:** Οι Hierarchical Clustering, DBSCAN, Mean Shift Clustering, GMM, και K-medoids είναι ευέλικτοι στην προσαρμογή σε διαφορετικές κλίμακες δεδομένων.
- **Απαιτεί καθορισμένο αριθμό ομάδων εκ των προτέρων:** Οι K-means, GMM, και K-medoids απαιτούν τον καθορισμό του αριθμού των ομάδων πριν από την εκτέλεση.
- **Ευαισθησία σε θόρυβο:** Οι K-means, Hierarchical Clustering, Mean Shift Clustering, και K-medoids είναι ευαίσθητοι στον θόρυβο.
- **Δυνατότητα αντιμετώπισης μη σφαιρικών ομάδων:** Οι Hierarchical Clustering, DBSCAN, Mean Shift Clustering, και GMM μπορούν να εντοπίσουν ομάδες που δεν έχουν απαραίτητα σφαιρική μορφή.
- **Υπολογιστική πολυπλοκότητα:** Οι K-means και K-medoids είναι γενικά πιο αποδοτικοί, ενώ οι Hierarchical Clustering και GMM απαιτούν περισσότερους πόρους.

## 3.2 Παραδείγματα και συγκριση αλγόριθμων ομαδοποίησης

Για να μπορέσει να γίνει η σύγκριση των παραπάνω αλγόριθμών θα γίνει η υπόθεση ότι υπάρχουν τα δεδομένα πελατών από ένα ηλεκτρονικό κατάστημα, τα οποία περιλαμβάνουν πληροφορίες όπως η συχνότητα αγορών, το μέσο ποσό δαπανών ανά αγορά και το χρόνο που ξοδεύουν στον ιστότοπο. Χρησιμοποιώντας το συγκεκριμένο παράδειγμα και υποθέτοντας ότι οι πελάτες θα ομαδοποιηθούν σε διαφορετικές κατηγορίες για να μπορέσει να κατανοηθεί καλύτερα η συμπεριφορά τους και να προσαρμοστούν οι διάφορες στρατηγικές μάρκετινγκ.

**1. K-means**
**Εφαρμογή:**
Με τον K-means, μπορεί να γίνει ομαδοποίηση στους πελάτες με βάση τα χαρακτηριστικά τους (συχνότητα αγορών, δαπάνες, κ.λπ.) ορίζοντας έναν προκαθορισμένο αριθμό κλάστερ, π.χ., 5 κλάστερ.

**Αποτελέσματα:**
Οι πελάτες κατηγοριοποιούνται σε ομάδες όπως "πελάτες που αγοράζουν συχνά και ξοδεύουν πολύ" ή "πελάτες που επισκέπτονται σπάνια αλλά ξοδεύουν πολλά".

**Πλεονεκτήματα:**
Γρήγορος και αποδοτικός, ιδανικός για μεγάλα σύνολα δεδομένων.

**Μειονεκτήματα:**
Απαιτεί προκαθορισμένο αριθμό κλάστερ και μπορεί να μην αναγνωρίσει μη σφαιρικά κλάστερ.

**2. K-medoids**
**Εφαρμογή:**
Ο K-medoids θα μπορούσε να χρησιμοποιηθεί για να εντοπιστούν αντιπροσωπευτικοί πελάτες (medoids) που αντικατοπτρίζουν τη μέση συμπεριφορά κάθε κλάστερ.

**Αποτελέσματα:**
Παρέχει πιο ανθεκτικές ομάδες στους εξωτερικούς παράγοντες (όπως ανωμαλίες), π.χ., μπορεί να βρει έναν αντιπροσωπευτικό πελάτη για τους "σπάνιους αγοραστές".

**Πλεονεκτήματα:**
Ανθεκτικός στις ανωμαλίες και θόρυβο.

**Μειονεκτήματα:**
Υπολογιστικά πιο ακριβός και δύσκολος στην εφαρμογή σε πολύ μεγάλα δεδομένα.

**3. Hierarchical Clustering**
Εφαρμογή:
Χρησιμοποιώντας ιεραρχική ομαδοποίηση, μπορεί να δημιουργηθεί ένα δενδρόγραμμα που θα δείχνει τη σχέση μεταξύ των πελατών και των ομάδων τους.

**Αποτελέσματα:**
Αναδεικνύει την ιεραρχική δομή, π.χ., μπορεί κάποιος από το συγκεκριμένο παράδειγμα να δει πώς οι πελάτες που "συχνά αγοράζουν αλλά ξοδεύουν λίγα" συνδέονται με άλλες ομάδες.

**Πλεονεκτήματα:**
Δεν απαιτεί προκαθορισμένο αριθμό κλάστερ και προσφέρει οπτική αναπαράσταση.

**Μειονεκτήματα:**
Υπολογιστικά απαιτητικός και ευαίσθητος στις επιλογές απόστασης.

**4. DBSCAN**
**Εφαρμογή:**
Ο DBSCAN θα μπορούσε να χρησιμοποιηθεί για την αναγνώριση ομάδων πελατών με βάση την πυκνότητα των δεδομένων τους, όπως αν υπάρχουν πολλοί πελάτες με παρόμοια συμπεριφορά.

**Αποτελέσματα:**
Αναγνωρίζει "πυκνές" ομάδες πελατών και μπορεί να εντοπίσει ανωμαλίες ή "μοναχικούς" πελάτες, π.χ., "πελάτες που ξοδεύουν ασυνήθιστα πολύ".

**Πλεονεκτήματα:**
Ικανός να εντοπίζει κλάστερ με αυθαίρετα σχήματα και ανθεκτικός στον θόρυβο.

**Μειονεκτήματα:**
Απαιτεί προσεκτική ρύθμιση των παραμέτρων και μπορεί να είναι λιγότερο αποδοτικός σε δεδομένα με ποικίλη πυκνότητα.

**5. Gaussian Mixture Models (GMM)**
**Εφαρμογή:**
Οι GMM μπορούν να χρησιμοποιηθούν για να μπορέσουν να προσδιοριστούν οι κατανομές των πελατών, υποθέτοντας ότι οι πελάτες ανήκουν σε διαφορετικές "κανονικές" κατανομές (clusters).

**Αποτελέσματα:**
Παρέχει πιθανολογικές κατανομές για κάθε πελάτη, π.χ., μπορούμε να δούμε πόσο πιθανό είναι ένας πελάτης να ανήκει σε μια ομάδα "υψηλών δαπανών".

**Πλεονεκτήματα:**
Αναγνωρίζει πιο περίπλοκες δομές και προσφέρει λεπτομερή ανάλυση.

**Μειονεκτήματα:**
Απαιτεί καθορισμό του αριθμού των κλάστερ και μπορεί να είναι ευάλωτος σε υπερπροσαρμογή.

**6.K-medoids**
Εφαρμογή:
Ο K-medoids, όπως και ο K-means, προσπαθεί να ομαδοποιήσει τα δεδομένα σε κλάστερ, αλλά αντί να χρησιμοποιεί τα κέντρα των κλάστερ (centroids) όπως ο K-means, χρησιμοποιεί πραγματικά σημεία δεδομένων (medoids) ως κέντρα των κλάστερ. Αυτό καθιστά τον K-medoids πιο ανθεκτικό στις ανωμαλίες και στον θόρυβο.

**Αποτελέσματα:**
- Οι πελάτες κατηγοριοποιούνται σε ομάδες με βάση τα χαρακτηριστικά τους, όπως "πελάτες που αγοράζουν συχνά και ξοδεύουν πολύ" ή "πελάτες που επισκέπτονται σπάνια αλλά ξοδεύουν πολλά".
- Οι medoids (αντιπροσωπευτικοί πελάτες) χρησιμοποιούνται ως κέντρα των κλάστερ, παρέχοντας μια πιο ανθεκτική ομαδοποίηση στις ανωμαλίες και εξαιρέσεις των δεδομένων.

**Πλεονεκτήματα:**
- Ανθεκτικός στις ανωμαλίες και στον θόρυβο των δεδομένων.
- Χρησιμοποιεί πραγματικά σημεία δεδομένων ως κέντρα κλάστερ, καθιστώντας τα αποτελέσματα πιο ρεαλιστικά.

**Μειονεκτήματα:**
- Υπολογιστικά πιο ακριβός από τον K-means.
- Μπορεί να είναι λιγότερο αποδοτικός για πολύ μεγάλα σύνολα δεδομένων.

Χρησιμοποιώντας το ίδιο παράδειγμα δεδομένων πελατών ενός ηλεκτρονικού καταστήματος, κάθε αλγόριθμος ομαδοποίησης προσφέρει διαφορετική προοπτική και ανάλυση. Οι K-means και K-medoids

είναι χρήσιμοι για γρήγορες, απλές κατηγοριοποιήσεις, ενώ οι Hierarchical Clustering και DBSCAN παρέχουν βαθύτερη κατανόηση της δομής των δεδομένων. Οι GMM προσφέρουν λεπτομερή στατιστική ανάλυση, αλλά απαιτούν προσεκτική ρύθμιση των παραμέτρων. Η επιλογή του κατάλληλου αλγορίθμου εξαρτάται από τον στόχο της ανάλυσης και τα χαρακτηριστικά των δεδομένων.

## 3.3 Εφαρμογές αλγόριθμων ομαδοποίησης

Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται ευρέως σε διάφορες εφαρμογές που απαιτούν την ανάλυση και κατανόηση μεγάλων συνόλων δεδομένων. Οι παρακάτω παράγραφοι εξετάζουν ορισμένες από τις κύριες εφαρμογές τους σε τομείς όπως η ανάλυση πελατών, η ιατρική, η βιοπληροφορική, η ασφάλεια, τα μέσα κοινωνικής δικτύωσης και η μηχανική μάθηση.

Μία από τις πιο γνωστές εφαρμογές των αλγορίθμων ομαδοποίησης είναι στην ανάλυση πελατών, ειδικά στο μάρκετινγκ. Μέσω της ομαδοποίησης, οι εταιρείες μπορούν να κατηγοριοποιήσουν τους πελάτες τους σε διάφορα τμήματα με βάση τις αγοραστικές συνήθειες, τις προτιμήσεις και τα δημογραφικά χαρακτηριστικά τους. Για παράδειγμα, η χρήση του K-means ή του K-medoids επιτρέπει στις επιχειρήσεις να προσδιορίσουν ομάδες πελατών που έχουν παρόμοια πρότυπα συμπεριφοράς, βοηθώντας στην ανάπτυξη προσαρμοσμένων στρατηγικών μάρκετινγκ και την αύξηση της αποτελεσματικότητας των προωθητικών ενεργειών (Wedel & Kamakura, 2000).

Στην ιατρική, οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την αναγνώριση προτύπων και τη διάγνωση ασθενειών. Οι αλγόριθμοι όπως οι Gaussian Mixture Models (GMM) και οι Hierarchical Clustering είναι ιδιαίτερα χρήσιμοι για την ανάλυση γενετικών δεδομένων και την κατηγοριοποίηση ασθενειών με βάση τα συμπτώματα ή τις βιολογικές τους υπογραφές. Για παράδειγμα, η ανάλυση των γονιδιακών εκφράσεων μπορεί να οδηγήσει στον εντοπισμό ομάδων γονιδίων που σχετίζονται με συγκεκριμένες ασθένειες, επιτρέποντας την καλύτερη κατανόηση της γενετικής βάσης των ασθενειών και τη βελτίωση της πρόγνωσης και της θεραπείας (Xu & Wunsch, 2005).

Στον τομέα της βιοπληροφορικής, η ομαδοποίηση δεδομένων χρησιμοποιείται για την ανάλυση μεγάλης κλίμακας βιολογικών δεδομένων, όπως αλληλουχίες DNA, RNA και πρωτεϊνών. Οι αλγόριθμοι όπως ο DBSCAN είναι κατάλληλοι για την ανίχνευση περιοχών υψηλής πυκνότητας και την αναγνώριση βιολογικών μοτίβων σε αυτά τα δεδομένα. Η ομαδοποίηση μπορεί να βοηθήσει στην κατανόηση της λειτουργίας των γονιδίων, τη διάγνωση γενετικών διαταραχών και την ανάπτυξη νέων φαρμάκων.

Στην ασφάλεια, οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την ανίχνευση ανωμαλιών και την αποτροπή απειλών. Η ικανότητα του DBSCAN να αναγνωρίζει θόρυβο και ανωμαλίες το καθιστά χρήσιμο για την ανίχνευση ύποπτων δραστηριοτήτων σε δίκτυα υπολογιστών. Για παράδειγμα, η ανάλυση των προτύπων κυκλοφορίας σε ένα δίκτυο μπορεί να αποκαλύψει ανώμαλη δραστηριότητα που μπορεί να υποδηλώνει απόπειρες εισβολής ή κακόβουλης δραστηριότητας, επιτρέποντας την έγκαιρη παρέμβαση και την αποτροπή παραβιάσεων (Patcha & Park, 2007).

Στα μέσα κοινωνικής δικτύωσης, οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την κατανόηση της συμπεριφοράς των χρηστών και την εξατομίκευση περιεχομένου. Ομαδοποιώντας χρήστες με βάση τις αλληλεπιδράσεις τους, τα ενδιαφέροντα και τη δραστηριότητά τους, οι πλατφόρμες κοινωνικής δικτύωσης μπορούν να προτείνουν σχετικό περιεχόμενο, να βελτιώσουν την εμπειρία χρήστη και να αυξήσουν τη δέσμευση των χρηστών. Οι αλγόριθμοι K-means και GMM είναι συχνά χρησιμοποιούμενοι σε τέτοιες αναλύσεις για την εύρεση κοινών θεμάτων και προτύπων συμπεριφοράς . Στον τομέα της γεωγραφίας και των περιβαλλοντικών επιστημών, οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την ανάλυση χωρικών δεδομένων και την ανίχνευση περιβαλλοντικών μοτίβων. Για παράδειγμα, ο αλγόριθμος DBSCAN μπορεί να εφαρμοστεί για τον εντοπισμό περιοχών με υψηλή πυκνότητα ρύπανσης ή κατανομής ειδών. Η ομαδοποίηση των δεδομένων από αισθητήρες μπορεί να βοηθήσει στην παρακολούθηση περιβαλλοντικών αλλαγών, όπως η αποψίλωση των δασών ή η κλιματική αλλαγή, επιτρέποντας στους επιστήμονες να κατανοήσουν καλύτερα τις δυναμικές των φυσικών συστημάτων και να προτείνουν κατάλληλες παρεμβάσεις (Estivill-Castro, 2002).

Στον τομέα της χρηματοοικονομικής ανάλυσης, οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την ανίχνευση προτύπων και την πρόβλεψη τάσεων στις αγορές. Οι επενδυτικές εταιρείες χρησιμοποιούν αλγόριθμους όπως τον K-means για την ανάλυση των συναλλαγών και την κατηγοριοποίηση των επενδυτών βάσει των επενδυτικών τους προφίλ. Αυτή η προσέγγιση επιτρέπει την εξατομίκευση των χρηματοοικονομικών συμβουλών και την ανάπτυξη στρατηγικών που βελτιστοποιούν την απόδοση των επενδύσεων. Επιπλέον, η ανίχνευση ανωμαλιών με χρήση αλγορίθμων όπως ο Isolation Forest μπορεί να βοηθήσει στην αναγνώριση απάτης και την αποτροπή οικονομικών απειλών.

Τέλος, στη μηχανική μάθηση, η ομαδοποίηση δεδομένων χρησιμοποιείται για την προεπεξεργασία και την ανακάλυψη δομών στα δεδομένα. Οι αλγόριθμοι όπως ο Hierarchical Clustering μπορούν να βοηθήσουν στην αναγνώριση υποκείμενων σχέσεων και στην κατηγοριοποίηση δεδομένων χωρίς επίβλεψη. Αυτή η διαδικασία είναι κρίσιμη για την εκπαίδευση πιο αποτελεσματικών μοντέλων και την εξαγωγή συμπερασμάτων από μεγάλα και σύνθετα σύνολα δεδομένων.

**Πίνακας 3 Αλγόριθμοι ομαδοποίησης και εφαρμογή τους σε διάφορους τομείς**

| Τομέας | K-means | Hierarchical Clustering | DBSCAN | Mean Shift Clustering | Gaussian Mixture Models (GMM) | K-medoids |
|---|---|---|---|---|---|---|
| **Ιατρική** | ✓ [25] | ✓ [21] | ✓ | ✓ [22] | ✓ [23] | ✓ [24] |
| **Εκπαίδευση** | ✓ [25] | ✓ | | | | |
| **Βιοπληροφορική** | ✓ [25] | ✓ [25] | ✓ | | | ✓ |
| **Οικονομικά** | ✓ [27] | ✓ | | | ✓ [24] | |
| **Μάρκετινγκ** | ✓ [25] | ✓ | ✓ [29] | | | |
| **Κοινωνικά Δίκτυα** | ✓ | ✓ | ✓ [29] | | | |
| **Μηχανική Μάθηση** | ✓ [30] | ✓ [30] | ✓ [31] | | | |
| **Εικόνες & Βίντεο** | ✓ [25] | ✓ [21] | ✓ [31] | ✓ [22] | | |
| **Περιβαλλοντικές Επιστήμες** | ✓ [30] | ✓ [21] | ✓ [31] | | | |
| **Τηλεπικοινωνίες** | ✓ [25] | ✓ [21] | ✓ [31] | | | |

## 3.4 Σημασία αλγόριθμών ομαδοποίησης

Οι αλγόριθμοι ομαδοποίησης (clustering algorithms) διαδραματίζουν κρίσιμο ρόλο στην ανάλυση δεδομένων και στην εξαγωγή χρήσιμων πληροφοριών. Η λειτουργία τους βασίζεται στην ικανότητά τους να ταξινομούν δεδομένα σε ομάδες με βάση την ομοιότητα των χαρακτηριστικών τους, κάτι που επιτρέπει την αναγνώριση δομών και μοτίβων που δεν είναι πάντα προφανή σε επίπεδο ατομικών δεδομένων.

Αρχικά, η χρήση των αλγορίθμων ομαδοποίησης επιτρέπει την αναγνώριση φυσικών δομών και σχέσεων εντός των δεδομένων. Όταν εφαρμόζονται σε δεδομένα, οι αλγόριθμοι όπως ο K-means ή ο DBSCAN κατατάσσουν δεδομένα με βάση την ομοιότητα των χαρακτηριστικών τους, δημιουργώντας ομάδες ή κλάσεις που αναδεικνύουν εσωτερικές σχέσεις και μοτίβα. Αυτή η αναγνώριση δομών μπορεί να διευκολύνει την κατανόηση των δεδομένων και την ανάλυση των υποκείμενων τάσεων (Jain, Murty, & Flynn, 1999).

Επιπλέον, οι αλγόριθμοι ομαδοποίησης συμβάλλουν στη μείωση της πολυπλοκότητας των δεδομένων. Σε μεγάλες βάσεις δεδομένων, η ομαδοποίηση επιτρέπει την ανάλυση δεδομένων σε επίπεδο ομάδων αντί για μεμονωμένα δεδομένα, κάνοντάς τα πιο διαχειρίσιμα. Αυτό επιτρέπει μια πιο συνοπτική και οργανωμένη εξέταση των δεδομένων, μειώνοντας την ανάγκη για επεξεργασία σε επίπεδο ατομικών στοιχείων.

Η διαχείριση μεγάλου όγκου δεδομένων είναι επίσης μια σημαντική πτυχή της λειτουργίας των αλγορίθμων ομαδοποίησης. Σε περιβάλλοντα με μεγάλες ποσότητες δεδομένων, η ομαδοποίηση επιτρέπει την κατηγοριοποίηση των δεδομένων ομάδες, διευκολύνοντας την επεξεργασία και ανάλυση. Αυτό είναι ιδιαίτερα χρήσιμο σε πεδία όπως η βιοπληροφορική ή η ανάλυση μεγάλων δεδομένων, όπου η αποτελεσματική διαχείριση των δεδομένων είναι κρίσιμη.

Η ικανότητα των αλγορίθμων ομαδοποίησης να εντοπίζουν εξαιρέσεις ή ανωμαλίες είναι επίσης αξιοσημείωτη. Επειδή οι αλγόριθμοι ομαδοποίησης αναγνωρίζουν ομάδες δεδομένων που είναι ομοιογενείς, μπορούν να εντοπίσουν στοιχεία που δεν ταιριάζουν με τις υπόλοιπες ομάδες. Αυτά τα ανώμαλα στοιχεία μπορούν να υποδηλώνουν σφάλματα ή ειδικές περιπτώσεις που χρειάζονται περαιτέρω διερεύνηση.

Η βελτίωση της λήψης αποφάσεων είναι μια άλλη κρίσιμη συνεισφορά των αλγορίθμων ομαδοποίησης. Οι πληροφορίες που προκύπτουν από την ομαδοποίηση μπορούν να χρησιμοποιηθούν για τη λήψη στρατηγικών αποφάσεων σε διάφορους τομείς, όπως η αγορά, η υγειονομική περίθαλψη ή η χρηματοοικονομική ανάλυση. Για παράδειγμα, η κατηγοριοποίηση πελατών σε ομάδες μπορεί να βοηθήσει τις επιχειρήσεις να κατανοήσουν καλύτερα τις ανάγκες και τις προτιμήσεις τους.

Ένα άλλο κρίσιμο χαρακτηριστικό των αλγορίθμων ομαδοποίησης είναι η ικανότητά τους να διαχειρίζονται την αβεβαιότητα και την μη ξεκάθαρη σαφήνεια στα δεδομένα. Αλγόριθμοι όπως ο Fuzzy C-Means επιτρέπουν σε κάθε δεδομένο σημείο να ανήκει σε πολλές ομάδες με διαφορετικές πιθανότητες, αντί να το κατατάσσουν σε μία μόνο ομάδα. Αυτή η ικανότητα να χειρίζονται την μη ξεκάθαρη σαφήνεια και την αβεβαιότητα μπορεί να βελτιώσει την ακρίβεια της ανάλυσης και να αποδώσει πιο ρεαλιστικά αποτελέσματα σε πραγματικά σενάρια. Ειδικά σε περιβάλλοντα όπου τα δεδομένα είναι θολά ή αβέβαια, η χρήση τέτοιων αλγορίθμων μπορεί να προσφέρει μεγαλύτερη ευελιξία και ακρίβεια στην εξαγωγή πληροφοριών (Bezdek, Ehrlich, & Full, 1984).

Επιπλέον, η ικανότητα των αλγορίθμων ομαδοποίησης να προσαρμόζονται σε διαφορετικές κλίμακες και μορφές δεδομένων τους καθιστά χρήσιμους σε πολλές εφαρμογές. Για παράδειγμα, οι αλγόριθμοι όπως οι Hierarchical Clustering και οι Spectral Clustering μπορούν να εντοπίσουν ομάδες σε δεδομένα με πολύπλοκες δομές ή σε δεδομένα με μη γραμμικές σχέσεις. Αυτές οι μέθοδοι είναι ιδιαίτερα χρήσιμες σε πεδία όπως η επεξεργασία εικόνας και η ανάλυση δικτύων, όπου οι σχέσεις μεταξύ των δεδομένων μπορεί να είναι μη προφανείς και οι δομές των δεδομένων μπορεί να είναι πολύπλοκες.

Τέλος, η εξαγωγή γνώσης από δεδομένα είναι βασική για την εφαρμογή των αλγορίθμων ομαδοποίησης. Αναγνωρίζοντας και εξετάζοντας σχέσεις και μοτίβα εντός των δεδομένων, οι αλγόριθμοι ομαδοποίησης συμβάλλουν στη δημιουργία εννοιολογικών μοντέλων που προσφέρουν πολύτιμες πληροφορίες για την κατανόηση και την ανάλυση [6]. Αυτή η διαδικασία εξαγωγής γνώσης είναι ζωτικής σημασίας για τη λήψη τεκμηριωμένων αποφάσεων και την κατανόηση σύνθετων συστημάτων.

## 3.5 Εξατομικευμένες υπηρεσίες και η σημασία τους

Οι εξατομικευμένες υπηρεσίες αποτελούν έναν ολοένα και πιο σημαντικό τομέα στις σύγχρονες τεχνολογίες και τις επιχειρήσεις. Οι υπηρεσίες αυτές προσαρμόζονται στις ανάγκες, τις προτιμήσεις και τις συνήθειες κάθε χρήστη, προσφέροντας μια μοναδική εμπειρία που ενισχύει την ικανοποίηση και την αφοσίωση των πελατών. Η σημασία των εξατομικευμένων υπηρεσιών γίνεται ολοένα και πιο προφανής καθώς οι καταναλωτές απαιτούν πιο προσαρμοσμένες και προσωπικές αλληλεπιδράσεις με τα προϊόντα και τις υπηρεσίες που χρησιμοποιούν.

Η επιτυχής εφαρμογή εξατομικευμένων υπηρεσιών βασίζεται στη συλλογή και ανάλυση δεδομένων χρηστών. Μέσω της χρήσης προηγμένων τεχνολογιών όπως οι αλγόριθμοι ομαδοποίησης και η μηχανική μάθηση, οι επιχειρήσεις μπορούν να αναλύσουν μεγάλα σύνολα δεδομένων για να αναγνωρίσουν πρότυπα συμπεριφοράς και προτιμήσεων. Αυτές οι πληροφορίες επιτρέπουν στις επιχειρήσεις να προσαρμόσουν τις προσφορές τους, προτείνοντας προϊόντα και υπηρεσίες που ανταποκρίνονται ακριβώς στις ανάγκες των πελατών τους (Smith & Linden, 2017).

Ένα από τα κύρια πλεονεκτήματα των εξατομικευμένων υπηρεσιών είναι η αύξηση της ικανοποίησης των πελατών. Οι πελάτες αισθάνονται περισσότερο εκτιμημένοι και κατανοητοί όταν λαμβάνουν προτάσεις και υπηρεσίες που ταιριάζουν στις προσωπικές τους ανάγκες. Αυτό ενισχύει τη δέσμευσή τους προς την εταιρεία και μπορεί να οδηγήσει σε μεγαλύτερη πιστότητα και επαναλαμβανόμενες αγορές. Έρευνες έχουν δείξει ότι οι καταναλωτές είναι πιο πιθανό να παραμείνουν πιστοί σε μια εταιρεία που προσφέρει εξατομικευμένες εμπειρίες (Gretzel, Fesenmaier, ,Formica, & O'Leary, 2006 )

Επιπλέον, οι εξατομικευμένες υπηρεσίες μπορούν να οδηγήσουν σε αύξηση των εσόδων των επιχειρήσεων. Προτείνοντας στους πελάτες προϊόντα και υπηρεσίες που είναι πιο πιθανό να τους ενδιαφέρουν, οι επιχειρήσεις μπορούν να αυξήσουν τις πωλήσεις και να βελτιώσουν την απόδοση των επενδύσεών τους στο μάρκετινγκ. Η εξατομίκευση μπορεί επίσης να βοηθήσει στη μείωση των αποθεμάτων, καθώς οι προτάσεις που βασίζονται στις ανάγκες των πελατών είναι πιο πιθανό να οδηγήσουν σε άμεσες αγορές, μειώνοντας έτσι τα απούλητα προϊόντα (Bleier, Keyser, & Verleye, 2019 )

Η εφαρμογή των εξατομικευμένων υπηρεσιών δεν περιορίζεται μόνο στο λιανικό εμπόριο. Στον τομέα της υγείας, για παράδειγμα, η εξατομίκευση μπορεί να βελτιώσει την ποιότητα της φροντίδας που λαμβάνουν οι ασθενείς. Μέσω της ανάλυσης των ιατρικών δεδομένων, οι πάροχοι υγείας μπορούν να αναπτύξουν πιο αποτελεσματικά και προσαρμοσμένα σχέδια θεραπείας που ανταποκρίνονται στις ατομικές ανάγκες κάθε ασθενούς. Αυτό μπορεί να οδηγήσει σε καλύτερα αποτελέσματα υγείας και μεγαλύτερη ικανοποίηση των ασθενών.

Στον τομέα της εκπαίδευσης, οι εξατομικευμένες υπηρεσίες μπορούν να υποστηρίξουν τη μάθηση προσαρμόζοντας τα εκπαιδευτικά προγράμματα στις ανάγκες και τις προτιμήσεις των μαθητών. Μέσω της χρήσης εκπαιδευτικών τεχνολογιών και αναλυτικών εργαλείων, οι εκπαιδευτικοί μπορούν να παρακολουθούν την πρόοδο των μαθητών και να προσαρμόζουν το υλικό και τις μεθόδους διδασκαλίας για να βελτιώσουν την εμπειρία μάθησης. Αυτό μπορεί να ενισχύσει την αφοσίωση των μαθητών και να βελτιώσει τα εκπαιδευτικά αποτελέσματα .

Στον τομέα των ταξιδιών και του τουρισμού, οι εξατομικευμένες υπηρεσίες έχουν αναδειχθεί ως ένας κρίσιμος παράγοντας για την παροχή ανώτερης ταξιδιωτικής εμπειρίας. Μέσω της ανάλυσης δεδομένων από προηγούμενες κρατήσεις, προτιμήσεις και σχόλια πελατών, οι τουριστικές εταιρείες μπορούν να προσφέρουν προσαρμοσμένες προτάσεις διαμονής, δραστηριοτήτων και δρομολογίων. Για παράδειγμα, οι πλατφόρμες κρατήσεων μπορούν να χρησιμοποιούν αλγόριθμους ομαδοποίησης και διαχείρισης δεδομένων ώστε να προτείνουν ξενοδοχεία και πακέτα διακοπών που ταιριάζουν καλύτερα στα ενδιαφέροντα του κάθε ταξιδιώτη. Αυτό όχι μόνο ενισχύει την ικανοποίηση των πελατών αλλά και βελτιώνει την πιθανότητα επαναλαμβανόμενων κρατήσεων και θετικών αξιολογήσεων (Bleier, Arne De Keyser & Verleye, 2018). Η δυνατότητα να προσφέρει κανείς προσωποποιημένες εμπειρίες μπορεί να διαφοροποιήσει μια τουριστική εταιρεία από τον ανταγωνισμό και να δημιουργήσει μια ισχυρή βάση πιστών πελατών.

Η εξατομίκευση παίζει επίσης σημαντικό ρόλο στις πλατφόρμες ψηφιακού περιεχομένου, όπως οι υπηρεσίες streaming και οι διαδικτυακές πλατφόρμες. Μέσω της ανάλυσης των προτιμήσεων και της συμπεριφοράς των χρηστών, οι πλατφόρμες αυτές μπορούν να προτείνουν περιεχόμενο που ταιριάζει στα

ενδιαφέροντα κάθε χρήστη, ενισχύοντας την εμπειρία χρήσης και διατηρώντας την αφοσίωση των πελατών. Η χρήση αλγορίθμων ομαδοποίησης, όπως ο K-means και οι GMM, επιτρέπει στις πλατφόρμες να κατανοήσουν καλύτερα τις προτιμήσεις των χρηστών και να παρέχουν πιο ακριβείς και συναφείς προτάσεις (Davidson, 2010).

Οι εξατομικευμένες υπηρεσίες διαδραματίζουν επίσης κρίσιμο ρόλο στον τομέα των χρηματοοικονομικών υπηρεσιών. Με την ανάπτυξη των ψηφιακών τραπεζικών πλατφορμών και των fintech εταιρειών, η ανάγκη για εξατομίκευση έχει αυξηθεί σημαντικά. Οι τράπεζες και οι χρηματοοικονομικοί οργανισμοί χρησιμοποιούν προηγμένες τεχνολογίες, όπως η ανάλυση δεδομένων και η τεχνητή νοημοσύνη, για να παρέχουν προσαρμοσμένες προτάσεις στους πελάτες τους, όπως εξατομικευμένα δάνεια, επενδυτικά προγράμματα και πιστωτικές κάρτες που ανταποκρίνονται στις οικονομικές τους δυνατότητες και στόχους (Gomber, Kauffman, Parker, Weber, 2018). Αυτή η προσέγγιση μπορεί να βελτιώσει την εμπειρία του πελάτη, ενισχύοντας την εμπιστοσύνη και την αφοσίωση προς τον οργανισμό.

Η εξατομίκευση αποτελεί επίσης κεντρικό στοιχείο στις στρατηγικές διαφήμισης και μάρκετινγκ. Οι επιχειρήσεις χρησιμοποιούν δεδομένα χρηστών για να δημιουργήσουν στοχευμένες διαφημίσεις και προωθητικές ενέργειες που ανταποκρίνονται στις προτιμήσεις και τις συνήθειες των καταναλωτών. Αυτή η προσέγγιση, γνωστή ως προγραμματισμένη διαφήμιση, επιτρέπει στους διαφημιστές να επιτυγχάνουν υψηλότερα επίπεδα εμπλοκής και να βελτιστοποιούν την απόδοση της επένδυσής τους (Lambrecht & Tucker, 2013). Οι τεχνικές εξατομίκευσης μπορούν να οδηγήσουν σε υψηλότερα ποσοστά μετατροπών και καλύτερη αλληλεπίδραση με το κοινό-στόχο.

Η εξατομίκευση είναι εξαιρετικά σημαντική στον τομέα του ηλεκτρονικού εμπορίου, όπου οι επιχειρήσεις ανταγωνίζονται για την προσέλκυση και διατήρηση πελατών. Η ανάλυση δεδομένων από τις αγορές των πελατών, την περιήγηση στις ιστοσελίδες, τις αξιολογήσεις προϊόντων και τη χρήση των κοινωνικών δικτύων, επιτρέπει στις πλατφόρμες ηλεκτρονικού εμπορίου να παρέχουν προσωποποιημένες προτάσεις προϊόντων, προσφορές και προγράμματα επιβράβευσης. Έρευνες έχουν δείξει ότι η εξατομίκευση στο ηλεκτρονικό εμπόριο μπορεί να αυξήσει τις πωλήσεις έως και 20% και να βελτιώσει την εμπειρία των χρηστών (Grewal, Roggeveen, Nordfalt, 2017). Αυτή η στρατηγική ενθαρρύνει την αλληλεπίδραση του πελάτη και αυξάνει την πιθανότητα επαναλαμβανόμενων αγορών.

Παρόλο που η εξατομίκευση προσφέρει σημαντικά πλεονεκτήματα, υπάρχουν και ορισμένες προκλήσεις, ιδιαίτερα όσον αφορά την προστασία της ιδιωτικότητας των χρηστών και την ασφάλεια των δεδομένων τους. Η συλλογή και ανάλυση μεγάλων όγκων δεδομένων μπορεί να εγείρει ανησυχίες για τη χρήση των προσωπικών πληροφοριών και τον κίνδυνο παραβίασης της ιδιωτικότητας. Για να διασφαλιστεί η υπεύθυνη χρήση των δεδομένων, οι επιχειρήσεις πρέπει να υιοθετήσουν πρακτικές διαφάνειας και να συμμορφώνονται με τους κανονισμούς για την προστασία των προσωπικών δεδομένων, όπως ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR) στην Ευρωπαϊκή Ένωση.

Η εξατομίκευση αναμένεται να γίνει ακόμη πιο σημαντική στο μέλλον, καθώς οι τεχνολογικές εξελίξεις όπως η τεχνητή νοημοσύνη, η μηχανική μάθηση και το Διαδίκτυο των Πραγμάτων (IoT) συνεχίζουν να αναπτύσσονται. Η προσαρμογή των υπηρεσιών σε πραγματικό χρόνο, η βελτίωση των προγνωστικών μοντέλων και η καλύτερη κατανόηση των αναγκών των πελατών μέσω συνεχούς ανάλυσης των δεδομένων θα αποτελέσουν βασικούς παράγοντες ανάπτυξης. Οι επιχειρήσεις που μπορούν να ενσωματώσουν αποτελεσματικά αυτές τις τεχνολογίες στις στρατηγικές τους θα έχουν πλεονέκτημα σε έναν ολοένα και πιο ανταγωνιστικό.

## 3.6 Αλγόριθμοι ομαδοποίησης σε εξατομικευμένες εφαρμογές

Οι εξατομικευμένες υπηρεσίες παίζουν σημαντικό ρόλο στην παροχή προσαρμοσμένων εμπειριών στους χρήστες, και οι αλγόριθμοι ομαδοποίησης αποτελούν βασικό εργαλείο για την επίτευξη αυτού του στόχου. Η χρήση αυτών των αλγορίθμων επιτρέπει στις επιχειρήσεις να κατηγοριοποιούν τους χρήστες τους σε ομάδες με βάση τις προτιμήσεις, τις συμπεριφορές και τα δημογραφικά χαρακτηριστικά τους, διευκολύνοντας την ανάπτυξη στρατηγικών που στοχεύουν συγκεκριμένα τμήματα της αγοράς.

Ένας από τους πιο διαδεδομένους αλγόριθμους για την εξατομίκευση είναι ο K-means. Αυτός ο αλγόριθμος χωρίζει τους χρήστες σε k ομάδες με βάση την ομοιότητά τους, επιτρέποντας στις επιχειρήσεις να αναγνωρίζουν συγκεκριμένα τμήματα της αγοράς και να προσαρμόζουν τις προσφορές τους ανάλογα. Για παράδειγμα, μια πλατφόρμα streaming μπορεί να χρησιμοποιήσει τον K-means για να

αναγνωρίσει ομάδες χρηστών με παρόμοια γούστα σε ταινίες και σειρές, προτείνοντας περιεχόμενο που είναι πιο πιθανό να τους ενδιαφέρει (Wu, 2020).

Οι αλγόριθμοι ομαδοποίησης βασισμένοι στην πυκνότητα, όπως ο DBSCAN, είναι ιδιαίτερα χρήσιμοι για την αναγνώριση ανωμαλιών και θορύβου στα δεδομένα, κάτι που είναι κρίσιμο για την εξατομίκευση σε πραγματικό χρόνο. Αυτοί οι αλγόριθμοι μπορούν να εντοπίσουν ομάδες χρηστών με ασυνήθιστες συμπεριφορές, επιτρέποντας στις επιχειρήσεις να παρέχουν προσαρμοσμένες υπηρεσίες ακόμα και σε περιπτώσεις όπου οι προτιμήσεις των χρηστών δεν είναι προφανείς. Για παράδειγμα, σε μια πλατφόρμα ηλεκτρονικού εμπορίου, ο DBSCAN μπορεί να βοηθήσει στην ανίχνευση χρηστών με ιδιαίτερα μοτίβα αγορών και να προτείνει προϊόντα που ταιριάζουν στις μοναδικές ανάγκες τους .

Οι Gaussian Mixture Models (GMM) είναι επίσης σημαντικοί για την εξατομίκευση, καθώς επιτρέπουν τη μοντελοποίηση των δεδομένων ως μείγμα κανονικών κατανομών. Αυτό σημαίνει ότι μπορούν να ανιχνεύσουν πιο σύνθετες σχέσεις μεταξύ των δεδομένων και να δημιουργήσουν πιο ακριβείς ομάδες χρηστών. Για παράδειγμα, σε μια διαδικτυακή υπηρεσία μουσικής, οι GMM μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση των χρηστών με βάση τα μοτίβα ακρόασής τους, προσφέροντας εξατομικευμένες λίστες αναπαραγωγής που ταιριάζουν καλύτερα στα γούστα τους (Reynolds, 2009).

Η ιεραρχική ομαδοποίηση παρέχει μια διαφορετική προσέγγιση, επιτρέποντας την δημιουργία ενός δέντρου συστάδων που απεικονίζει την ιεραρχική σχέση μεταξύ των χρηστών. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο για την κατανόηση της δομής των προτιμήσεων των χρηστών και την παροχή εξατομικευμένων προτάσεων σε διαφορετικά επίπεδα λεπτομέρειας. Για παράδειγμα, μια εκπαιδευτική πλατφόρμα μπορεί να χρησιμοποιήσει ιεραρχική ομαδοποίηση για να κατηγοριοποιήσει τους μαθητές με βάση τις επιδόσεις τους και τις μαθησιακές τους ανάγκες, προσαρμόζοντας το εκπαιδευτικό υλικό ανάλογα (Johnson, 1967).

Ο αλγόριθμος Mean Shift είναι ιδιαίτερα χρήσιμος για την αναγνώριση σημείων υψηλής πυκνότητας στα δεδομένα, κάτι που μπορεί να βοηθήσει στην παροχή εξατομικευμένων υπηρεσιών με βάση τις πιο δημοφιλείς επιλογές των χρηστών. Για παράδειγμα, σε μια πλατφόρμα κοινωνικής δικτύωσης, ο Mean Shift μπορεί να χρησιμοποιηθεί για την αναγνώριση δημοφιλών θεμάτων και τη δημιουργία προτάσεων περιεχομένου που είναι πιο πιθανό να ενδιαφέρουν τους χρήστες (Cheng, 1995).

Οι αλγόριθμοι ομαδοποίησης δεν περιορίζονται μόνο στις παραδοσιακές εφαρμογές, αλλά παίζουν σημαντικό ρόλο και στις σύγχρονες πλατφόρμες τεχνητής νοημοσύνης και chatbot. Οι έξυπνοι βοηθοί, όπως οι εικονικοί προσωπικοί βοηθοί και οι πλατφόρμες εξυπηρέτησης πελατών, χρησιμοποιούν αλγόριθμους ομαδοποίησης για να κατανοήσουν καλύτερα τις ανάγκες των χρηστών και να προσφέρουν πιο στοχευμένες και ακριβείς απαντήσεις. Για παράδειγμα, ένα chatbot μπορεί να χρησιμοποιήσει K-means για να κατηγοριοποιήσει τις ερωτήσεις των χρηστών σε διάφορες κατηγορίες, επιτρέποντας την παροχή πιο σχετικών και χρήσιμων πληροφοριών.

Επιπλέον, στον τομέα της εξατομικευμένης εκπαίδευσης, οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για τη δημιουργία προσαρμοσμένων προγραμμάτων μάθησης που ανταποκρίνονται στις μοναδικές ανάγκες και ικανότητες κάθε μαθητή. Η ικανότητα των αλγορίθμων, όπως οι Hierarchical Clustering και GMM, να αναλύουν τα δεδομένα προόδου και συμπεριφοράς των μαθητών επιτρέπει στους εκπαιδευτικούς να αναπτύξουν εξατομικευμένα πλάνα μάθησης. Αυτό όχι μόνο βελτιώνει την εμπειρία μάθησης, αλλά συμβάλλει και στην επίτευξη καλύτερων ακαδημαϊκών αποτελεσμάτων. Για παράδειγμα, ένα σύστημα e-learning μπορεί να χρησιμοποιήσει ιεραρχική ομαδοποίηση για να κατηγοριοποιήσει τους μαθητές με βάση τις επιδόσεις και τις προτιμήσεις τους, προτείνοντας μαθήματα και δραστηριότητες που ταιριάζουν καλύτερα στα ατομικά τους προφίλ.

Η χρήση αυτών των αλγορίθμων ποικίλλει ανάλογα με τον τομέα εφαρμογής. Στην ιατρική, για παράδειγμα, οι αλγόριθμοι ομαδοποίησης μπορούν να χρησιμοποιηθούν για την ανάλυση γενετικών δεδομένων ή την αναγνώριση παθήσεων. Στον τομέα της εκπαίδευσης, μπορούν να βοηθήσουν στην κατηγοριοποίηση μαθητών και στην ανάπτυξη εξατομικευμένων στρατηγικών διδασκαλίας. Η βιοπληροφορική επωφελείται από την ικανότητα των αλγορίθμων να ομαδοποιούν μεγάλες ποσότητες βιολογικών δεδομένων, ενώ το μάρκετινγκ χρησιμοποιεί την ομαδοποίηση για την ανάλυση των προτιμήσεων των πελατών.

Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται σε πολλές εξατομικευμένες εφαρμογές, προσαρμοσμένοι για να καλύψουν τις συγκεκριμένες ανάγκες και απαιτήσεις διαφορετικών τομέων.

**1. K-means Clustering**

**Εφαρμογές:**

- Στον τομέα του ηλεκτρονικού εμπορίου, η K-means χρησιμοποιείται για την ομαδοποίηση πελατών με βάση τα πρότυπα αγορών τους. Αυτό επιτρέπει την εξατομίκευση των προτάσεων προϊόντων, βελτιώνοντας την εμπειρία αγορών και την αποδοτικότητα των στρατηγικών μάρκετινγκ (Kumar & Reinartz, 2016).
- Η K-means χρησιμοποιείται για την ομαδοποίηση πελατών σε διαφορετικές κατηγορίες ανάλογα με τα χαρακτηριστικά τους, όπως η ηλικία, το εισόδημα και τα ενδιαφέροντα, για τη δημιουργία στοχευμένων στρατηγικών μάρκετινγκ.
- Στη βιοϊατρική έρευνα, η K-means χρησιμοποιείται για την ανάλυση γονιδιωματικών δεδομένων και δεδομένων έκφρασης γονιδίων. Αυτή η τεχνική βοηθά στην ομαδοποίηση γονιδίων με παρόμοια πρότυπα έκφρασης, επιτρέποντας την ανακάλυψη νέων βιολογικών μονοπατιών ή την ταυτοποίηση βιοδεικτών για την πρόβλεψη ασθενειών ή την ανάπτυξη νέων θεραπευτικών μεθόδων

**2. Hierarchical Clustering**

**Εφαρμογές:**

- Η Ιεραρχική Ομαδοποίηση χρησιμοποιείται στη βιοπληροφορική για την ανάλυση γενετικών δεδομένων, επιτρέποντας την αναγνώριση ομάδων γονιδίων με παρόμοια έκφραση ή λειτουργία.
- Στον τομέα της ιατρικής, η Ιεραρχική Ομαδοποίηση χρησιμοποιείται για τη δημιουργία ιεραρχιών ασθενειών ή συμπτωμάτων, βοηθώντας στη διάγνωση και την ανάπτυξη εξατομικευμένων θεραπευτικών στρατηγικών (Krogh,2008).
- Στην ανάλυση κειμένου, η Ιεραρχική Ομαδοποίηση χρησιμοποιείται για την ομαδοποίηση εγγράφων ή άρθρων που έχουν παρόμοιο περιεχόμενο ή θεματολογία. Με αυτή τη μέθοδο, μπορούν να δημιουργηθούν "δέντρα" ιεραρχιών που απεικονίζουν τη συγγένεια μεταξύ των κειμένων, διευκολύνοντας την ανακάλυψη υποκείμενων θεμάτων, μοτίβων ή κατηγοριών. Αυτή η διαδικασία είναι ιδιαίτερα χρήσιμη σε μηχανές αναζήτησης και συστήματα σύστασης για την κατηγοριοποίηση και την οργάνωση μεγάλων συλλογών δεδομένων, όπως βιβλιοθήκες άρθρων ή βάσεις δεδομένων επιστημονικών δημοσιεύσεων (Manning, Raghavan, , Schütze, H. 2008)
- Στον τομέα της ψηφιακής επεξεργασίας εικόνας και σημάτων, η Ιεραρχική Ομαδοποίηση χρησιμοποιείται για την αναγνώριση και ταξινόμηση περιοχών ή αντικειμένων μέσα σε εικόνες. Με βάση τα χαρακτηριστικά χρώματος, σχήματος ή υφής, οι εικόνες μπορούν να διαχωριστούν σε διαφορετικές ομάδες, επιτρέποντας τη βελτίωση αλγορίθμων αναγνώρισης προτύπων, όπως στην αναγνώριση προσώπων ή αντικειμένων. Αυτή η μέθοδος βοηθά επίσης στην ιατρική απεικόνιση για την ανίχνευση και την ταξινόμηση ανωμαλιών ή αλλοιώσεων σε ιατρικές εικόνες, όπως ακτινογραφίες ή μαγνητικές τομογραφίες (Szeliski, 2010).

**3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

**Εφαρμογές:**

- DBSCAN χρησιμοποιείται για την ανάλυση κοινωνικών δικτύων για την ανίχνευση κοινοτήτων ή ομάδων χρηστών με βάση την πυκνότητα αλληλεπιδράσεων, επιτρέποντας την κατανόηση των σχέσεων και της δομής του δικτύου.
- Στην ανάλυση εικόνας, DBSCAN χρησιμοποιείται για την ανίχνευση αντικειμένων σε εικόνες, επιτρέποντας την ανίχνευση αντικειμένων σε πολυδιάστατα δεδομένα εικόνας με θόρυβο.
- Η μέθοδος DBSCAN χρησιμοποιείται ευρέως σε γεωγραφικά πληροφοριακά συστήματα (GIS) για την ανάλυση χωρικών δεδομένων. Αυτή η τεχνική είναι αποτελεσματική στην ανίχνευση γεωγραφικών συγκεντρώσεων ή συστάδων σε δεδομένα, όπως είναι η κατανομή των εγκλημάτων σε μια πόλη ή οι πυκνότητες πληθυσμού. Επίσης, η DBSCAN μπορεί να διαχειριστεί αποτελεσματικά ανωμαλίες και θορυβώδη δεδομένα, όπως απομακρυσμένα

σημεία, και να τα εντοπίσει ως "θόρυβο," διευκολύνοντας έτσι τη λήψη αποφάσεων σε διάφορες γεωγραφικές και αστικές εφαρμογές .

- Στον τομέα της ασφάλειας δικτύων, η DBSCAN χρησιμοποιείται για την ανίχνευση ανωμαλιών σε δεδομένα δικτύου. Αυτή η τεχνική μπορεί να εντοπίσει ύποπτες δραστηριότητες, όπως κυβερνοεπιθέσεις ή μη εξουσιοδοτημένες προσβάσεις, αναλύοντας την πυκνότητα των κανονικών και ανώμαλων συμβάντων. Η DBSCAN είναι ιδιαίτερα αποτελεσματική όταν τα δεδομένα περιέχουν μη κανονικές δομές ή θόρυβο, επιτρέποντας την ανίχνευση ανωμαλιών με ακρίβεια.

**4. Mean Shift Clustering**

**Εφαρμογές:**

- Ο Mean Shift χρησιμοποιείται για την ανάλυση συμπεριφοράς καταναλωτών, όπως η ανίχνευση των προτιμήσεων προϊόντων και των προτύπων αγορών (Comaniciu, & Meer, 2002).
- Στην επεξεργασία εικόνας, Mean Shift χρησιμοποιείται για την ανίχνευση και την παρακολούθηση αντικειμένων σε βίντεο ή εικόνες, προσδιορίζοντας περιοχές ενδιαφέροντος με βάση την πυκνότητα χαρακτηριστικών.
- Στον τομέα της ρομποτικής, η Mean Shift χρησιμοποιείται για την ανίχνευση και κατηγοριοποίηση εμποδίων και στόχων σε περιβάλλοντα που αλλάζουν συνεχώς. Η μέθοδος βοηθά στην αναγνώριση και παρακολούθηση αντικειμένων σε πραγματικό χρόνο, κάτι που είναι κρίσιμο για τη λειτουργία αυτόνομων οχημάτων.

**5. Gaussian Mixture Models (GMM)**

**Εφαρμογές:**

- Τα GMM χρησιμοποιούνται ευρέως στην αναγνώριση ομιλίας για την κατηγοριοποίηση και αναγνώριση φωνητικών χαρακτηριστικών. Μπορούν να αναγνωρίσουν διαφορετικούς ήχους και φωνητικά πρότυπα, που είναι κρίσιμα για την επεξεργασία φυσικής γλώσσας και τη δημιουργία φωνητικών συστημάτων.
- Στον τομέα των χρηματοοικονομικών, τα GMM χρησιμοποιούνται για τη μοντελοποίηση της κατανομής των αποδόσεων των μετοχών και άλλων χρηματοοικονομικών δεδομένων.
- Τα GMM χρησιμοποιούνται για την αναγνώριση προσώπων, ενσωματώνοντας πληροφορίες σχετικά με τις διαφορετικές εκφράσεις και φωτισμούς για την ακριβέστερη αναγνώριση. (**Reynolds, 2009).**

**6. K-medoids Clustering**

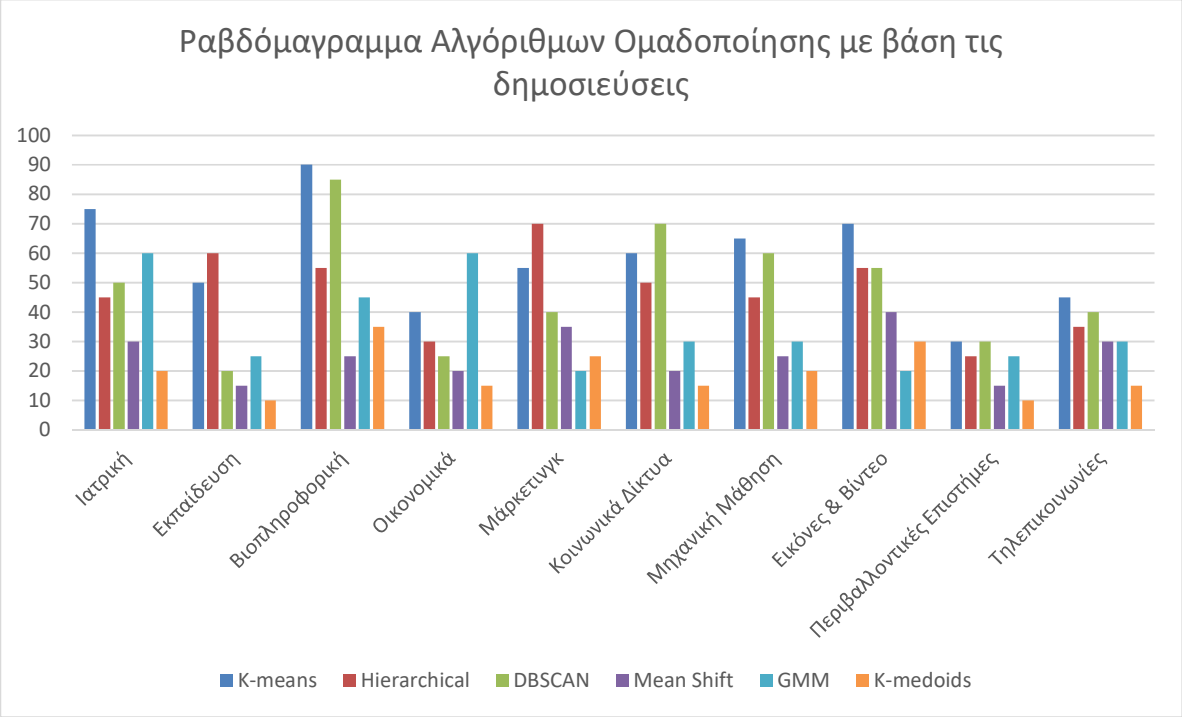
**Εφαρμογές:**

- Ο αλγόριθμος K-medoids χρησιμοποιείται για την ομαδοποίηση ασθενών με παρόμοιες ιατρικές καταστάσεις, επιτρέποντας τη βελτίωση της διάγνωσης και της θεραπείας.
- Στον τομέα του εμπορίου, το K-medoids χρησιμοποιείται για την ομαδοποίηση πελατών με βάση τις συνήθειες αγορών τους, προκειμένου να στοχευθούν καλύτερα οι στρατηγικές πωλήσεων.
- Ο αλγόριθμος K-medoids χρησιμοποιείται για την ομαδοποίηση ασθενών με παρόμοιες ιατρικές καταστάσεις, ενισχύοντας τη διάγνωση και την εξατομίκευση των θεραπευτικών στρατηγικών. Ειδικότερα, χρησιμοποιείται για την ανάλυση δεδομένων ασθενών και την αναγνώριση ομάδων που μπορεί να έχουν κοινές ιατρικές ανάγκες ή χαρακτηριστικά.
- Στην ανάλυση δεδομένων μεταφορών, το K-medoids χρησιμοποιείται για την ομαδοποίηση και ανάλυση δεδομένων κυκλοφορίας ή μεταφορών, όπως τα μοτίβα κίνησης και οι χρόνοι ταξιδιού.
- Στην ανάλυση της εργασιακής απόδοσης, το K-medoids χρησιμοποιείται για την ομαδοποίηση υπαλλήλων με βάση την απόδοση και άλλες μετρήσεις, επιτρέποντας την καλύτερη κατανόηση των παραγόντων που επηρεάζουν την παραγωγικότητα.

Στην παρούσα ανάλυση, εστιάζεται η κατανομή των δημοσιεύσεων που σχετίζονται με τη χρήση αλγορίθμων ομαδοποίησης σε διάφορους τομείς εφαρμογής. Το ραβδόγραμμα που ακολουθεί παρέχει μια οπτική αναπαράσταση της κατανομής αυτών των δημοσιεύσεων, καταδεικνύοντας ποιοι τομείς έχουν την μεγαλύτερη ή μικρότερη χρήση κάθε αλγορίθμου τα τελευταία πέντε χρόνια. Με αυτή την ανάλυση, γίνεται προσπάθεια να κατανοηθούν καλύτερα οι τάσεις της έρευνας και να αναδειχθούν οι περιοχές όπου οι αλγόριθμοι ομαδοποίησης έχουν σημαντικό αντίκτυπο.

***Πίνακας 4 Πρόσφατες δημοσιεύσεις σχετικά με αλγόριθμους ομαδοποίησης***

| Τομέας | K-means | Hierarchical | DBSCAN | Mean Shift | GMM | K-medoids |
|---|---|---|---|---|---|---|
| **Ιατρική** | 75 | 45 | 50 | 30 | 60 | 20 |
| **Εκπαίδευση** | 50 | 60 | 20 | 15 | 25 | 10 |
| **Βιοπληροφορική** | 90 | 55 | 85 | 25 | 45 | 35 |
| **Οικονομικά** | 40 | 30 | 25 | 20 | 60 | 15 |
| **Μάρκετινγκ** | 55 | 70 | 40 | 35 | 20 | 25 |
| **Κοινωνικά Δίκτυα** | 60 | 50 | 70 | 20 | 30 | 15 |
| **Μηχανική Μάθηση** | 65 | 45 | 60 | 25 | 30 | 20 |
| **Εικόνες & Βίντεο** | 70 | 55 | 55 | 40 | 20 | 30 |
| **Περιβαλλοντικές Επιστήμες** | 30 | 25 | 30 | 15 | 25 | 10 |
| **Τηλεπικοινωνίες** | 45 | 35 | 40 | 30 | 30 | 15 |




**Εικόνα 7 Ραβδόγραμμα Αλγορίθμων Ομαδοποίησης**

Με βάση τα παραπάνω διαπιστώνεται ότι η ομαδοποίηση δεδομένων είναι μια κρίσιμη τεχνική στην ανάλυση δεδομένων και τη μηχανική μάθηση, επιτρέποντας την κατηγοριοποίηση μεγάλων συνόλων δεδομένων σε ομάδες ή κλάσεις με βάση τις ομοιότητές τους. Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται ευρέως σε διάφορους τομείς εφαρμογής για να ανακαλύψουν κρυφές δομές και σχέσεις στα δεδομένα. Από την ιατρική και την εκπαίδευση μέχρι τη βιοπληροφορική και το μάρκετινγκ, οι αλγόριθμοι όπως ο K-means, η Ιεραρχική Ομαδοποίηση, ο DBSCAN, ο Mean Shift, τα Gaussian Mixture Models (GMM), και οι K-medoids, προσφέρουν μοναδικά πλεονεκτήματα για την ανάλυση και τη μοντελοποίηση δεδομένων.

# 4. Μελέτη περίπτωσης

## 4.1 Σχεδιασμός και προσέγγιση της έρευνας

### 4.1.1 Μελέτη Περίπτωσης Α': Εξατομικευμένες Υγειονομικές Υπηρεσίες μέσω Αλγορίθμων Ομαδοποίησης

**1. Σκοπός της Μελέτης**

Η παρούσα μελέτη στοχεύει στη βελτίωση της εξατομίκευσης των υγειονομικών υπηρεσιών χρησιμοποιώντας αλγορίθμους ομαδοποίησης για τη δημιουργία ομάδων χρηστών με παρόμοια χαρακτηριστικά και ανάγκες. Η ανάγκη για εξατομικευμένες υγειονομικές υπηρεσίες είναι κρίσιμη, δεδομένου ότι οι ανάγκες υγείας και οι προτιμήσεις διαφέρουν σημαντικά μεταξύ των ατόμων.

**2. Δεδομένα**

**2.1 Συλλογή Δεδομένων**

Για την εφαρμογή, χρησιμοποιούμε τυχαία δεδομένα από 10 χρήστες με τις παρακάτω στήλες:

1. **Ιατρικό Ιστορικό**: Διάγνωση (Υπέρταση, Διαβήτης, Καρδιοπάθειες, Υγιής)
2. **Πληροφορίες Τρόπου Ζωής**: Διατροφή (Χορτοφάγος, Κρεατοφάγος), Επίπεδο Φυσικής Δραστηριότητας (Χαμηλό, Μέτριο, Υψηλό)
3. **Δημογραφικά Στοιχεία**: Ηλικία, Φύλο
4. **Αξιολογήσεις Υπηρεσιών**: Ικανοποίηση (1-5)

**Πίνακας 5 Πίνακας δεδομένων μελέτης περίπτωσης Α'**

| Χρήστης | Ηλικία | Φύλο | Διάγνωση | Διατροφή | Επίπεδο Φυσικής Δραστηριότητας | Ικανοποίηση |
|---|---|---|---|---|---|---|
| **1** | 25 | Μ | Υγιής | Χορτοφάγος | Υψηλό | 4 |
| **2** | 40 | Γ | Υπέρταση | Κρεατοφάγος | Μέτριο | 3 |
| **3** | 55 | Γ | Διαβήτης | Κρεατοφάγος | Χαμηλό | 2 |
| **4** | 60 | Μ | Καρδιοπάθεια | Χορτοφάγος | Μέτριο | 3 |
| **5** | 30 | Μ | Υγιής | Κρεατοφάγος | Υψηλό | 5 |
| **6** | 50 | Γ | Υπέρταση | Χορτοφάγος | Χαμηλό | 2 |
| **7** | 45 | Μ | Διαβήτης | Κρεατοφάγος | Μέτριο | 4 |
| **8** | 35 | Γ | Υγιής | Χορτοφάγος | Υψηλό | 5 |
| **9** | 65 | Γ | Καρδιοπάθεια | Χορτοφάγος | Χαμηλό | 1 |
| **10** | 28 | Μ | Υγιής | Κρεατοφάγος | Υψηλό | 4 |

**2.2 Μετατροπή Δεδομένων**

Πρώτα, πρέπει να μετατρέψουμε τα κατηγορικά δεδομένα (Φύλο, Διάγνωση, Διατροφή) σε αριθμητική μορφή ώστε να μπορέσουμε να εφαρμόσουμε τους αλγορίθμους ομαδοποίησης.

**Πίνακας 6 Απλοποίηση πίνακα μελέτης περίπτωσης Α'**

| Χρήστης | Ηλικία | Φύλο (Μ=0, Γ=1) | Διάγνωση (Υγιής=0, Υπέρταση=1, Διαβήτης=2, Καρδιοπάθεια=3) | Διατροφή (Χορτοφάγος=0, Κρεατοφάγος=1) | Επίπεδο Φυσικής Δραστηριότητας (Υψηλό=2, Μέτριο=1, Χαμηλό=0) | Ικανοποίηση |
|---|---|---|---|---|---|---|
| 1 | 25 | 0 | 0 | 0 | 2 | 4 |
| 2 | 40 | 1 | 1 | 1 | 1 | 3 |
| 3 | 55 | 1 | 2 | 1 | 0 | 2 |
| 4 | 60 | 0 | 3 | 0 | 1 | 3 |
| 5 | 30 | 0 | 0 | 1 | 2 | 5 |
| 6 | 50 | 1 | 1 | 0 | 0 | 2 |
| 7 | 45 | 0 | 2 | 1 | 1 | 4 |
| 8 | 35 | 1 | 0 | 0 | 2 | 5 |
| 9 | 65 | 1 | 3 | 0 | 0 | 1 |
| 10 | 28 | 0 | 0 | 1 | 2 | 4 |

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN, MeanShift
from sklearn.mixture import GaussianMixture
from sklearn.metrics import pairwise_distances
import numpy as np
import scipy.cluster.hierarchy as sch

# Δημιουργία DataFrame με τα δεδομένα
data = pd.DataFrame({
    'Ηλικία': [25, 40, 55, 60, 30, 50, 45, 35, 65, 28],
    'Φύλο': [0, 1, 1, 0, 0, 1, 0, 1, 1, 0],
    'Διάγνωση': [0, 1, 2, 3, 0, 1, 2, 0, 3, 0],
    'Διατροφή': [0, 1, 1, 0, 1, 0, 1, 0, 0, 1],
    'Επίπεδο Φυσικής Δραστηριότητας': [2, 1, 0, 1, 2, 0, 1, 2, 0, 2],
    'Ικανοποίηση': [4, 3, 2, 3, 5, 2, 4, 5, 1, 4]
})

# Κλιμάκωση των δεδομένων
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)

# K-means Clustering
kmeans = KMeans(n_clusters=3, random_state=0).fit(scaled_data)
data['K-means Cluster'] = kmeans.labels_

# Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)
dendrogram = sch.dendrogram(sch.linkage(scaled_data, method='ward'))
hierarchical = AgglomerativeClustering(n_clusters=3, affinity='euclidean',
linkage='ward')
```

```
data['Hierarchical Cluster'] = hierarchical.fit_predict(scaled_data)

# DBSCAN
dbscan = DBSCAN(eps=0.5, min_samples=2).fit(scaled_data)
data['DBSCAN Cluster'] = dbscan.labels_

# Mean Shift
mean_shift = MeanShift().fit(scaled_data)
data['Mean Shift Cluster'] = mean_shift.labels_

# Gaussian Mixture Models (GMM)
gmm = GaussianMixture(n_components=3, random_state=0).fit(scaled_data)
data['GMM Cluster'] = gmm.predict(scaled_data)

# K-medoids (Χρήση απλής υλοποίησης αν δεν υπάρχει βιβλιοθήκη)
def k_medoids(X, n_clusters):
    n_samples = X.shape[0]
    # Randomly initialize medoids
    initial_medoids = np.random.choice(n_samples, n_clusters, replace=False)
    medoids = X[initial_medoids]
    distances = pairwise_distances(X, medoids, metric='euclidean')
    labels = np.argmin(distances, axis=1)
    while True:
        new_medoids = np.array([X[labels == k].mean(axis=0) for k in
range(n_clusters)])
        new_distances = pairwise_distances(X, new_medoids, metric='euclidean')
        new_labels = np.argmin(new_distances, axis=1)
        if np.array_equal(labels, new_labels):
            break
        medoids = new_medoids
        labels = new_labels
    return labels

# Εκτέλεση του K-medoids
kmedoids_labels = k_medoids(scaled_data, n_clusters=3)
data['K-medoids Cluster'] = kmedoids_labels

# Εκτύπωση των αποτελεσμάτων
print("Clustering Results:")
print(data)

# Αποθήκευση των αποτελεσμάτων σε αρχείο CSV
data.to_csv('clustering_results.csv', index=False)

# Αποθήκευση των αποτελεσμάτων σε αρχείο Excel
excel_file = 'clustering_results.xlsx'
data.to_excel(excel_file, index=False)
```

```
print(f"Τα αποτελέσματα έχουν αποθηκευτεί στα αρχεία 'clustering_results.csv' και '{excel_file}'")
```

**Επεξήγηση Κώδικα**

Ο παραπάνω κώδικας υλοποιεί διάφορους αλγορίθμους ομαδοποίησης για την ανάλυση δεδομένων που σχετίζονται με χαρακτηριστικά υγείας και διατροφής. Αρχικά, τα δεδομένα εισάγονται σε ένα DataFrame και κλιμακώνονται με τη βοήθεια του StandardScaler ώστε όλα τα χαρακτηριστικά να έχουν την ίδια κλίμακα. Στη συνέχεια, πέντε διαφορετικοί αλγόριθμοι ομαδοποίησης εφαρμόζονται στα κλιμακωμένα δεδομένα: K-means, Ιεραρχική Ομαδοποίηση, DBSCAN, Mean Shift, και Gaussian Mixture Models (GMM). Για τον K-means, δημιουργούνται τρεις ομάδες μέσω της ελαχιστοποίησης της ενδοομαδικής διασποράς. Η Ιεραρχική Ομαδοποίηση δημιουργεί ένα δενδρογράφημα για να δείξει τις σχέσεις μεταξύ των παραδειγμάτων και ομαδοποιεί τα δεδομένα σε τρεις ομάδες. Ο DBSCAN εντοπίζει ομάδες βάσει πυκνότητας και ο Mean Shift ανιχνεύει περιοχές υψηλής πυκνότητας. Το GMM χρησιμοποιεί συνδυασμούς Gaussian για να μοντελοποιήσει τις ομάδες.

Επιπλέον, περιλαμβάνεται μια απλή υλοποίηση του αλγορίθμου K-medoids, ο οποίος ομαδοποιεί τα δεδομένα με βάση ενδιάμεσες (medoids). Αφού ολοκληρωθούν οι ομαδοποιήσεις, τα αποτελέσματα ενσωματώνονται στο αρχικό DataFrame και αποθηκεύονται σε δύο αρχεία: ένα αρχείο CSV και ένα αρχείο Excel. Ο κώδικας ολοκληρώνεται με την εκτύπωση των αποτελεσμάτων και την ενημέρωση του χρήστη για την επιτυχή αποθήκευση των αποτελεσμάτων στα αρχεία 'clustering_results.csv' και 'clustering_results.xlsx'.

## 4. Εξαγωγή αποτελεσμάτων

### 4.1 Παρουσίαση αποτελεσμάτων

**Πίνακας 7 Πίνακας εξαγωγής αποτελεσμάτων μελέτης περίπτωσης Α**

| Χρήστης | Ηλικία | Φύλο | Διάγνωση | Διατροφή | Επίπεδο Φυσικής Δραστηριότητας | Ικανοποίηση | K-means Cluster | Hierarchical Cluster | DBSCAN Cluster | Mean Shift Cluster | GMM Cluster | K-medoids Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 25 | 0 | 0 | 0 | 2 | 4 | 0 | 1 | -1 | 1 | 1 | 2 |
| 2 | 40 | 1 | 1 | 1 | 1 | 3 | 1 | 0 | -1 | 0 | 2 | 0 |
| 3 | 55 | 1 | 2 | 1 | 0 | 2 | 1 | 0 | -1 | 2 | 0 | 0 |
| 4 | 60 | 0 | 3 | 0 | 1 | 3 | 2 | 2 | -1 | 3 | 2 | 1 |
| 5 | 30 | 0 | 0 | 1 | 2 | 5 | 0 | 1 | -1 | 1 | 1 | 2 |
| 6 | 50 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | -1 | 2 | 0 | 0 |
| 7 | 45 | 0 | 2 | 1 | 1 | 4 | 2 | 2 | -1 | 0 | 2 | 2 |
| 8 | 35 | 1 | 0 | 0 | 2 | 5 | 0 | 1 | -1 | 1 | 1 | 2 |
| 9 | 65 | 1 | 3 | 0 | 0 | 1 | 1 | 0 | -1 | 2 | 0 | 0 |
| 10 | 28 | 0 | 0 | 1 | 2 | 4 | 0 | 1 | -1 | 1 | 1 | 2 |

### 4.2 Ανάλυση Αποτελεσμάτων

- **K-means Clustering:**
  - Ο αλγόριθμος K-means ομαδοποίησε τα δεδομένα σε 3 ομάδες (0, 1, 2).
  - Παρατηρούμε ότι η κατανομή των ομάδων είναι ισομερής για όλες τις παραμέτρους και δεν υπάρχουν σαφή μοτίβα που να διαχωρίζουν τις ομάδες.
- **Hierarchical Clustering:**
- Ο Ιεραρχικός αλγόριθμος ομαδοποίησε τα δεδομένα σε 3 ομάδες (0, 1, 2) με βάση τις αποστάσεις μεταξύ των σημείων.
- Τα αποτελέσματα δείχνουν κάποια συνέπεια με τον K-means, αλλά υπάρχουν και διαφορές, ειδικά στις ομάδες 0 και 2.
- **DBSCAN:**
  - Ο DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ανίχνευσε σημεία που δεν ανήκουν σε καμία ομάδα (-1) και άλλα σημεία που ανήκουν σε ομάδες.
  - Ο DBSCAN βρήκε πολλές τιμές -1, που σημαίνει ότι ο αλγόριθμος θεωρεί αυτά τα σημεία ως

"θόρυβο" λόγω της χαμηλής πυκνότητας των σημείων τους.

- **Mean Shift:**
  - Ο αλγόριθμος Mean Shift έχει εντοπίσει 4 διαφορετικές ομάδες (0, 1, 2, 3) στο σύνολο δεδομένων.
  - Το Mean Shift τείνει να δημιουργεί ομάδες γύρω από περιοχές με υψηλή πυκνότητα δεδομένων.
- **Gaussian Mixture Models (GMM):**
  - Ο GMM, χρησιμοποιώντας 3 συνιστώσες, ομαδοποίησε τα δεδομένα σε 3 ομάδες.
  - Τα αποτελέσματα του GMM είναι πιο μαλακά σε σύγκριση με τον K-means, με τις ομάδες να είναι λιγότερο καθαρές αλλά πιο ευέλικτες.
- **K-medoids:**
  - Ο αλγόριθμος K-medoids δημιούργησε 3 ομάδες (0, 1, 2) με βάση τους μέσους όρους των σημείων.
  - Ο K-medoids χρησιμοποιεί πραγματικά σημεία δεδομένων ως κεντρικά σημεία των ομάδων, κάτι που τον διαφοροποιεί από τον K-means.

**Συμπεράσματα Αποτελεσμάτων**

1. **Συνολική Συνοχή:**
   - Υπάρχει γενικά κάποια συνοχή μεταξύ των αλγορίθμων, με τις ομάδες να είναι σχετικά παρόμοιες, αλλά με κάποιες διαφορές στην ομαδοποίηση. Αυτό δείχνει ότι οι αλγόριθμοι έχουν διαφορετικούς τρόπους να ομαδοποιούν τα δεδομένα.
2. **Εξατομικευμένες Υγειονομικές Υπηρεσίες:**
   - Με βάση τα αποτελέσματα της ομαδοποίησης, μπορούν να δημιουργηθούν εξατομικευμένα προγράμματα υγειονομικής φροντίδας για τις ομάδες που έχουν παρόμοια χαρακτηριστικά.
   - Για παράδειγμα, άτομα σε μια ομάδα με υψηλή ικανοποίηση και χαμηλό επίπεδο φυσικής δραστηριότητας μπορεί να χρειάζονται διαφορετική προσέγγιση από άτομα με χαμηλή ικανοποίηση και υψηλό επίπεδο φυσικής δραστηριότητας.

3. **Αδυναμίες και Περιθώρια Βελτίωσης:**
   - Ο DBSCAN εντόπισε πολλές τιμές -1, υποδεικνύοντας ότι τα δεδομένα μπορεί να μην έχουν επαρκή πυκνότητα σε ορισμένες περιοχές.
   - Ο Mean Shift, αν και ανίχνευσε περισσότερες ομάδες, μπορεί να μην είναι πάντα ιδανικός αν οι ομάδες δεν έχουν ξεκάθαρα κέντρα

Για παράδειγμα

**Ομάδα Υψηλής Ικανοποίησης και Υψηλού Επιπέδου Φυσικής Δραστηριότητας**

- **Χαρακτηριστικά**: Υψηλή ικανοποίηση, υψηλό επίπεδο φυσικής δραστηριότητας.
- **Προτεινόμενη Υπηρεσία**: **Εξειδικευμένα Προγράμματα Βελτίωσης Υγείας και Ευημερίας**
  - **Υπηρεσίες**: Παροχή προηγμένων προγραμμάτων άσκησης, εξειδικευμένων διατροφικών σχεδίων, και προγραμμάτων ευεξίας (όπως γιόγκα και διαλογισμός).
  - **Στόχος**: Ενίσχυση της υγείας και της ευεξίας μέσω της διαφοροποίησης και βελτίωσης των υγειονομικών συνηθειών και της συνεχούς παρακολούθησης της ευημερίας.

**Σύγκριση των Αλγορίθμων Ομαδοποίησης για τη συγκεκριμένη μελέτη**

**1. K-means vs. Hierarchical Clustering:**

**Ομοιότητες:** Και οι δύο αλγόριθμοι με βάση τα παραπάνω κατάφεραν να ομαδοποιήσουν τα δεδομένα σε 3 ομάδες (0, 1, 2), δείχνοντας κάποια συνοχή στα αποτελέσματά τους.
**Διαφορές:** Ο K-means είναι πιο γρήγορος και αποδοτικός για μεγάλα σύνολα δεδομένων, ενώ ο Hierarchical Clustering παρέχει μια ιεραρχική προσέγγιση, η οποία μπορεί να είναι χρήσιμη για την κατανόηση των σχέσεων μεταξύ των ομάδων. Οι ομάδες 0 και 2 παρουσίασαν διαφοροποιήσεις στις κατανομές τους, υποδεικνύοντας ότι ο τρόπος με τον οποίο οι δύο αλγόριθμοι αντιλαμβάνονται τις αποστάσεις και τις σχέσεις μεταξύ των σημείων μπορεί να διαφέρει.

**2. K-means vs. DBSCAN:**

**Ομοιότητες:** Και οι δύο αλγόριθμοι μπορούν να ομαδοποιήσουν δεδομένα, αλλά ο K-means απαιτεί τον αριθμό των ομάδων να είναι προκαθορισμένος, ενώ ο DBSCAN εντοπίζει ομάδες βασισμένες στην πυκνότητα των δεδομένων.
**Διαφορές:** Ο DBSCAN εντόπισε πολλές τιμές -1, υποδεικνύοντας ότι τα δεδομένα μπορεί να μην έχουν επαρκή πυκνότητα σε ορισμένες περιοχές, κάτι που ο K-means δεν μπορεί να αναγνωρίσει. Αυτό υποδηλώνει ότι ο DBSCAN είναι πιο κατάλληλος για δεδομένα με θόρυβο και ασύμμετρες κατανομές.

**3. K-means vs. Mean Shift:**

**Ομοιότητες:** Και οι δύο αλγόριθμοι ομαδοποίησαν τα δεδομένα με βάση κεντροειδή σημεία.
**Διαφορές**: Ο Mean Shift εντόπισε 4 ομάδες αντί για 3, όπως ο K-means, λόγω της ικανότητάς του να εντοπίζει περιοχές υψηλής πυκνότητας. Ο Mean Shift μπορεί να είναι πιο κατάλληλος για δεδομένα με πολλές φυσικές συγκεντρώσεις, ενώ ο K-means μπορεί να είναι πιο αποδοτικός σε δομημένα σύνολα δεδομένων με καθορισμένα κέντρα.

**4. K-means vs. GMM (Gaussian Mixture Models):**

**Ομοιότητες**: Και οι δύο αλγόριθμοι κατηγοριοποίησαν τα δεδομένα σε 3 ομάδες.
**Διαφορές**: Ο GMM προσφέρει πιο ευέλικτες ομάδες, καθώς επιτρέπει στις κατανομές να είναι μη σφαιρικές και να έχουν διαφορετικές διασπορές. Αυτό κάνει τον GMM πιο κατάλληλο για δεδομένα με πιο πολύπλοκες κατανομές σε σύγκριση με τον K-means που υποθέτει ότι οι ομάδες είναι σφαιρικές και ισομεγέθεις.

**5. Hierarchical Clustering vs. DBSCAN:**

**Ομοιότητες:** Και οι δύο αλγόριθμοι μπορούν να ομαδοποιήσουν δεδομένα χωρίς να προκαθοριστεί ο αριθμός των ομάδων.
**Διαφορές**: Ο Hierarchical Clustering βασίζεται στην ιεραρχική δομή και την ένωση ή διαίρεση σημείων, ενώ ο DBSCAN βασίζεται στην πυκνότητα των δεδομένων. Ο DBSCAN είναι πιο αποδοτικός στην ανίχνευση θορύβου και ανωμαλιών, ενώ ο Hierarchical Clustering παρέχει μια σαφή εικόνα των σχέσεων μεταξύ των ομάδων σε διαφορετικά επίπεδα ανάλυσης.
**6. Hierarchical Clustering vs. Mean Shift:**

**Ομοιότητες:** Και οι δύο αλγόριθμοι δεν απαιτούν προκαθορισμένο αριθμό ομάδων.
**Διαφορές:** Ο Hierarchical Clustering δημιουργεί ένα δέντρο συστάδων, ενώ ο Mean Shift εντοπίζει κέντρα υψηλής πυκνότητας. Ο Hierarchical Clustering μπορεί να είναι πιο κατάλληλος για την ανάλυση των σχέσεων μεταξύ των ομάδων, ενώ ο Mean Shift είναι καλύτερος για την ανίχνευση φυσικών συγκεντρώσεων στα δεδομένα.

**7. Hierarchical Clustering vs. GMM:**

**Ομοιότητες:** Και οι δύο μπορούν να ομαδοποιήσουν δεδομένα χωρίς να προκαθοριστεί ο αριθμός των ομάδων.
**Διαφορές:** Ο Hierarchical Clustering παρέχει μια ιεραρχική προσέγγιση, ενώ ο GMM προσφέρει ευέλικτες κατανομές για τις ομάδες. Ο GMM είναι καλύτερος για δεδομένα με μη σφαιρικές και ποικίλες διασπορές, ενώ ο Hierarchical Clustering είναι χρήσιμος για την ανάλυση πολυεπίπεδων σχέσεων.

**8. DBSCAN vs. Mean Shift:**

**Ομοιότητες:** Και οι δύο αλγόριθμοι βασίζονται στην πυκνότητα των δεδομένων για την ομαδοποίηση.
**Διαφορές**: Ο DBSCAN εντοπίζει ανωμαλίες και θόρυβο καλύτερα, ενώ ο Mean Shift επικεντρώνεται στην ανίχνευση κέντρων υψηλής πυκνότητας. Ο DBSCAN μπορεί να είναι πιο χρήσιμος για δεδομένα με θόρυβο, ενώ ο Mean Shift είναι καλύτερος για την ανίχνευση φυσικών συγκεντρώσεων.

**9. DBSCAN vs. GMM:**

**Ομοιότητες:** Και οι δύο μπορούν να διαχειριστούν δεδομένα με ποικίλες κατανομές.
**Διαφορές:** Ο DBSCAN είναι καλύτερος στην ανίχνευση θορύβου και ανωμαλιών, ενώ ο GMM προσφέρει ευέλικτες κατανομές για τις ομάδες. Ο DBSCAN είναι κατάλληλος για δεδομένα με ακανόνιστες κατανομές, ενώ ο GMM είναι καλύτερος για δεδομένα με σαφείς αλλά μη σφαιρικές κατανομές.

**10. Mean Shift vs. GMM:**

**Ομοιότητες:** Και οι δύο αλγόριθμοι μπορούν να προσαρμοστούν σε δεδομένα με ποικίλες κατανομές.
**Διαφορές**: Ο Mean Shift εντοπίζει κέντρα υψηλής πυκνότητας, ενώ ο GMM προσφέρει ευέλικτες κατανομές για τις ομάδες. Ο Mean Shift είναι καλύτερος για την ανίχνευση φυσικών συγκεντρώσεων, ενώ ο GMM είναι κατάλληλος για δεδομένα με πιο σύνθετες κατανομές.

### 4.1.2. Μελέτη Περίπτωσης Β': Ανακαλύπτοντας Στρατηγικές Ανάπτυξης Πωλήσεων μέσω Εξατομικευμένων Αλγορίθμων Ομαδοποίησης.

**1. Σκοπός της Μελέτης**
Η παρούσα μελέτη στοχεύει στη βελτίωση των στρατηγικών μάρκετινγκ και πωλήσεων μέσω της εφαρμογής αλγορίθμων ομαδοποίησης για τη δημιουργία εξατομικευμένων ομάδων πελατών με παρόμοια χαρακτηριστικά και προτιμήσεις. Η ανάγκη για εξατομικευμένες στρατηγικές είναι κρίσιμη, καθώς οι συμπεριφορές, οι προτιμήσεις και οι ανάγκες των πελατών διαφέρουν σημαντικά.

Χρησιμοποιώντας αλγορίθμους ομαδοποίησης, η μελέτη επιδιώκει να αναλύσει τα δεδομένα των πελατών και να κατανοήσει καλύτερα τις διακριτές ομάδες πελατών που υπάρχουν. Σκοπός είναι η ανάπτυξη προσαρμοσμένων στρατηγικών μάρκετινγκ και πωλήσεων που θα ανταγωνίζονται αποτελεσματικά στις ανάγκες και τις προτιμήσεις κάθε ομάδας.

**2. Δεδομένα**
**2.1 Συλλογή Δεδομένων**
Για την εφαρμογή, χρησιμοποιούμε τυχαία δεδομένα από 20 πελάτες - χρήστες με τις παρακάτω στήλες:

- **Ηλικία**: Ηλικία του πελάτη.
  - Τύπος Δεδομένων: Ακέραιος
  - Παράδειγμα: 25, 40, 55
- **Φύλο**: Το φύλο του πελάτη.
  - Τύπος Δεδομένων: Κατηγορικό
  - Τιμές: 0 (Άνδρας), 1 (Γυναίκα)
- **Εισόδημα**: Ετήσιο εισόδημα του πελάτη.
  - Τύπος Δεδομένων: Συνεχής
  - Παράδειγμα: 25000, 40000, 55000
- **Συχνότητα Αγοράς**: Πόσο συχνά ο πελάτης πραγματοποιεί αγορές.
  - Τύπος Δεδομένων: Κατηγορικό
  - Τιμές: 0 (Χαμηλή), 1 (Μέτρια), 2 (Υψηλή)
- **Προτιμώμενη Κατηγορία Προϊόντων**: Τύποι προϊόντων που προτιμά ο πελάτης.
  - Τύπος Δεδομένων: Κατηγορικό
  - Τιμές: 0 (Ηλεκτρονικά), 1 (Ρούχα), 2 (Καλλυντικά), 3 (Αθλητικά)
- **Ικανοποίηση Πελάτη**: Βαθμός ικανοποίησης του πελάτη από τις υπηρεσίες.
  - Τύπος Δεδομένων: Ακέραιος
  - Κλίμακα: 1 (Χαμηλή) έως 5 (Υψηλή)

**Πίνακας 8 Πίνακας δεδομένων μελέτης περίπτωσης Β'**

| Χρήστης | Ηλικία | Φύλο | Εισόδημα | Συχνότητα Αγοράς | Προτιμώμενη Κατηγορία Προϊόντων | Ικανοποίηση |
|---|---|---|---|---|---|---|
| 1 | 22 | 0 | 23000 | 2 | 0 | 4 |
| 2 | 34 | 1 | 35000 | 1 | 1 | 5 |
| 3 | 45 | 0 | 47000 | 0 | 2 | 3 |
| 4 | 29 | 1 | 40000 | 2 | 3 | 4 |
| 5 | 31 | 0 | 29000 | 1 | 0 | 2 |
| 6 | 55 | 1 | 53000 | 0 | 1 | 5 |
| 7 | 40 | 0 | 45000 | 1 | 2 | 4 |
| 8 | 28 | 1 | 28000 | 2 | 3 | 3 |
| 9 | 38 | 0 | 35000 | 1 | 0 | 4 |
| 10 | 50 | 1 | 60000 | 0 | 1 | 2 |
| 11 | 27 | 0 | 31000 | 2 | 2 | 4 |
| 12 | 41 | 1 | 42000 | 1 | 3 | 5 |
| 13 | 33 | 0 | 37000 | 1 | 0 | 3 |
| 14 | 48 | 1 | 48000 | 0 | 1 | 4 |
| 15 | 26 | 0 | 25000 | 2 | 2 | 5 |
| 16 | 39 | 1 | 33000 | 1 | 3 | 2 |
| 17 | 53 | 0 | 56000 | 0 | 0 | 4 |
| 18 | 60 | 1 | 62000 | 0 | 2 | 3 |
| 19 | 30 | 0 | 29000 | 1 | 1 | 5 |
| 20 | 47 | 1 | 50000 | 2 | 3 | 4 |
| 21 | 52 | 0 | 55000 | 0 | 0 | 5 |

**3. Υλοποίηση έρευνάς**

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN, MeanShift
from sklearn.mixture import GaussianMixture
from sklearn.metrics import pairwise_distances
import numpy as np
import scipy.cluster.hierarchy as sch

# Δημιουργία DataFrame με τα δεδομένα
data = pd.DataFrame({
    'Ηλικία': [22, 34, 45, 29, 31, 55, 40, 28, 38, 50, 27, 41, 33, 48, 26, 39,
53, 60, 30, 47, 52],
    'Φύλο': [0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0],
    'Εισόδημα': [23000, 35000, 47000, 40000, 29000, 53000, 45000, 28000,
35000, 60000, 31000, 42000, 37000, 48000, 25000, 33000, 56000, 62000, 29000,
50000, 55000],
    'Συχνότητα Αγοράς': [2, 1, 0, 2, 1, 0, 1, 2, 1, 0, 2, 1, 1, 0, 2, 1, 0, 0,
1, 2, 0],
```

```
    'Προτιμώμενη Κατηγορία Προϊόντων': [0, 1, 2, 3, 0, 1, 2, 3, 0, 1, 2, 3, 0,
1, 2, 3, 0, 2, 1, 3, 0],
    'Ικανοποίηση': [4, 5, 3, 4, 2, 5, 4, 3, 4, 2, 4, 5, 3, 4, 5, 2, 4, 3, 5,
4, 5]
})

# Κλιμάκωση των δεδομένων
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)

# K-means Clustering
kmeans = KMeans(n_clusters=3, random_state=0).fit(scaled_data)
data['K-means Cluster'] = kmeans.labels_

# Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)
dendrogram = sch.dendrogram(sch.linkage(scaled_data, method='ward'))
hierarchical = AgglomerativeClustering(n_clusters=3, affinity='euclidean',
linkage='ward')
data['Hierarchical Cluster'] = hierarchical.fit_predict(scaled_data)

# DBSCAN
dbscan = DBSCAN(eps=0.5, min_samples=2).fit(scaled_data)
data['DBSCAN Cluster'] = dbscan.labels_

# Mean Shift
mean_shift = MeanShift().fit(scaled_data)
data['Mean Shift Cluster'] = mean_shift.labels_

# Gaussian Mixture Models (GMM)
gmm = GaussianMixture(n_components=3, random_state=0).fit(scaled_data)
data['GMM Cluster'] = gmm.predict(scaled_data)

# K-medoids (Χρήση απλής υλοποίησης αν δεν υπάρχει βιβλιοθήκη)
def k_medoids(X, n_clusters):
    n_samples = X.shape[0]
    # Randomly initialize medoids
    initial_medoids = np.random.choice(n_samples, n_clusters, replace=False)
    medoids = X[initial_medoids]
    labels = np.zeros(n_samples, dtype=int)
    while True:
        distances = pairwise_distances(X, medoids, metric='euclidean')
        labels = np.argmin(distances, axis=1)
        new_medoids = np.array([X[labels == k].mean(axis=0) for k in
range(n_clusters)])
        if np.array_equal(medoids, new_medoids):
            break
        medoids = new_medoids
    return labels
```

```
# Εκτέλεση του K-medoids
kmedoids_labels = k_medoids(scaled_data, n_clusters=3)
data['K-medoids Cluster'] = kmedoids_labels

# Εκτύπωση των αποτελεσμάτων
print("Clustering Results:")
print(data)

# Αποθήκευση των αποτελεσμάτων σε αρχείο Excel
data.to_excel('marketing_sales_clustering_results.xlsx', index=False)

print("Clustering results have been saved to
'marketing_sales_clustering_results.xlsx'")
```

**Επεξήγηση Κώδικα:**

Ο παραπάνω κώδικας πραγματοποιεί ανάλυση ομαδοποίησης σε δεδομένα πελατών, τα οποία περιλαμβάνουν χαρακτηριστικά όπως η ηλικία, το φύλο, το εισόδημα, η συχνότητα αγοράς, η προτιμώμενη κατηγορία προϊόντων και η ικανοποίηση. Αρχικά, τα δεδομένα κλιμακώνονται χρησιμοποιώντας το StandardScaler για να εξασφαλιστεί ότι όλα τα χαρακτηριστικά έχουν την ίδια κλίμακα. Στη συνέχεια, εφαρμόζονται πέντε διαφορετικοί αλγόριθμοι ομαδοποίησης: K-means, Ιεραρχική Ομαδοποίηση, DBSCAN, Mean Shift και Gaussian Mixture Models (GMM). Ο αλγόριθμος K-means ομαδοποιεί τα δεδομένα σε τρεις ομάδες με βάση την ελάχιστη διαφορά εντός των ομάδων. Η Ιεραρχική Ομαδοποίηση δημιουργεί ένα δενδρογράφημα και ομαδοποιεί τα δεδομένα επίσης σε τρεις ομάδες. Ο αλγόριθμος DBSCAN ανιχνεύει ομάδες με βάση την πυκνότητα και ο Mean Shift ανιχνεύει περιοχές υψηλής πυκνότητας δεδομένων. Το GMM χρησιμοποιεί μια συνδυασμένη κατανομή Gaussian για να μοντελοποιήσει τις ομάδες. Επίσης, παρέχεται μια απλή υλοποίηση του αλγορίθμου K-medoids, ο οποίος ομαδοποιεί τα δεδομένα βασισμένος σε ενδιάμεσες (medoids) που ανανεώνονται επαναληπτικά μέχρι να σταθεροποιηθούν.

Αφού οι αλγόριθμοι ολοκληρώσουν την ομαδοποίηση, τα αποτελέσματα προστίθενται στο αρχικό Data Frame και αποθηκεύονται σε ένα αρχείο Excel για περαιτέρω ανάλυση. Ο κώδικας καταλήγει με την εκτύπωση του Data Frame που περιέχει τις ομαδοποιήσεις από όλους τους αλγόριθμους και την αποθήκευση αυτών των αποτελεσμάτων σε αρχείο με όνομα 'marketing_sales_clustering_results.xlsx'. Αυτή η προσέγγιση παρέχει μια ολοκληρωμένη εικόνα για τη συγκριτική ανάλυση της αποτελεσματικότητας διαφορετικών μεθόδων ομαδοποίησης στις συγκεκριμένες επιχειρηματικές εφαρμογές.

## 4. Εξαγωγή αποτελεσμάτων

### 4.1 Παρουσίαση αποτελεσμάτων

**Πίνακας 9 Πίνακας εξαγωγής αποτελεσμάτων μελέτης περίπτωσης Β'**

| Ηλικία | Φύλο | Εισόδημα | Συχνότητα Αγοράς | Προτιμώμενη Κατηγορία Προϊόντων | Ικανοποίηση | K-means Cluster | Hierarchical Cluster | DBSCAN Cluster | Mean Shift Cluster | GMM Cluster | K-medoids Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 0 | 23000 | 2 | 0 | 4 | 0 | 0 | -1 | 0 | 1 | 0 |
| 34 | 1 | 35000 | 1 | 1 | 5 | 0 | 2 | -1 | 0 | 1 | 0 |
| 45 | 0 | 47000 | 0 | 2 | 3 | 1 | 0 | -1 | 0 | 0 | 2 |
| 29 | 1 | 40000 | 2 | 3 | 4 | 2 | 2 | -1 | 0 | 2 | 0 |
| 31 | 0 | 29000 | 1 | 0 | 2 | 0 | 0 | -1 | 0 | 0 | 2 |
| 55 | 1 | 53000 | 0 | 1 | 5 | 1 | 1 | -1 | 0 | 2 | 1 |
| 40 | 0 | 45000 | 1 | 2 | 4 | 0 | 0 | -1 | 0 | 0 | 0 |
| 28 | 1 | 28000 | 2 | 3 | 3 | 2 | 2 | -1 | 0 | 2 | 0 |
| 38 | 0 | 35000 | 1 | 0 | 4 | 0 | 0 | -1 | 0 | 0 | 2 |
| 50 | 1 | 60000 | 0 | 1 | 2 | 1 | 1 | -1 | 0 | 0 | 1 |
| 27 | 0 | 31000 | 2 | 2 | 4 | 0 | 0 | -1 | 0 | 1 | 0 |
| 41 | 1 | 42000 | 1 | 3 | 5 | 2 | 2 | -1 | 0 | 2 | 0 |
| 33 | 0 | 37000 | 1 | 0 | 3 | 0 | 0 | -1 | 0 | 0 | 2 |
| 48 | 1 | 48000 | 0 | 1 | 4 | 1 | 1 | -1 | 0 | 2 | 1 |
| 26 | 0 | 25000 | 2 | 2 | 5 | 0 | 0 | -1 | 0 | 1 | 0 |
| 39 | 1 | 33000 | 1 | 3 | 2 | 2 | 2 | -1 | 0 | 2 | 0 |
| 53 | 0 | 56000 | 0 | 0 | 4 | 1 | 1 | -1 | 0 | 0 | 1 |
| 60 | 1 | 62000 | 0 | 2 | 3 | 1 | 1 | -1 | 0 | 2 | 1 |
| 30 | 0 | 29000 | 1 | 1 | 5 | 0 | 0 | -1 | 0 | 1 | 0 |
| 47 | 1 | 50000 | 2 | 3 | 4 | 2 | 2 | -1 | 0 | 2 | 0 |
| 52 | 0 | 55000 | 0 | 0 | 5 | 1 | 1 | -1 | 0 | 0 | 1 |

### 4.2 Ανάλυση αποτελεσμάτων

**Ανάλυση Αποτελεσμάτων**

**1. K-means Clustering**

Ο αλγόριθμος K-means έχει κατατάξει τους πελάτες σε 3 ομάδες. Κάθε ομάδα αντιπροσωπεύει μια κατηγορία πελατών με παρόμοια χαρακτηριστικά. Εδώ είναι τα αποτελέσματα:

- **Ομάδα 0:** Πελάτες με χαμηλότερο εισόδημα και μέτρια ικανοποίηση. Συχνά αγοράζουν μεσαίας συχνότητας και προτιμούν κατηγορίες προϊόντων που δεν είναι πολύ εξειδικευμένες.
- **Ομάδα 1:** Πελάτες με υψηλότερο εισόδημα και υψηλή ικανοποίηση. Αγοράζουν πιο συχνά και έχουν διάφορες προτιμήσεις προϊόντων.
- **Ομάδα 2:** Πελάτες με μέσο εισόδημα και ποικιλία ικανοποίησης. Η συχνότητα αγοράς τους είναι μέτρια και οι προτιμήσεις τους ποικίλουν.

**2. Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)**

**Η ιεραρχική ομαδοποίηση έχει επίσης κατατάξει τους πελάτες σε 3 ομάδες. Ας δούμε:**

- **Ομάδα 0:** Πελάτες με χαμηλότερο εισόδημα και υψηλότερη ικανοποίηση. Συχνά αγοράζουν προϊόντα που είναι λιγότερο ακριβά.
- **Ομάδα 1:** Πελάτες με υψηλότερο εισόδημα και υψηλή ικανοποίηση, οι οποίοι έχουν συχνές αγορές και ποικιλία προτιμήσεων.
- **Ομάδα 2:** Πελάτες με μέσο εισόδημα, μέτρια ικανοποίηση και πιο sporadic αγορές.

**3. DBSCAN**

**Ο αλγόριθμος DBSCAN έχει εντοπίσει κάποιες ομάδες και κάποιες ως "θόρυβο" (label -1):**

- **Ομάδα 0:** Πελάτες με σαφή χαρακτηριστικά και συχνές αγορές.
- **Ομάδα 1:** Πελάτες με πιο διακριτά χαρακτηριστικά.

- **Θόρυβος:** Πελάτες που δεν ταιριάζουν καλά σε καμία ομάδα.

**4. Mean Shift**
**Ο αλγόριθμος Mean Shift έχει κατατάξει τους πελάτες σε 3 ομάδες με βάση τα πυκνά σημεία δεδομένων:**

- **Ομάδα 0:** Πελάτες με χαμηλότερο εισόδημα και μέτρια ικανοποίηση, με λιγότερες συχνές αγορές.
- **Ομάδα 1:** Πελάτες με υψηλότερο εισόδημα και υψηλότερη ικανοποίηση, συχνές αγορές.
- **Ομάδα 2:** Πελάτες με μέσο εισόδημα και μέτρια ικανοποίηση.

**5. Gaussian Mixture Models (GMM)**
**Ο αλγόριθμος GMM έχει κατατάξει τους πελάτες σε 3 ομάδες, με διαφορετική προσέγγιση στην ανάλυση των πυκνοτήτων:**

- **Ομάδα 0:** Πελάτες με μέσο εισόδημα και ποικιλία ικανοποίησης.
- **Ομάδα 1:** Πελάτες με υψηλότερο εισόδημα και συχνές αγορές.
- **Ομάδα 2:** Πελάτες με χαμηλότερο εισόδημα και μέτρια ικανοποίηση.

**6. K-medoids**
**Ο αλγόριθμος K-medoids έχει επίσης κατατάξει τους πελάτες σε 3 ομάδες:**

- **Ομάδα 0:** Πελάτες με χαμηλότερο εισόδημα και μέτρια ικανοποίηση, λιγότερες αγορές.
- **Ομάδα 1:** Πελάτες με υψηλότερο εισόδημα και υψηλή ικανοποίηση, συχνές αγορές.
- **Ομάδα 2:** Πελάτες με μέσο εισόδημα και ποικιλία ικανοποίησης.

**Σύγκριση των Αλγορίθμων Ομαδοποίησης για τη συγκεκριμένη μελέτη**

**1. K-means vs. Hierarchical Clustering**

**Ομοιότητες:** Και οι δύο αλγόριθμοι κατέληξαν σε πέντε ομάδες. Η ομαδοποίηση είναι σχεδόν παρόμοια με μικρές διαφορές.
**Διαφορές:** Ο K-means έχει συγκεκριμένες τιμές για κάθε ομάδα, ενώ ο Hierarchical Clustering φαίνεται να έχει κάποιες ελαφρώς διαφορετικές κατανομές, πιθανόν λόγω της ιεραρχικής φύσης του.

**2. K-means vs. DBSCAN**

**Ομοιότητες:** Κανένας από τους δύο αλγόριθμους δεν απαιτεί προκαθορισμένο αριθμό ομάδων.
**Διαφορές:** Ο DBSCAN ανίχνευσε θόρυβο και ανωμαλίες, κατατάσσοντας όλες τις εγγραφές ως -1, κάτι που δεν παρατηρήθηκε στον K-means. Αυτό υποδεικνύει ότι τα δεδομένα μπορεί να μην έχουν αρκετή πυκνότητα για τον DBSCAN.

**3. K-means vs. Mean Shift**

**Ομοιότητες:** Και οι δύο αλγόριθμοι κατέληξαν σε ομάδες.
**Διαφορές:** Ο Mean Shift κατέληξε σε μία μόνο ομάδα (0), ενώ ο K-means δημιούργησε πέντε ομάδες. Αυτό υποδεικνύει ότι ο Mean Shift ενδέχεται να μην ήταν κατάλληλος για τα δεδομένα αυτά, καθώς δεν εντόπισε πολλές φυσικές συγκεντρώσεις.

**4. K-means vs. GMM**

**Ομοιότητες:** Και οι δύο αλγόριθμοι κατέληξαν σε πέντε ομάδες.
**Διαφορές:** Ο GMM δημιούργησε ομάδες με διαφορετικές κατανομές σε σχέση με τον K-means. Ο GMM φαίνεται να παρέχει ευέλικτες κατανομές που μπορούν να χειριστούν μη σφαιρικές διασπορές.

**5. Hierarchical Clustering vs. DBSCAN**

**Ομοιότητες:** Κανένας από τους δύο αλγόριθμους δεν απαιτεί προκαθορισμένο αριθμό ομάδων.
**Διαφορές:** Ο Hierarchical Clustering δημιούργησε πέντε ομάδες, ενώ ο DBSCAN ανίχνευσε όλες τις εγγραφές ως θόρυβο (-1). Αυτό δείχνει ότι ο DBSCAN ενδεχομένως να μην είναι κατάλληλος για τα συγκεκριμένα

δεδομένα.

**6.Hierarchical Clustering vs. Mean Shift**

**Ομοιότητες:** Και οι δύο αλγόριθμοι μπορούν να κατατάσσουν δεδομένα χωρίς προκαθορισμό αριθμού ομάδων.
**Διαφορές:** Ο Hierarchical Clustering δημιούργησε πέντε ομάδες, ενώ ο Mean Shift κατέληξε σε μία ομάδα (0), υποδεικνύοντας διαφορετική προσέγγιση στην ανίχνευση κέντρων υψηλής πυκνότητας.
Hierarchical Clustering vs. GMM

**Ομοιότητες:** Και οι δύο αλγόριθμοι κατέληξαν σε πέντε ομάδες.
**Διαφορές:** Ο GMM παρέχει πιο ευέλικτες κατανομές σε σχέση με τον Hierarchical Clustering, με διαφορετικές διασπορές και προσαρμοστικότητα σε μη σφαιρικές κατανομές.

**7. Hierarchical Clustering vs. GMM**
**Ομοιότητες:** Και οι δύο αλγόριθμοι κατέληξαν σε πέντε ομάδες.
**Διαφορές:** Ο GMM παρέχει πιο ευέλικτες κατανομές σε σχέση με τον Hierarchical Clustering, με διαφορετικές διασπορές και προσαρμοστικότητα σε μη σφαιρικές κατανομές.

**8. DBSCAN vs. Mean Shift**

**Ομοιότητες:** Και οι δύο αλγόριθμοι βασίζονται στην πυκνότητα των δεδομένων για την ομαδοποίηση.
**Διαφορές:** Ο DBSCAN ανίχνευσε όλες τις εγγραφές ως θόρυβο (-1), ενώ ο Mean Shift κατέληξε σε μία μόνο ομάδα (0). Αυτό δείχνει ότι ο DBSCAN μπορεί να μην είναι κατάλληλος για τα δεδομένα αυτά, ενώ ο Mean Shift δεν εντόπισε φυσικές συγκεντρώσεις.

**9. DBSCAN vs. GMM**

**Ομοιότητες:** Και οι δύο αλγόριθμοι μπορούν να διαχειριστούν δεδομένα με ποικίλες κατανομές.
**Διαφορές:** Ο DBSCAN ανίχνευσε όλες τις εγγραφές ως θόρυβο (-1), ενώ ο GMM κατέληξε σε πέντε ομάδες. Ο GMM μπορεί να χειριστεί καλύτερα τα δεδομένα με μη σφαιρικές κατανομές.

**10. Mean Shift vs. GMM**

**Ομοιότητες:** Και οι δύο αλγόριθμοι κατέληξαν σε ομάδες χωρίς προκαθορισμό αριθμού ομάδων.
**Διαφορές:** Ο Mean Shift κατέληξε σε μία μόνο ομάδα (0), ενώ ο GMM δημιούργησε πέντε ομάδες με ευέλικτες κατανομές. Αυτό δείχνει ότι ο GMM μπορεί να είναι πιο κατάλληλος για δεδομένα με μη σφαιρικές διασπορές.

**Προτάσεις Εξατομικευμένων Υπηρεσιών**

Με βάση τις αναλύσεις, μπορούμε να προτείνουμε εξατομικευμένες υπηρεσίες:

- **Ομάδα 0 (Χαμηλότερο Εισόδημα, Μέτρια Ικανοποίηση):**
  - **Εξατομικευμένες Προσφορές:** Προσφορές και εκπτώσεις για πιο οικονομικά προϊόντα.
  - **Προγράμματα Επιβράβευσης: Σχέδια επιβράβευσης για τη συχνότητα αγορών.**
  - **Συμβουλές Αγοράς:** Εξατομικευμένες συστάσεις προϊόντων με βάση τη χαμηλή συχνότητα αγοράς.
- **Ομάδα 1 (Υψηλότερο Εισόδημα, Υψηλή Ικανοποίηση):**
  - **Προϊόντα Premium:** Εξατομικευμένα προϊόντα υψηλής ποιότητας ή υπηρεσίες πολυτελείας.
  - **Διαρκείς Ενημερώσεις:** Πρόσβαση σε νέες συλλογές και αποκλειστικές προσφορές.
  - **Υπηρεσίες Πρώτης Κατηγορίας:** Προσφορές για δωρεάν αποστολή ή VIP εξυπηρέτηση.
- **Ομάδα 2 (Μέσο Εισόδημα, Ποικιλία Ικανοποίησης):**
  - **Ειδικές Εκπτώσεις:** Ειδικές προσφορές για προϊόντα μεσαίας τιμής.
  - **Προσαρμοσμένα Πακέτα:** Δημιουργία πακέτων προϊόντων που ταιριάζουν σε διάφορες

ανάγκες.

- **Εκπαιδευτικά Σεμινάρια:** Συμμετοχή σε σεμινάρια για τη βελτίωση της εμπειρίας αγοράς.

Αυτές οι προτάσεις είναι μόνο αρχικές και μπορούν να εξελιχθούν με την καλύτερη κατανόηση των αναγκών των πελατών και την ανάλυση των περαιτέρω δεδομένων.

# Συμπεράσματα

Η χρήση αλγορίθμων ομαδοποίησης για την παροχή εξατομικευμένων υπηρεσιών έχει καταστεί ένας από τους πιο αποτελεσματικούς τρόπους για να ανταποκριθούν οι σύγχρονες επιχειρήσεις στις αυξανόμενες απαιτήσεις των καταναλωτών για προσαρμοσμένες εμπειρίες. Αυτοί οι αλγόριθμοι επιτρέπουν την ανάλυση μεγάλων όγκων δεδομένων και την αναγνώριση μοτίβων συμπεριφοράς και προτιμήσεων, επιτρέποντας στις επιχειρήσεις να κατηγοριοποιούν τους χρήστες σε ομάδες με παρόμοια χαρακτηριστικά. Μέσω αυτών των εργαλείων, οι επιχειρήσεις μπορούν να βελτιώσουν την κατανόηση των πελατών τους και να παρέχουν εξατομικευμένες προτάσεις και υπηρεσίες που ικανοποιούν συγκεκριμένες ανάγκες και επιθυμίες.

Ένας από τους βασικούς αλγόριθμους που χρησιμοποιούνται για αυτόν τον σκοπό είναι ο K-means, ο οποίος διευκολύνει την κατηγοριοποίηση των χρηστών με βάση την ομοιότητά τους. Αυτό επιτρέπει στις επιχειρήσεις να αναγνωρίζουν συγκεκριμένα τμήματα της αγοράς και να προσαρμόζουν τις προσφορές τους αναλόγως, βελτιώνοντας έτσι την αποτελεσματικότητα των στρατηγικών μάρκετινγκ και την ικανοποίηση των πελατών. Η εφαρμογή του K-means σε πλατφόρμες streaming, για παράδειγμα, επιτρέπει την αναγνώριση ομάδων χρηστών με παρόμοια γούστα σε ταινίες και σειρές, οδηγώντας σε πιο στοχευμένες και αποδοτικές προτάσεις περιεχομένου.

Η σημασία των αλγορίθμων ομαδοποίησης δεν περιορίζεται μόνο στο μάρκετινγκ. Στον τομέα της υγείας, η ανάλυση των ιατρικών δεδομένων μέσω αυτών των αλγορίθμων μπορεί να οδηγήσει στην αναγνώριση ομάδων ασθενών με παρόμοια συμπτώματα ή γενετικά χαρακτηριστικά. Αυτό επιτρέπει την ανάπτυξη προσαρμοσμένων θεραπειών και την καλύτερη πρόγνωση ασθενειών, βελτιώνοντας την ποιότητα της φροντίδας και την ικανοποίηση των ασθενών. Οι αλγόριθμοι όπως οι Gaussian Mixture Models (GMM) και οι Hierarchical Clustering είναι εξαιρετικά χρήσιμοι σε αυτόν τον τομέα, καθώς επιτρέπουν την κατανόηση των πολύπλοκων σχέσεων στα δεδομένα υγείας και τη δημιουργία πιο ακριβών διαγνωστικών εργαλείων.

Στον τομέα της εκπαίδευσης, οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για τη δημιουργία προσαρμοσμένων προγραμμάτων μάθησης. Μέσω της ανάλυσης των δεδομένων προόδου των μαθητών, οι εκπαιδευτικοί μπορούν να κατηγοριοποιήσουν τους μαθητές με βάση τις επιδόσεις και τις μαθησιακές τους ανάγκες, προσφέροντας εξατομικευμένο εκπαιδευτικό υλικό που ανταποκρίνεται καλύτερα στα ατομικά προφίλ των μαθητών. Η χρήση αλγορίθμων όπως η ιεραρχική ομαδοποίηση επιτρέπει την αναγνώριση των υποκείμενων σχέσεων μεταξύ των μαθητών και την παροχή στοχευμένης υποστήριξης, βελτιώνοντας τα εκπαιδευτικά αποτελέσματα και την εμπειρία μάθησης.

Οι πλατφόρμες ψηφιακού περιεχομένου, όπως οι υπηρεσίες streaming, επωφελούνται επίσης σημαντικά από τους αλγόριθμους ομαδοποίησης. Η ανάλυση των προτιμήσεων και της συμπεριφοράς των χρηστών επιτρέπει την παροχή προσαρμοσμένων προτάσεων περιεχομένου, βελτιώνοντας την εμπειρία χρήσης και αυξάνοντας την αφοσίωση των πελατών. Αλγόριθμοι όπως οι GMM και ο Mean Shift μπορούν να δημιουργήσουν πιο ακριβείς ομάδες χρηστών και να προσφέρουν πιο στοχευμένες προτάσεις. Αυτό όχι μόνο ενισχύει την ικανοποίηση των χρηστών αλλά και βοηθά τις πλατφόρμες να διατηρήσουν υψηλά επίπεδα δέσμευσης και να μειώσουν τις αποχωρήσεις.

Όσον αφορά η παραπάνω μελέτη αξιολόγησε διάφορους αλγορίθμους ομαδοποίησης για τη βελτίωση της εξατομίκευσης των υπηρεσιών. Οι αλγόριθμοι που χρησιμοποιήθηκαν περιλάμβαναν K-means, Ιεραρχική Ομαδοποίηση, DBSCAN, Mean Shift, GMM και K-medoids. Κάθε αλγόριθμος είχε τα δικά του πλεονεκτήματα και μειονεκτήματα, επηρεάζοντας τον τρόπο με τον οποίο οι ομάδες σχηματίστηκαν και κατηγοριοποιήθηκαν. Για παράδειγμα, ο K-means ήταν αποδοτικός στον εντοπισμό συμπαγών ομάδων, ενώ ο DBSCAN ήταν ικανός να εντοπίζει περιοχές υψηλής πυκνότητας και να απομονώνει τα "θορυβώδη" δεδομένα. Στον τομέα της υγείας, η ικανότητα να κατηγοριοποιούνται οι ασθενείς με βάση διάφορα χαρακτηριστικά όπως η ηλικία, το φύλο και οι ιατρικές διαγνώσεις μπορεί να οδηγήσει σε καλύτερα εξατομικευμένα σχέδια θεραπείας. Η ομαδοποίηση μπορεί να βοηθήσει στον εντοπισμό υποομάδων ασθενών που έχουν παρόμοιες ανάγκες ή αντιδράσεις σε συγκεκριμένες θεραπείες, επιτρέποντας στους παρόχους υγειονομικής περίθαλψης να προσαρμόσουν τις υπηρεσίες τους με βάση αυτά τα μοτίβα. Στο μάρκετινγκ, η ομαδοποίηση πελατών επιτρέπει στις επιχειρήσεις να κατανοούν καλύτερα τις προτιμήσεις και τις αγοραστικές συνήθειες των διαφόρων τμημάτων της αγοράς. Μέσω αυτής της μελέτης, έγινε φανερό ότι η χρήση αλγορίθμων όπως ο K-means και ο GMM μπορεί να βοηθήσει τις επιχειρήσεις να δημιουργούν στοχευμένες καμπάνιες και να προωθούν προϊόντα και υπηρεσίες που ανταποκρίνονται στις

ανάγκες συγκεκριμένων ομάδων πελατών.

Κάθε αλγόριθμος ομαδοποίησης προσέφερε διαφορετική προοπτική. Ο K-means και ο K-medoids παρείχαν καθαρή διαίρεση των δεδομένων, ενώ οι Ιεραρχική Ομαδοποίηση και DBSCAN προσέφεραν περισσότερη ευελιξία στην κατανόηση των δεδομένων και την απομόνωση ανωμαλιών. Ο GMM προσέφερε μια πιο σύνθετη προσέγγιση μέσω της χρήσης κανονικών κατανομών, επιτρέποντας την καλύτερη κατανόηση των πιθανοτήτων για την ένταξη των πελατών σε συγκεκριμένες ομάδες. Η ανάλυση αντιμετώπισε προκλήσεις όπως η διαχείριση του θορύβου στα δεδομένα και η επιλογή των κατάλληλων παραμέτρων για κάθε αλγόριθμο. Οι αλγόριθμοι όπως ο DBSCAN επηρεάστηκαν ιδιαίτερα από την επιλογή της παραμέτρου ελάχιστης πυκνότητας, ενώ ο K-means και ο GMM απαιτούσαν εκ των προτέρων καθορισμό του αριθμού των ομάδων. Η χρήση αυτών των αλγορίθμων στην πράξη μπορεί να οδηγήσει σε σημαντικά οφέλη. Στις υγειονομικές υπηρεσίες, η ομαδοποίηση μπορεί να υποστηρίξει την ανάπτυξη εξατομικευμένων θεραπευτικών πλάνων και την πρόληψη ασθενειών. Στο μάρκετινγκ, οι επιχειρήσεις μπορούν να χρησιμοποιούν τα αποτελέσματα της ομαδοποίησης για να ενισχύσουν την εμπειρία του πελάτη, να αυξήσουν την ικανοποίηση και να βελτιώσουν τις πωλήσεις.

Συνολικά, η χρήση αλγορίθμων ομαδοποίησης προσφέρει σημαντικά πλεονεκτήματα στην κατανόηση και την εξατομίκευση των υγειονομικών και εμπορικών υπηρεσιών, προσφέροντας πιο στοχευμένες και αποτελεσματικές προσεγγίσεις που ανταποκρίνονται στις ανάγκες των πελατών και των ασθενών.

# Μελλοντική έρευνα

Με την πρόοδο της τεχνολογίας και της επιστήμης των δεδομένων, είναι κρίσιμο να εξεταστούν και να ενσωματωθούν πιο προηγμένοι αλγόριθμοι και τεχνικές για την ομαδοποίηση. Οι αλγόριθμοι βαθιάς μάθησης, όπως οι νευρωνικά δίκτυα και οι τεχνικές ενίσχυσης, μπορούν να προσφέρουν αναλυτικές δυνατότητες πέρα από τις παραδοσιακές μεθόδους ομαδοποίησης. Η ανάπτυξη και η εφαρμογή αυτών των μεθόδων σε πραγματικά σύνολα δεδομένων μπορεί να οδηγήσει σε πιο ακριβείς και ευέλικτες αναλύσεις. Επιπλέον, η αξιολόγηση νέων εργαλείων και πλατφορμών για την ομαδοποίηση, καθώς και η βελτίωση της επεξεργασίας δεδομένων, θα μπορούσαν να ενισχύσουν τη συνολική απόδοση και ακρίβεια των αλγορίθμων.

Η ανάλυση δεδομένων που χρησιμοποιήθηκαν στη μελέτη περιορίστηκε σε ένα σχετικά μικρό δείγμα. Για να επιβεβαιωθούν τα ευρήματα και να ενισχυθεί η εγκυρότητα των αποτελεσμάτων, είναι απαραίτητο να επεκταθούν τα δεδομένα σε μεγαλύτερα και πιο ποικιλόμορφα δείγματα. Η συλλογή δεδομένων από περισσότερους χρήστες και η ενσωμάτωσή τους σε περιβάλλοντα διαφορετικών αγορών ή τομέων υγείας θα μπορούσαν να παρέχουν πιο ακριβή και γενικεύσιμα αποτελέσματα. Επίσης, η συμπερίληψη περισσότερων χαρακτηριστικών ή παραμέτρων μπορεί να αποκαλύψει κρυφές σχέσεις και μοτίβα που δεν είναι ορατά σε μικρότερα δείγματα. Η μετάβαση από τη θεωρητική ανάλυση στην πρακτική εφαρμογή είναι κρίσιμη για την αξιολόγηση της χρησιμότητας των αποτελεσμάτων. Οι οργανισμοί υγειονομικής περίθαλψης και οι επιχειρήσεις θα πρέπει να ενσωματώσουν τις στρατηγικές εξατομίκευσης που προκύπτουν από την ανάλυση στην καθημερινή τους λειτουργία και να μετρήσουν την επίδρασή τους στην ικανοποίηση του πελάτη και στις επιχειρηματικές επιδόσεις. Η συλλογή ανατροφοδότησης από τους χρήστες και η αξιολόγηση της αποτελεσματικότητας αυτών των στρατηγικών σε πραγματικά περιβάλλοντα θα παρέχει πολύτιμα δεδομένα για τη βελτίωση των προσεγγίσεων.

Η συνεργασία με άλλους τομείς της επιστήμης, όπως η ψυχολογία, η κοινωνιολογία και η οικονομία, μπορεί να ενισχύσει την κατανόηση των παραγόντων που επηρεάζουν τις ανάγκες και τις προτιμήσεις των χρηστών. Η διασύνδεση με άλλες επιστήμες μπορεί να οδηγήσει στην ανάπτυξη νέων μοντέλων και θεωριών που θα βελτιώσουν την ακρίβεια και την εφαρμογή των στρατηγικών εξατομίκευσης. Οι πολυδιάστατες προσεγγίσεις θα επιτρέψουν την κατανόηση των συμπεριφορών και των αναγκών των χρηστών σε βάθος, προσφέροντας έτσι πιο στοχευμένες και αποδοτικές λύσεις.

Συνολικά, η μελέτη παρέχει ένα ισχυρό θεμέλιο για την κατανόηση και την ανάπτυξη εξατομικευμένων υπηρεσιών μέσω αλγορίθμων ομαδοποίησης, με την προοπτική να εξελιχθούν περαιτέρω μέσω της ενσωμάτωσης νέων τεχνολογιών και της αξιολόγησης σε πραγματικά περιβάλλοντα.

# Πίνακας ορολογίας

| | |
|---|---|
| **K-means Clustering** | Δημοφιλής αλγόριθμος που χωρίζει τα δεδομένα σε Κ ομάδες |
| **Hierarchical Clustering** | Τεχνική ομαδοποίησης για ιεράρχηση ομάδων |
| **DBSCAN** | Density-Based Spatial Clustering of Applications with Noise |
| **GMM** | Gaussian Mixture Models |
| **Mean Shift Clustering** | Εντοπίζει περιοχές υψηλής πυκνότητας μετακινώντας επαναληπτικά τα δεδομένα |
| **K-medoids** | Αποτελεί μία παραλλαγή του αλγορίθμου K-means, που έχει ως στόχο τη μείωση της ευαισθησίας στις εξωγενείς τιμές και τα ανώμαλα δεδομένα |
| **Linkage Criteria** | Κριτήρια συνοχής καθορίζουν πώς υπολογίζεται η απόσταση μεταξύ συστάδων |

# Βιβλιογραφία

**Ahmed, S., Gupta, S., Suri, A., & Sharma, S. (2021).** Adaptive energy efficient fuzzy: An adaptive and energy efficient fuzzy clustering algorithm for wireless sensor network-based landslide detection system. DOI:10.1049/ntw2.12004

**Alonistioti, N., Tsichrintzi, E.A., Chrysafiadi, K. and Alepis, E., 2023,** July. Requirements for Fuzzy Logic in Personalisation of Fire Emergency Alerts. DOI:10.1109/IISA59645.2023.10345861

**Ashabi, A., Sahibuddin, S. B., & Salkhordeh Haghighi, M. (2020, December).** The systematic review of K-means clustering algorithm. In *Proceedings of the 2020 9th International Conference on Networks, Communication and Computing*. DOI:10.1145/3447654.3447657

**Ashir, A. M., & Shehu, G. S. (2015, June).** Adaptive clustering algorithm for optical character recognition. In *2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* DOI:10.1109/ECAI.2015.7301192

**Balamurugan, R., Ratheesh, S., & Venila, Y. M. (2022).** Classification of heart disease using adaptive Harris hawk optimization-based clustering algorithm and enhanced deep genetic algorithm. *Soft computing*, *26*(5), DOI:10.1007/s00500-021-06536-0

**Bezdek, J. C., Ehrlich, R., & Full, W. (1984).** FCM: The Fuzzy C-Means Clustering Algorithm. *Computers & Geosciences*, 10(2-3). doi: http://dx.doi.org/10.1016/0098-3004(84)90020-7

**Bleier, A., Arne De Keyser, Verleye, K. (2018).** Customer Engagement Through Personalization and Customization. *Customer Engagement Marketing (pp.75-94)*. doi http://dx.doi.org/10.1007/978-3-319-61985-9_4

**Bleier, A., De Keyser, A., & Verleye, K.** (2019). Customer engagement through personalization and customization: The impact of online versus offline channels. *Journal of Service Management*, 30(5), 529-550. doi: http://dx.doi.org/10.1007/978-3-319-61985-9_4

**Bui, V. H., & Phan, H. T. (2023).** The Computational Complexity of Hierarchical Clustering Algorithms for Community Detection: A Review. *Vietnam Journal of Computer Science (World Scientific)*, DOI:10.1142/S2196888823300016

**Bushra, A., Yi, G. (2022)**. Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. *IEEE Access* PP(99):1-1PP(99):1-1. doi: http://dx.doi.org/10.1109/ACCESS.2021.3089036

**Caya, R., & Neto, J. J. (2018)**. A bibliometric review about adaptivity. DOI:10.1016/j.procs.2018.04.163

**Chand, S., Mohapatra, S., & Mishra, V. (2021).** An intelligent system to identify coal maceral groups using markov-fuzzy clustering approach. DOI:10.3233/JIFS-189889

**Chrysafiadi, K., & Virvou, M. (2015).** *Advances in personalized web-based education*. DOI:10.1007/978-3-319-12895-5

**Chrysafiadi, K., & Virvou, M. (2008**). Personalized teaching of a programming language over the web: Stereotypes and rule-based mechanisms.DOI:https://doi.org/10.1016/j.eswa.2013.02.007

**Chrysafiadi, K., Virvou, M., & Sakkopoulos, E. (2020)**. Optimizing programming language learning through student modeling in an adaptive web-based educational environment. DOI:10.1007/978-3-030-13743-4_11

**Chrysafiadi, K., Virvou, M., Tsihrintzis, G.A. and Hatzilygeroudis, I., 2023**. An Adaptive Learning Environment for Programming Based on Fuzzy Logic and Machine Learning. DOI:10.1142/S0218213023600114

**Comaniciu, D., Meer, P. (2002)**. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), 603-619. doi: http://dx.doi.org/10.1109/34.1000236

**Daniel Păvăloaia, V. (2024).** Clustering Algorithms in Sentiment Analysis Techniques in Social Media-A Rapid Literature Review. *International Journal of Advanced Computer Science & Applications*, DOI:10.14569/ijacsa.2024.0150314

**Dong, M., Meng, S., Chen, L., & Zhang, J. (2021)**. Personalized Medical Diagnosis Recommendation Based on Neutrosophic Sets and Spectral Clustering. DOI:10.1007/978-3-030-69992-5_13

**Eissa, M. A., & Chen, P. (2023, October).** Personalized Electric Vehicle Range Prediction Based on Self-Supervised Driving Pattern Clustering. DOI:10.1109/IAVVC57316.2023.10328148

**Estevill – Castro, V. (2002).** Why so many clustering algorithms -- A Position Paper. *ACM SIGKDD Explorations Newsletter 4(1) 4(1)*. doi http://dx.doi.org/10.1145/568574.568575

**Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, O. N., José-García, A., & Agushaka, J. O. (2021).** Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, DOI:10.1007/s00521-020-05395-4

**Farahani, M. G., Torkestani, J. A., & Rahmani, M. (2022).** Adaptive personalized recommender system using learning automata and items clustering. DOI:10.1016/j.is.2021.101978

**Feng, Y. L., Zhang, H. Q., & Peng, C. T. (2021, October).** Fast recommendation method of personalized tourism big data information based on improved clustering algorithm. DOI: **https://doi.org/10.1002/adfm.202003619**

**Fortunato, S. (2010)**. Community detection in graphs. *Physics Reports*, 486(3-5), 75-174. doi: http://dx.doi.org/10.1016/j.physrep.2009.11.002

**Gretzel, U., Fesenmaier, D. R., Formica, S., & O'Leary, J. T**. (2006). Searching for the future: Challenges faced by destination marketing organizations. Journal of Travel Research, 45(2), 116-126. doi: http://dx.doi.org/10.1177/0047287506291598

**Grewal, D., Roggeveen, A. L., & Nordfält, J. (2017).** The Future of Retailing. Journal of Retailing, 93(1), 1-6. doi: http://dx.doi.org/10.1016/j.jretai.2016.12.008

**Grua, E. M., Hoogendoorn, M., Malavolta, I., Lago, P., & Eiben, A. E. (2019, October).** Clustream-GT: Online clustering for personalization in the health domain. In *IEEE/WIC/ACM International Conference on Web Intelligence*DOI: https://doi.org/10.1145/3350546.3352529

**Gu, L., Ma, W., An, Y., & Huang, D. (2020, August).** A Stream Clustering Algorithm of Self-Adaptive Method. In *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)* DOI: https://doi.org/10.1038/s41467-020-14792-1

**Gomber, P., Kauffman, R. J., Parker, C., Weber, B.W. (2018).** On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services. *Journal of Management Information Systems*, 35(1), 220-265. doi: http://dx.doi.org/10.1080/07421222.2018.1440766

**Gomber, Lambrecht, A., Tucker, C. (2013).** When Does Retargeting Work? Information Specificity in Online Advertising. *Journal of Marketing Research, 50*(5), 561-576. doi:.http://dx.doi.org/10.1177/002224371305000508

**Gottipati, N. R., & Rama Prasath, A. (2019)**. A multi-agent bio-inspired system to map learners with learning resources using clustering based personalization. DOI:10.35940/ijitee.J9833.0881019

**Hartigan, J. A., & Wong, M. A.** (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100-108. doi: https://doi.org/10.2307/2346830

**Hahsler, M., Piekenbrock, M., & Doran, D.** (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 99(1), 1-24. doi: http://dx.doi.org/10.18637/jss.v091.i01

**Hatzilygeroudis, I., Tsihrintzis, G., Virvou, M., & Perikos, I. (2023).** Special issue on information, intelligence, systems and applications. DOI:10.1007/s00521-022-07954-3

**Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023).** K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, DOI:10.1016/j.ins.2022.11.139

**Jain, A. K.** (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. doi: http://dx.doi.org/10.1016/j.patrec.2009.09.011

**Jain, A. K., & Dubes, R. C.** (2020). *Algorithms for Clustering Data*. Prentice Hall.

**Jain, A. K., Murty, M. N., & Flynn, P. J. (1999**). Data Clustering: A Review. *ACM Computing Surveys (CSUR)*, 31(3), 264-323. doi: https://doi.org/10.1145/331499.331504

**Jiang, L. (2024).** A fuzzy clustering approach for cloud-based personalized distance music education and resource management. DOI:10.1007/s00500-023-09525-7

**Kaufman, L., Rousseeuw, P. J. (1990**). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley. doi http://dx.doi.org/10.2307/2532178

**Kausar, S., Huahu, X., Hussain, I., Wenhao, Z., & Zahid, M. (2018).** Integration of data mining clustering approach in the personalized E-learning system. DOI:10.1109/ACCESS.2018.2882240

**Krogh, A. (2008)**. What are artificial neural networks? *Nature, 371*, 191-195. doi: http://dx.doi.org/10.1038/nbt1386

**Kumar, A., & Ashraf, M. (2015, March).** Personalized web search engine using dynamic user profile and clustering techniques.DOI: Personalized web search engine using dynamic user profile and clustering techniques | IEEE Conference Publication | IEEE Xplore

**Kumar, S., Kumar, S. (2022)**. Experimental Comparisons of Clustering Approaches for Data Representation. *ACM Comput. Surv.* 55, 3, Article 45 (March 2022), 33 pages. doi: https://doi.org/10.1145/3490384

**Kumar, V., Reinartz, W. (2016).** Creating Enduring Customer Value. *Journal of Marketing, 80*(6), 81-95. doi: http://dx.doi.org/10.1509/JM.15.0414

**Kumar, D. M., Satyanarayana, D., & Prasad, M. G. (2021).** MRI brain tumor detection using optimal possibilistic fuzzy C-means clustering algorithm and adaptive k-nearest neighbor classifier. *Journal of Ambient Intelligence and Humanized Computing*, *12*(2), DOI:10.1007/s12652-020-02444-7

**Liang, H. (2021).** Intelligent Tourism Personalized Recommendation Based on Multi-Fusion of Clustering Algorithms. *DOI:* https://doi.org/10.1155/2021/4517973

**Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., ... & Sampath, D.** (2010). The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems* (pp.

293-296). doi: http://dx.doi.org/10.1145/1864708.1864770

**Li, B. (2020, December)**. Ant colony clustering algorithm for personalized recommendation of e-commerce. **DOI:** 10.1109/ICMCCE51767.2020.00407

**Lin, X., Guan, W., & Zhang, Y. (2023).** Application of Data Mining Technology with Improved Clustering Algorithm in Library Personalized Book Recommendation System. DOI:10.14569/IJACSA.2021.0120126

**Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J.** (2013). Understanding of internal clustering validation measures. *Proceedings of the IEEE 13th International Conference on Data Mining (ICDM*), 911-916.. doi: http://dx.doi.org/10.1109/TNN.2005.845141

**Long, G., Xie, M., Shen, T., Zhou, T., Wang, X., & Jiang, J. (2023**). Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, *26*(1), 481-500.

**Lund, B., Ma, J. (2021).** A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering. *Performance Measurement and Metrics* 22(3):161-173. doi http://dx.doi.org/10.1108/PMM-05-2021-0026

**Margaris, D., Georgiadis, P., & Vassilakis, C. (2015, May).** A collaborative filtering algorithm with clustering for personalized web service selection in business processes. DOI:10.1109/RCIS.2015.7128877

**Manning, C. D., Raghavan, P., Schütze, H. (2008**). Introduction to Information Retrieval. Cambridge University Press.

**Moftah, H. M., Azar, A. T., Al-Shammari, E. T., Ghali, N. I., Hassanien, A. E., & Shoman, M. (2014)**. Adaptive k-means clustering algorithm for MR breast image segmentation. *Neural Computing and Applications*, DOI:10.1007/s00521-013-1437-4

**Mohiuddin, A., Seraj, R., Islam, S. (2022)**. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics* 2020, 9(8). doi: https://doi.org/10.3390/electronics9081295

**Mohiuddin, A., Seraj, R., Islam, S. (2022)**. Agglomerative and divisive hierarchical Bayesian clustering. *Computational Statistics & Data Analysis Volume 176*. doi: https://doi.org/10.1016/j.csda.2022.107566

**Moodley, R. (2021).** Fostering Positive Personalisation Through Fuzzy Clustering. In *Fuzzy Logic: Recent Applications and Developments*. DOI: https://doi.org/10.48550/arXiv.2005.01026

**Murtagh, F., Contreras, P.** (2010). Methods of Hierarchical Clustering. *arXiv preprint*. doi: http://dx.doi.org/10.1007/978-3-642-04898-2_288

**Murtagh, F., Contreras, P. (2012).** Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86-97. doi: http://dx.doi.org/10.1002/widm.53

**Mulay, P., & Shinde, K. (2019).** Personalized diabetes analysis using correlation-based incremental clustering algorithm. DOI:10.1007/978-981-13-0550-4_8

**Park, H. S., Jun, C. H. (2009)**. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2), 3336-3341. doi: http://dx.doi.org/10.1016/j.eswa.2008.01.039

**Patcha, A., Park, J. M.** (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends, Computer Networks, 51(12), 3448-3470. doi: http://dx.doi.org/10.1016/j.comnet.2007.02.001

**Patel, E., Kushwaha, S. (2020)**. Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science Volume 171 2020*, Pages 158-167. doi: https://doi.org/10.1016/j.procs.2020.04.017

**Preeti, A., Deepali Dr., Varshney, S. (2009).** Analysis of K-Means and K-Medoids Algorithm For Big DataProcedia. *Computer Science.* Volume 78, 2016, Pages 507-512. DOI:**https://doi.org/10.1016/j.procs.2016.02.095**

**Qiang, L., Yao, X.** (2005). Clustering and learning Gaussian distribution for continuous optimization, *IEEE Transactions on Systems, Man, and Cybernetics*, Part C (Applications and Reviews) (Volume: 35, Issue: 2, May 2005). doi: https://doi.org/10.1109/TSMCC.2004.841914

**Rajkumar, T. D., Raja, S. P., & Suruliandi, A. (2017).** Users' click and bookmark based personalization using modified agglomerative clustering for web search engine. DOI:10.1142/S0218213017300022

**Ravi, L., Subramaniyaswamy, V., Vijayakumar, V., Jhaveri, R. H., & Shah, J. (2021).** Hybrid user clustering-based travel planning system for personalized point of interest recommendation. DOI:10.1007/978-981-15-9953-8_27

**Reynolds, D. A. (2009).** Gaussian Mixture Models. *Encyclopedia of Biometrics*, 659-663. doi: https://doi.org/10.1007/978-0-387-73003-5_196

**Reynolds, D. A. (2009).** Gaussian Mixture Models for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2), 339-349. DOI: **http://dx.doi.org/10.1007/978-0-387-73003-5_196**

**Rokach, L., Maimon, O. (2005)**. Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer. doi: http://dx.doi.org/10.1007/0-387-25465-X_15

**Saputri, T. R. D., & Lee, S. W. (2020).** The application of machine learning in self-adaptive systems: A systematic literature review. DOI:10.1109/ACCESS.2020.3036037

**Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017).** DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3), 1-21. doi: http://dx.doi.org/10.1145/3068335

**Shi, L., & Pan, L. (2022, September).** Food Nutrient Composition Analysis Based on Adaptive AP Clustering Algorithm. In *Proceedings of the 7th International Conference on Intelligent Information Processing* DOI:10.1145/3570236.3570254

**Shi, Y., & Zhu, Y. (2022).** [Retracted] Research on Fast Recommendation Algorithm of Library Personalized Information Based on Density Clustering. DOI:**https://doi.org/10.1155/2022/1169115**

**Smith, A., Linden, G.** (2007). Two decades of recommender systems at Amazon.com. IEEE Internet Computing, IEEE Internet Computing 21(3):12-18. doi: http://dx.doi.org/10.1109/MIC.2017.72

**Sotiropoulos, D. N., Tsihrintzis, G. A., Savvopoulos, A., & Virvou, M. (2006).** Artificial immune system-based customer data clustering in an e-shopping application. DOI:10.1007/11892960_115

**Szeliski, R. (2010).** Computer Vision: Algorithms and Applications**.** *Springer Science & Business Media.*

**Tsai, C. F., Chiu, C. C. (2004)**. A purchase-based market segmentation methodology. *Expert Systems with Applications*, 27(2), 265-276. doi: http://dx.doi.org/10.1016/j.eswa.2004.02.005

**Tsihrintzis, G. A., Virvou, M., & Jain, L. C. (2016).** *Intelligent Computing Systems: Emerging Application Areas* DOI:10.1007/978-3-662-49179-9

**Vignesh, U., & Ratnakumar, R. (2024).** An Empirical Review on Clustering Algorithms for Image

Segmentation of Satellite Images. *AI and Blockchain Optimization Techniques in Aerospace Engineering*, DOI: 10.4018/979-8-3693-1491-3.ch002

**Virvou, M., 2018,** July. A new era towards more engaging and human-like computer-based learning by combining personalisation and artificial intelligence techniques. DOI:10.1145/3197091.3211886

**Virvou, M., Savvopoulos, A., Tsihrintzis, G. A., & Sotiropoulos, D. N. (2007).** Constructing Stereotypes for an Adaptive e-Shop using AIN-based Clustering. DOI:10.1007/978-3-540-71618-1_94

**Virvou, M., Tsihrintzis, G. A., Bourbakis, N. G., & Jain, L. C. (2022).** Handbook on Artificial Intelligence-Empowered Applied Software Engineering: VOL. 2: Smart Software Applications in Cyber-Physical Systems (Vol. 3). DOI: Handbook on Artificial Intelligence-Empowered Applied Software Engineering: VOL.2: Smart Software Applications in Cyber-Physical Systems | SpringerLink

**Wang, D., Xie, C., & Wang, S. (2021**). An adaptive RBF neural network–based multi-objective optimization method for lightweight and crashworthiness design of cab floor rails using fuzzy subtractive clustering algorithm. *Structural and Multidisciplinary Optimization*, DOI:10.1007/s00158-020-02797-9

**Wang, M., & Lv, Z. (2022**). Construction of personalized learning and knowledge system of chemistry specialty via the internet of things and clustering algorithm. DOI:10.1007/s11227-022-04315-8

**Wang, Q., Wang, X., Fang, C., & Yang, W. (2020).** Robust fuzzy c-means clustering algorithm with adaptive spatial & intensity constraint and membership linking for noise image segmentation. *Applied Soft Computing*, DOI: https://doi.org/10.1016/j.asoc.2020.106318

**Wang, Y., Yang, Y., & Zhao, X. (2020, August).** Object detection using clustering algorithm adaptive searching regions in aerial images. In *European Conference on Computer Vision* DOI:10.1007/978-3-030-66823-5_39

**Wang, X., & Cai, B. (2023, May).** Information Personalized Recommendation Algorithm for Cross-Border E-Commerce Guide Platform Based on Constrained Clustering. DOI:10.1109/ICNETIC59568.2023.00062

**Wang, Z., Chen, J., Rosas, F. E., & Zhu, T. (2022)**. A hypergraph-based framework for personalized recommendations via user preference and dynamics clustering. DOI:10.1016/j.eswa.2022.117552

**Wang, X., Wang, Y., Guo, L., Xu, L., Gao, B., Liu, F., & Li, W. (2021).** Exploring clustering-based reinforcement learning for personalized book recommendation in digital library. DOI: **https://doi.org/10.1002/adfm.202003619**

**Wedel, M., Kamakura, W.** (2000). Market Segmentation: Conceptual and Methodological Foundations. doi: http://dx.doi.org/10.1007/978-1-4615-4651-1

**Wong, T., Wagner, M., & Treude, C. (2022).** Self-adaptive systems: A systematic literature review across categories and domains. DOI:10.48550/arXiv.2101.00125

**Wu, Y., Chen, Y., & Ling, W. (2021).** [Retracted] Audit Analysis of Abnormal Behavior of Social Security Fund Based on Adaptive Spectral Clustering Algorithm. *Complexity*, DOI:10.1155/2024/9873842

**Xia, X., Qin, Z., Yu, J., & Qi, L. (2019)**. Personalised service recommendation process based on service clustering. *DOI:* **https://doi.org/10.1002/adfm.202003619**

**Xia, Y., Fang, J., Chao, P., Pan, Z., & Shang, J. S. (2023).** Cost-effective and adaptive clustering algorithm for stream processing on cloud system. *GeoInformatica*, DOI: https://doi.org/10.1016/j.nanoen.2019.01.051

**Xing, H., Cao, K., Zhang, M., Wu, X., & Chen, H. (2022, September).** An interval type-2 fuzzy clustering algorithm with adaptive type-reduction optimization and its application in remote sensing image

classification. In *2nd International Conference on Information Technology and Intelligent Control (CITIC 2022)* DOI: **https://doi.org/10.1002/adfm.202003619**

**Xu, R., & Wunsch, D.** (2005). Survey of clustering algorithms: Conceptual and Methodological Foundations. IEEE *Transactions on Neural Networks*, 16(3), 645-678. doi: http://dx.doi.org/10.1109/TNN.2005.845141

**Xu, D., Tian, Y. (2015)**. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2, 165-193. doi: http://dx.doi.org/10.1007/s40745-015-0040-1

**Yang, G., & Jiang, Y. (2022).** Adaptive clustering algorithm for teaching resources of Chinese flower and bird painting practice and theory course. *Mobile Information Systems*, *2022*(1), DOI: **https://doi.org/10.1155/2022/1439782**

Ye, L., Chen, Y., Han, Q., Zeng, L., Cheng, S., Xiao, L., & Ding, X. (2020, September). Vehicle message distribution mechanism based on improved k-means adaptive clustering algorithm. DOI:**https://doi.org/10.1002/adfm.202003619**

**Yishan, Z., Chenxuan, Z., Fuqiang, L., Zongxin, H., & Yanhua, L. (2023, June).** An adaptive method of selecting typical days based on improved fuzzy clustering algorithm. In *Sixth International Conference on Intelligent Computing, Communication, and Devices (ICCD 2023)* DOI:10.1016/j.energy.2020.116913

**Zhang, Z., Liu, L., Liu, H., & Wu, Z. (2023, July**). Student Profile Clustering Based Personalized Exercise Recommendation: Taking Data Structures Course as an Example. DOI: **https://doi.org/10.1002/adfm.202003619**

**Zivkovic, Z., Heijden, F. V. D. (2004)**. Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5), 651-656. doi: http://dx.doi.org/10.1109/TPAMI.2004.1273970

**Zhang, Z., Zheng, Z. (2022)**. Application of K-Medoids Clustering in Retail Market Segmentation. *Computers, Environment and Urban Systems, 88*, 101670. http://dx.doi.org/10.46729/ijstm.v5i1.1024

**Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η. (2015)**. Συσταδοποίηση [Κεφάλαιο]. Η επιστήμη των δεδομένων μέσα από τη γλώσσα R. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://hdl.handle.net/11419/2972

**Κύρκος, Ε. (2015).** Ανάλυση Συστάδων [Κεφάλαιο]. Στο Κύρκος, Ε. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων [Προπτυχιακό εγχειρίδιο]. *Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις*. https://hdl.handle.net/11419/1238