ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Συγκριτική Μελέτη Αλγορίθμων Ομαδοποίησης με Εφαρμογή σε Ιατρικά Δεδομένα

Διπλωματική Εργασία

Νικολαΐδης Γιώργος

Επιβλέπων

Πέτρος Στεφανέας

Αθήνα, Φεβρουάριος 2020

Περιεχόμενα

Αλγόριθμος K means

Δεδομένα: Σύνολο των παρατηρήσεων Π , αριθμός συστάδων K, διαστάσεις m

Έξοδος : Διαμέριση $(C'_1, C'_2, \ldots, C'_k)$ του συνόλου Π

{ C_i είναι η i-οστή συστάδα }

1: Έστω $(C_1, C_2, \ldots, C_k)$ μια τυχαία αρχική διαμέριση του Π

2: Επανάλαβε

3: d_{ij} = Απόσταση παρατήρησης i απο το μέσο της συστάδας j

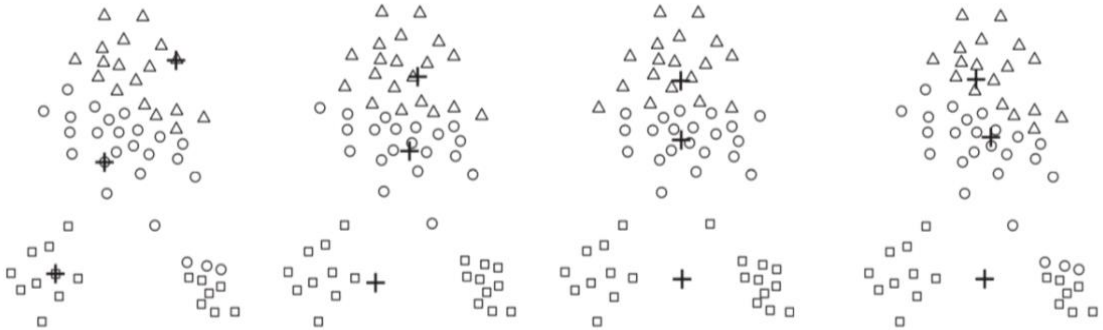
4: $n_i = argmin_{1 \leqslant j \leqslant K} d_{ij}$

5: Τοποθέτησε την παρατήρηση i στη συστάδα n_i

6: Υπολόγισε ξανά τους μέσους κάθε συστάδας

7: Μέχρι να μην συμβεί καμία αλλαγή στις συστάδες σε μία πλήρη επανάληψη

8: Επιστροφή του αποτελέσματος, δηλαδή της διαμέρισης $(C'_1, C'_2, \ldots, C'_k)$



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

$$\hat{z}_{ki} := \frac{1}{|C_k|} \sum_{l=1}^{|C_k|} x_{li}$$

$k = 1,\ldots,K$

$i = 1,..,m$

Συνάρτηση σφάλματος

$$f_\Sigma(C) := \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{m} (x_{ij} - \hat{z}_{kj})^2$$

Elbow Method

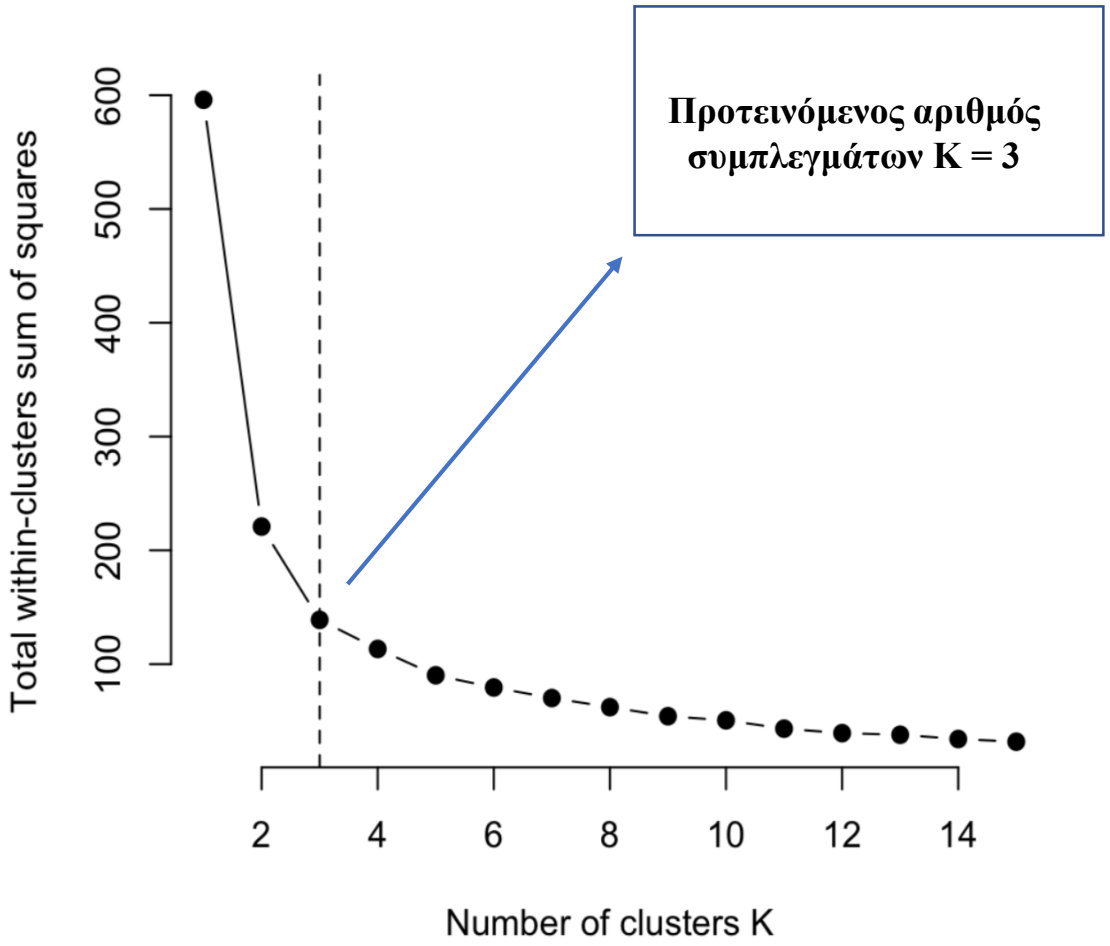
1. Εφαρμογή του K- means για τιμές του K σε κάποιο διάστημα π.χ. από 1 έως 15

2. Για κάθε K, υπολογισμός της ποσότητας tot. withiness

$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - z_k)^2$$

3. Σχεδιασμός γραφικής παράστασης tot. withiness - K

4. Το σημείο στην καμπύλη που κάνει απότομη στροφή και μοιάζει με αγκώνα, συνήθως θεωρείται ένδειξη υποψήφιου αριθμού K για τον αριθμό των συμπλεγμάτων.



Silhouette Method

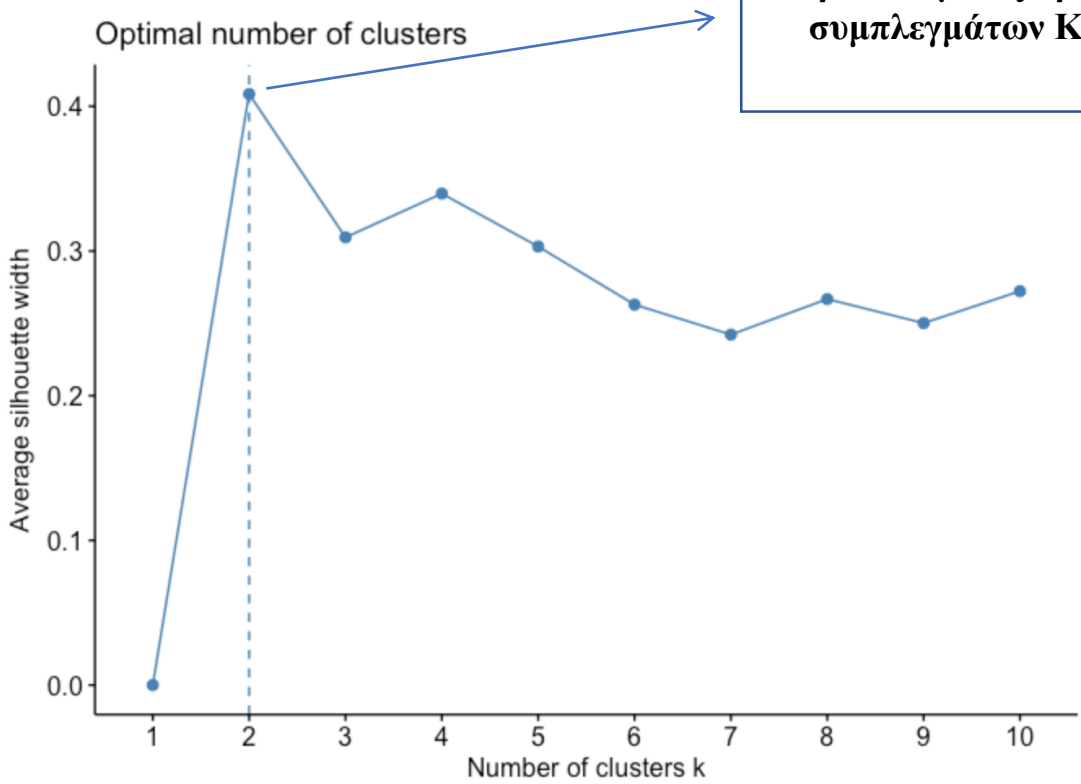
1. Εφαρμογή του K-means για τιμές του K σε κάποιο διάστημα, π.χ. απο 1 έως 10.

2. Για κάθε τιμή του k υπολογίζουμε τη μέση silhouette όλων των παρατηρήσεων (avg.sil). Για κάθε παρατήρηση i ο δείκτης αυτός υπολογίζεται ως εξής:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad , -1 \leq s(i) \leq 1$$

3. Σχεδιασμός γραφικής παράστασης avg.sil - k

4. Το σημείο στο οποίο εμφανίζεται μέγιστο θεωρείται ως ο ιδανικός αριθμός συμπλεγμάτων k.



Προτεινόμενος αριθμός συμπλεγμάτων K = 2

Gap Statistic Method

1. Τοποθετούμε τις παρατηρήσεις σε συμπλέγματα για διάφορες τιμές του k = 1,….,*kmax* και υπολογίζουμε την αντίστοιχη συνολική εσω-συμπλεγματική απόκλιση *W(k)* = tot. withiness

2. Παράγουμε B σύνολα δεδομένων με μία τυχαία ομοιόμορφη κατανομή. Ομαδοποιούμε τα δεδομένα αυτά σε συμπλέγματα για κάθε μία από τιμές του k = 1,....,*kmax* και υπολογίζουμε την αντίστοιχη συνολική εσω-συμπλεγματική απόκλιση *Wkb*.

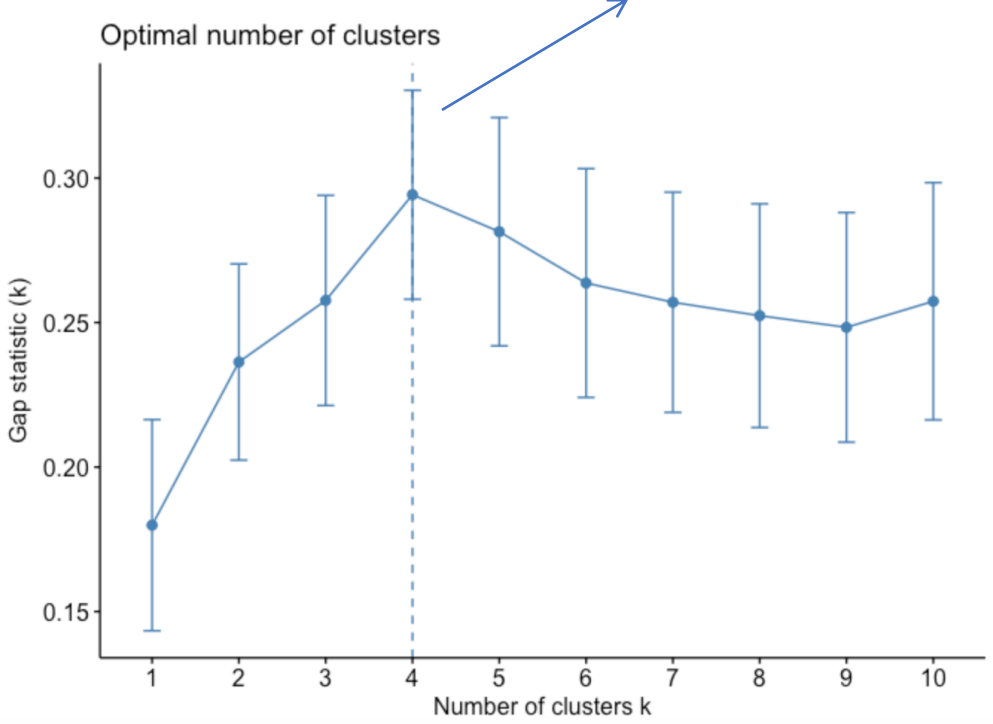
3. Έστω $\bar{w} = \frac{1}{B}\sum_b log(W_{kb})$, υπολογισμός της τυπικής απόκλισης

$$sd(k) = \sqrt{(1/b)\sum_b (log(W_{kb}) - \bar{w})^2} \quad \text{και} \quad s_k = sd(k) \times \sqrt{1 + 1/B}$$

4. Επιλέγουμε τον μικρότερο αριθμό k για τον οποίο ισχύει:

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

Προτεινόμενος αριθμός συμπλεγμάτων K = 4



Γενίκευση K means

Συνάρτηση Σφάλματος

$$f_{\Sigma}(C) := \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{m} dist(x_{ij}, \hat{z}_{kj})$$

Μετρική	Μέσος	Αντικειμενική Συνάρτηση
Manhattan L_1 $d(x,y) = L_1 = \sum_{i=1}^{m} \lvert x_i - y_i \rvert$	Διάμεσος	Ελαχιστοποίηση αθροίσματος της L_1 απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο
Τετραγωνική Ευκλείδεια L_2 $d(x,y) = L_2 = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$	Μέση Τιμή	Ελαχιστοποίηση αθροίσματος της L_2 απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο
cosine $sin(x,y) = cos(x,y) = \frac{\sum_{i=1}^{m} x_i * y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} * \sqrt{\sum_{i=1}^{m} y_i^2}}$	Μέση Τιμή	Ελαχιστοποίηση αθροίσματος της cosine απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο
Απόκλιση του Bregman $D_F(p,q) = F(p) - F(q) - < \nabla F(q), p - q >$	Μέση Τιμή	Ελαχιστοποίηση αθροίσματος της απόκλισης του Bregman απόστασης κάθε παρατήρησης από τον αντίστοιχο μέσο

Πλεονεκτήματα – Μειονεκτήματα K means

+

✓ Απλότητα

Απλότητα τόσο στην κατανόηση, όσο και στην υλοποίηση.

✓ Αποδοτικότητα

Χρονική πολυπλοκότητα : $O(LKnm) \approx O(n)$
Χωρική Πολυπλοκότητα : $O((n+K)m) \approx O(n)$

✓ Σφαιρικά Συμπλέγματα

Καλά αποτελέσματα όταν η μορφή των συμπλεγμάτων είναι σφαιρική, δηλαδή όταν χαρακτηριστικά των παρατηρήσεων εντός κάθε συμπλέγματος έχουν κοινή διασπορά και είναι ανεξάρτητα μεταξύ τους.

−

✓ Τοπικό Ελάχιστο

Η διαμέριση που επιστρέφει αποτελεί τοπικό ελάχιστο και εξαρτάται από την αρχικοποίηση των συμπλεγμάτων. Διαφορετικά τρεξίματα του αλγορίθμου επιστρέφουν διαφορετικά αποτελέσματα.

✓ Ακραίες Παρατηρήσεις - Μέγεθος Κλίμακας

Το αποτέλεσμα του αλγορίθμου επηρεάζεται αρκετά από ακραίες παρατηρήσεις, καθώς επίσης και από τη διαφορά στην κλίμακα των χαρακτηριστικών.

✓ Διαφορετικό μέγεθος, πυκνότητα και μη σφαιρικό σχήμα

Η ακρίβεια της διαμέρισης δεν είναι ιδιαίτερα καλή όταν τα συμπλέγματα διαφέρουν σε μέγεθός, πυκνότητα αλλά και όταν έχουν σφαιρικό σχήμα.

✓ Πολυδιάστατα δεδομένα

Η απόδοση του αλγορίθμου επηρεάζεται αρνητικά, όταν το πλήθος των χαρακτηριστικών m για κάθε παρατήρηση είναι μεγάλο (Keim, Hinneburg, 1999).

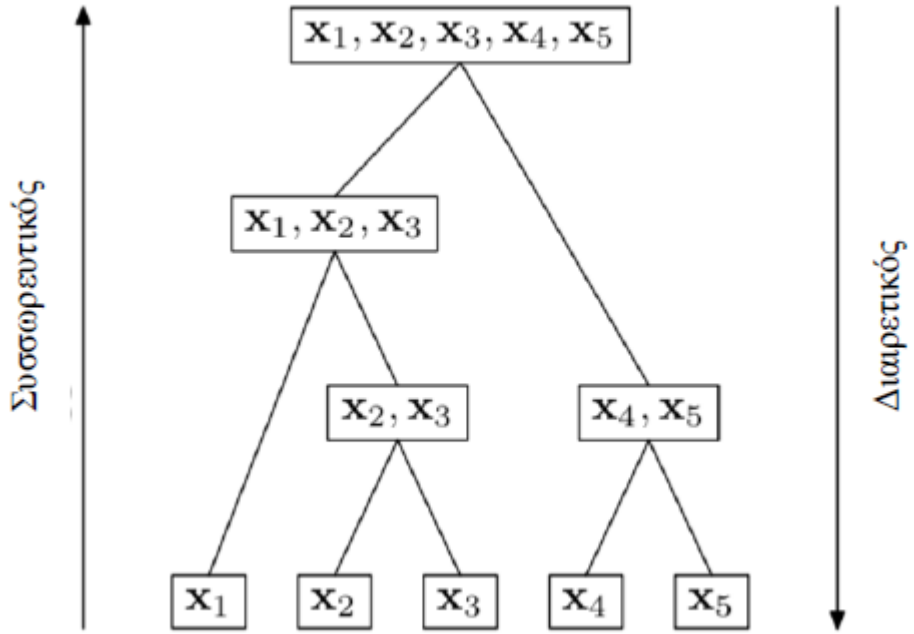
✓ Ομοιόμορφο αποτέλεσμα

Ομοιομορφία συμπλεγμάτων, ακόμα και για σύνολα παρατηρήσεων για τα οποία αυτό δεν ισχύει στη πραγματικότητα.

Αλγόριθμος Single Link

Δεδομένα : Σύνολο των παρατηρήσεων $\Pi = (x_1, x_2, \ldots, x_n)$
Έξοδος : Ιεραρχία συμπλεγμάτων

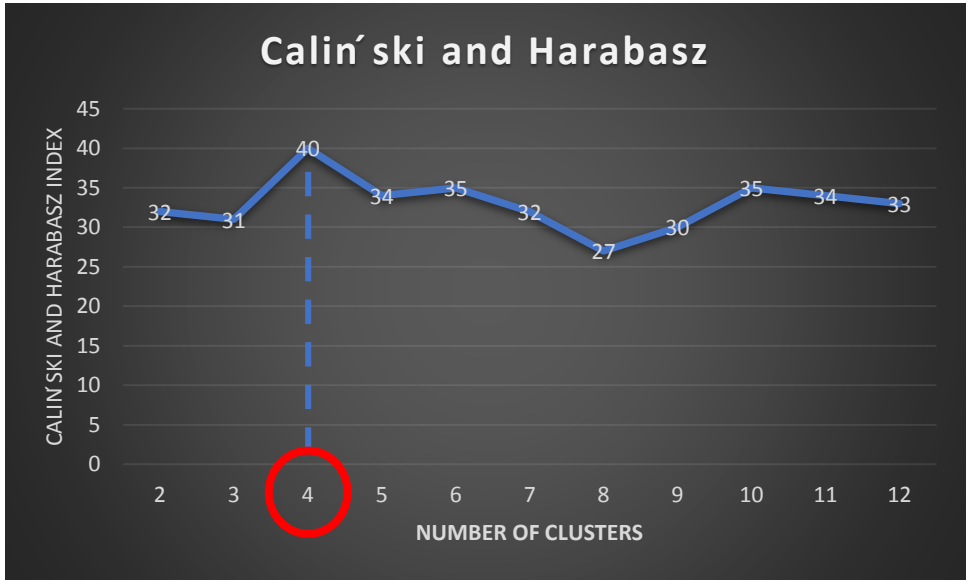
1. Για κάθε $x_i \varepsilon \Pi$ επανάλαβε
2. όρισε ως σύμπλεγμα $C_i = (x_i)$
3. Έστω $C = (C_1, C_2, \ldots, C_n)$
4. Όσο $|C| \neq 1$ κάνε
5. Για όλα τα ζευγάρια συμπλεγμάτων $< C_i, C_{j \neq i} > \varepsilon C x C$ επανάλαβε
6. υπολόγισε $d(C_i, C_j)$
7. Όρισε $best(C_i, C_j) = \forall < C_{k \neq i}, C_{l \neq k,j} > \varepsilon C x C : [d(C_i, C_j) \leq d(C_k, C_l)]$
8. Για $best(C_i, C_j)$ κάνε
9. όρισε $C_{ij} = C_i \cup C_j$
10. όρισε $C^{new} = C - (C_i, C_j)$
11. όρισε $C = C^{new} \cup C_{ij}$
12. Τέλος

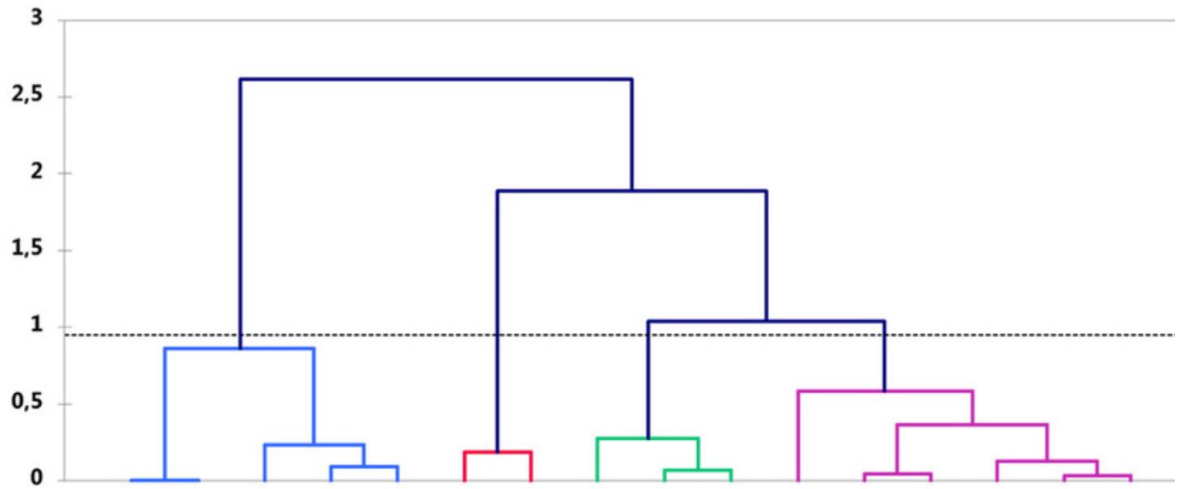


Απόσταση συμπλεγμάτων

$$d(C_i, C_j) = d_{min}(C_i, C_j) = min_{x \varepsilon C_i, y \varepsilon C_j} f(x, y)$$

Προσδιορισμός Διαμέρισης





- Υπολογισμός της ποσότητας για τιμές του K σε κάποιο διάστημα, π.χ. απο 1 έως 12.

$$VRC_k = \frac{trace(B)}{trace(W)} x \frac{n-k}{k-1}$$

$$W = \sum_{i=1}^{k} \sum_{l=1}^{n_i} (\vec{x}_i(l) - \bar{x}_i)(\vec{x}_i(l) - \bar{x}_i)' \qquad B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

- Γραφική παράσταση VRC – k

- Επιλογή σημείου μεγίστου

Υποσύνολα που ενώνονται σε απόσταση κάτω από αυτή την τιμή τοποθετούνται στο ίδιο σύμπλεγμα. Αντιθέτως, υποσύνολα που ενώνονται σε απόσταση μεγαλύτερη από αυτή την τιμή τοποθετούνται σε διαφορετικά συμπλέγματα.

Πλεονεκτήματα – Μειονεκτήματα Single Link

+

✓ Απλή υλοποίηση και εφαρμοφή

Δεν προϋποθέτει τον προσδιορισμό του αριθμού των συμπλεγμάτων εκ των προτέρων.

✓ Έλλειψη Τυχαιότητας

Διαδοχικές εφαρμογές της μεθόδου έχουν το ίδιο αποτέλεσμα.

✓ Δημιουργία εμφωλευμένων συσχετίσεων

Εκτός της τελικής διαμέρισης, εξετάζει και τις επιμέρους συσχετίσεις μεταξύ των παρατηρήσεων.

✓ Καλή εφαρμογή στα μη ελλειπτικά συμπλέγματα

Κατάλληλος για την αναγνώριση μη ελλειπτικών δομών.

−

✓ Χρονική πολυπλοκότητα

Χρονική πολυπλοκότητα : $O(n^2)$
Χωρική πολυπλοκότητα : $O(n^2)$

✓ Φαινόμενο της αλυσίδας

Έχει την τάση να συνδυάζει σε σχετικά χαμηλές τιμές , παρατηρήσεις που συνδέονται με μια σειρά στενών ενδιάμεσων παρατηρήσεων.

✓ Οι αποφάσεις συγχώνευσης είναι τελικές

Αυτή η προσέγγιση εμποδίζει ένα τοπικό κριτήριο βελτιστοποίησης να γίνει ένα κριτήριο ολικής βελτιστοποίησης.

Συμπλέγματα με βάση την πυκνότητα

- **Ορισμός (Άμεσα πυκνά-προσβάσιμη)**

Μία παρατήρηση x θα είναι άμεσα πυκνά-προσβάσιμη από μία παρατήρηση y (με βάση κάποιο Eps και N_{min}) εάν :

1. $x \varepsilon N_{Eps}(y)$

2. $|N_{Eps}(y)| \geq N_{min}$ (συνθήκη κεντρικής παρατήρησης)

- **Ορισμός (Πυκνά-προσβάσιμη)**

Μία παρατήρηση x θα είναι πυκνά-προσβάσιμη από μία παρατήρηση y, εάν υπάρχει ακολουθία παρατηρήσεων $x = x_1, x_2, \ldots, x_i = y$ τέτοια, ώστε κάθε x_l να είναι άμεσα πυκνά-προσβάσιμη από τη x_{l+1} για κάθε $l = 1,2,\ldots,i-1$

- **Ορισμός (Πυκνά-συνδεδεμένη)**

Δύο παρατηρήσεις x και y λέγεται ότι είναι πυκνά-συνδεδεμένες σε σχέση με Eps και N_{min} εάν υπάρχει μία παρατήρηση z τέτοια, ώστε τόσο η x, όσο και η y είναι πυκνά-προσβάσιμες από τη z σε σχέση με Eps κα N_{min} .

- **Ορισμός (Σύμπλεγμα)**

Εάν $\Pi = \{x_1, x_2, \ldots, x_n\}$ το σύνολο των παρατηρήσεων, ένα μη κενό υποσύνολο C του Π θα καλείται σύμπλεγμα, αν ικανοποιεί τις ακόλουθες προϋποθέσεις:

1. $\forall x, y \varepsilon \Pi$, εάν $x \varepsilon C$ και y είναι πυκνά-προσβάσιμη από τη x δεδομένου των Eps και N_{min}, τότε $y \varepsilon C$

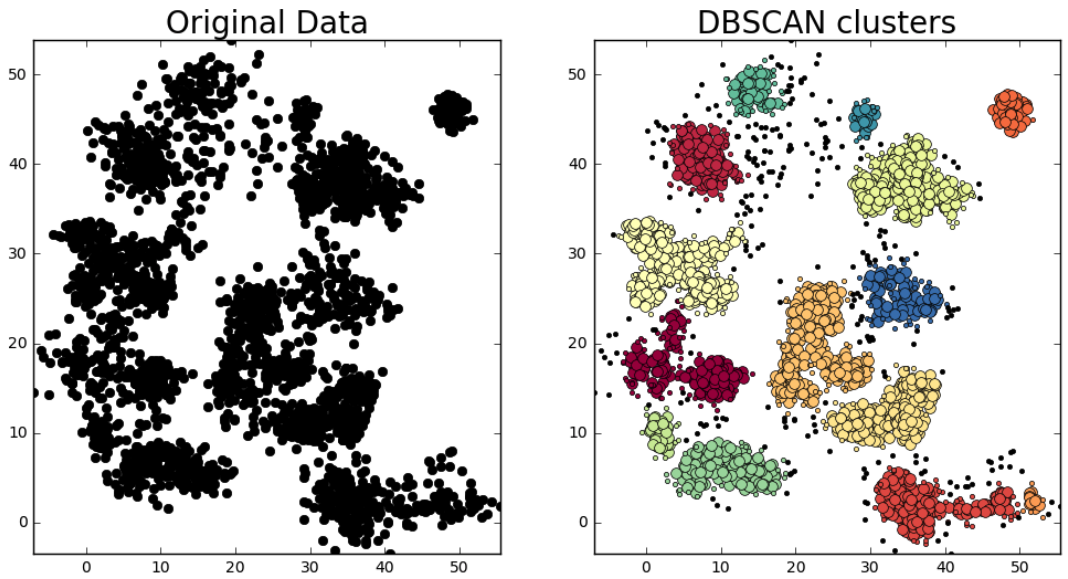
2. $\forall x, y \varepsilon C$, x και y είναι πυκνά-συνδεδεμένες δεδομένου των Eps και N_{min},

Αλγόριθμος DBSCAN

Δεδομένα : Σύνολο των παρατηρήσεων Π , ακτίνα Eps, ελ.αριθμός παρατηρήσεων N_{min}

Έξοδος : ClusterId η οποία εκχωρεί μια ετικέτα συμπλέγματος σε κάθε παρατήρηση

1. ClusterId = ετικέτα για το πρώτο σύμπλεγμα
2. Για κάθε παρατήρηση p στο Π επανάλαβε
3. Εάν (p.Clusterid = UNCLASSIFIED) τότε
4. Εάν ExpandCluster(Π,p,ClusterId,Eps,N_{min}) τότε
5. ClusterId = NextId(ClusterId)
6. τέλος
7. τέλος
8. τέλος



Συνάρτηση ExpandCluster

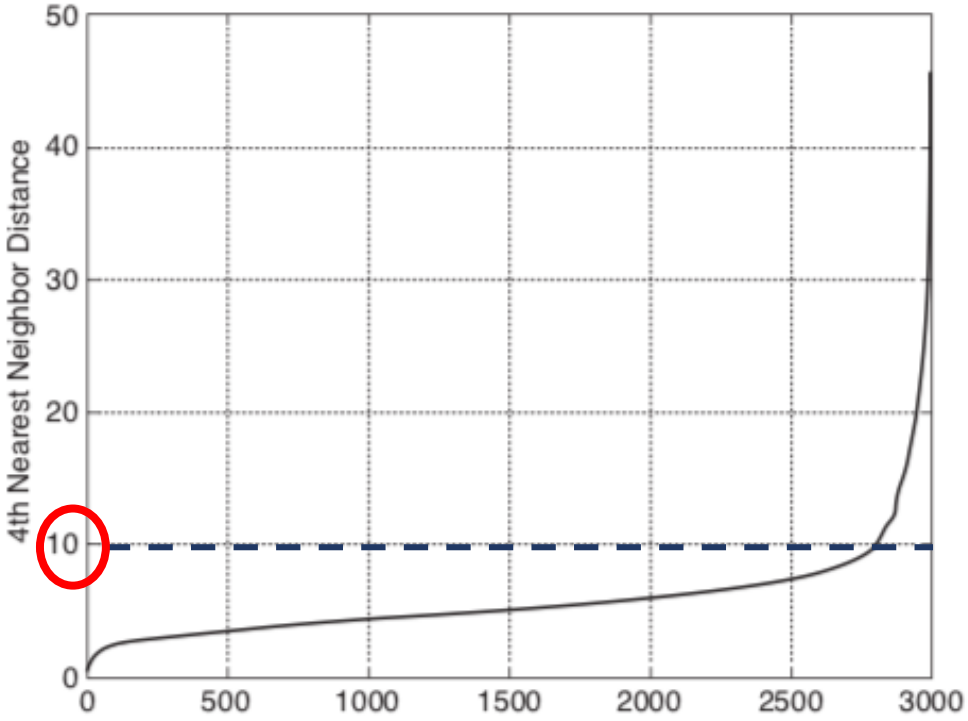
Δεδομένα : Σύνολο Π ,ακτίνα Eps, ελ.αριθμός παρατηρήσεων N_{min}, παρατήρηση p, τρέχον σύμπλεγμα ClId

Έξοδος : Αληθής ή Ψευδής

1. Σ (ουρά προτεραιότητας) = $N_{Eps}(p)$
2. Εάν $|\Sigma| < N_{min}$ τότε
3. p.Clusterid = NOISE
4. επέστρεψε Ψευδής
5. αλλιώς
6. για κάθε q στο Σ επανάλαβε // μαζί με το p
7. q.ClusterId= ClId
8. τέλος
9. διάγραψε p από το Σ
10. Όσο $|\Sigma| > 0$ επανάλαβε
11. curPoint = πρώτη παρατήρηση στο Σ
12. Σ' = $N_{Eps}(curPoint)$
13. Εάν $|\Sigma'| >= N_{min}$ τότε
14. Για κάθε q στο Σ' επανάλαβε
15. Εάν (q.ClusterId = UNCLASSIFIED) τότε
16. q.ClusterId = ClId
17. πρόσθεσε q στο Σ
18. Αλλίως εάν (q.ClusterId = NOISE) τότε
19. q.ClusterId = ClId
20. τέλος
21. τέλος
22. τέλος
23. διάγραψε curPoint από το Σ
24. τέλος

Προσδιορισμός Παραμέτρων Nmin & Eps

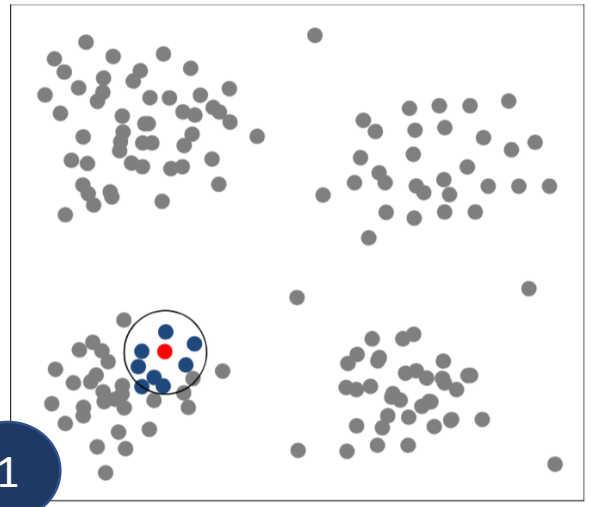
1. Για δεδομένα 2 διαστάσεων επιλέγεται η τιμή Nmin = 4, ενώ για μεγαλύτερες διαστάσεις θέτουμε Nmin = m + 1, όπου m το σύνολο των χαρακτηριστικών κάθε παρατήρησης
2. Υπολογίζουμε το K-dist για κάθε παρατήρηση όπου K = Nmin
3. Ταξινομούμε με αύξουσα ή φθίνουσα σειρά τις παραπάνω τιμές και στη συνέχεια τις σχεδιάζουμε.
4. Απότομη αλλαγή στην τιμή του k-dist αντιστοιχεί σε μία τιμή για την παράμετρο Eps



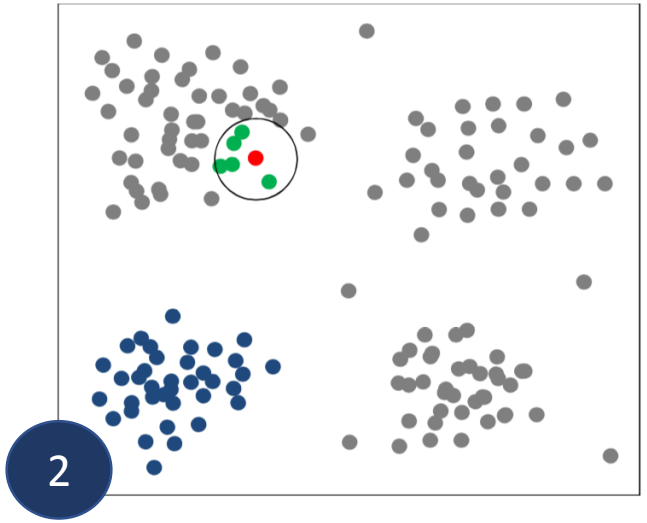
Οι παρατηρήσεις για τις οποίες το k-dist είναι μικρότερο από το Eps=10 θα επισημαίνονται ως κεντρικές, ενώ οι υπόλοιπες θα επισημαίνονται ως θόρυβος ή συνοριακές.

Παράδειγμα εφαρμογής DBSCAN

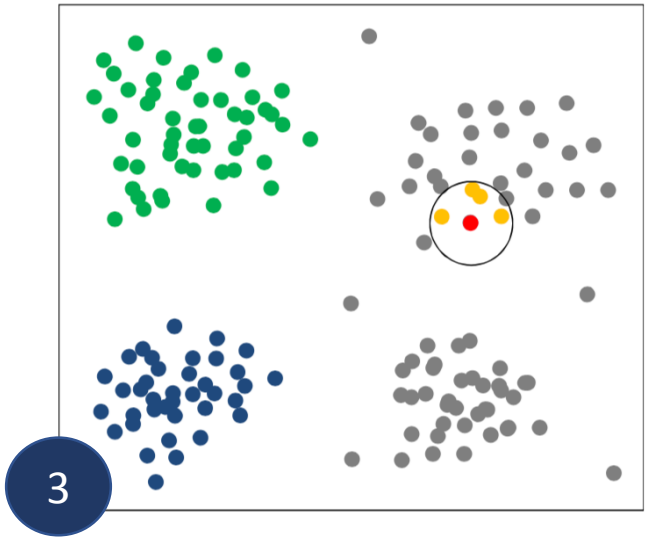
Η γειτονιά της πρώτης κεντρικής παρατήρησης εισάγεται στο σύμπλεγμα.



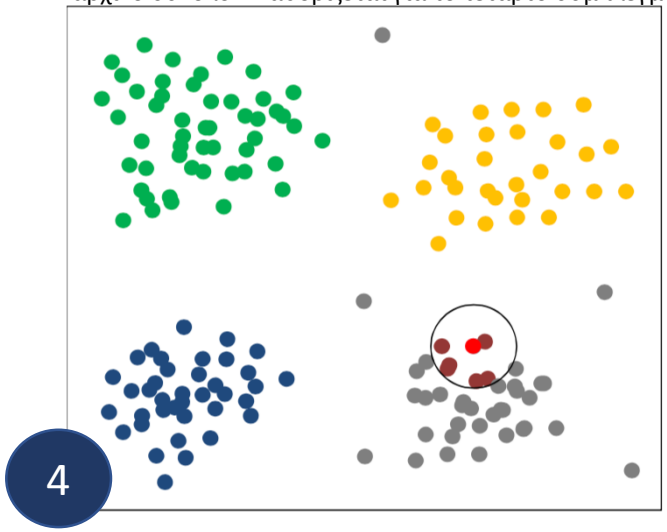
Η μεταγενέστερη αντιστοίχιση πυκνά προσβάσιμων παρατηρήσεων αποτελεί την πρώτη συστάδα. Το αρχικό σύνολο Σ καθορίζεται για το δεύτερο σύμπλεγμα.



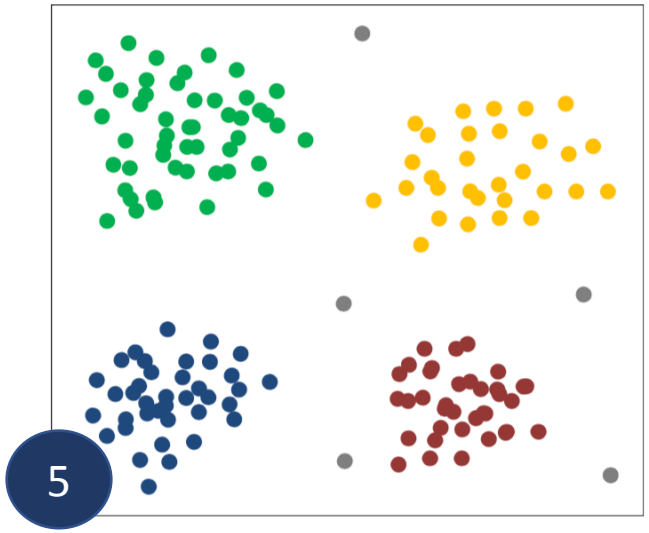
Το δεύτερο σύμπλεγμα φτάνει στο μέγιστο του μέγεθος. Το αρχικό σύνολο Σ καθορίζεται για το τρίτο σύμπλεγμα.



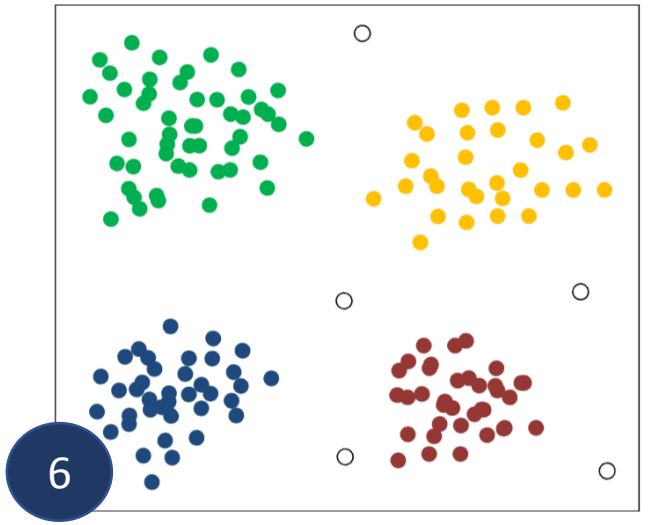
Το τρίτο σύμπλεγμα φτάνει στο μέγιστο του μέγεθος. Το αρχικό σύνολο Σ καθορίζεται για το τέταρτο σύμπλεγμα.



Το τελικό αποτέλεσμα ομαδοποίησης με DBSCAN



Θόρυβος (κενές τελείες)



Πλεονεκτήματα – Μειονεκτήματα DBSCAN

+

✓ Εύρεση συμπλεγμάτων με αυθαίρετα σχήματα και μεγέθη

Έχει τη δυνατότητα εντοπισμού συμπλεγμάτων που περιβάλονται από άλλα, αλλά και συμπλέγματα ενσωματωμένα μέσα σε θόρυβο.

✓ Απλότητα στην εφαρμογή

Η εφαρμογή του αλγορίθμου δεν προαπαιτεί τη γνώση του αριθμού των συμπλεγμάτων.

−

✓ Συμπλέγματα διαφορετικής πυκνότητας

Μπορεί να έχει προβλήματα με την πυκνότητα, εάν η πυκνότητα των συμπλεγμάτων διαφέρει αρκετά

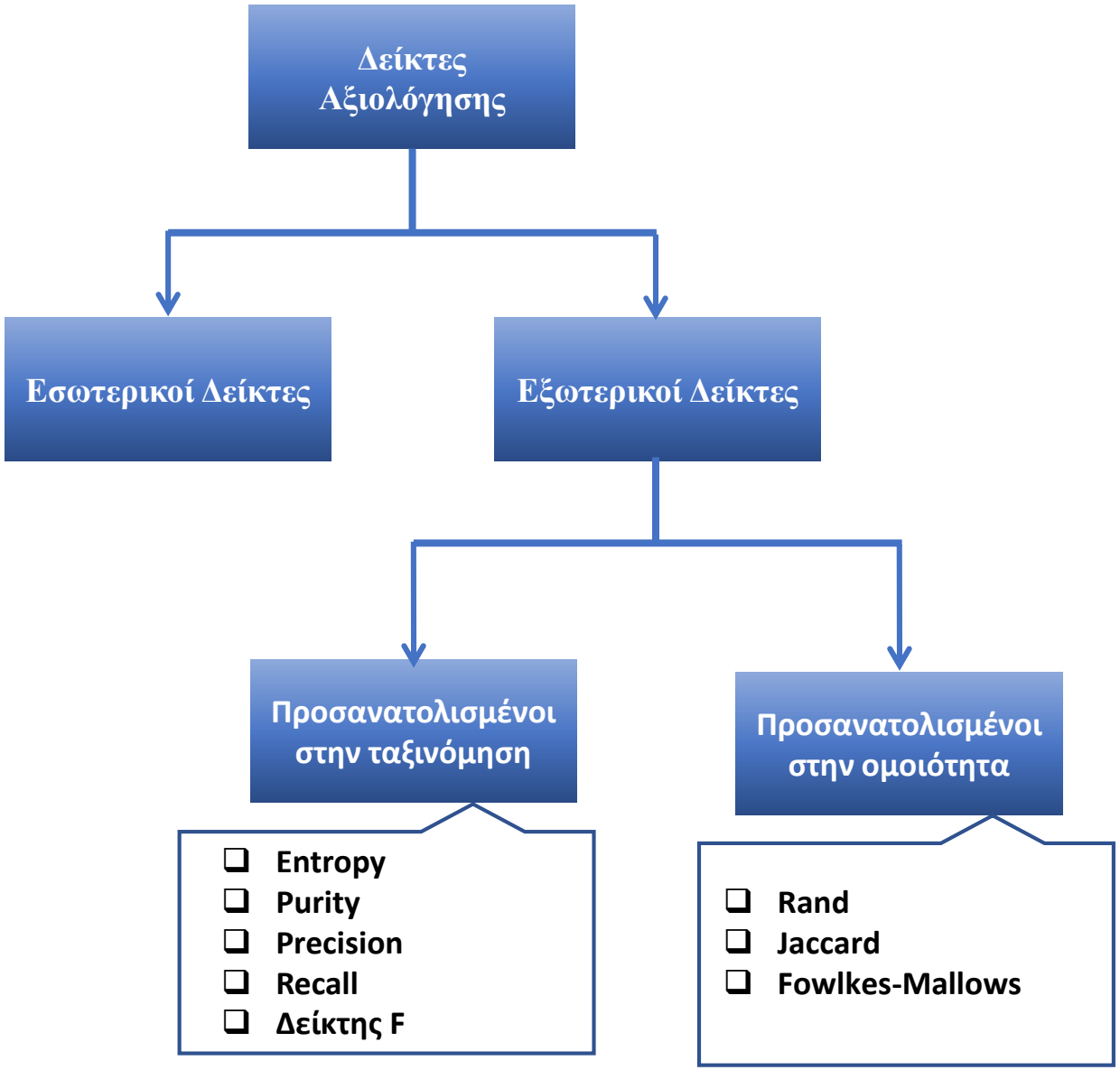
✓ Δεν είναι εξ ολοκλήρου ντετερμινιστικός

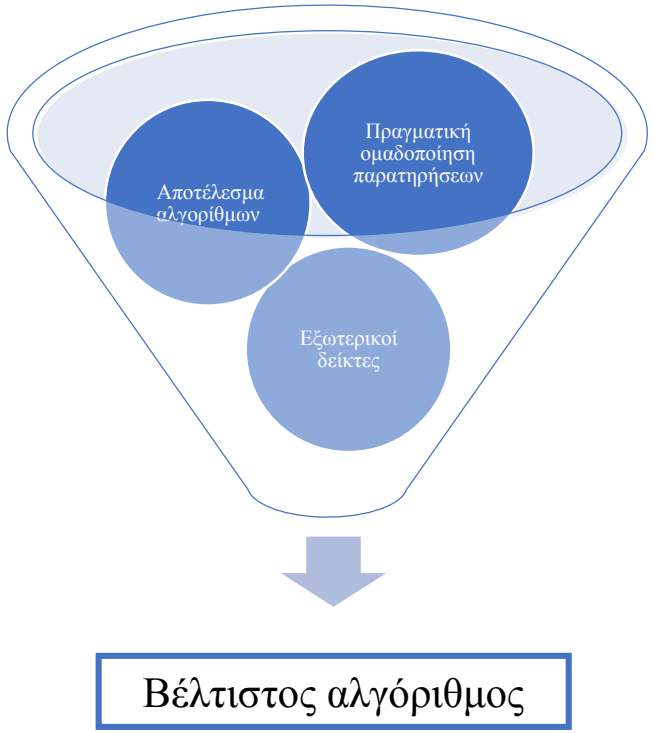
Οι συνοριακές παρατηρήσεις μπορούν να προσεγγιστούν από περισσότερα από ένα συμπλέγματα.

✓ Χρονική πολυπλοκότητα

Η χρονική πολυπλοκότητα της μεθόδου είναι: $O(n^2)$
Η χωρική πολυπλοκότητα της μεθόδου είναι : $O(n)$

Δείκτες Αξιολόγησης

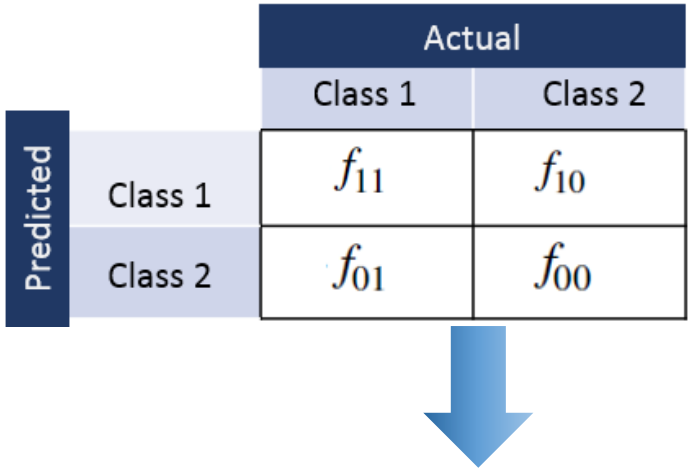




Δείκτες ταξινόμησης

Δείκτης	Τρόπος υπολογισμού	Ερμηνεία
Entropy	$e = \sum_{i=1}^{K} \frac{n_i}{n} e_i \qquad e_i = -\sum_{j=1}^{L} p_{ij} log_2 p_{ij} \qquad p_{ij} = \frac{n_{ij}}{n_i}$	Ο βαθμός στον οποίο κάθε σύμπλεγμα αποτελείται από παρατηρήσεις μιας κλάσης.
Purity	$purity = \sum_{i=1}^{K} \frac{n_i}{n} p_i \qquad p_i = max(p_{ij}) \qquad j = 1,2,...,L$	Ο βαθμός κατά τον οποίο ένα σύμπλεγμα περιέχει παρατηρήσεις μόνο από μία κλάση.
Precision	$precision(i,j) = p_{ij} \qquad p_{ij} = \frac{n_{ij}}{n_i}$	Το μέγεθος του τμήματος ενός συμπλέγματος που αποτελείται από αντικείμενα συγκεκριμένης κλάσης.
Recall	$recall(i,j) = \frac{n_{ij}}{n_j}$	Ο βαθμός στον οποίο ένα σύμπλεγμα περιέχει όλες τις παρατηρήσεις συγκεκριμένης κλάσης.
Δείκτης F	$F(i,j) = \frac{2 \times precision(i,j) \times recall(i,j)}{precision(i,j) + recall(i,j)}$	Ο βαθμός στον οποίο ένα σύμπλεγμα περιέχει μόνο παρατηρήσεις συγκεκριμένης κλάσης και όλες τις παρατηρήσεις αυτής της κλάσης.

Δείκτες ομοιότητας

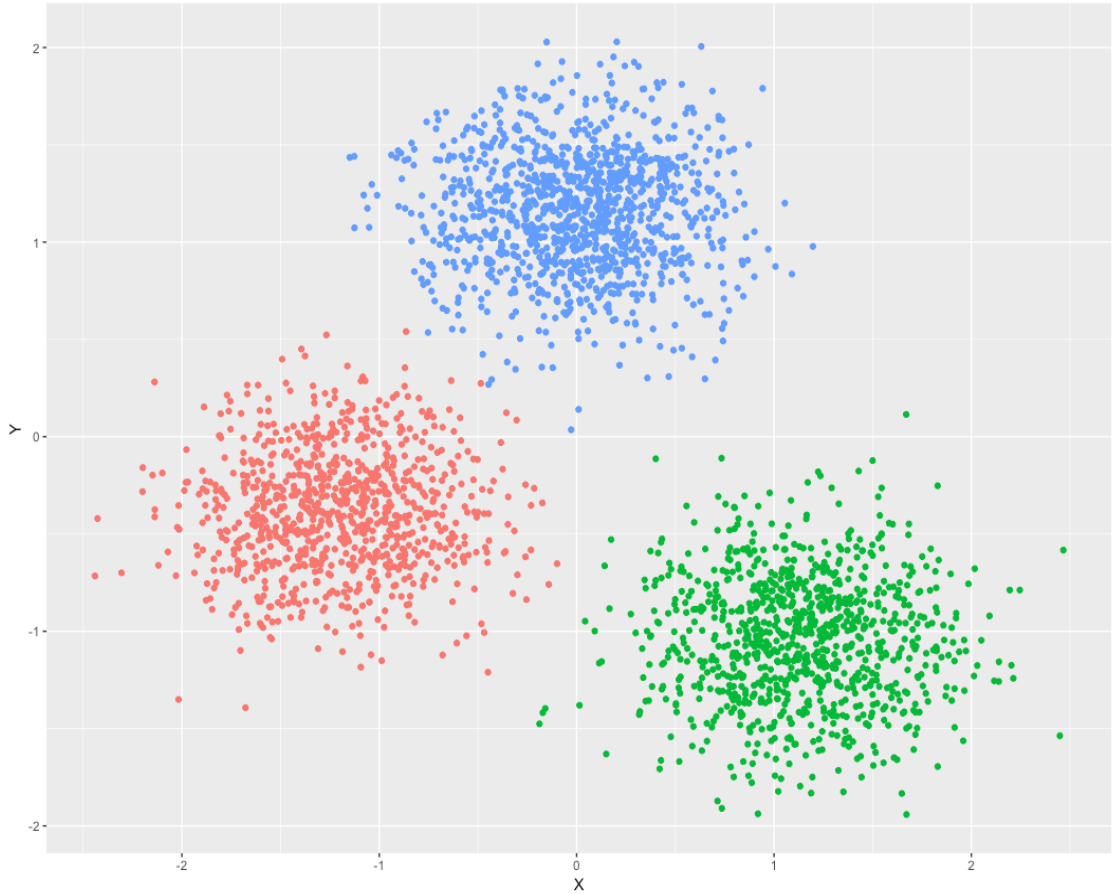


Δείκτης	Τρόπος υπολογισμού	Ερμηνεία
Rand	$Rand = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$	Το ποσοστό των ορθών αποφάσεων που έχουν παραχθεί από τον αλγόριθμο.
Jaccard	$Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$	Ομοιότητα μεταξύ δύο συνόλων.
Fowlkes-Mallows	$FM = \sqrt{\frac{f_{11}}{f_{11} + f_{10}} \times \frac{f_{11}}{f_{11} + f_{01}}}$	Ομοιότητα μεταξύ δύο συνόλων.

Εφαρμογή σε 2d – K means, Single Link

K means

✓ Elbow, Gap Statistic και Silhouette υπέδειξαν K = 3



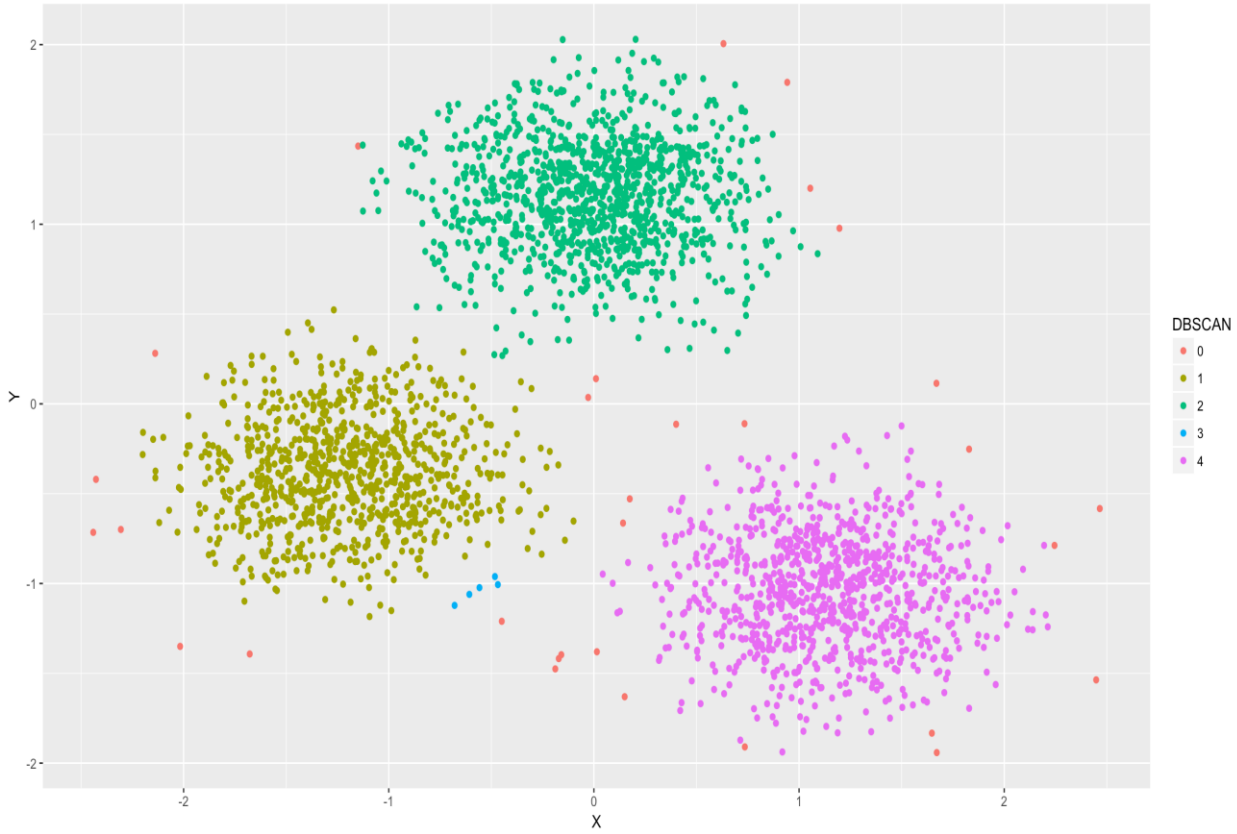
Single Link

✓ Διάγραμμα Calin´ski και Harabasz υπέδειξε K = 3

Εφαρμογή σε 2d – DBSCAN

DBSCAN (Nmin = 4 , Eps = 0.17)

✓ Διάγραμμα 4-NN υπέδειξε Eps = 0.1



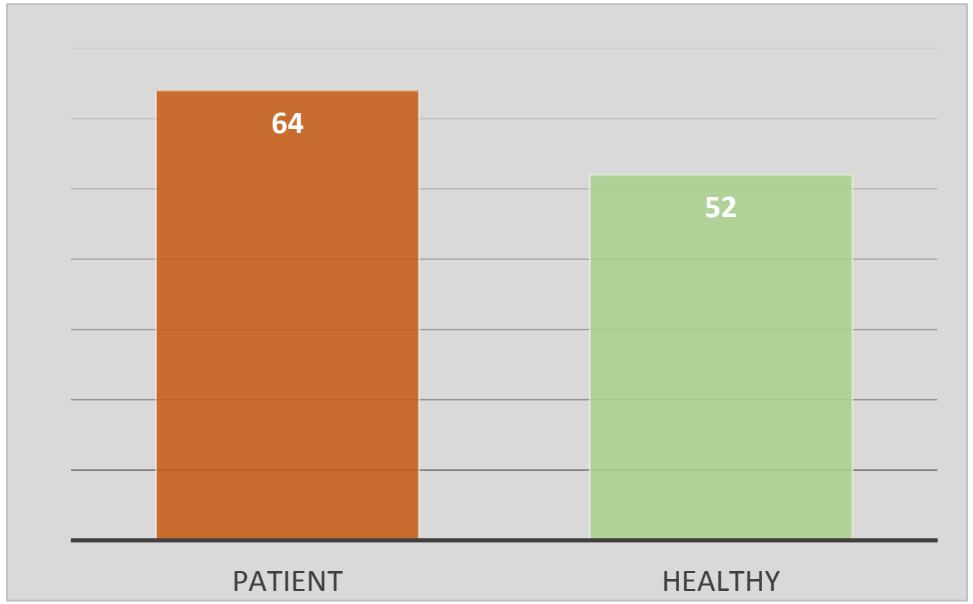
Θόρυβος

Κλάση	K means	Single link	DBSCAN
0	0	0	31
1	897	897	883
2	1151	1151	1146
3	0	0	5
4	952	952	935

Παρουσίαση ιατρικών δεδομένων (1/2)

Οι ανεξάρτητες μεταβλητές αποτελούν ανθρωπομετρικά δεδομένα και παραμέτρους που μπορούν να συγκεντρωθούν με μία απλή ανάλυση αίματος.

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Class
48.0	23.5	70.0	2.71	0.467	8.81	9.70	8.00	417	Healthy
83.0	20.7	92.0	3.12	0.707	8.84	5.43	4.06	469	Healthy
82.0	23.1	91.0	4.50	1.01	17.9	22.4	9.28	555	Healthy
68.0	21.4	77.0	3.23	0.613	9.88	7.17	12.8	928	Healthy
86.0	21.1	92.0	3.55	0.805	6.70	4.82	10.6	774	Healthy
49.0	22.9	92.0	3.23	0,732	6.83	13.7	10.3	530	Healthy

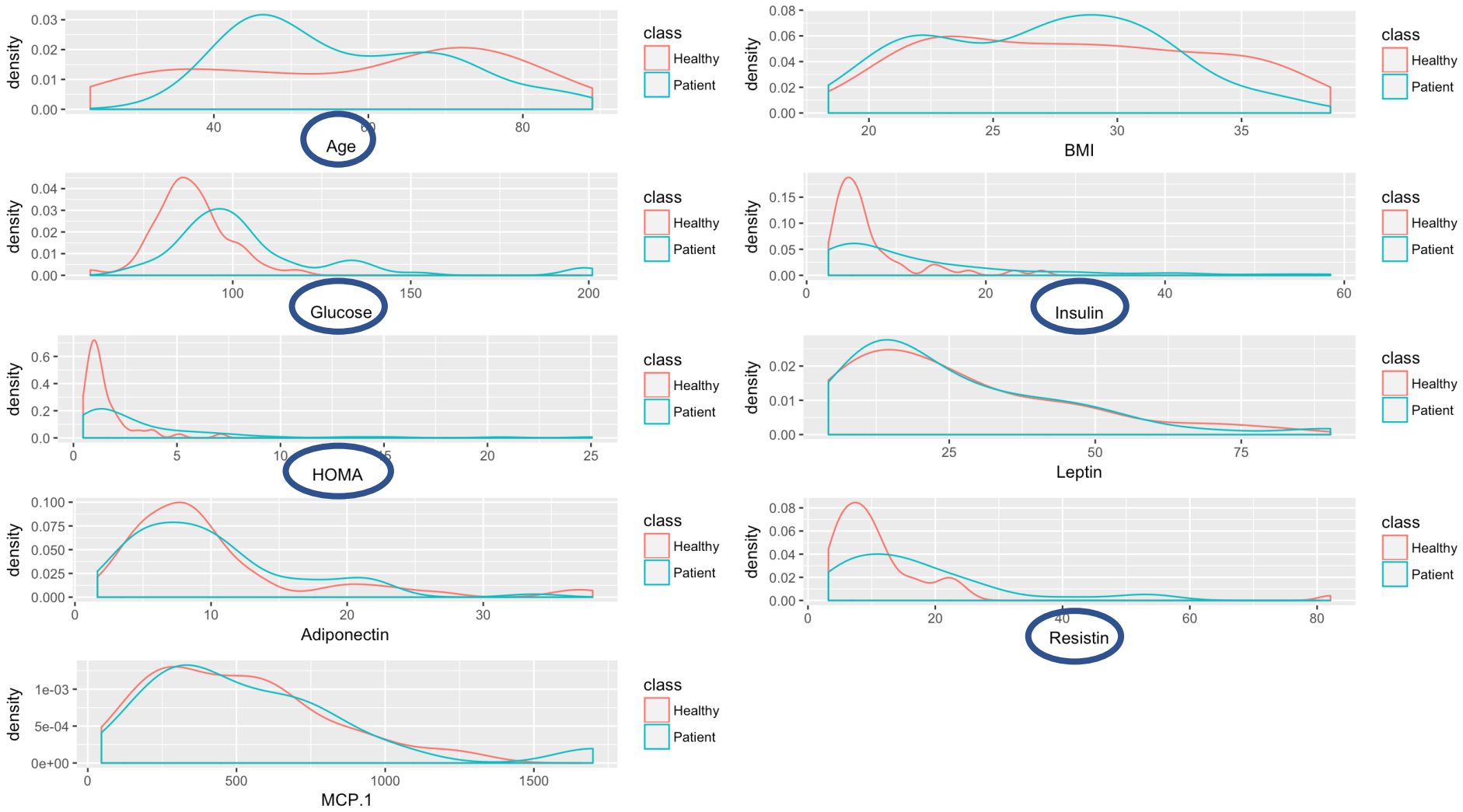


Age		BMI		Glucose		Insulin		HOMA	
Min.	: 24.0	Min.	: 18.37	Min.	: 60.00	Min.	: 2.432	Min.	: 0.4674
1st Qu.	: 45.0	1st Qu.	: 22.97	1st Qu.	: 85.75	1st Qu.	: 4.359	1st Qu.	: 0.9180
Median	: 56.0	Median	: 27.66	Median	: 92.00	Median	: 5.925	Median	: 1.3809
Mean	: 57.3	Mean	: 27.58	Mean	: 97.79	Mean	: 10.012	Mean	: 2.6950
3rd Qu.	: 71.0	3rd Qu.	: 31.24	3rd Qu.	: 102.00	3rd Qu.	: 11.189	3rd Qu.	: 2.8578
Max.	: 89.0	Max.	: 38.58	Max.	: 201.00	Max.	: 58.460	Max.	: 25.0503

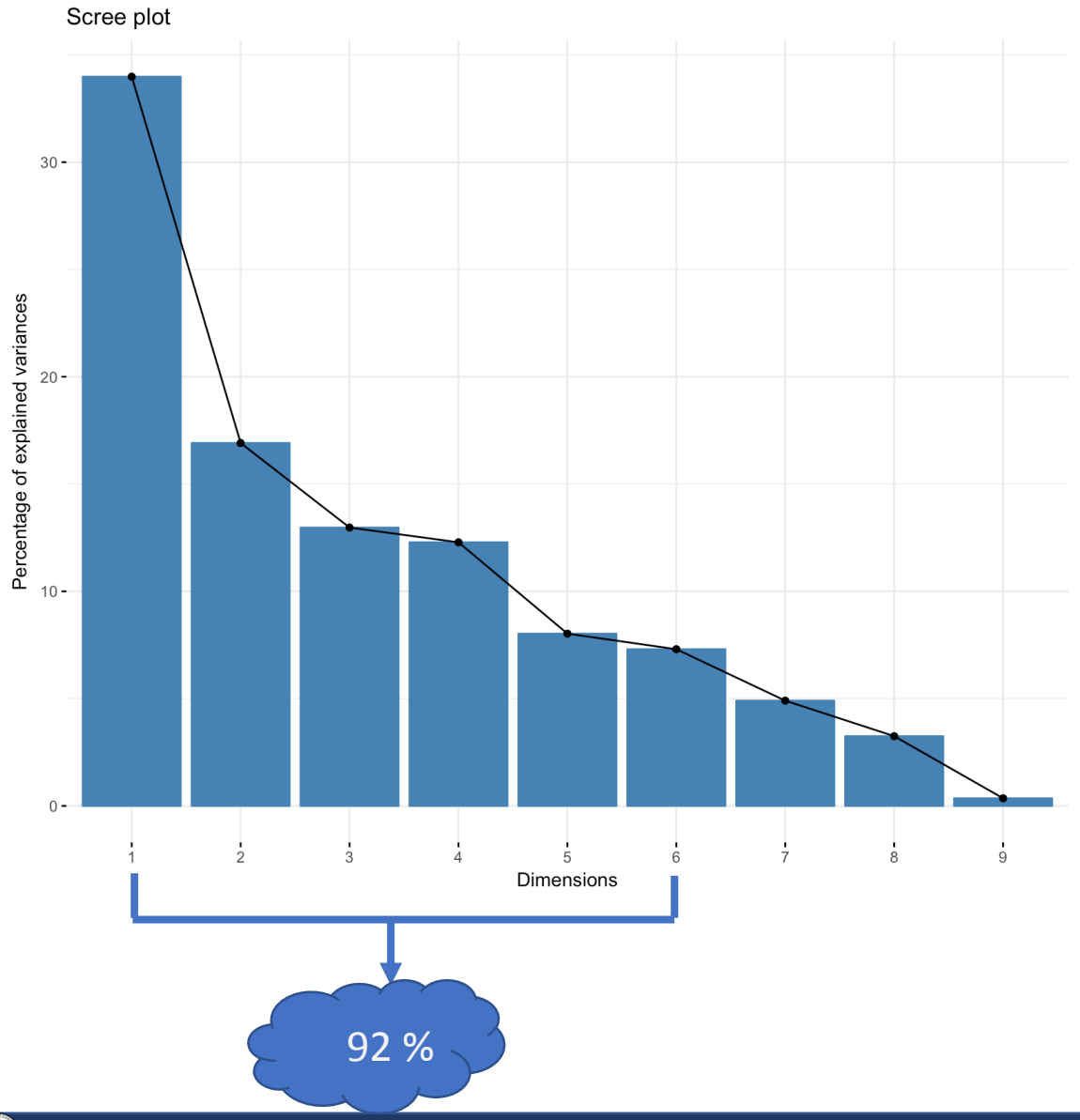
Leptin		Adiponectin		Resistin		MCP.1	
Min.	: 4.311	Min.	: 1.656	Min.	: 3.210	Min.	: 45.84
1st Qu.	: 12.314	1st Qu.	: 5.474	1st Qu.	: 6.882	1st Qu.	: 269.98
Median	: 20.271	Median	: 8.353	Median	: 10.828	Median	: 471.32
Mean	: 26.615	Mean	: 10.181	Mean	: 14.726	Mean	: 534.65
3rd Qu.	: 37.378	3rd Qu.	: 11.816	3rd Qu.	: 17.755	3rd Qu.	: 700.09
Max.	: 90.280	Max.	: 38.040	Max.	: 82.100	Max.	: 1698.44

Παρουσίαση ιατρικών δεδομένων (2/2)

Εκτίμηση των κατανομών των μεταβλητών για κάθε κλάση ξεχωριστά με τη μέθοδο των πυρήνων, χρησιμοποιώντας ως συνάρτηση πυρήνα τη γκαουσιανή κατανομή.



Μείωση Διαστάσεων



Ποσοστό μεταβλητότητας για κάθε βασική συνιστώσα

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.7489	1.2338	1.0805	1.0515	0.85002	0.81073	0.66449	0.54095	0.17894
Proportion of Variance	0.3398	0.1691	0.1297	0.1229	0.08028	0.07303	0.04906	0.03251	0.00356
Cumulative Proportion	0.3398	0.5090	0.6387	0.7615	0.84184	0.91487	0.96393	0.99644	1.00000

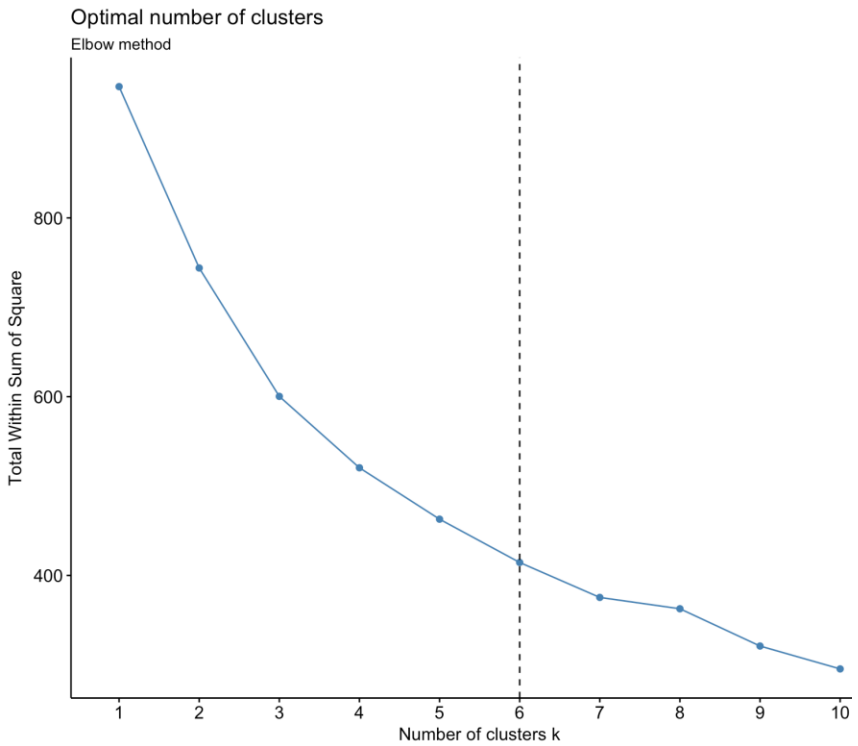
$$x' = \frac{x - \bar{x}}{\sigma}$$

$$\Upsilon_{116\times6} = X_{116\times9} V_{9\times6}$$

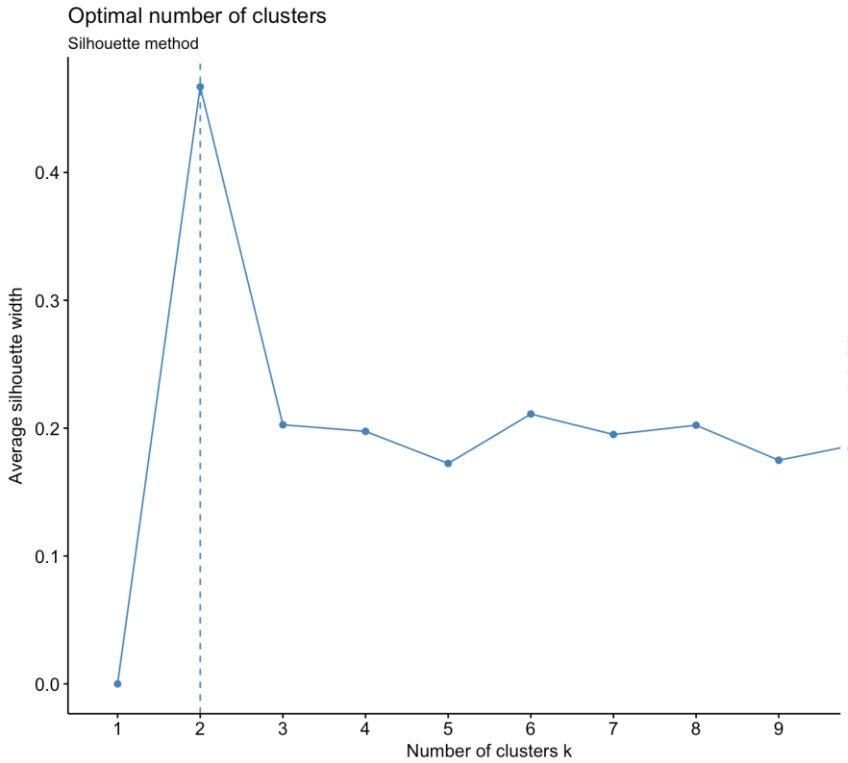
Πίνακας με τα δεξιά ιδιάζουσα ιδιοδιανύσματα της μεθόδου SVD

Εφαμορφή K means – Προσδιορισμός K

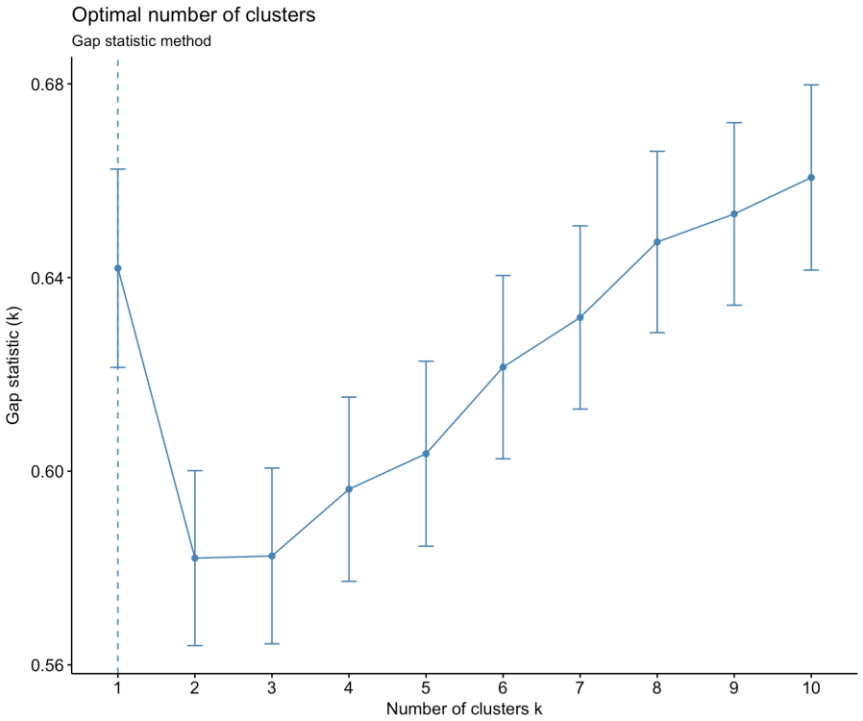
Elbow method



Silhouette method



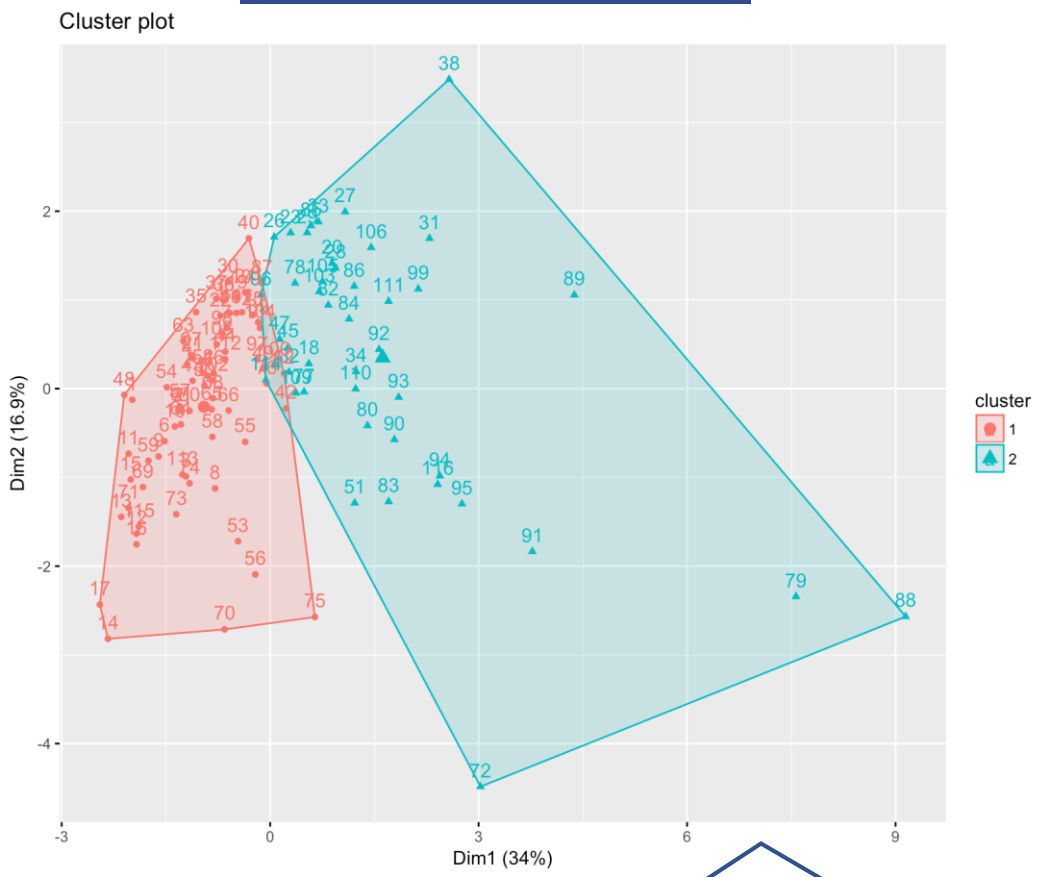
Gap statistic method



Μόνο η Silhouette method εντόπισε τον πραγματικό αριθμό συμπλεγμάτων K

Αποτελέσματα K means

Συμπλέγματα K Means



Προβολή παρατηρήσεων στις 2 πρώτες κύριες συνιστώσες

Αριθμητικοί μέσοι μεθόδου K means

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Healthy	56.64	25.19	90.27	6.080	1.377	16.66	11.07	11.28	462.8
Patient	58.42	31.65	110.6	16.69	4.933	43.52	8.664	20.59	656.7

Πραγματικές τιμές αριθμητικών μέσων

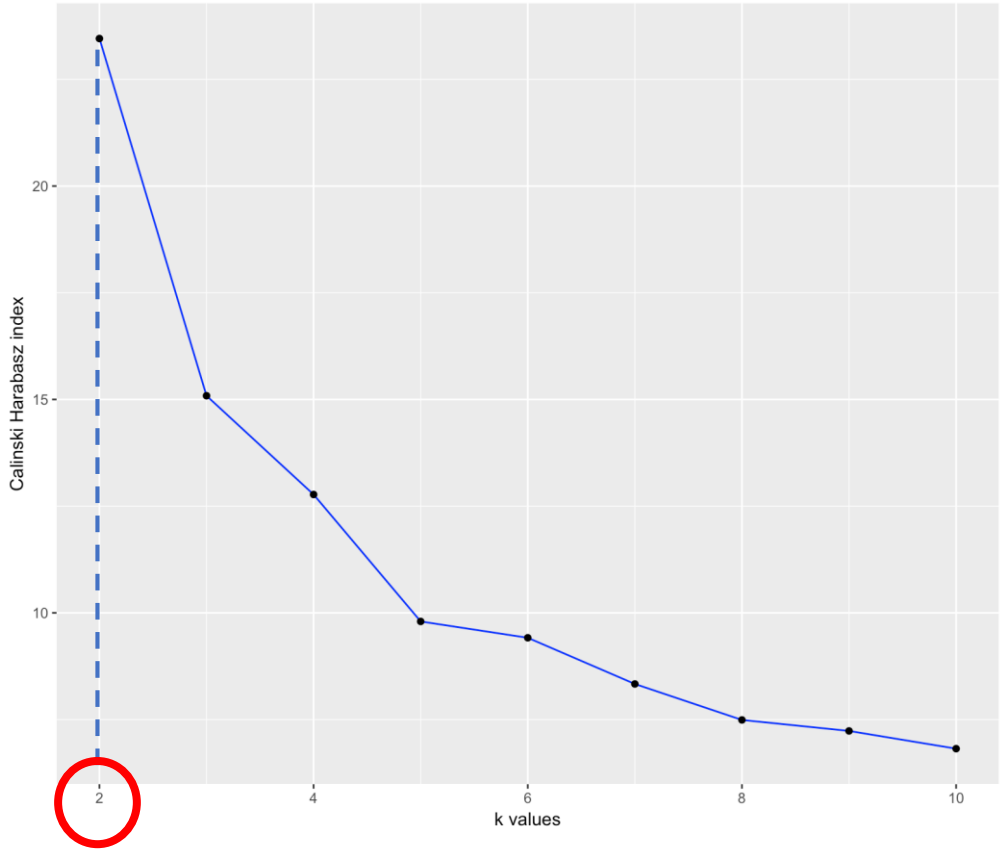
	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Healthy	58.08	28.32	88.23	6.934	1.552	26.64	10.33	11.62	499.7
Patient	56.67	26.98	105.6	12.51	3.623	26.60	10.06	17.25	563.0

		Actual Values		
		Healthy	Patient	
Predicted values	Healthy	$f_{11} = 37$	$f_{10} = 36$	73
	Patient	$f_{01} = 15$	$f_{00} = 28$	43
		52	64	116

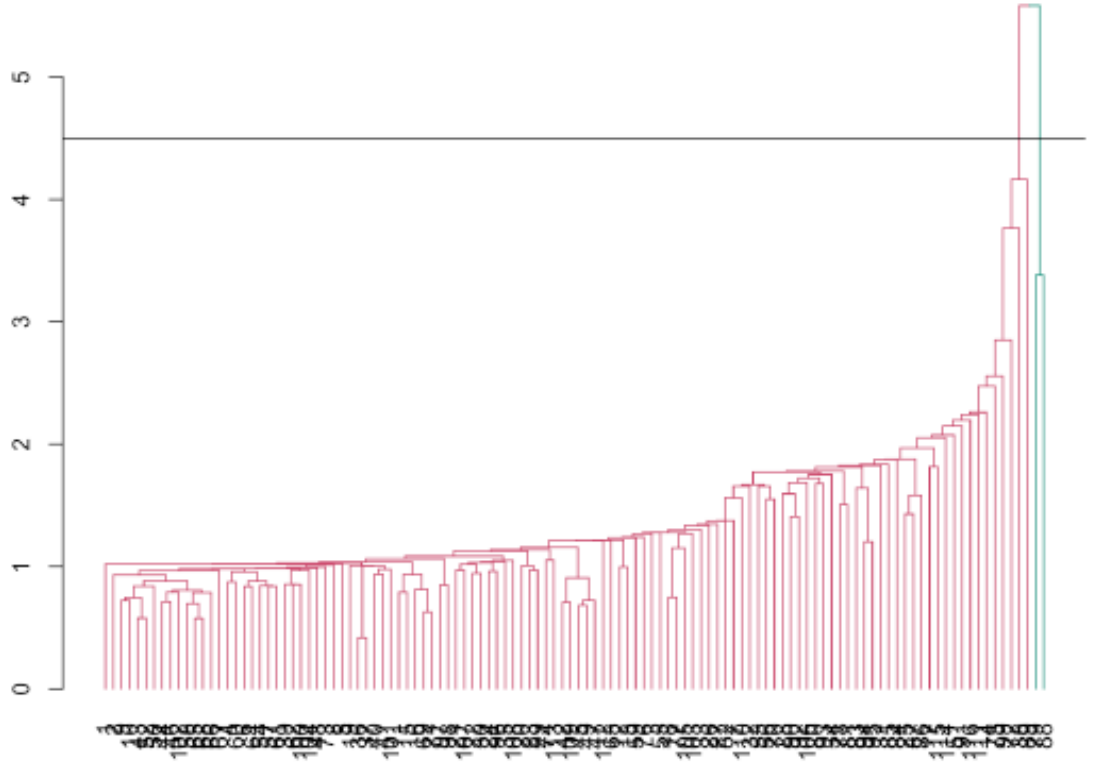
Rand = 0.5603

Εφαμορφή Single Link

Διάγραμμα Calin´ski και Harabasz για κάθε k



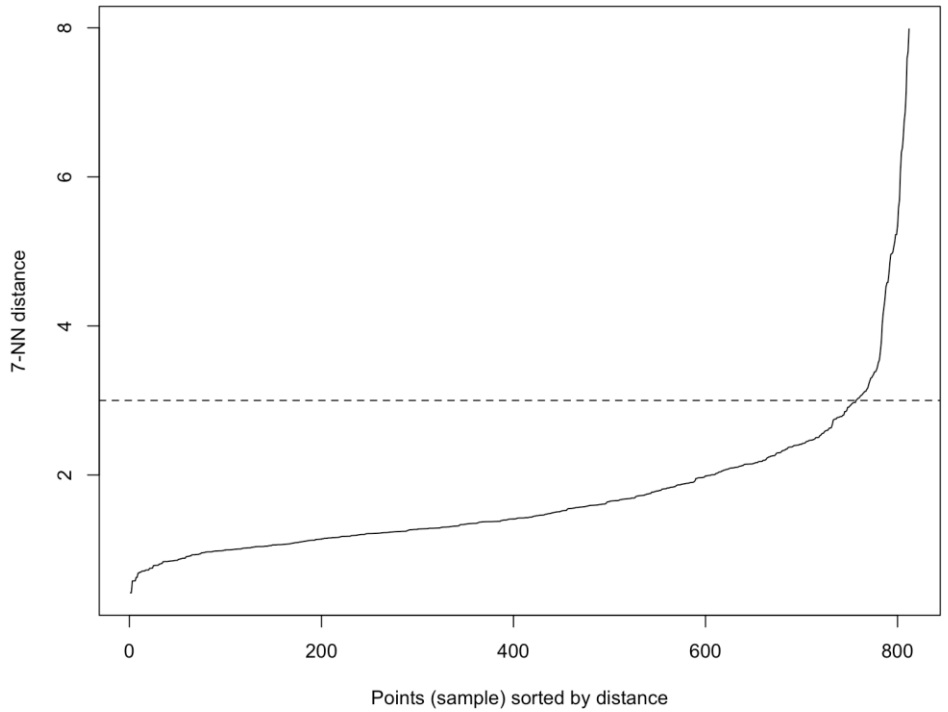
Δενδρογράφημα αλγορίθμου Single Link



Παραγόμενα συμπλέγματα αλγορίθμου Single Link για K = 2

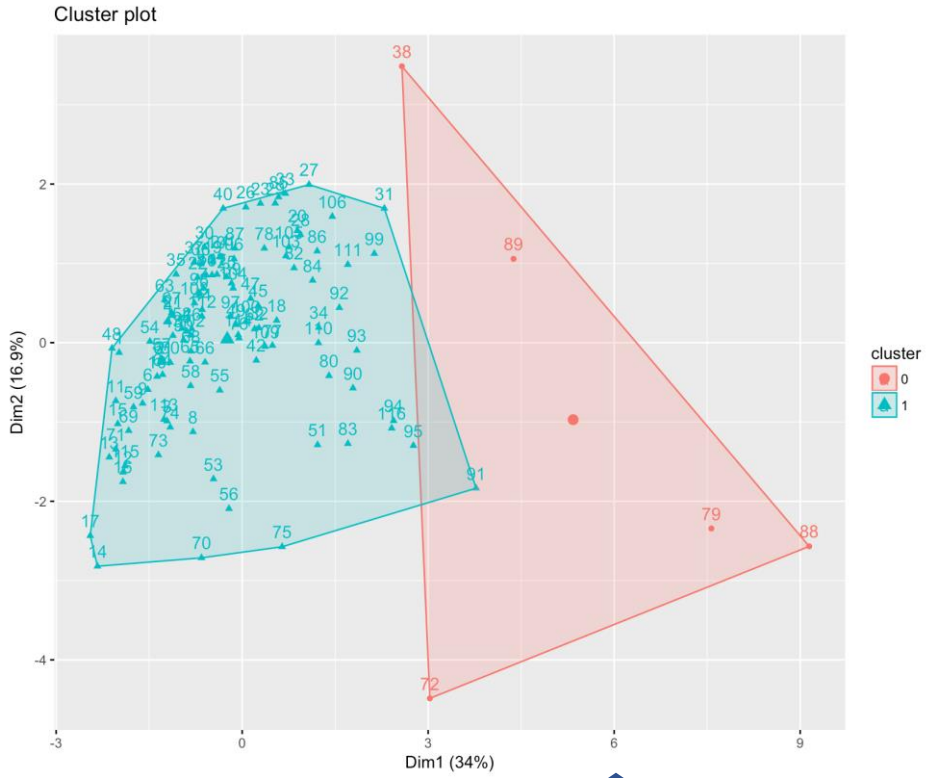
Εφαμορφή DBSCAN (1/2)

Διάγραμμα 7-NN



Προτεινόμενη τιμή
Eps=3

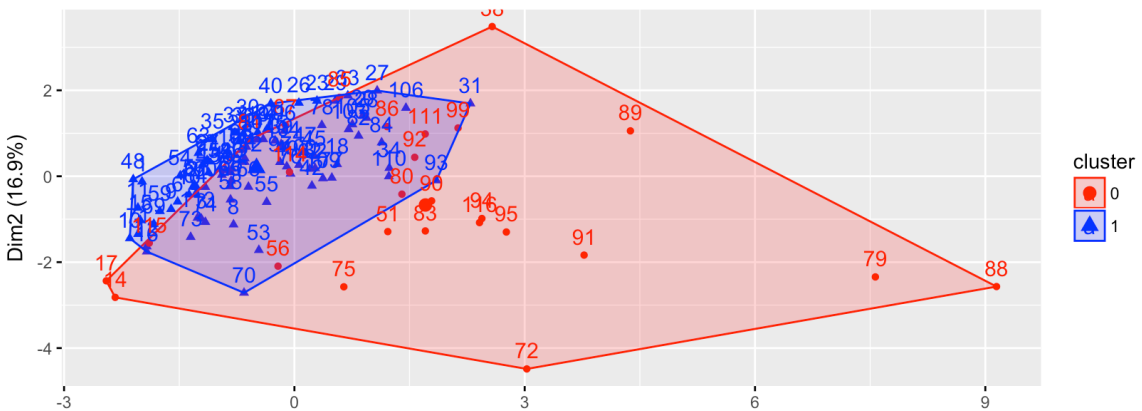
Συμπλέγματα DBSCAN για Nmin=7 & Eps=3



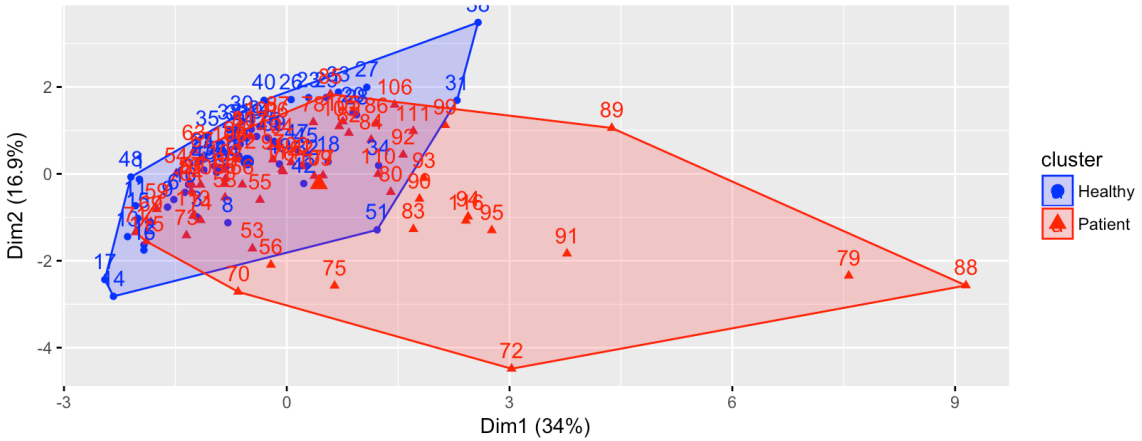
Προβολή παρατηρήσεων στις 2 πρώτες κύριες συνιστώσες

Εφαμορφή DBSCAN (2/2)

Συμπλέγματα DBSCAN για Nmin=7 & Eps=1.9



Πραγματική Ομαδοποίηση των παρατηρήσεων



Ο αλγόριθμος τοποθέτησε την πλειοψηφία των παρατηρήσεων στην κλάση “Healthy”

		Actual Values		
		Healthy	Patient	
Predicted values	Healthy	$f_{11} = 48$	$f_{10} = 42$	90
	Patient	$f_{01} = 4$	$f_{00} = 22$	26
		52	64	116

Rand = 0.6034

Σύγκριση Αποτελεσμάτων

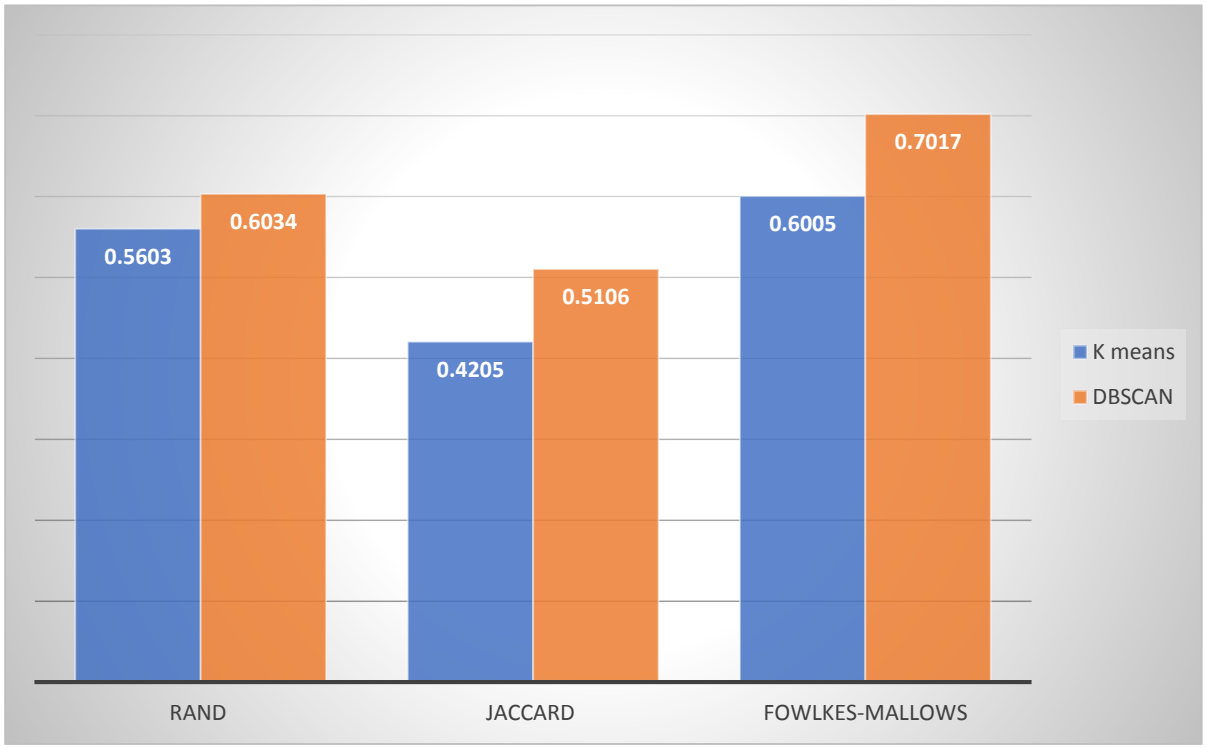
	K means		DBSCAN	
	Entropy	Purity	Entropy	Purity
Healthy	0.9999	0.5070	0.9968	0.5333
Patient	0.9332	0.6510	0.6194	0.8462
Συνολικά	0.9751	0.5604	0.9122	0.6034

	K means		DBSCAN	
	Precision		Precision	
	Healthy	Patient	Healthy	Patient
Healthy	0.507	0.493	0.533	0.467
Patient	0.349	0.651	0.154	0.846

	K means		DBSCAN	
	Recall		Recall	
	Healthy	Patient	Healthy	Patient
Healthy	0.7115	0.5625	0.923	0.656
Patient	0.2885	0.4375	0.077	0.344

	K means		DBSCAN	
	Δείκτης F		Δείκτης F	
	Healthy	Patient	Healthy	Patient
Healthy	0.5921	0.5255	0.6758	0.5454
Patient	0.3159	0.5233	0.1026	0.4891

Ο αλγόριθμος DBSCAN υπερέχει με βάση όλους τους δείκτες.



Συμπεράσματα

1. Οι αλγόριθμοι K means και Single Link πέτυχαν βέλτιστα αποτελέσματα στα δυσδιάστατα δεδομένα, ενώ ο αλγόριθμος DBSCAN δεν κατάφερε να εντοπίσει πλήρως τα 3 φυσικά συμπλέγματα τοποθετώντας μερικές παρατηρήσεις ως θόρυβο.

2. Οι αλγόριθμοι Single Link και K means αποδίδουν χειρότερα στα πολυδιάστατα δεδομένα σε σχέση με τον αλγόριθμο DBSCAN.

3. Η παραμετροποίηση του αλγορίθμου DBSCAN δίνει μεγαλύτερη ελευθερία κινήσεων στο χρήστη ως προς τα παραγόμενα συμπλέγματα.

4. Τα συνολικά αποτελέσματα και των 3 αλγορίθμων δεν είναι ιδιαίτερα ικανοποιητικά για την πρόβλεψη του καρκίνου του μαστού.

Σας Ευχαριστώ!

Ερωτήσεις;