ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

ΕΡΓΑΣΙΑ

Αλγόριθμοι Συσταδοποίησης – Ποιότητα Συσταδοποίησης

Εισαγωγή

Κεντρικό θέμα της εργασίας αποτελούν οι αλγόριθμοι συσταδοποίησης, και ειδικότερα οι αλγόριθμοι k-means και DBSCAN. Γενικά οι αλγόριθμοι συσταδοποίησης χωρίζονται σε 2 κατηγορίες: 1. Παραδοσιακοί(Διαμεριστικός:k-means) και 2.Μοντέρνοι(Πυκνότητας:DBSCAN).

Όσον αφορά τον k-means, κάθε συστάδα αντιπροσωπεύεται από το κέντρο της. Για να βρεθεί η βέλτιστη λύση πρέπει να εξεταστούν όλες οι περιπτώσεις. Ωστόσο, έχει ευαισθησία στα ακραία δεδομένα και ενδέχεται να συγκλίνει σε τοπικό ελάχιστο. Είναι απλός και γρήγορος. Αρχικά, επιλέγουμε k σημεία ως αρχικά κέντρα βάρους(τυχαία επιλογή). Έπειτα, και μέχρι να μη μεταβάλλονται πλέον τα κέντρα, σχηματίζουμε τις συστάδες και μετακινούμε το κέντρο κάθε συστάδας στο μέσο της.

Από την άλλη πλευρά, ο DBSCAN είναι κατάλληλος για ομάδες που έχουν υψηλή πυκνότητα σημείων, οι οποίες είναι διαχωρισμένες από άλλα σημεία χαμηλότερης πυκνότητας. Με τον όρο πυκνότητα εννοούμε τον αριθμό σημείων(min_samples ή MinPts) σε ακτίνα eps(ή Eps). Αρχικά, ο αλγόριθμος ορίζει τα σημεία του συνόλου δεδομένων σε πυρήνα(έχουν πυκνότητα μεγαλύτερη από min_samples), οριακά έχουν πυκνότητα μικρότερη από min_samples, αλλά απέχουν από ένα σημείο πυρήνα απόσταση μικρότερη από eps), και θορύβου(τα υπόλοιπα σημεία). Έπειτα, εξαλείφει τα σημεία θορύβου. Επίσης, κάθε ομάδα συνδεδεμένων σημείων πυρήνα αποτελεί χωριστή συστάδα, στην οποία εκχωρούνται τα σημεία ορίου που συμφωνούν με τον παραπάνω περιορισμό. Γενικότερα, ο συγκεκριμένος αλγόριθμος δεν επηρεάζεται από το θόρυβο και χειρίζεται δεδομένα με διαφορετικά σχήματα και μεγέθη. Αντίθετα, εμφανίζει μεγαλύτερη πολυπλοκότητα από τον αλγόριθμο k-means.

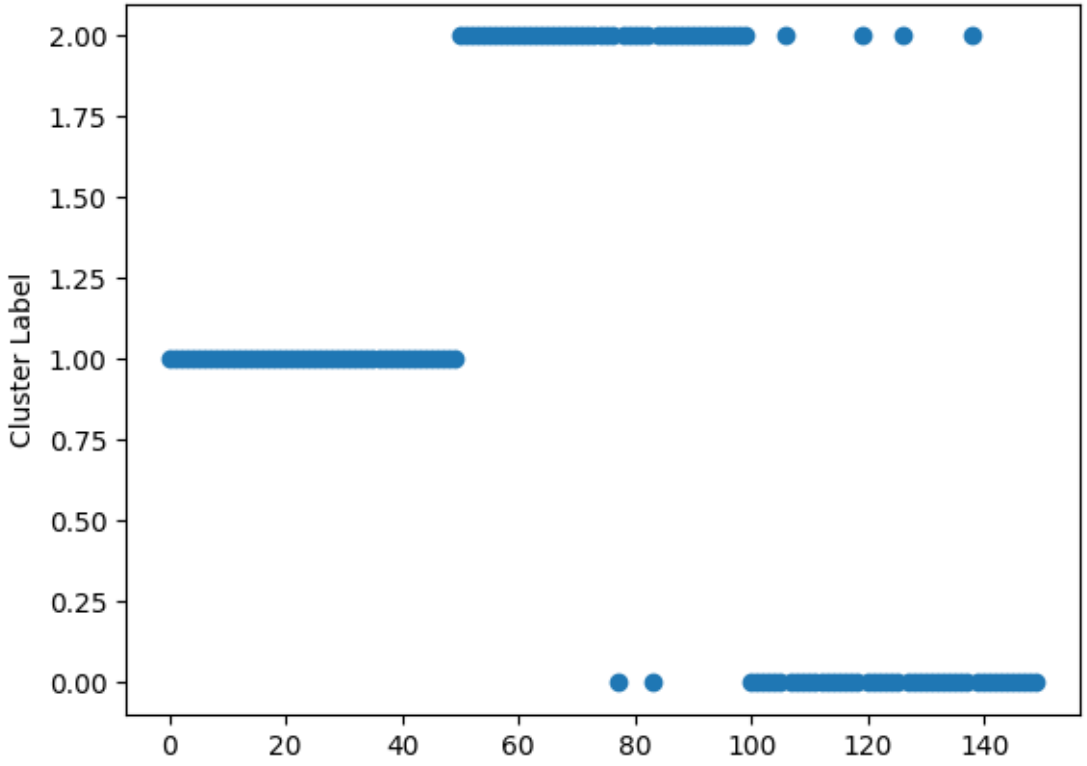
Σε αρκετά σημεία της εργασίας αναφέρονται το Μέσο Τετραγωνικό Σφάλμα(SSE) και ο Συντελεστής Περιγράμματος(Silhouette Coefficient), τα οποία αποτελούν τους τρόπους-μέτρα που ελέγχεται η ποιότητα συσταδοποίησης των παραπάνω αλγορίθμων.

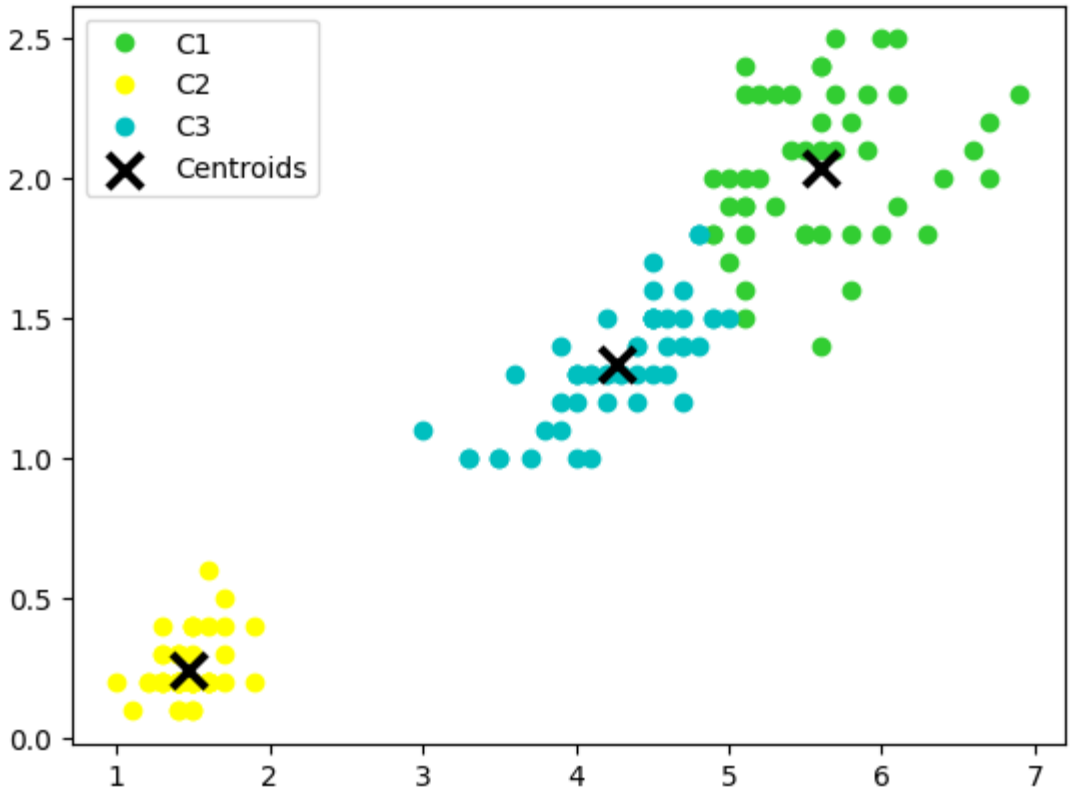
Τέλος, για τη συγγραφή του κώδικα της εργασίας επιλέχθηκε το περιβάλλον ανάπτυξης Dataspell(και το PyCharm για επαλήθευση των αποτελεσμάτων).

Μέρος 1ο: k-means(διαμεριστική συσταδοποίηση)

1.1 Εφαρμογή στο σύνολο iris

Στο πρώτο ερώτημα της εργασίας χρησιμοποιήσαμε το σύνολο iris, που αποτελεί μέρος του sklearn.datasets. Το συγκεκριμένο σύνολο δεδομένων περιέχει τυχαία δείγματα λουλουδιών, και συγκεκριμένα 3 είδη ίριδας. Για κάθε είδος υπαρχουν 50 παρατηρήσεις, οι οποίες αφορούν: μήκος, πλάτος σεπάλου και μήκος, πλάτος πετάλου. Όσον αφορά το κεντρικό θέμα του ερωτήματος, οι παραπάνω παρατηρήσεις ομαδοποιούνται σε 3 συστάδες(k=3). Γίνεται χρήση του αλγόριθμου k-means, που βασίζεται στην Ευκλείδεια Απόσταση, λόγω περιορισμού που θέτει η γλώσσα ανάπτυξης του προγράμματος(Python). Στη συνέχεια, όπως θα γίνει και σε κάθε βήμα της εργασίας, ακολουθούν φωτογραφίες-γραφήματα από την εκτέλεση του κώδικα και εξηγούνται παρακάτω.

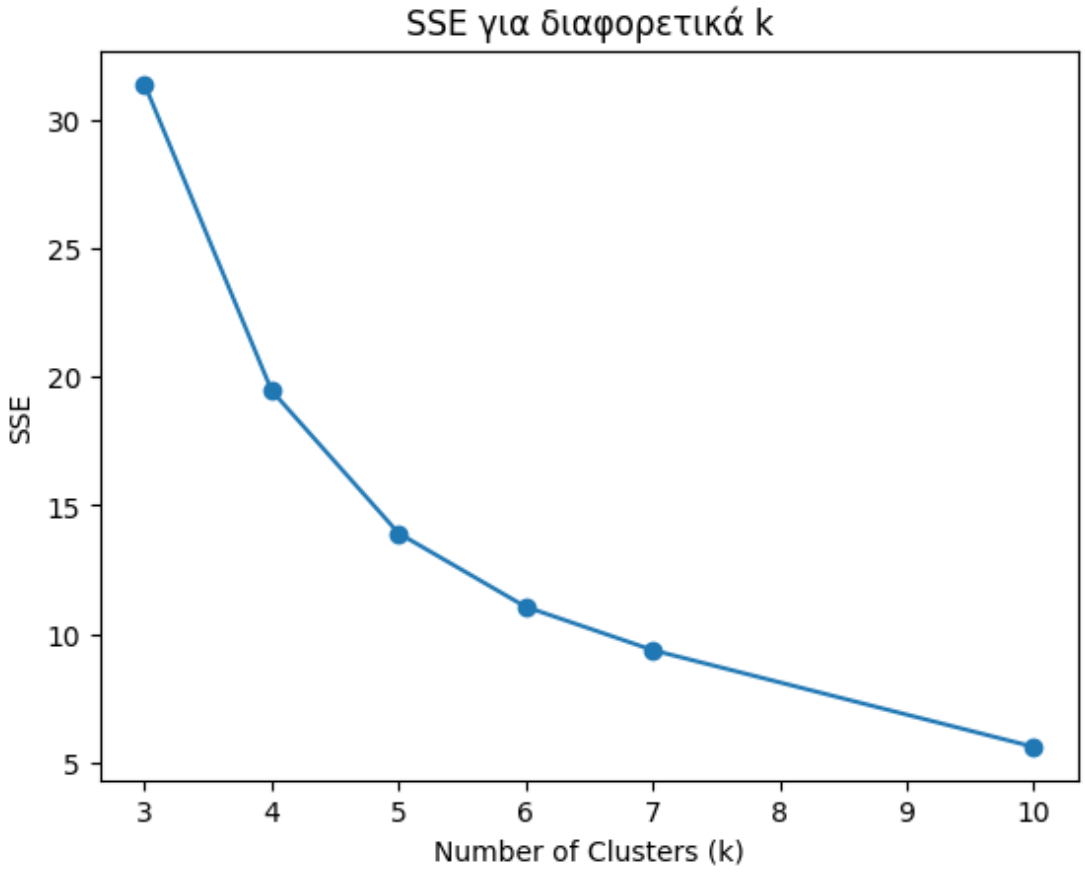




Στο 1ο διάγραμμα απεικονίζονται τα σημεία-δεδομένα, χωριζόμενα σε 3 οριζόντιες γραμμές. Οι τιμές των οριζόντιων αυτών γραμμών είναι οι εξής:0, 1 και 2. Οπότε, καταλαβαίνουμε ότι τα σημεία έχουν οριστεί ανάλογα τη συστάδα στην οποία ανήκουν.

Στο 2ο διάγραμμα παρουσιάζεται το σύνολο δεδομένων, μετά από χρήση του αλγόριθμου συσταδοποίησης k-means με k=3. Οι 3 συστάδες έχουν σχεδιαστεί με διαφορετικά χρώματα(πράσινο, κίτρινο και μπλε αντίστοιχα). Επιπλέον, στο γράφημα υπαρχουν και τα 3 κέντρα των συστάδων, στο μέσο της κάθε συστάδας. Αναλυτικότερα, τα δεδομένα από τη φύση του είναι χωρισμένα σε 2 σχήματα. Μετά την τμηματοποίηση, το κάτω σχήμα αποτελεί τη μία συστάδα. Ενώ, το πάνω σχήμα-πολυπληθέστερο , όπως βλέπουμε το διάγραμμα έχει χωριστεί σε 2 συστάδες.

Τα διαγράμματα και η ανάλυση του ακολουθεί αφορούν το Βήμα 4 του ερωτήματος 1.1. Συγκεκριμένα, γίνεται λόγος για 2 κριτήρια αποτίμησης της ποιότητας της τμηματοποίησης:SSE και Συντελεστής Περιγράμματος. Θέλουμε να μελετήσουμε την επίδραση του αριθμού των συστάδων στην τμηματοποίηση, δηλαδή το πως επηρεάζεται για τις διάφορες τιμές του k.



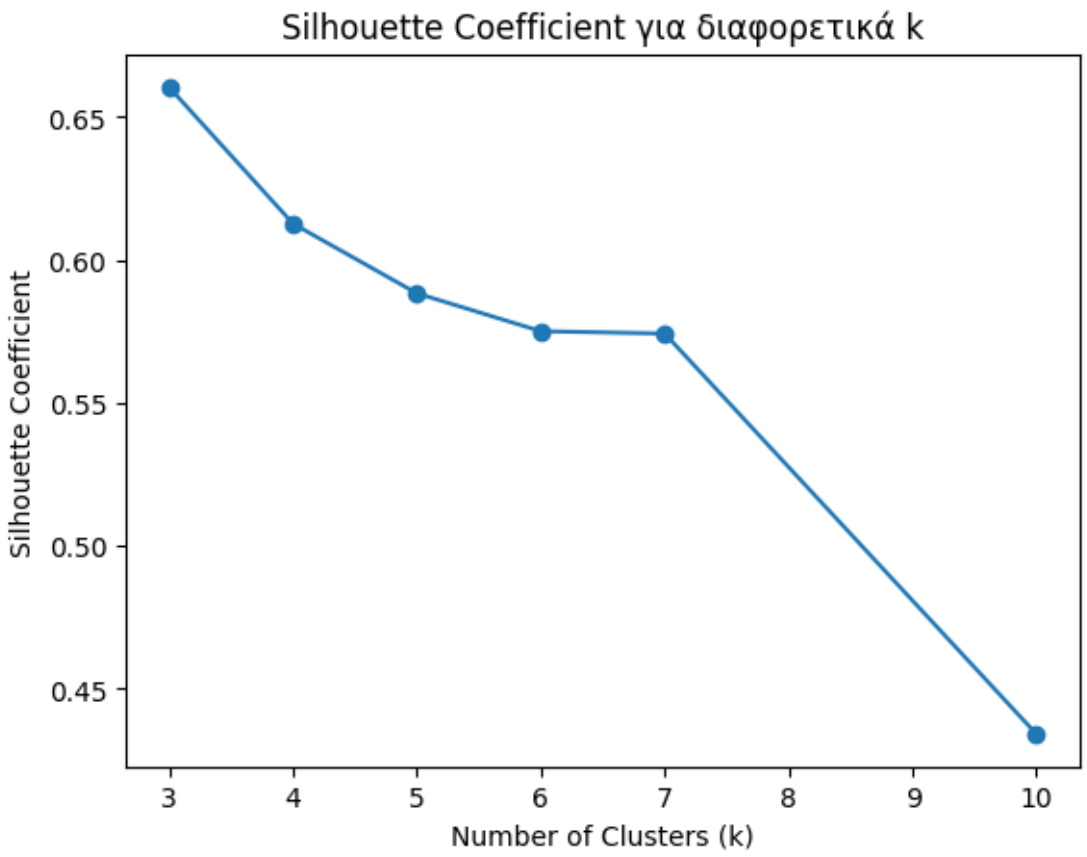
Γραφική παράσταση της σχέσης k - SSE

SSE: Συνολικό Τετραγωνικό Σφάλμα'

Έστω τα σημεία A,B,C και D και τα κέντρα X,Y →

SSE = $[(X-A)^2 + (X-B)^2\ ] + [(Y-C)^2 + (Y-D)^2]$ ($1^{η}$ αγκύλη: αναφέρεται στην πρώτη συστάδα)

Συνοπτικά, υπολογίζει το συνολικό άθροισμα των τετραγώνων των αποστάσεων μεταξύ κάθε δείγματος και του κέντρου της συστάδας στην οποία ανήκει το δείγμα. Αναμφίβολα, όσο μικρότερο είναι το SSE αντιστοιχεί σε καλύτερη συσταδοποίηση. Όσον αφορά τη σχέση του με το k, κατά την επιλογή του k λαμβάνεται σοβαρά υπόψιν ο αριθμός που οδηγεί σε μικρή τιμή του SSE.



Γραφική παράσταση της σχέσης k – Silhouette Coefficient

s_i : Συντελεστής Περιγράμματος του σημείου i

Στο διάγραμμα απεικονίζεται ο μέσος Συντελεστής Περιγράμματος.

$s_i = (b_i - a_i) / \max(b_i , a_i)$ (κυμαίνεται μεταξύ -1 και 1)

Για κάθε σημείο i, το a_i δηλώνει τη μέση απόσταση του i από τα σημεία της συστάδας. Από την άλλη πλευρά, το b_i αποτελεί τη μέση απόσταση κάθε σημείου i από όλα τα σημεία κάθε άλλης συστάδας. Προτιμάται η επιλογή της ελάχιστης τιμής b_i. Πρέπει να αναφερθεί ότι αρνητικές τιμές του συντελεστή είναι ανεπιθύμητες. Παράλληλα, για a_i=0 s_i=max.

Με λίγα λόγια, αποτελεί ένα μέτρο που χρησιμοποιείται για να παρέχει πληροφορίες για την τοποθέτηση ενός δείγματος στο χώρο, σε σχέση με τις συστάδες. Μεγάλες τιμές του συντελεστή σημαίνουν καλή συσταδοποίηση. Άρα, κατά την επιλογή του k υπολογίζεται ο Συντελεστής Περιγράμματος για διάφορες τιμές του k, και επιλέγεται το k που οδηγεί σε μεγάλο s_i , εάν αποτελεί το μοναδικό μέτρο.

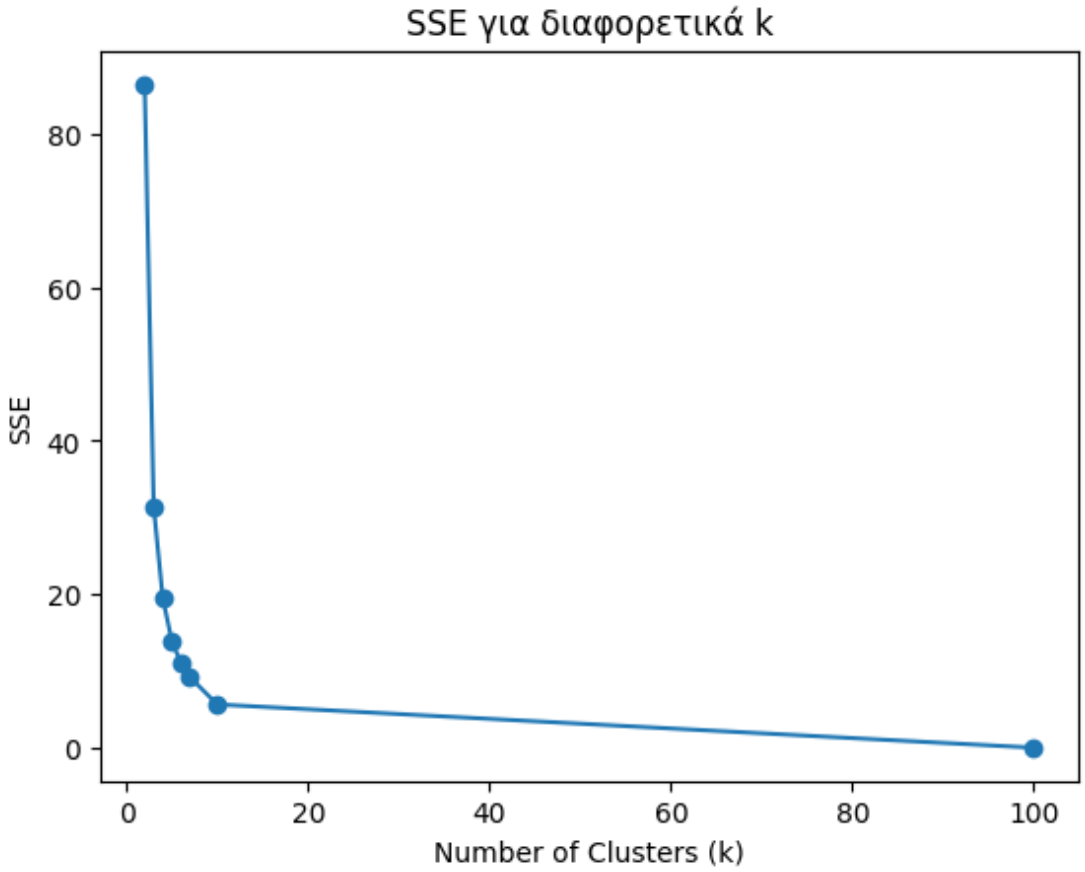
Σε αυτό το σημείο πρέπει να αναφέρουμε ότι σε πραγματικές εφαρμογές δε χρησιμοποιείται μόνο ένα μέτρο υπολογισμού της ποιότητας συσταδοποίησης, αλλά συνδυασμός τους και ανάλυση των αποτελεσμάτων του.

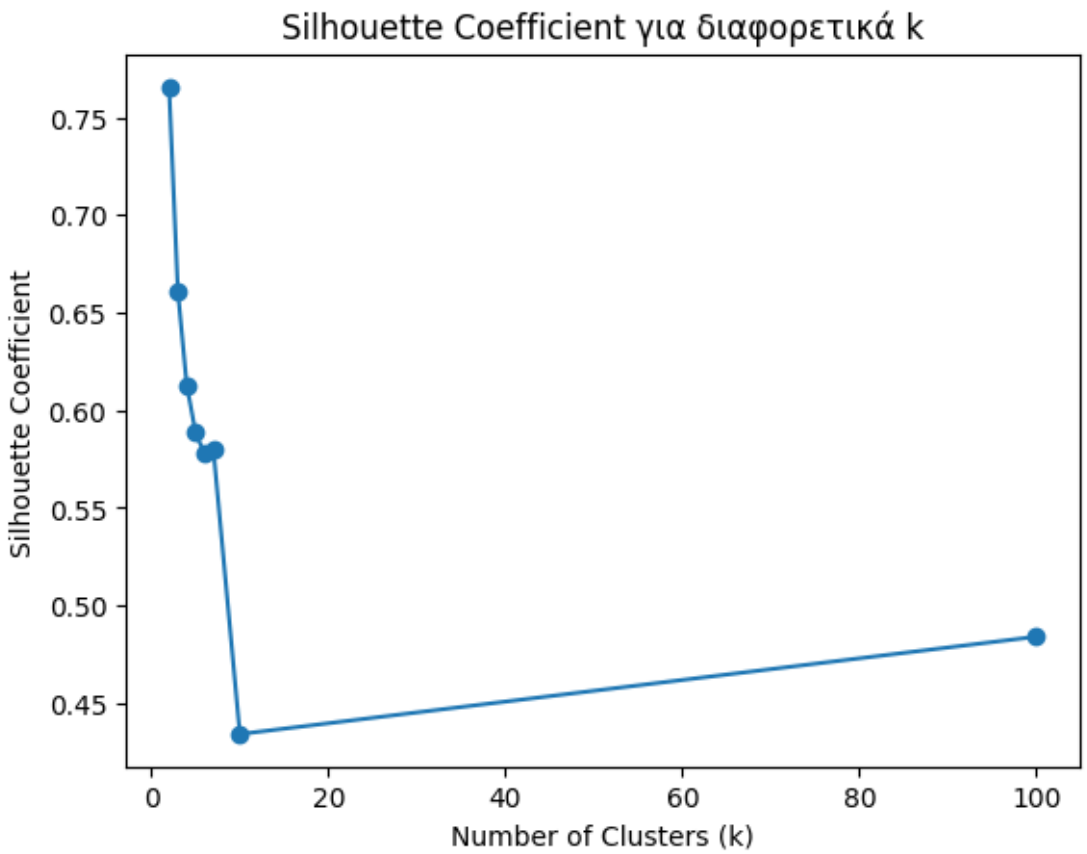
Αναλυτικότερα, όσον αφορά τα 2 παραπάνω διαγράμματα, για k=3: SSE=31.371, Συντελεστής Περιγράμματος=0.660.

Εξετάστηκαν οι τιμές των 2 μέτρων για k=3,4,5,6,7,10. Η τιμή του SSE μειώνεται με την αύξηση του k. Ελάχιστη τιμή σημειώνει για k=10(SSE=6).

Για το 2ο μέτρο: η τιμή του Συντελεστή Περιγράμματος μειώνεται με την αύξηση του k. Ελάχιστη τιμή(0.15) σημειώνει για k=10.

Συμπερασματικά, χρειάζεται αρκετή σκέψη για να αποφασιστεί ποιος είναι ο κατάλληλος αριθμός συστάδων για το συγκεκριμένο παράδειγμα, καθώς οι μικρές τιμές του SSE και τιμές του Συντελεστή Περιγράμματος που πλησιάζουν το 1, είναι συνώνυμες της καλής συσταδοποίησης.





Τέλος, αν στις τιμές του k συμπεριλάβουμε το k=2 και αρκετά μεγαλύτερες τιμές(μέχρι το 100) παρατηρούμε τα εξής:

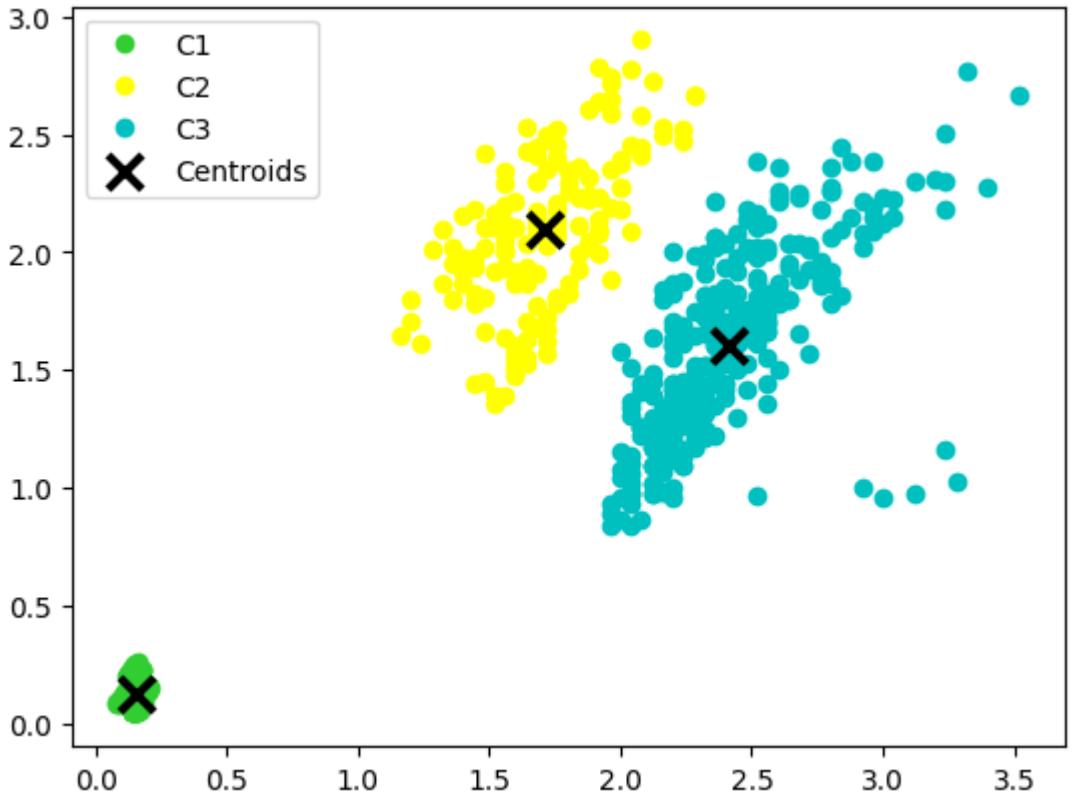
1. Για k=2 → SSE=90. Η συγκεκριμένη τιμή είναι λογική, καθώς τα σημεία βρίσκονται αρκετά μακριά σε σχέση με τα 2 κέντρα. Ενώ, παρατηρούμε ότι από την τιμή k=15 και όσο αυξάνεται το k, η τιμή του SSE τείνει στο 0.

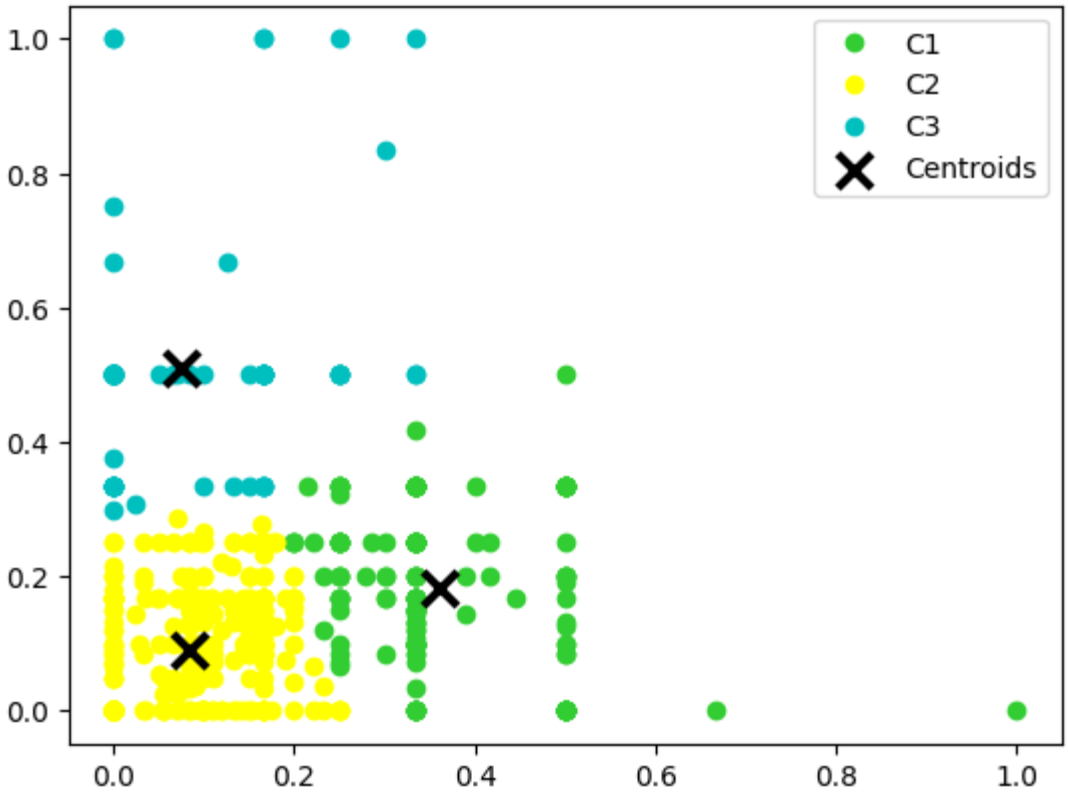
2. Για k=2 → Συντελεστής Περιγράμματος=0.77. Και αυτή η τιμή φαίνεται λογική, και αρκετά κοντά στο 1 σε σύγκριση με όλες τις υπόλοιπες. Για k=15 παρατηρείται η μικρότερη τιμή του συντελεστή, ενώ έπειτα αρχίζει να αυξάνεται. Ωστόσο, η τιμή του συντελεστή φτάνει μέχρι το 0.48 για k=100.

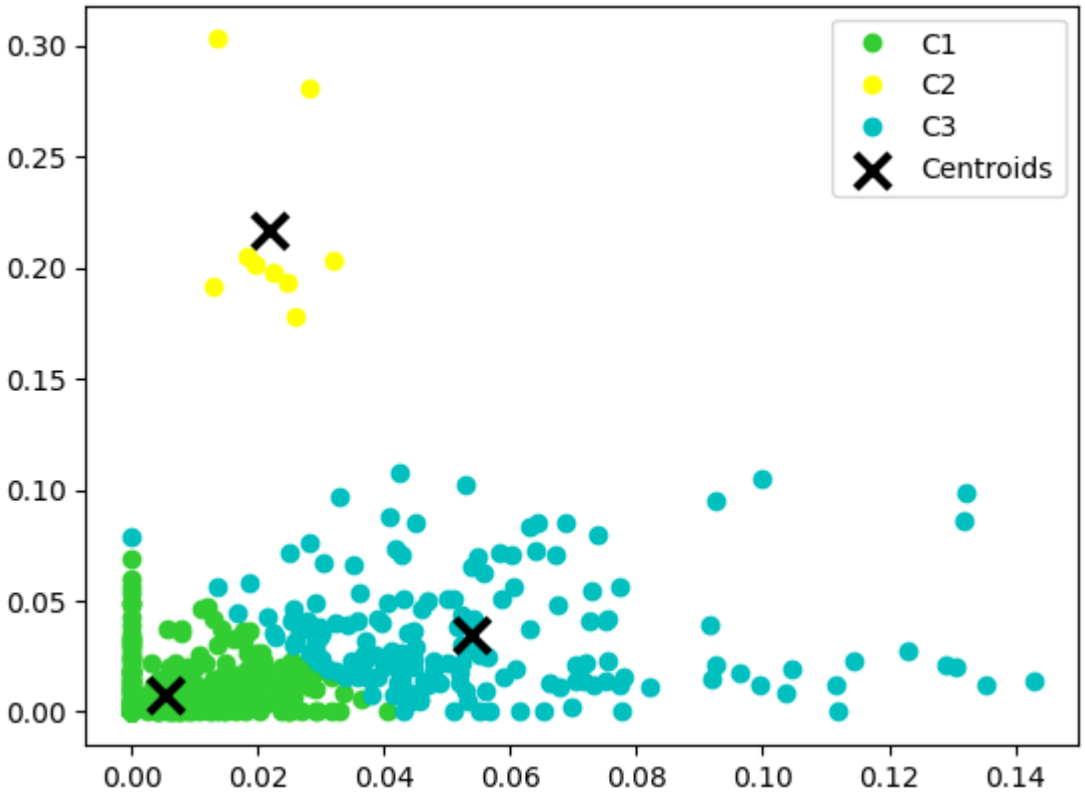
1.2 Εφαρμογή στο σύνολο δεδομένων xV.mat

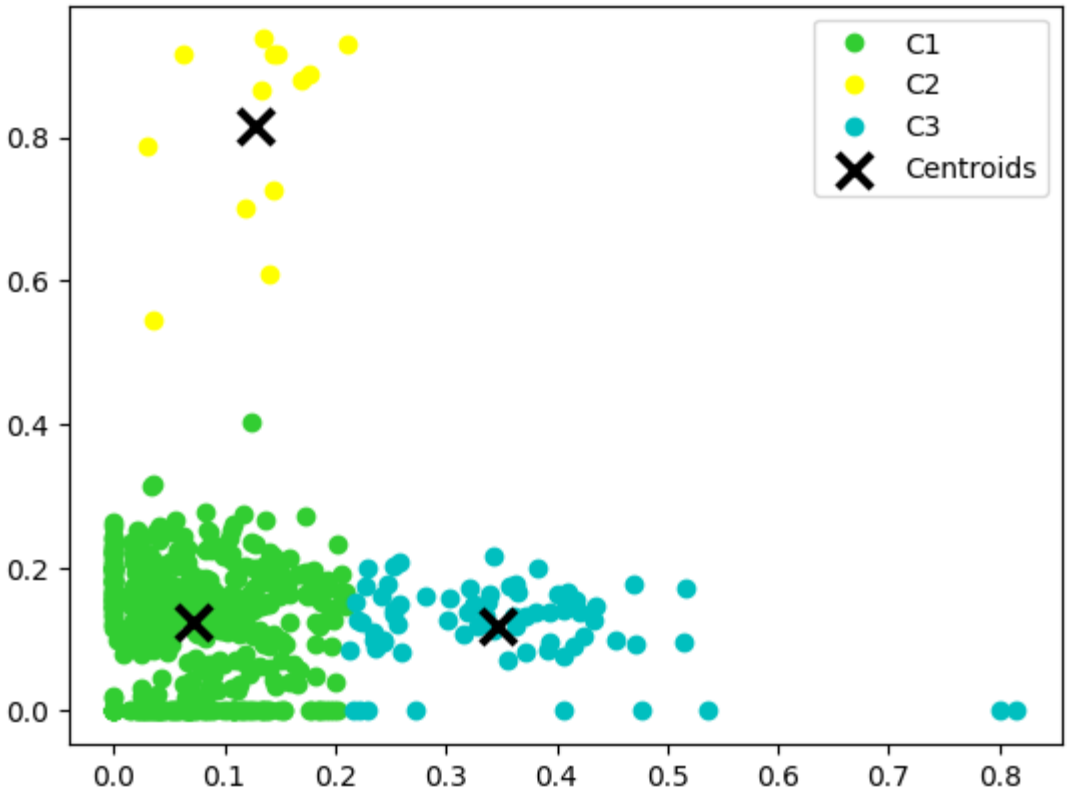
Για το 2ο ερώτημα του 1ου μέρους της εργασίας, ακολουθήσαμε τα βήματα της εκφώνησης και χρησιμοποιήσαμε τη γενική προσέγγιση της k-means συσταδοποίησης όπως στο ερώτημα 1.1 . Αρχικά, κάναμε load το αρχείο μορφής Matlab, xV.mat. Το χειριστήκαμε ως numpy(βιβλιοθήκη της Python) array και σε

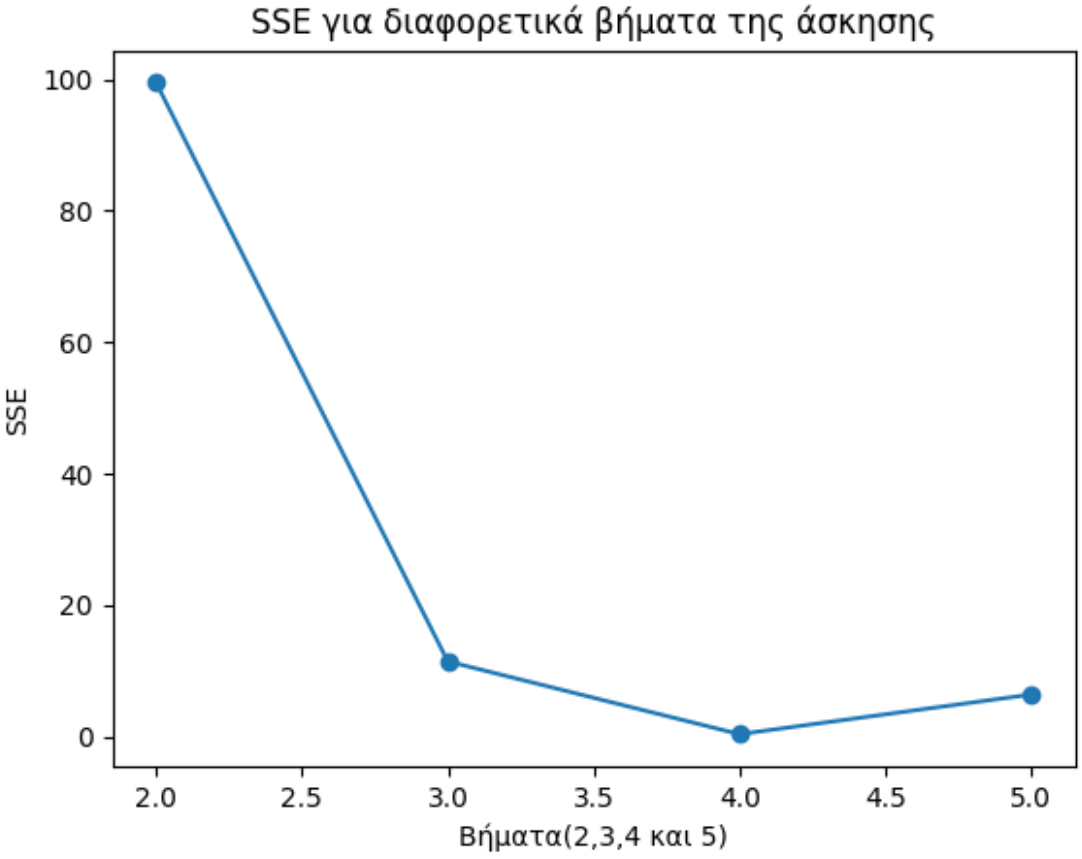
κάθε ένα από τα επόμενα βήματα επεξεργαστήκαμε δύο από τις στήλες του(σε κάθε βήμα διαφορετικές). Αξίζει να τονιστεί ότι σε όλο το ερώτημα ο αριθμός των συστάδων παραμένει 3 και γίνεται χρήση της Ευκλείδειας Απόστασης. Τα βήματα που ακολουθήσαμε υπάρχουν αναλυτικά σχολιασμένα στο αρχείο exercise1_2.py . Συνοπτικά, δημιουργήσαμε 2 συναρτήσεις: k_means_clustering και sse_calculation. Για κάθε ένα από τα εξής βήματα:2,3,4 και 5 γίνεται συσταδοποίηση για 2 συγκεκριμένες στήλες-χαρακτηριστικά του πίνακα xV, γραφική παράσταση των σημείων και των κέντρων των συστάδων στο χώρο των χαρακτηριστικών. Επίσης, σε κάθε βήμα γίνεται υπολογισμός του συνολικού SSE. Έτσι, για λόγους ανάγνωσης του αρχείου, αλλά και πολυπλοκότητας δημιουργήθηκαν οι 2 προαναφερθείσες συναρτήσεις.

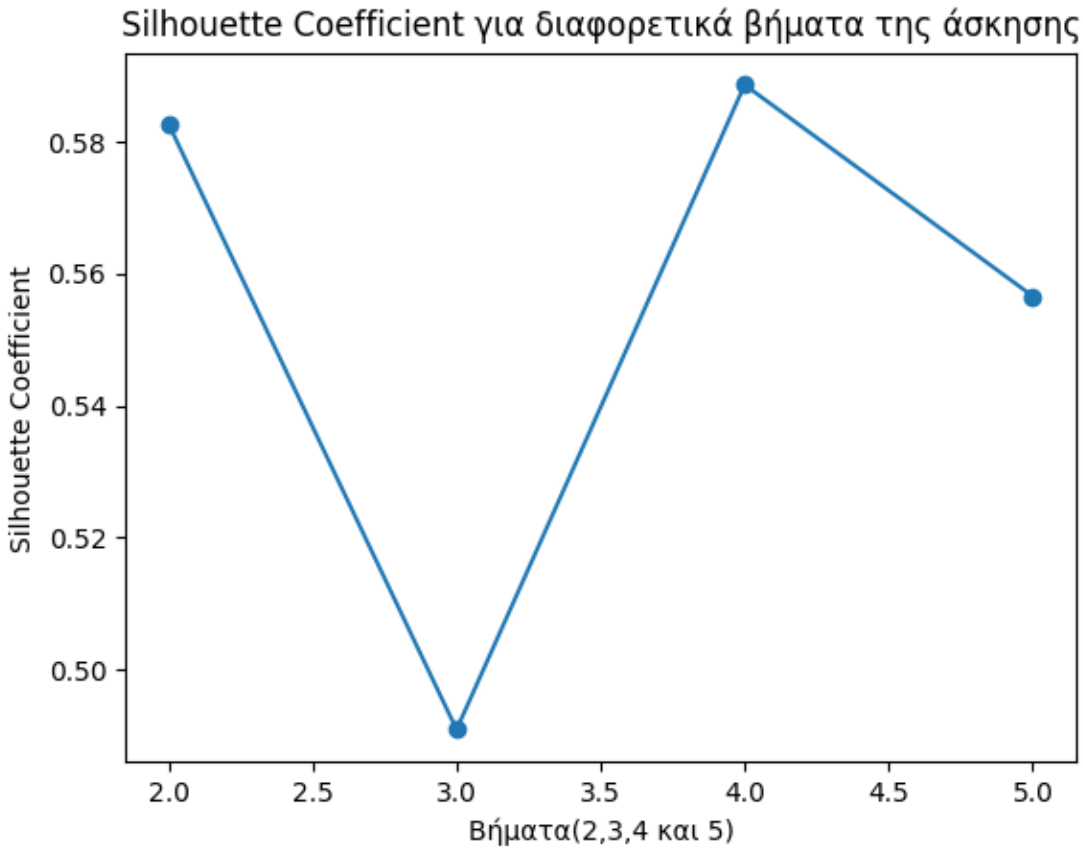



C1
C2
C3
Centroids
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


C1
C2
C3
Centroids
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14


C1
C2
C3
Centroids
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


SSE για διαφορετικά βήματα της άσκησης
SSE
100
80
60
40
20
0
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Βήματα(2,3,4 και 5)



Αναλυτικότερα, όπως παρατηρούμε και στο γράφημα του υπολογιζόμενου SSE σε συνάρτηση με τα βήματα της άσκησης(2,3,4 και 5), η τιμή του SSE για το βήμα 2 αποτελεί τη μέγιστη τιμή και πολύ μεγαλύτερη σε σύγκριση με όλες τις υπόλοιπες τιμές(SSE2=99.449). Καταρχάς, από το 1ο διάγραμμα παρατηρούμε ότι ο δισδιάστατος χώρος των χαρακτηριστικών είναι 3.5 x 3.0 . Το παραπάνω είναι αρκετά σημαντικό, καθώς ο χώρος χαρακτηριστικών της συγκεκριμένης συσταδοποίησης είναι ο μεγαλύτερος, σε σχέση με τα επόμενα διαγράμματα. Επιπλέον, πρέπει να σημειώσουμε ότι το συνολικό SSE κάθε βήματος προκύπτει από το άθροισμα 3 τετραγωνικών σφαλμάτων(1 για κάθε συστάδα). Όσον αφορά την κάθε συστάδα ξεχωριστά, βλέπουμε ότι τα σημεία που ανήκουν στο C1 βρίσκονται πολύ κοντά στο κέντρο της συστάδας και μοιάζουν στο γράφημα ως ένα μεγάλο σημείο. Οπότε, δεν προκύπτει ότι η πρώτη συστάδα δημιουργεί τόσο μεγάλο συνολικό SSE. Από την άλλη πλευρά, η $3^{η}$ συστάδα και σε μικρότερο βαθμό και η $2^{η}$ έχουν στον πληθυσμό τους αρκετά ακραία σημεία-outliers. Αναντίρρητα, αυξάνουν το συνολικό SSE, αφού η μέθοδος k-means είναι ευαίσθητη στα ακραία σημεία. Επιπροσθέτως, ειδικά στην $3^{η}$ συστάδα, τα ακραία σημεία βρίσκονται και αρκετά απομακρυσμένα μεταξύ τους, στον αρκετά μεγάλο χώρο των χαρακτηριστικών. Συμπερασματικά, η τιμή του SSE2 φανερώνει ότι δε βρέθηκε η βέλτιστη λύση και η συσταδοποίηση με k=3 δεν έχει γίνει με τον κατάλληλο τρόπο.

Η συσταδοποίηση του βήματος 3 οδηγεί σε SS3=11.402 .Το συνολικό τετραγωνικό σφάλμα είναι υποδεκαπλάσιο περίπου σε σύγκριση με του προηγούμενου βήματος. Αξίζει να τονιστεί ότι ο χώρος των χαρακτηριστικών είναι μικρότερος κατά 1/3

περίπου(1.0 x 1.0). Οπότε, τα δείγματα μιας συστάδας είναι σχετικά κοντά στο κέντρο. Ωστόσο, στην $2^{η}$, αλλά κυρίως στην $3^{η}$ συστάδα παρατηρούμε αρκετά ακραία σημεία της $3^{ης}$ συστάδας βρίσκονται ακριβώς πάνω στο κέντρο της, γεγονός που δικαιολογεί το μειωμένο SSE σε σύγκριση με το βήμα 2. Σε αυτό το σημείο πρέπει να γίνει αναφορά και στο διάγραμμα του Συντελεστή Περιγράμματος σε συνάρτηση με τα βήματα της άσκησης. Δηλαδή, για κάθε βήμα υπολογίζεται το s_i. Γενικά, όλες οι συσταδοποιήσεις παρουσιάζουν θετικό συντελεστή(s_i>0.45), οπότε τα δείγματα έχουν τοποθετηθεί στις σωστές συστάδες. Επίσης, παρατηρούμε ότι κατά τη $2^{η}$ συσταδοποίηση(Βήμα 3) ο Συντελεστής Περιγράμματος σημειώνει τη μικρότερη τιμή του. Άρα, για τη συγκεκριμένη συσταδοποίηση(αλλά και για τις υπόλοιπες, σε μικρότερο βαθμό όμως) συμπεραίνουμε ότι τα σημεία δεν είναι αρκετά καλά συσταδοποιημένα, και ότι δεν υπάρχει τεράστια απόσταση μεταξύ των συστάδων. Ακόμη, όσον αφορά τη συσταδοποίηση του βήματος 3(ο μικρότερος συντελεστής) εμφανίζεται μεγαλύτερη επικάλυψη των συστάδων σε σύγκριση με τις συσταδοποιήσεις των υπόλοιπων 3 βημάτων. Το παραπάνω προκύπτει από το γεγονός ότι ο συντελεστής βρίσκεται στο βήμα 3 πιο κοντά στο 0, σε σχέση με τα υπόλοιπα βήματα.

Επιστρέφοντας στην ανάλυση του SSE, συνεχίζουμε με τη συσταδοποίηση του $4^{ου}$ βήματος, όπου χρησιμοποιήσαμε τις 2 τελευταίες στήλες του πίνακα xV. Το SSE4 κυμαίνεται πολύ κοντά στο 0(SSE4=0.330) και με βάση αποκλειστικά το συγκεκριμένο μέτρο, η συσταδοποίηση του βήματος 4 είναι βέλτιστη, και καλύτερη με μεγάλη διαφορά από τις υπόλοιπες. Σε ένα μεγάλο βαθμό το παραπάνω συμπέρασμα δικαιολογείται από τον πολύ μικρό χώρο χαρακτηριστικών που παρουσιάζεται(0.14 x 0.30). Αρκετά σημεία φαίνονται με «γυμνό μάτι» ως outliers, ωστόσο πρέπει να σκεφτούμε ότι βρίσκονται σε μεγέθυνση 10 φορές περίπου, σε σχέση με το Βήμα 2. Επίσης, παρατηρούμε ότι τα κίτρινα σημεία($2^{η}$ συστάδα) είναι πληθυσμιακά πολύ λιγότερα σε σχέση με τον πληθυσμό των υπόλοιπων συστάδων στα βήματα που αναλύσαμε μέχρι τώρα. Ωστόσο, στο υποθετικό σενάριο όπου k=2 συστάδες, τότε όλα τα κίτρινα σημεία θα ήταν ακραία σημεία και το υπολογιζόμενο SSE θα ήταν πολύ μεγαλύτερο.

Τέλος, το SSE της $4^{ης}$ συσταδοποίησης($5^{ο}$ Βήμα) είναι το $3^{ο}$ μικρότερο, αλλά αρκετά μεγαλύτερο από το βέλτιστο SSE4(SSE5=6.363). Σε αυτό το σημείο πρέπει να αναφέρουμε τις διαστάσεις του χώρου των χαρακτηριστικών: 0.8 x 0.8. Άρα, είναι λογική η αύξηση του SSE, σε σχέση με το προηγούμενο βήμα. Όπως και στο προηγούμενο βήμα , το πλήθος των σημείων που ανήκει στην $2^{η}$ συστάδα είναι αρκετά μικρότερο από τα υπόλοιπα. Παράλληλα, κάθε συστάδα έχει outliers. Αναλυτικότερα, η $3^{η}$ συστάδα έχει 2 δείγματα που έχουν πολύ μεγάλη απόσταση από το X, δηλαδή το κέντρο της συστάδας. Επίσης, φαίνεται ότι ο πληθυσμός της $1^{ης}$ συστάδας είναι αρκετά μεγάλος και τα περισσότερα δείγματα βρίσκονται στα πέριξ του κέντρου της συστάδας. Το συγκεκριμένο γεγονός δικαιολογεί σε κάποιο βαθμό τη μη αύξηση της τιμής του SSE5 σε ανώτερα επίπεδα.

Συμπερασματικά, διαφορετικές στήλες του πίνακα xV έχουν διαφορετική επίδραση στη μορφή, αλλά και στην ποιότητα της συσταδοποίησης του εκάστοτε βήματος.

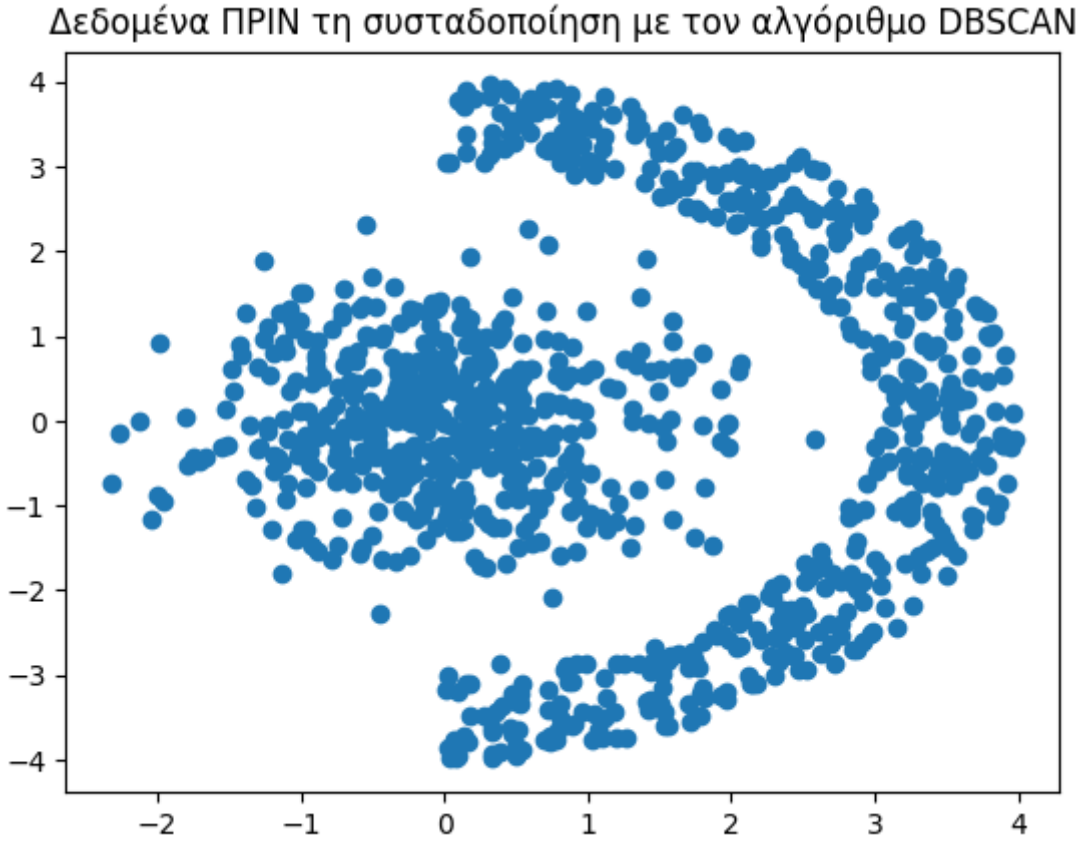
Σημείωση: Η κλήση της συνάρτησης k_means_clustering με τις παραμέτρους(n_init, max_ter, tol, …), που αναγράφονται στις πρώτες σελίδες της εκφώνησης της

άσκησης, δε μείωσε το SSE2. Για αυτό το λόγο και δεν υπάρχει στην τελική μορφή του αρχείου exercise1_2.py .

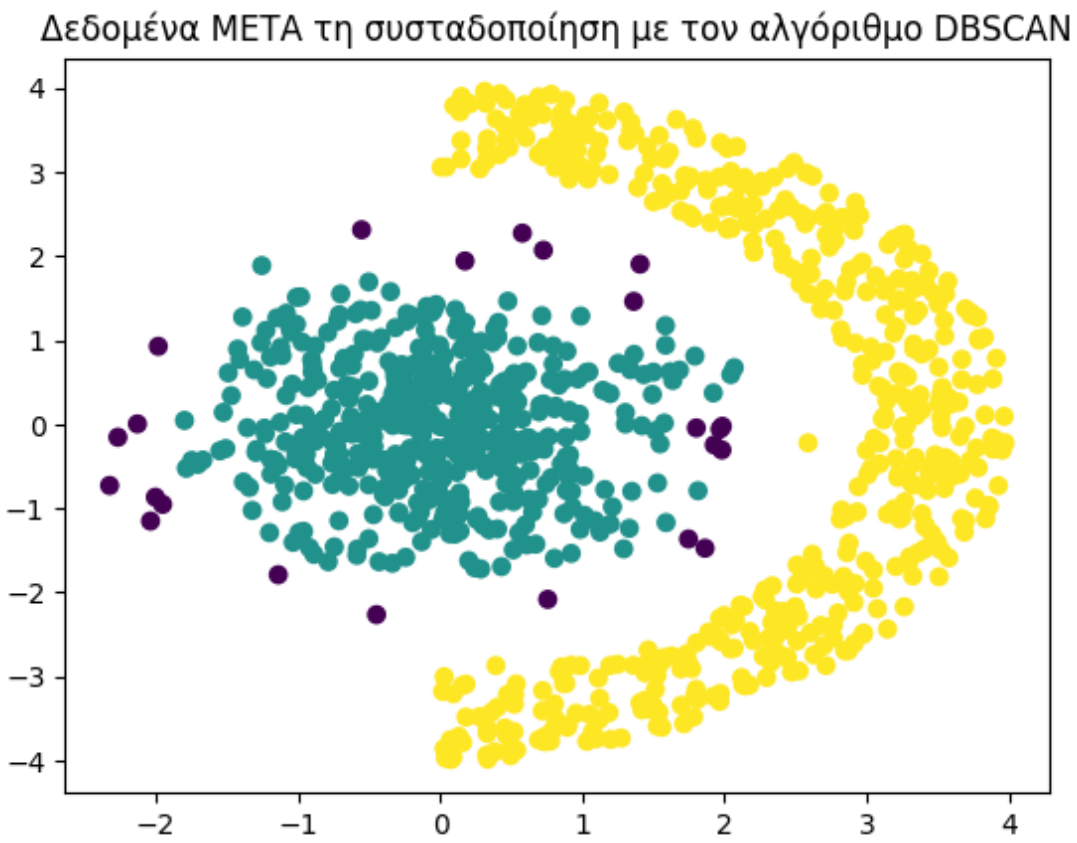
Μέρος 2ο: DBSCAN(Συσταδοποίηση βάσει πυκνότητας)

2.1 Εφαρμογή στο σύνολο δεδομένων mydata

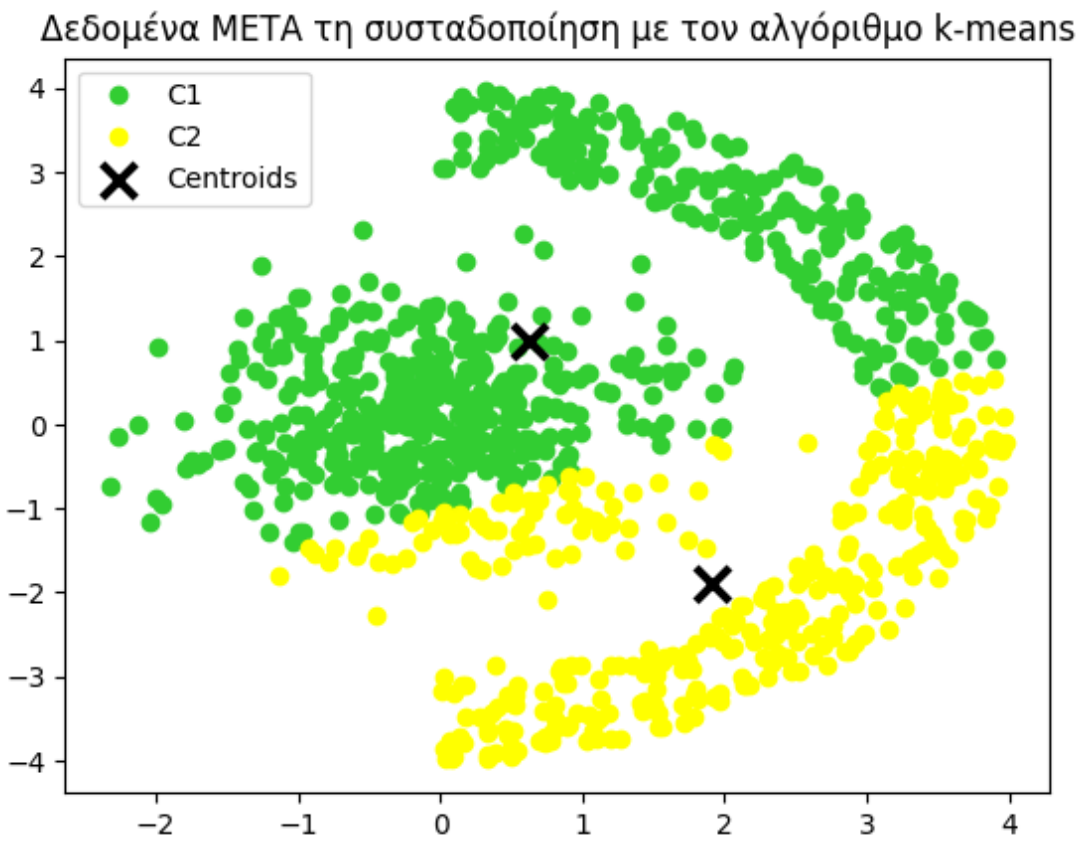
Καταρχάς, κοιτάζοντας το γράφημα που περιέχει τα δεδομένα πριν τη συσταδοποίηση πιθανολογούμε ότι θα χωριστούν σε 2 συστάδες. Είναι κατανοητό ότι υπάρχουν 2 διαφορετικά σχήματα(μισοφέγγαρο και έλλειψη) και ίσως κάποια σημεία θορύβου. Η χρήση της συσταδοποίησης DBSCAN για το συγκεκριμένο dataset είναι λογική, καθώς ο αλγόριθμος DBSCAN μπορεί να χειριστεί συστάδες με διαφορετικά σχήματα και μεγέθη.



Αξίζει να τονιστεί ότι χρησιμοποιήσαμε τη συνάρτηση fit, την εντολή IDX=dbscan.labels_ και τις εξής τιμές παραμέτρων: eps=0.5 και min_samples=15, για να προβούμε σε DBSCAN συσταδοποίηση των 2 πρώτων στηλών του πίνακα X(mydata.mat).



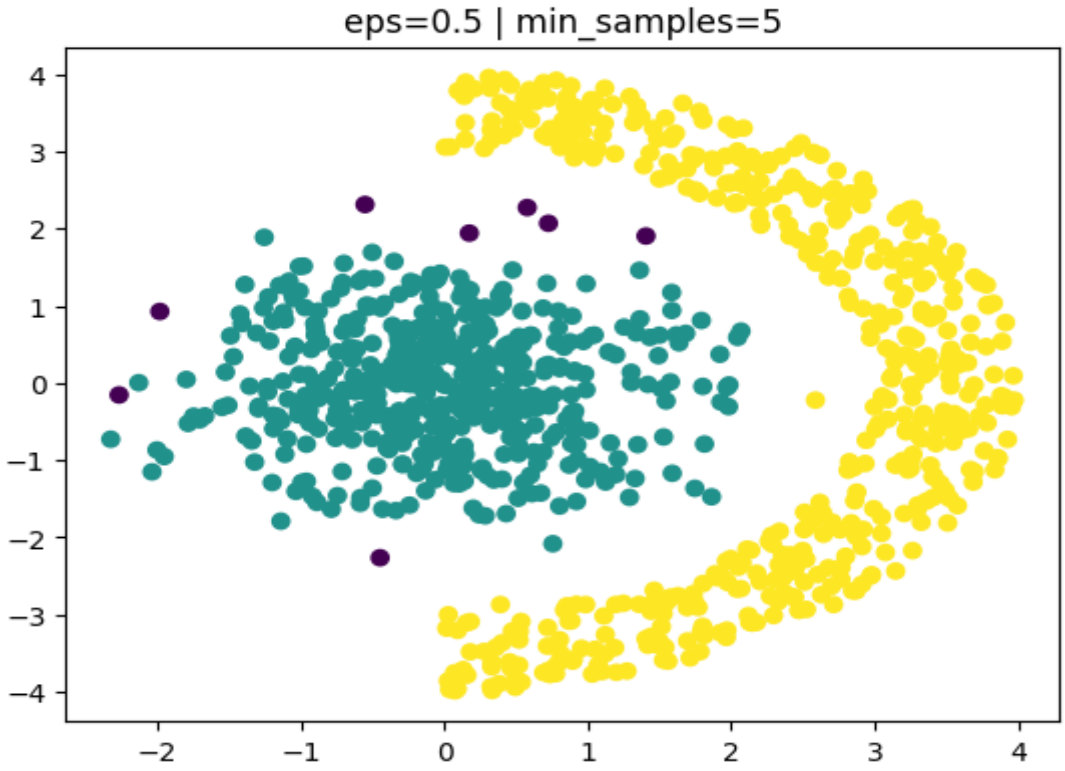
Έτσι, το σχήμα που προκύπτει από τη DBSCAN είναι αυτό που περιμένουμε. Τα σημεία πυρήνα, αλλά και τα οριακά βρίσκονται σε μία από τις 2 συστάδες. Ανάλογα τη θέση τους στο χώρο των χαρακτηριστικών. Επιπροσθέτως, το σύνολο των σημείων θορύβου είναι 22 και ξεχωρίζουν χρωματικά από τις 2 συστάδες όντας με σκούρο μπλε χρώμα.

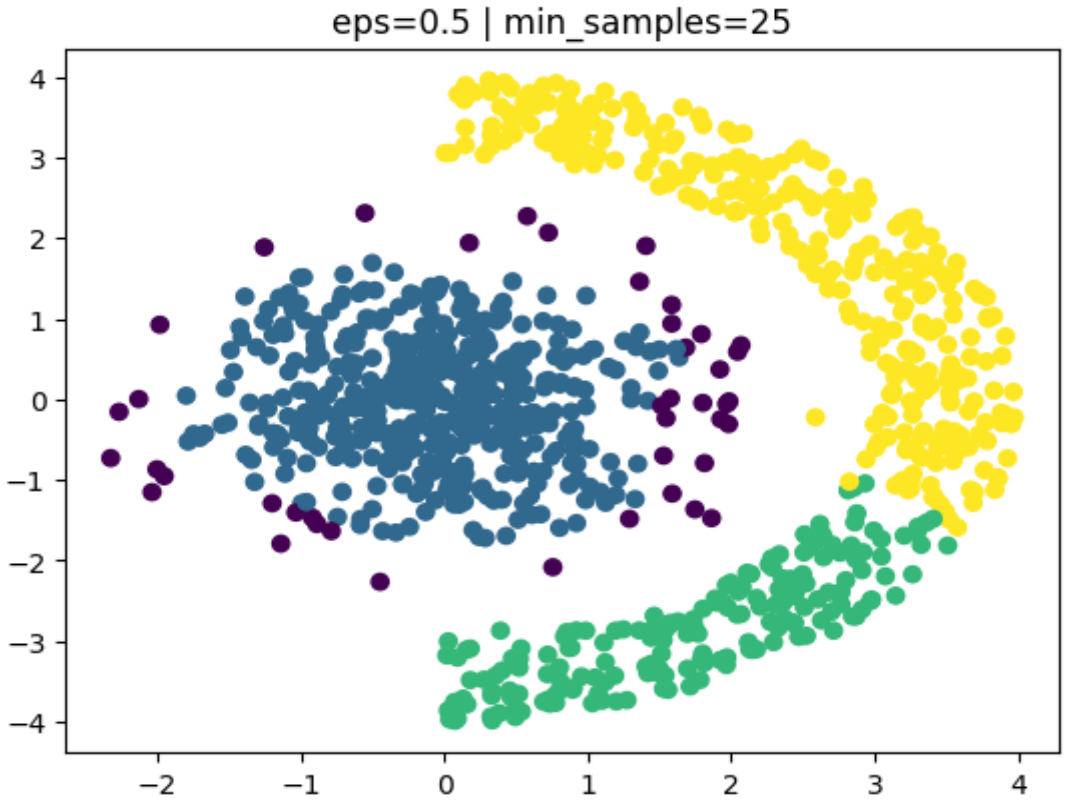


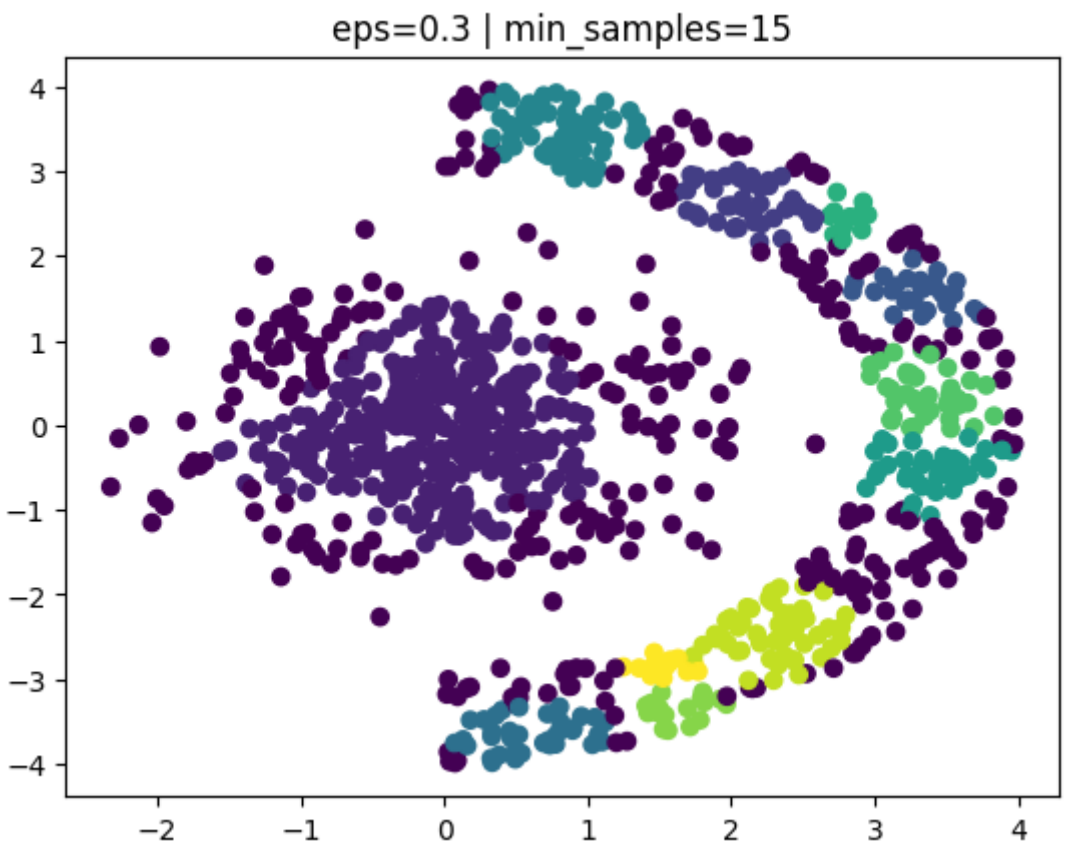
Στη συνέχεια, χρησιμοποιώντας κώδικα από το 1ο μέρος της εργασίας, επιλέξαμε να πειραματιστούμε με τη συσταδοποίηση k-means, πάνω στα ίδια δεδομένα. Χρησιμοποιήσαμε 2 συστάδες, καθώς θέλαμε να συγκρίνουμε τις 2 μορφές συσταδοποίησης υπό τις ίδιες συνθήκες. Στο διάγραμμα παρουσιάζονται οι 2 συστάδες(με πράσινο και κίτρινο χρώμα αντίστοιχα) αλλά και τα κέντρα τους. Παρατηρούμε ότι το 1ο σχήμα: έλλειψη έχει συσταδοποιηθεί κατά το μεγαλύτερο μέρος του στην πράσινη συστάδα. Επίσης, το 2ο σχήμα: μισοφέγγαρο έχει μοιραστεί ανάμεσα στις 2 συστάδες, με τα σημεία της κίτρινης συστάδας να υπερτερούν αριθμητικά στο ίδιο σχήμα. Το κέντρο της πράσινης συστάδας βρίσκεται στο ελλειπτικό σχήμα ψηλά και δεξιά. Ακόμα μεγαλύτερη εντύπωση προκαλεί η θέση του 2ου κέντρου. Βρίσκεται σε ένα σημείο που δεν υπάρχουν καθόλου σημεία.

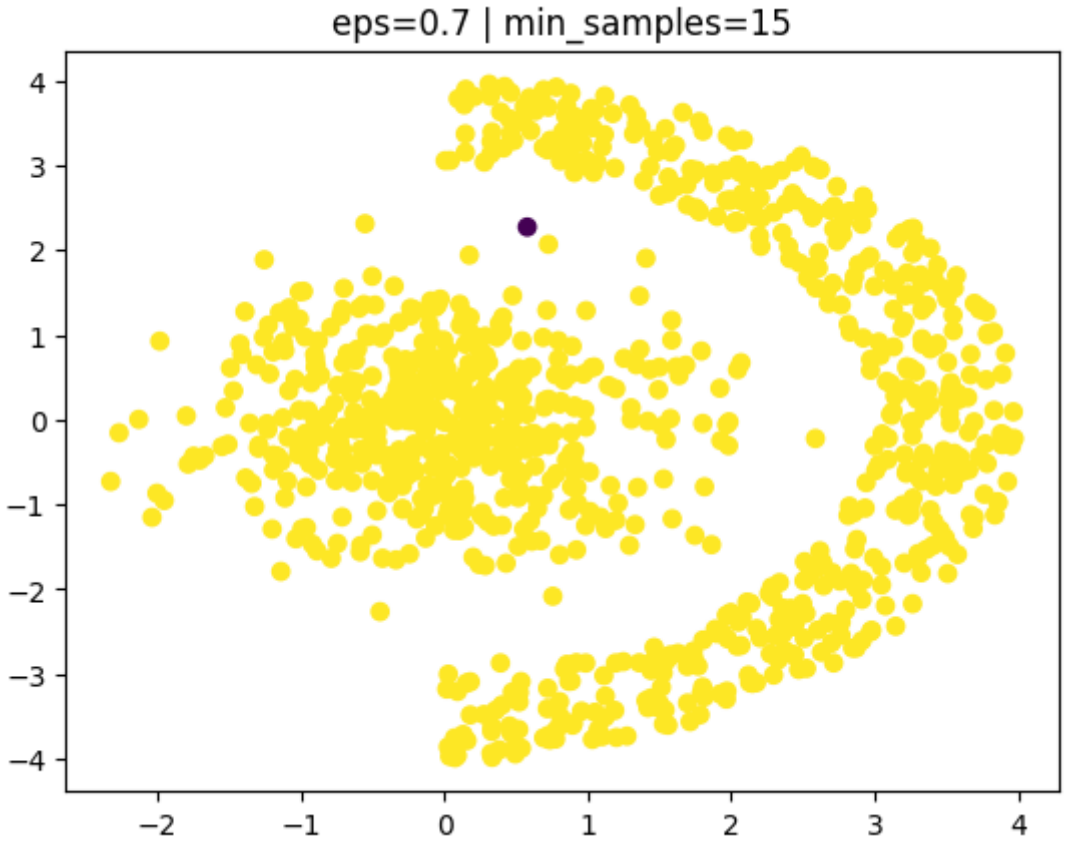
Οι παραπάνω παρατηρήσεις, σε συνάρτηση με τα αποτελέσματα των μέτρων: s_i(k-means)=0.380 και SSE(k-means)=3478.866 μας οδηγούν στο συμπέρασμα ότι ο αλγόριθμος k-means δεν είναι κατάλληλος για το συγκεκριμένο τύπο δεδομένων. Ειδικά η τιμή του SSE είναι πάρα πολύ υψηλή. Από την άλλη πλευρά, s_i(DBSCAN)=0.233 < s_i(k-means). Άρα, με βάση το μέτρο Συντελεστή Περιγράμματος οι αλγόριθμοι έχουν μέτρια διαχωριστική ικανότητα, με την k-means να έχει ελαφρώς καλύτερη ,απόδοση στα συγκεκριμένα δεδομένα. Ωστόσο, πρέπει να επαναληφθεί το γεγονός ότι το SSE του k-means είναι πολύ υψηλό.

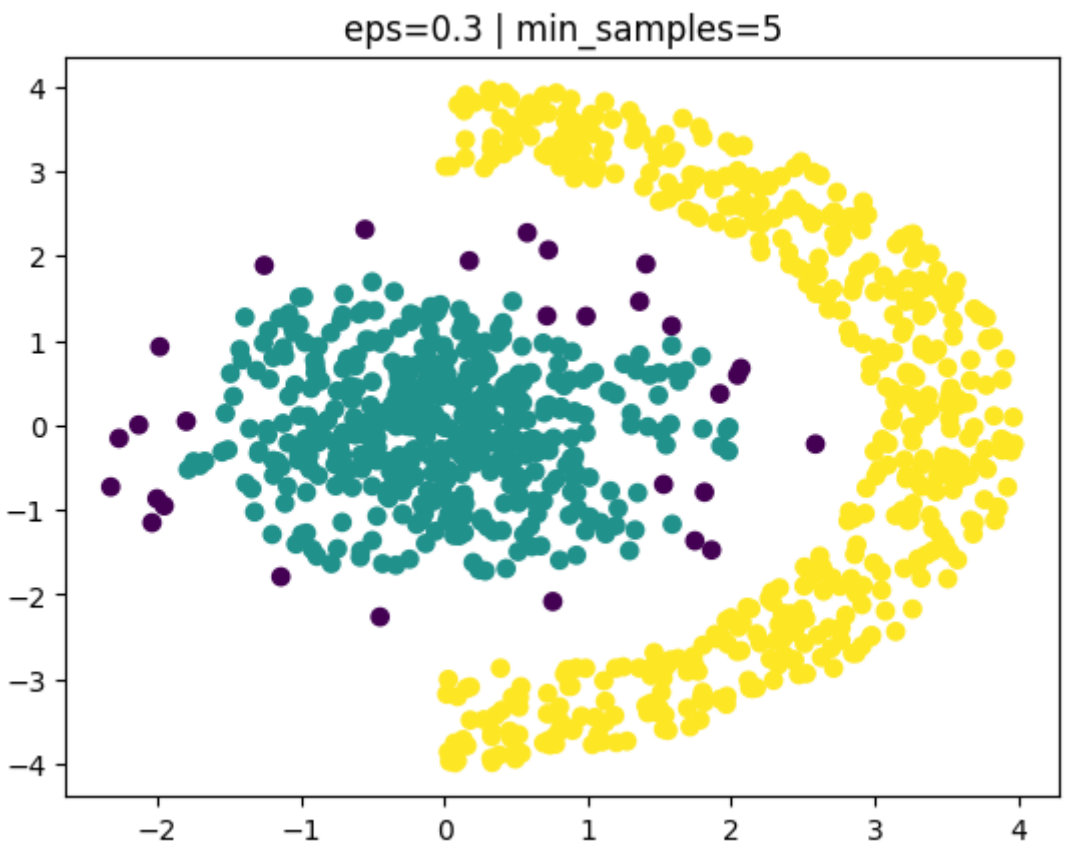
- **Αλλαγή στις παραμέτρους(min_samples, eps) του DBSCAN**

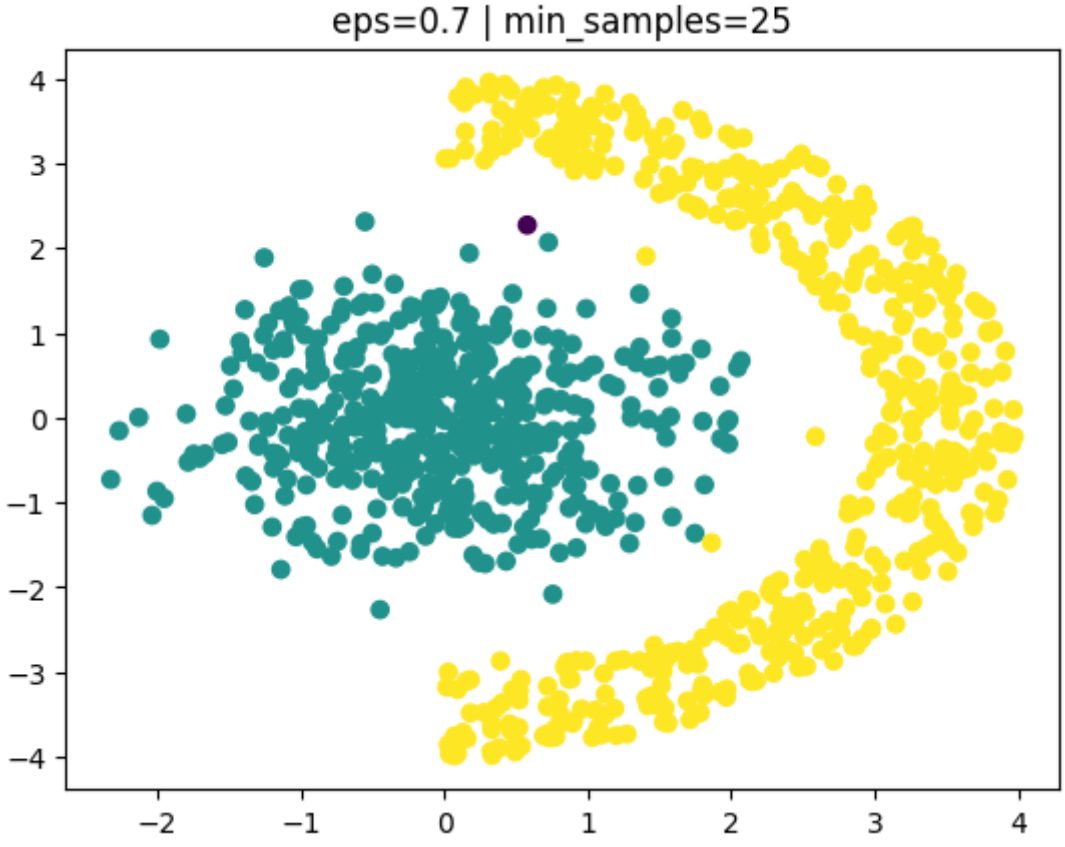


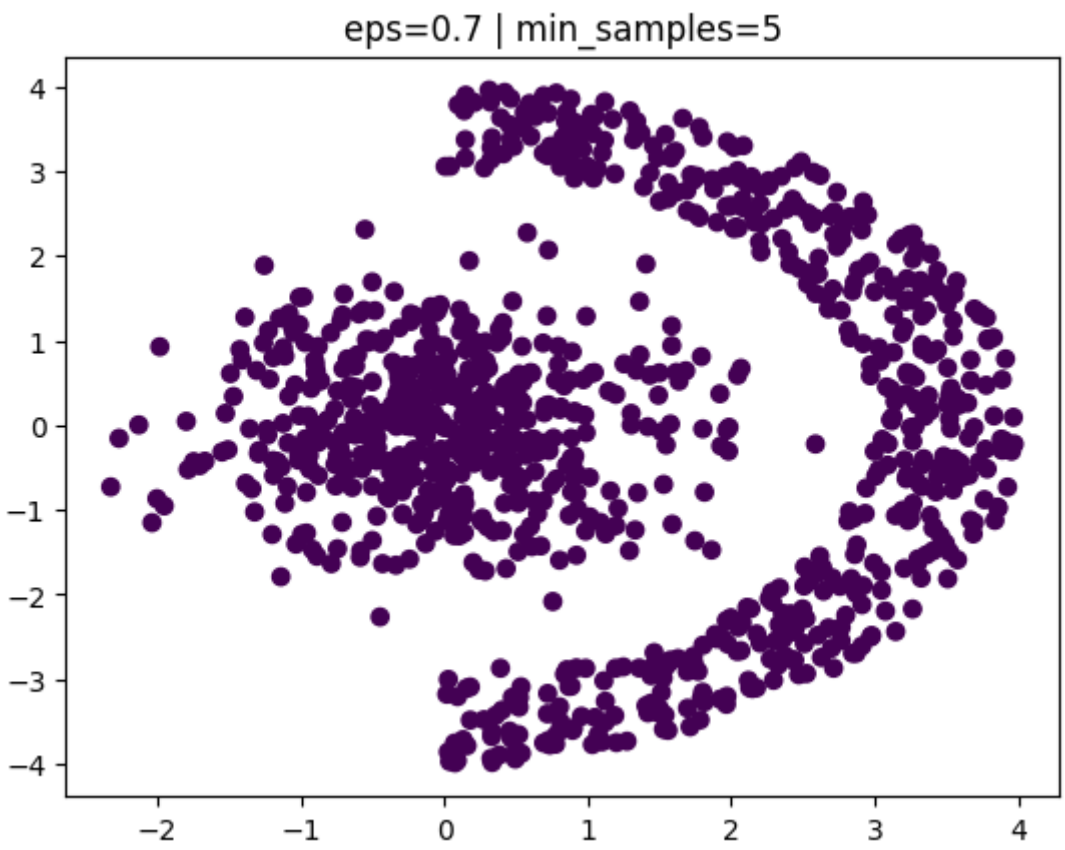


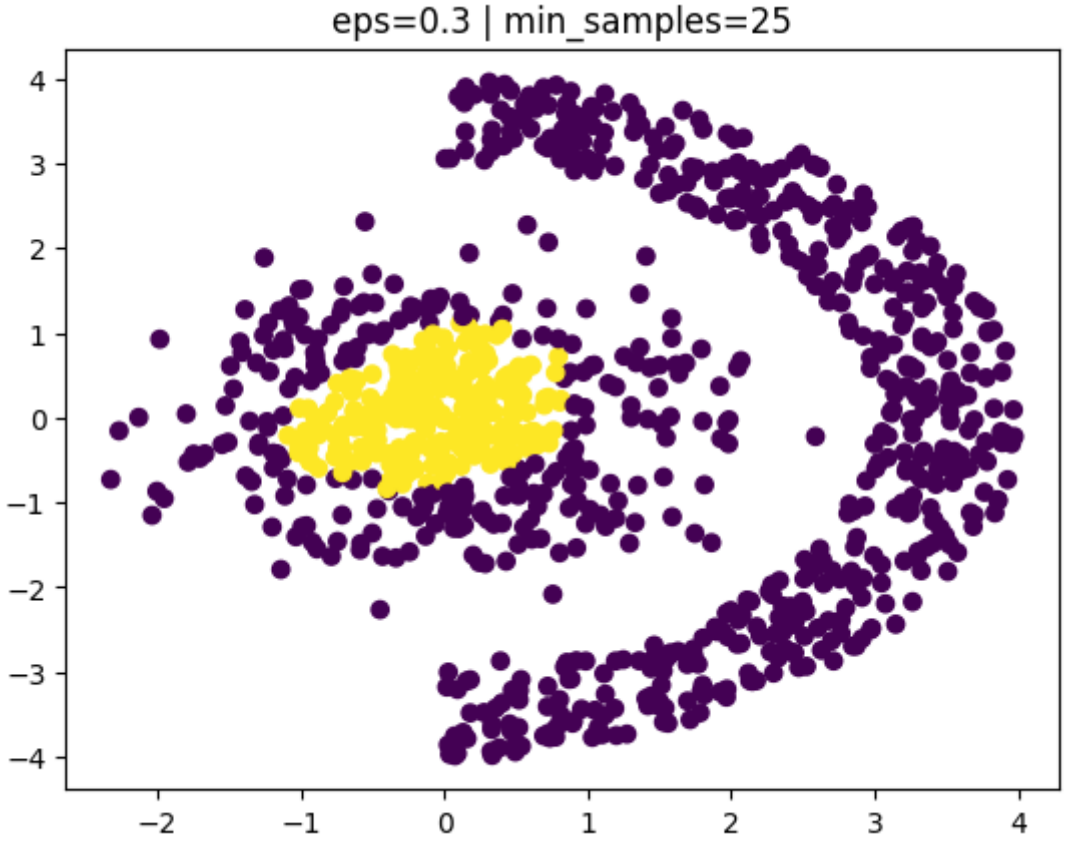

eps=0.3 | min_samples=15
4
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4


eps=0.7 | min_samples=15
4
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4


eps=0.3 | min_samples=5
4
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4


eps=0.7 | min_samples=25
4
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4


eps=0.7 | min_samples=5
4
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4


eps=0.3 | min_samples=25
4
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4

1) **Επίδραση του min_samples(eps:σταθερό)**

Default=15

Μείωση: Δημιουργία συστάδων με λιγότερα σημεία. Οι συστάδες παραμένουν 2. Επίσης, μειώθηκαν τα σημεία θορύβου.

Αύξηση: Απαιτούνται περισσότερα σημεία για τον ορισμό μιας συστάδας. Ωστόσο, στο συγκεκριμένο παράδειγμα, παρατηρούμε ότι δημιουργήθηκαν 3 συστάδες. Παράλληλα αυξήθηκαν τα σημεία θορύβου.

2) **Επίδραση του eps(min_samples:σταθερό)**

Default=0.5

Μείωση: Αύξηση της πυκνότητας των συστάδων. Από το γράφημα φαίνεται ότι έχουν δημιουργηθεί 11 συστάδες με λίγα σημεία πληθυσμού έκαστη. Τα περισσότερα σημεία του dataset θεωρούνται θόρυβος.

Αύξηση: Οι συστάδες είναι λιγότερο πυκνές. Άρα, θα μειωθεί και ο αριθμός τους, καθώς μπορούν να φιλοξενήσουν περισσότερα σημεία. Από το γράφημα παρατηρούμε ότι υπάρχει μόνο μία συστάδα, και μόνο ένα σημείο θορύβου.

3) **Ταυτόχρονη αλλαγή των παραμέτρων**

Με την παράλληλη μείωση, αλλά και με την παράλληλη αύξηση τους, δεν παρατηρήσαμε εντυπωσιακές διαφορές στα γραφήματα, σε σχέση με την αρχική τους μορφή. Αναλυτικότερα, ο αριθμός των συστάδων παραμένει 2. Στην $1^{η}$ περίπτωση αυξάνονται τα σημεία θορύβου, ενώ στη $2^{η}$ μειώνονται.

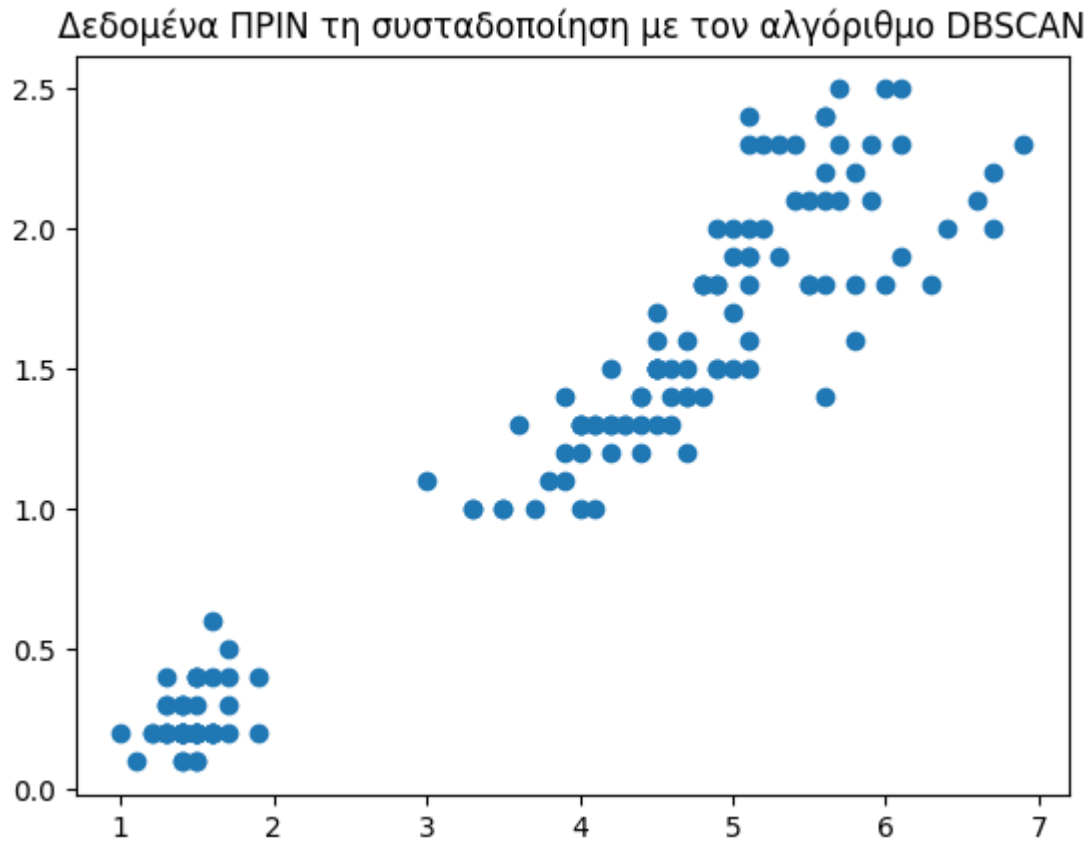
4) **Αντίστροφη αλλαγή των παραμέτρων**

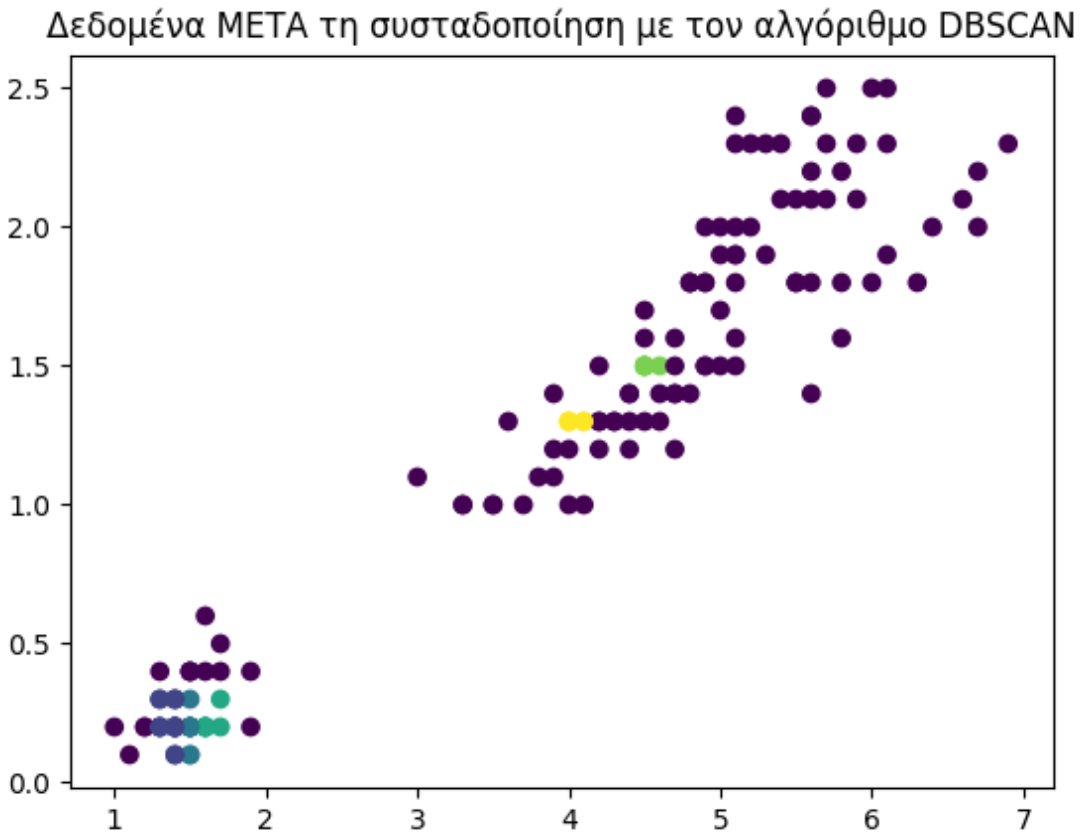
Αντίθετα, η οπτικοποίηση των 2 τελευταίων περιπτώσεων δε μοιάζει με όσα περιεγράφηκαν παραπάνω. Ο θόρυβος αυξήθηκε κατακόρυφα. Συγκεκριμένα, για eps=0.7 και min_samples=5 όλα τα σημεία ανιχνεύονται ως θόρυβος. Από την άλλη πλευρά, με eps=0.3 και min_samples=25, υπάρχει μόνο μία πυκνή συστάδα, και τα περισσότερα σημεία στο χώρο είναι θόρυβος.

2.2 Εφαρμογή στο σύνολο δεδομένων iris

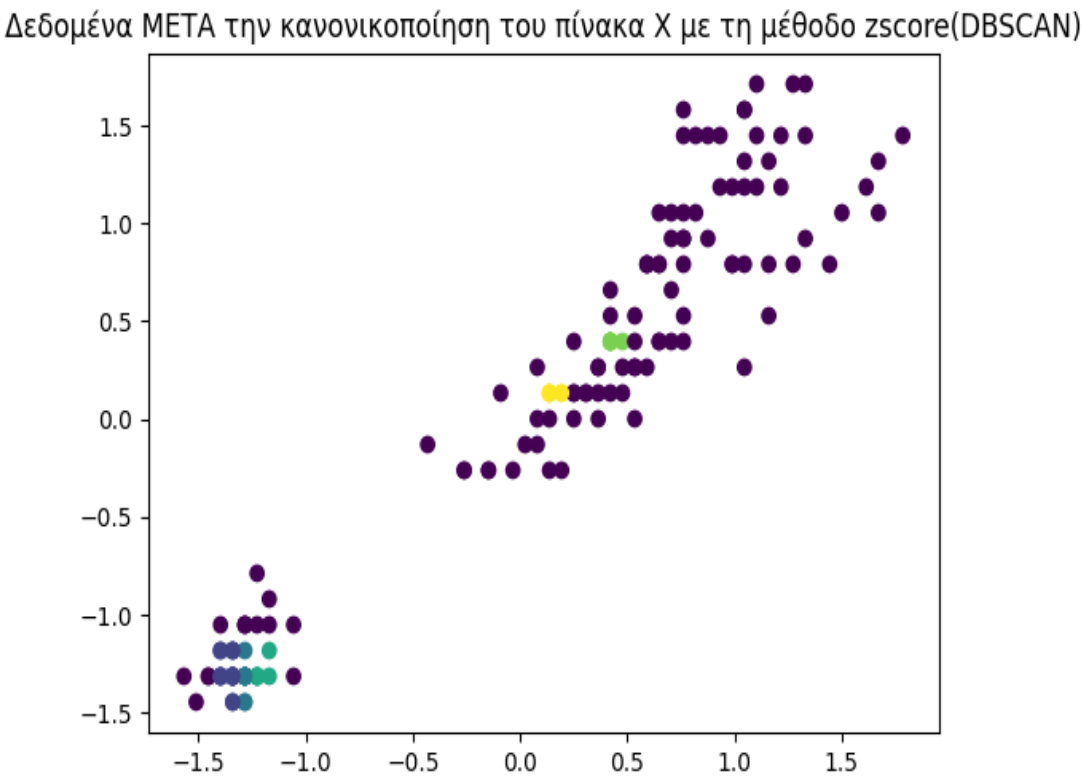
Από τη συγγραφή του κώδικα του αρχείου exercise2_2.py προκύπτουν 6 διαγράμματα. Αρχικά, παρουσιάζονται τα δεδομένα του συνόλου iris(στήλες 3 και 4) στο χώρο διαστάσεων 7 x 2.5 . Με τα συγκεκριμένα δεδομένα εργαστήκαμε και κατά την εκπόνηση του 1^{ου} μέρους της εργασίας, χρησιμοποιώντας τον αλγόριθμο συσταδοποίησης k-means.

Έπειτα, παρουσιάζουμε τα δεδομένα με την DBSCAN συσταδοποίηση. Με την πρώτη ματιά, παρατηρούμε ότι το μεγαλύτερο μέρος των σημείων είναι θόρυβος. Αξίζει να τονιστεί ότι eps=0.1 και min_samples=5. Δηλαδή, οι συστάδες αποτελούνται από λίγα σημεία η κάθε μία, τα οποία βρίσκονται πολύ κοντά μεταξύ τους. Συγκεκριμένα, υπάρχουν 5 συστάδες. Οι 3 συστάδες εντοπίζονται στο κάτω σχήμα(πυκνότερο αλλά με λιγότερα σημεία) και 2 στο πάνω σχήμα. Ο μέσος συντελεστής περιγράμματος της DBSCAN είναι -0.189. Το αρνητικό πρόσημο φανερώνει ότι η απόσταση μεταξύ των 5 συστάδων είναι μεγαλύτερη από την απόσταση εντός της εκάστοτε συστάδας. Γενικά, η συσταδοποίηση δεν είναι ικανοποιητικά ξεκάθαρη, αν λάβουμε υπόψιν μας το γεγονός ότι βρίσκεται αρκετά κοντά στο 0.

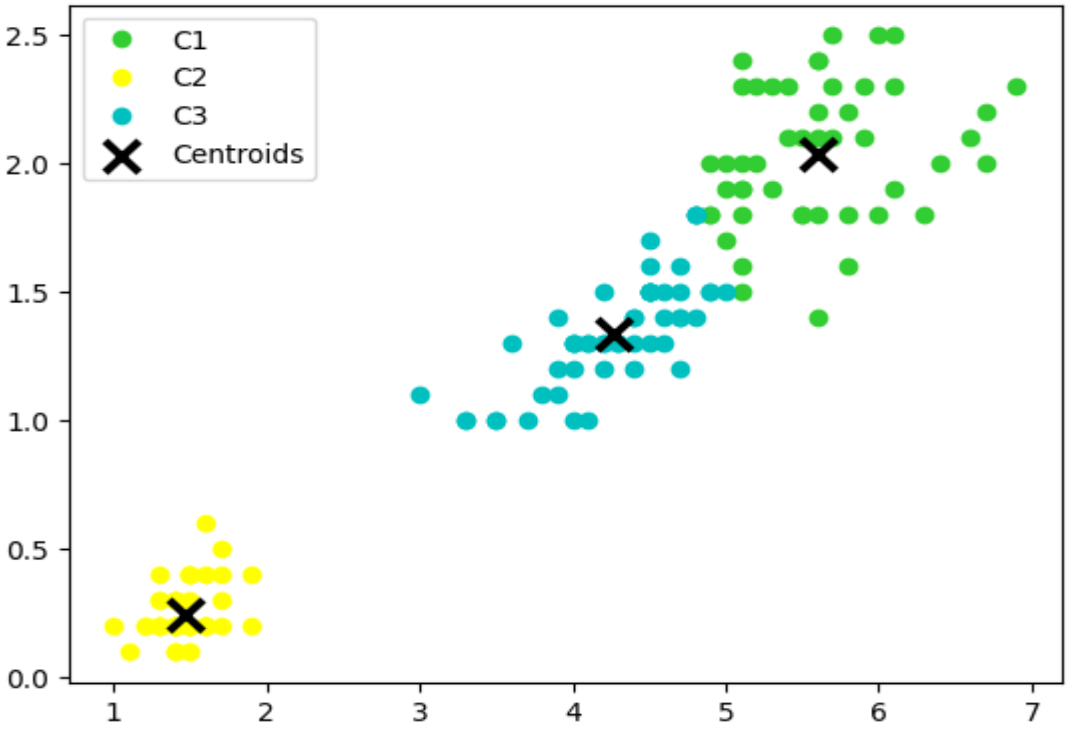




Στη συνέχεια, μετασχηματίζουμε τα δεδομένα με τη βοήθεια του μέτρου zscore. Αναλυτικότερα, ο χώρος των χαρακτηριστικών έχει διαστάσεις 1.6 x 1.6, με τις τυπικές αποκλίσεις να βρίσκονται γύρω από τη μέση τιμή. Όσον αφορά το διάγραμμα που ακολουθεί, οι συστάδες αλλά και ο θόρυβος έχουν παραμείνει αμετάβλητα. Οπότε, τα 2 διαγράμματα είναι ίδια, εκτός από τις διαστάσεις των αξόνων x και y.
Υπολογίζουμε και το νέο συντελεστή περιγράμματος. Παρατηρούμε ότι αυξήθηκε ελαφρώς(-0.133 από -0.189). Ωστόσο, η αρνητική τιμή παραμένει. Άρα, παρά την κανονικοποίηση, η συσταδοποίηση δεν είναι καλή.



Μετά από τα παραπάνω χρησιμοποιούμε τον αλγόριθμο k-means με k=3. Από τα αποτελέσματα των 2 μέτρων(SSE=31.371 και Συντελεστής Περιγράμματος=0.660) παρατηρούμε ότι η k-means είναι πιο αποτελεσματική σε σχέση με τη DBSCAN, πάνω στα ίδια δεδομένα. Το παραπάνω γεγονός εξηγείται από την κατανομή των δεδομένων, που φανερώνει ότι δε φημίζονται για την πυκνότητα τους.



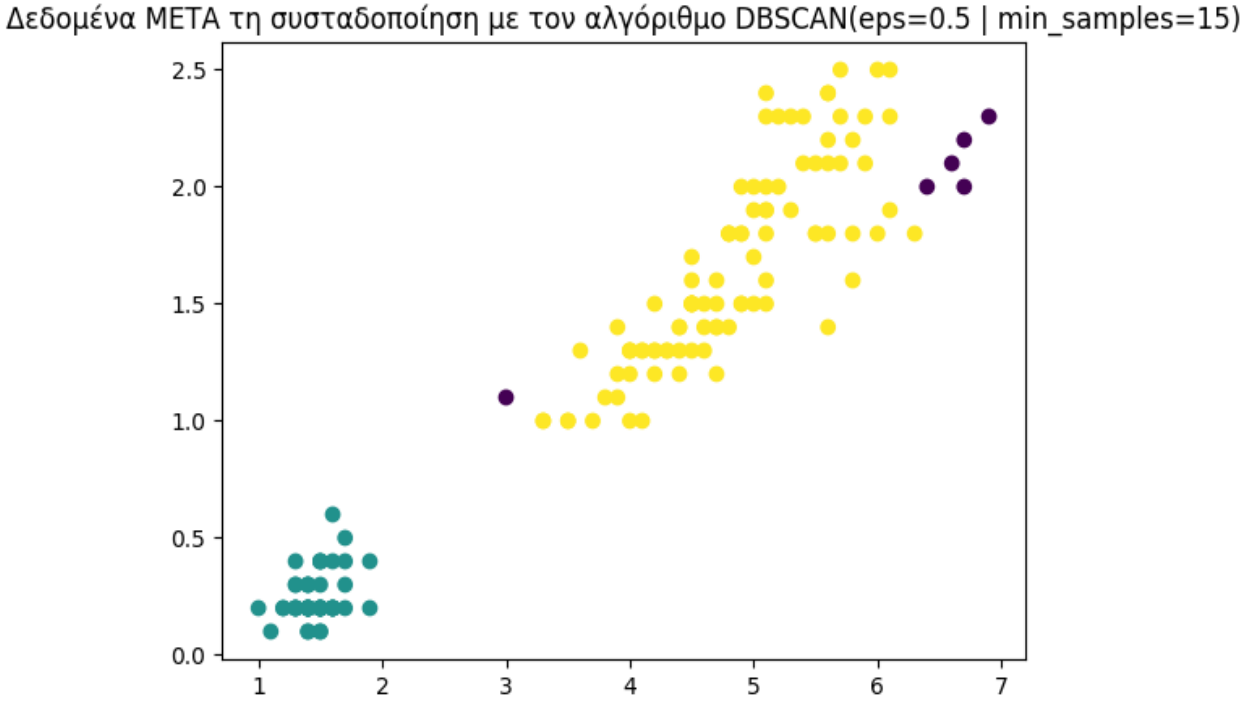
Τέλος, ακολουθώντας τα ίδια βήματα με την αρχή του ερωτήματος, χρησιμοποιούμε τον αλγόριθμο συσταδοποίησης DBSCAN ίδια δεδομένα, με διαφορετικές παραμέτρους. Χαρακτηριστικά, αυξάνουμε και τις 2 παραμέτρους: eps=0.5 και min_samples=15. Από το διάγραμμα, παρατηρούμε ότι τα 2 σχήματα αποτελούν 2 συστάδες.

Επιπροσθέτως, ο θόρυβος έχει μειωθεί εντυπωσιακά. Με την τροποποίηση των παραμέτρων, αυξήσαμε την απόσταση που απέχουν τα οριακά σημεία από ένα κεντρικό σημείο, αλλά και των αριθμό των σημείων που βρίσκονται σε ακτίνα eps και σχηματίζουν μία συστάδα. Όπως είναι λογικό, βελτιώθηκε η ποιότητα της συσταδοποίησης. Ειδικότερα, ο συντελεστής περιγράμματος ισούται με 0.591. Έχει δηλαδή πλέον θετικό πρόσημο και απέχει αρκετά μεγαλύτερη απόσταση από το 0, σε σύγκριση με τις προηγούμενες μετρήσεις για τον αλγόριθμο DBSCAN.

Η κανονικοποίηση των παραπάνω δεδομένων αποτελεί το τελευταίο βήμα του ερωτήματος. Όπως και στην προηγούμενη περίπτωση(διαφορετικοί παράμετροι) μειώθηκε ο χώρος των χαρακτηριστικών με τη χρήση του zscore. Το zscore για ένα σύνολο δεδομένων X: zscore=(X-μ)/σ . Αναλυτικότερα, αποτελεί ένα μέτρο που χρησιμοποιείται για να μετρήσει την απόσταση ενός σημείου από τον μέσο όρο των δεδομένων. Άρα, παίζει καθοριστικό ρόλο στην εύρεση ακραίων σημείων.

Όσον αφορά το γράφημα των δεδομένων μετά την κανονικοποίηση του πίνακα X με τη μέθοδο zscore και χρήση του DBSCAN, παρατηρούμε ότι το πάνω σχήμα αποτελεί τη μοναδική συστάδα. Σε αντίθεση με τις προηγούμενες περιπτώσεις, όλο το κάτω σχήμα ανιχνεύεται ως θόρυβος.

Ο Συντελεστής Περιγράμματος, μετά την κανονικοποίηση, αυξήθηκε ακόμη περισσότερο(0.743) όντας αρκετά κοντά στο 1. Συμπερασματικά, η κανονικοποίηση και στο 2^{ο} σενάριο(διαφορετικές παράμετροι) βελτίωσε την ποιότητα της συσταδοποίησης.



Δεδομένα ΜΕΤΑ την κανονικοποίηση του πίνακα X με τη μέθοδο zscore(DBSCAN(eps=0.5 | min_samples=15))