2η Εργασία Εξόρυξης Δεδομένων

Περιγραφή

Στην εργασία αυτή εφαρμόζονται δύο αλγόριθμοι συσταδοποίησης, K-means και DBSCAN, στα δεδομένα IRIS, xV.mat και mydata.mat. Δοκιμάζονται διαφορετικές παράμετροι για τους αλγορίθμους, γίνεται αξιολόγηση με βάση τον συντελεστή περιγράμματος και το SSE και επιλέγονται διαφορετικά χαρακτηριστικά των δεδομένων για την συσταδοποίησή τους. Με αυτόν τον τρόπο διαπιστώνεται η καλύτερη επιλογή παραμέτρων και χαρακτηριστικών για την επίτευξη καλύτερης συσταδοποίησης.

Στον κώδικα με τον K-means, διαδικασίες οι οποίες ήταν επαναλαμβανόμενες σε όλες τις δοκιμές υλοποιήθηκαν σε συναρτήσεις όπως:

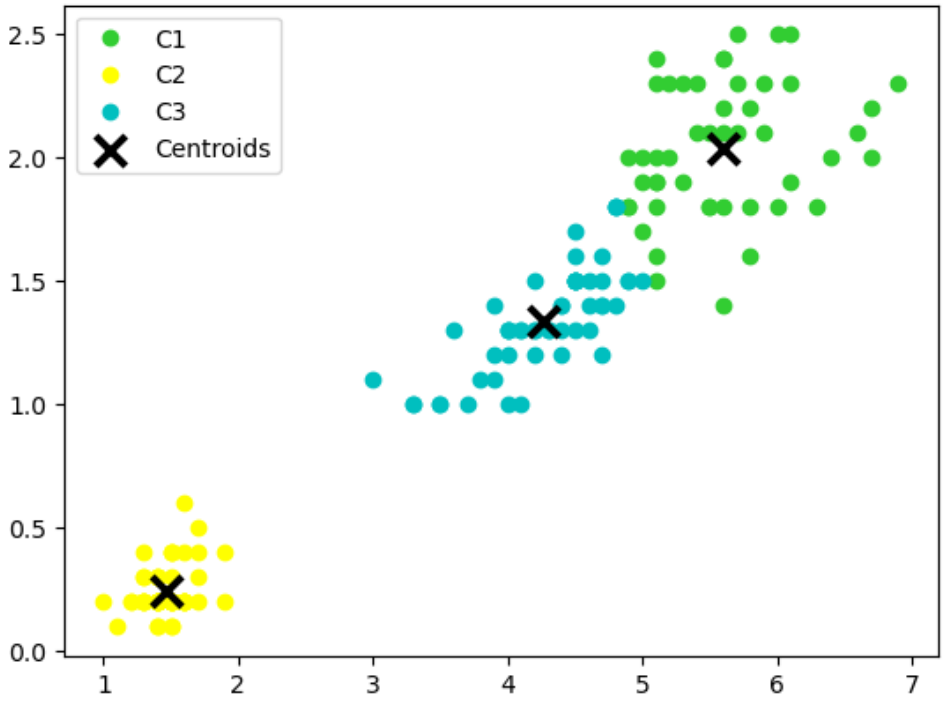
1. showPlot: όπου δέχεται σαν ορίσματα έναν πίνακα X με τα δεδομένα, έναν πίνακα IDX που περιέχει τα labels από την συσταδοποίηση, έναν πίνακα C που περιέχει τα κέντρα των συστάδων και εμφανίζει γράφημα με τις συστάδες με διαφορετικά χρώματα και τα κέντρα τους.
2. findSSE: όπου δέχεται σαν ορίσματα έναν πίνακα X με τα δεδομένα, τον αριθμό των κλάσεων, έναν πίνακα IDX που περιέχει τα labels από την συσταδοποίηση, έναν πίνακα C που περιέχει τα κέντρα των συστάδων και υπολογίζει την τιμή SSE.
3. findK: όπου δέχεται σαν ορίσματα έναν πίνακα X με τα δεδομένα και δοκιμάζει σε ένα εύρος τιμών από 2 έως και 10 για K κέντρα και εμφανίζει μία γραφική με την σχέση k – SSE και μία γραφική για κάθε με συντελεστή περιγράμματος.

Εφαρμογή K-means

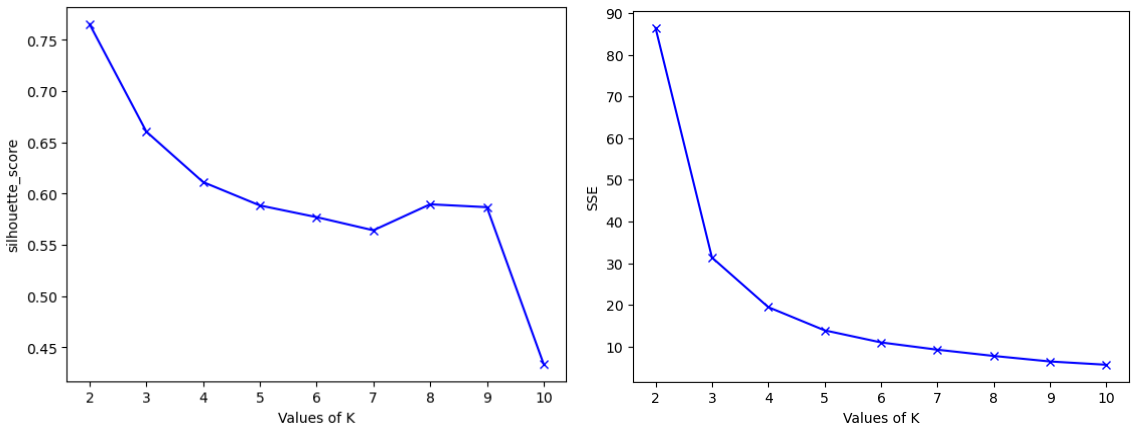
- Iris dataset

Για το Iris dataset θα εφαρμοστεί ο αλγόριθμος του K-means σε διαφορετικά χαρακτηριστικά και θα δοκιμάζεται ένα εύρος τιμών k από 2 έως και 10.

Για την 3η και 4η στήλη του dataset εμφανίστηκαν τα δεδομένα σε γράφημα με k=3 συστάδες.

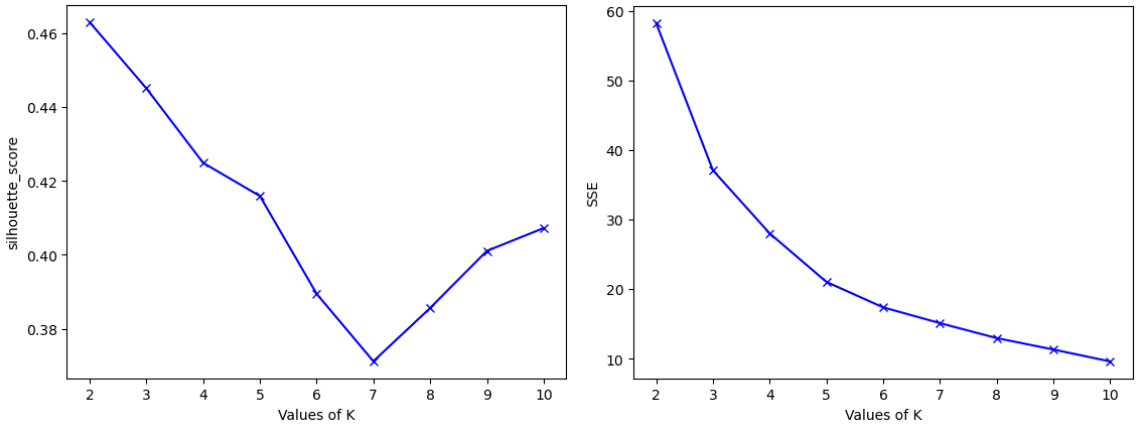


Για να εκτιμηθεί αν η επιλογή του k=3 ήταν σωστή θα πρέπει να δοκιμαστούν διαφορετικές τιμές και να παρατηρηθούν οι τιμές SEE και συντελεστή περιγράμματος όπως φαίνεται παρακάτω. Παρατηρείται ότι η επιλογή για k=3 είναι καλύτερη διότι με βάση το SSE θα διαλεγόταν μία τιμή από την 3 και 4 και με βάση τον συντελεστή περιγράμματος επιλέγεται η τιμή 3 που έχει και μεγαλύτερο συντελεστή.

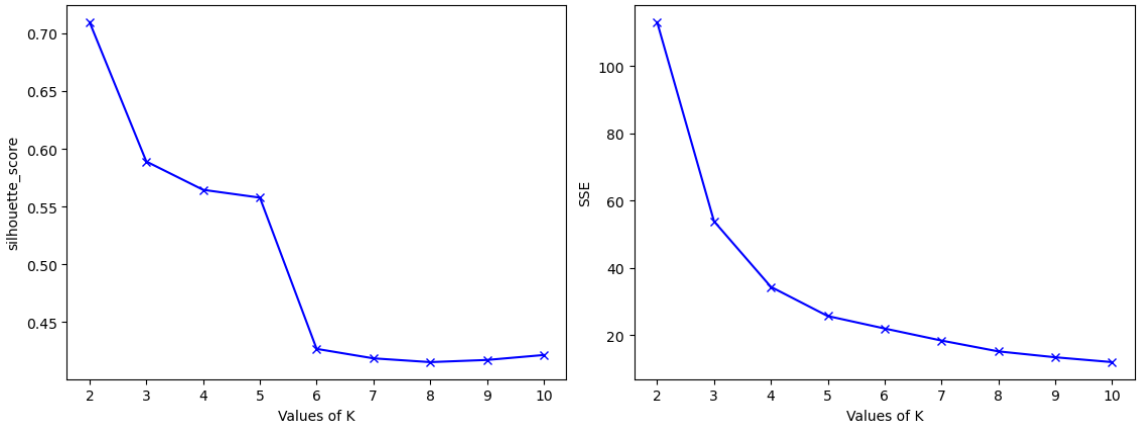


Αυτή η επιλογή τιμής k διατηρείται, όπως φαίνεται παρακάτω, σε όλες τις γραφικές για οποιαδήποτε επιλογή χαρακτηριστικών.

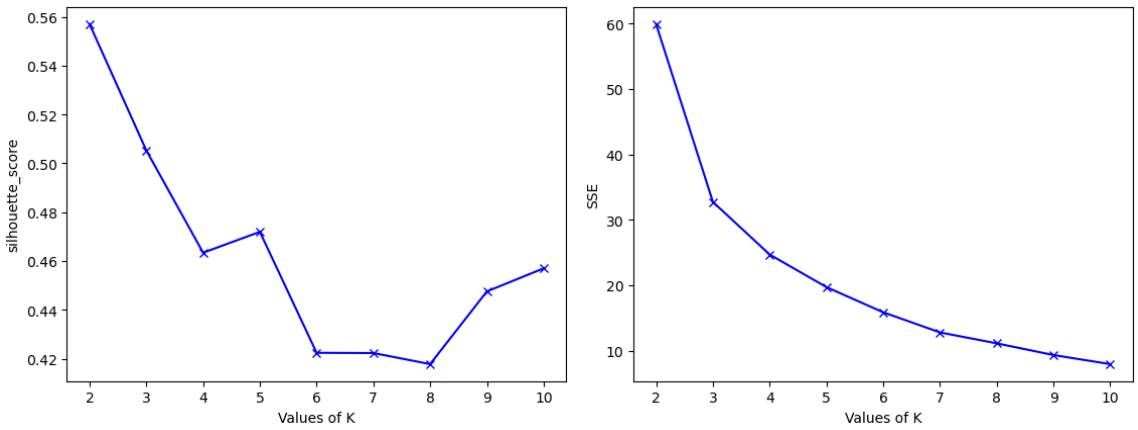
Για την 1η και 2η στήλη.



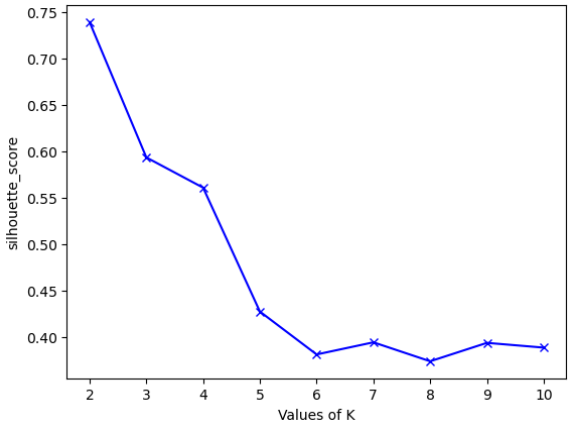
Για την 1η και 3η στήλη.

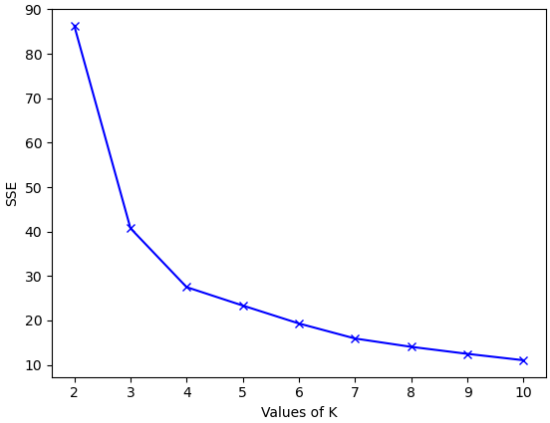


Για την 1η και 4η στήλη.

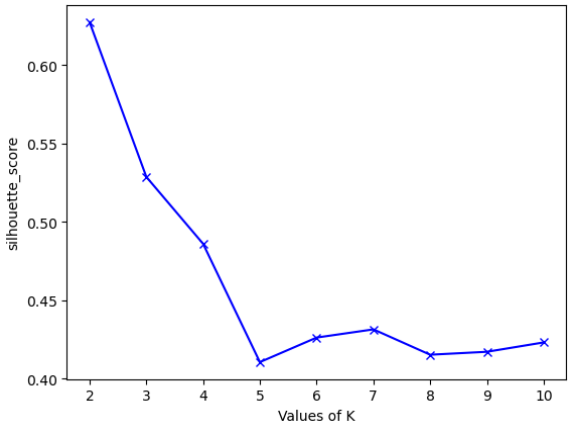


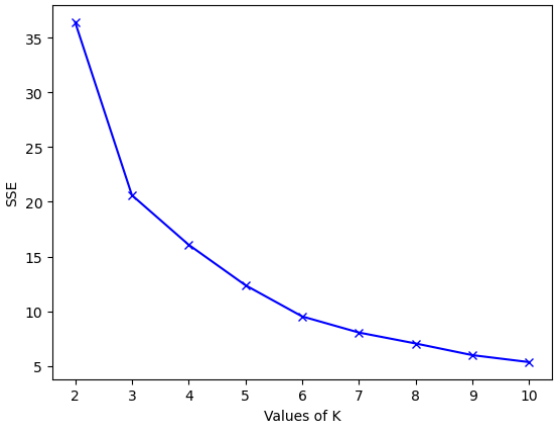
Για την 2η και 3η στήλη.





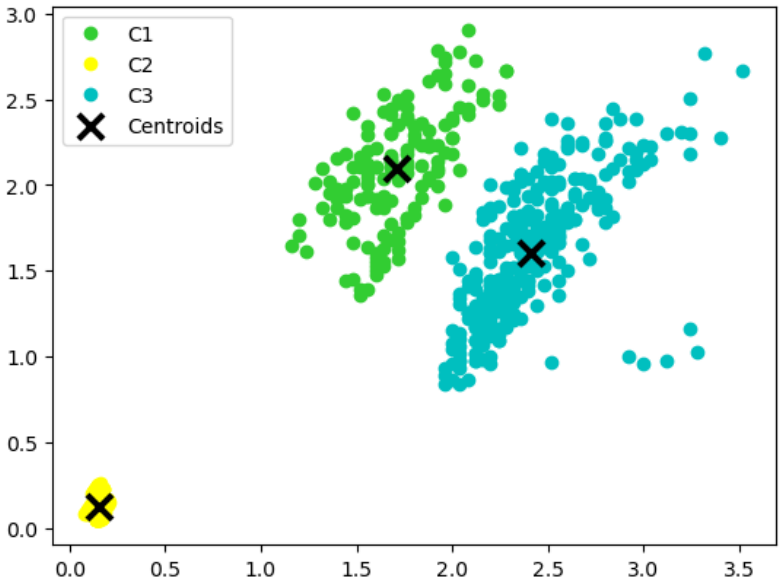
Και για 2η και 4η στήλη.



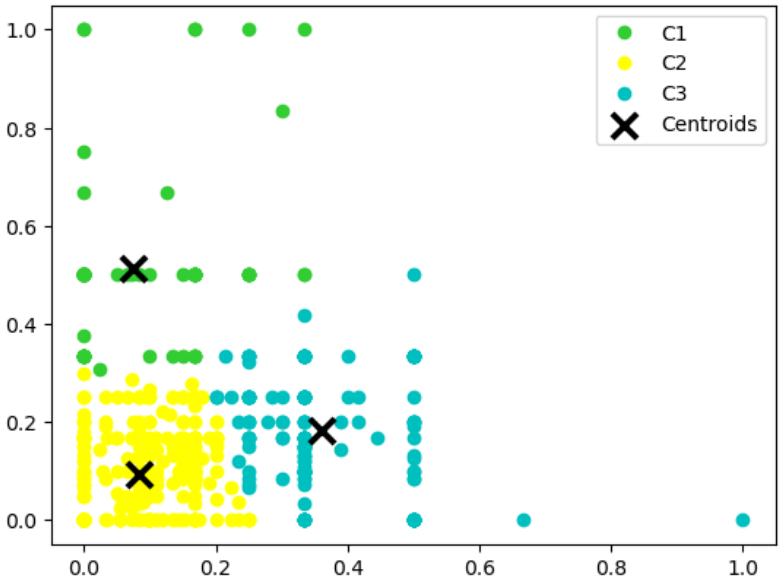


- <u>xV dataset</u>:

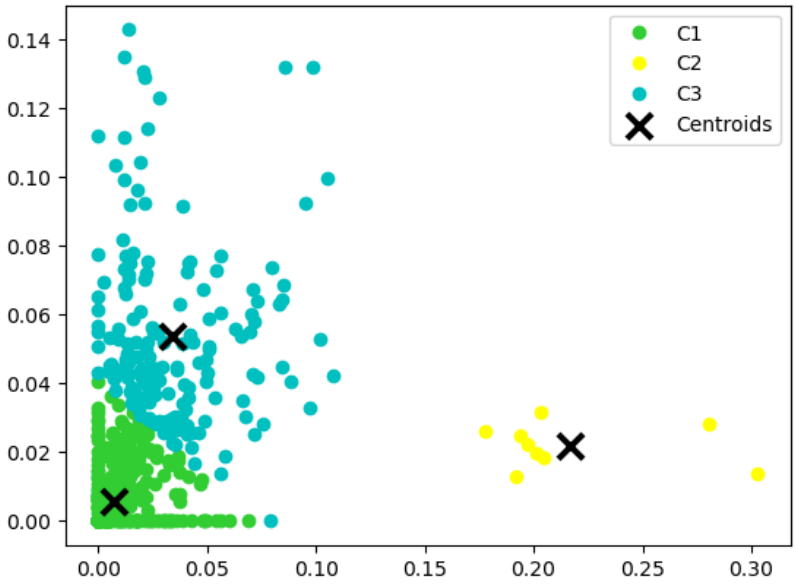
Για την συσταδοποίηση αυτού του dataset με k=3, θα χρησιμοποιηθούν οι πρώτες δύο στήλες με SSE= 99.449, οι στήλες 296 και 305 με SSE= 11.4018, οι δύο τελευταίες στήλες με SSE= 0.3300 και οι στήλες 205 – 175 με SSE= 6.3629.



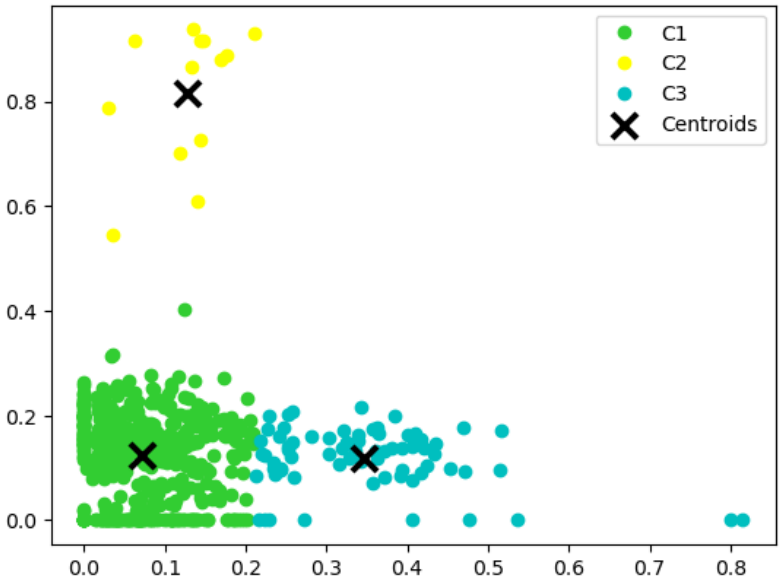
Εικόνα 1 Δεδομένα από τις πρώτες δυο στήλες



Εικόνα 2 Δεδομένα από τις στήλες 296 και 305



Εικόνα 3 Δεδομένα από τις δύο τελευταίες στήλες



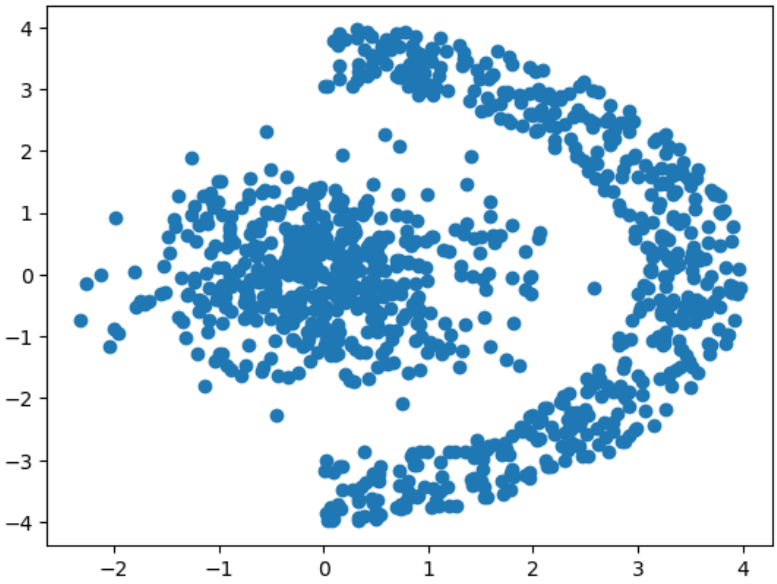
Εικόνα 4 Δεδομένα από τις στήλες 205 και 175

Συγκρίνοντας τα αποτελέσματα από τις πρώτες δύο στήλες, τις στήλες 205-175 και τις δυο τελευταίες στήλες παρατηρούμε ότι με τις δύο πρώτες στήλες γίνεται καλύτερη συσταδοποίηση καθώς τα δεδομένα σχηματίζουν ξεχωριστές και ξεκάθαρες συστάδες. Το SSE εμφανίζεται μεγαλύτερο όμως επειδή τα δεδομένα έχουν μεγαλύτερες τιμές στους άξονες x και y σε σύγκριση με τα δεδομένα των τελευταίων στηλών και των στηλών 205-175.

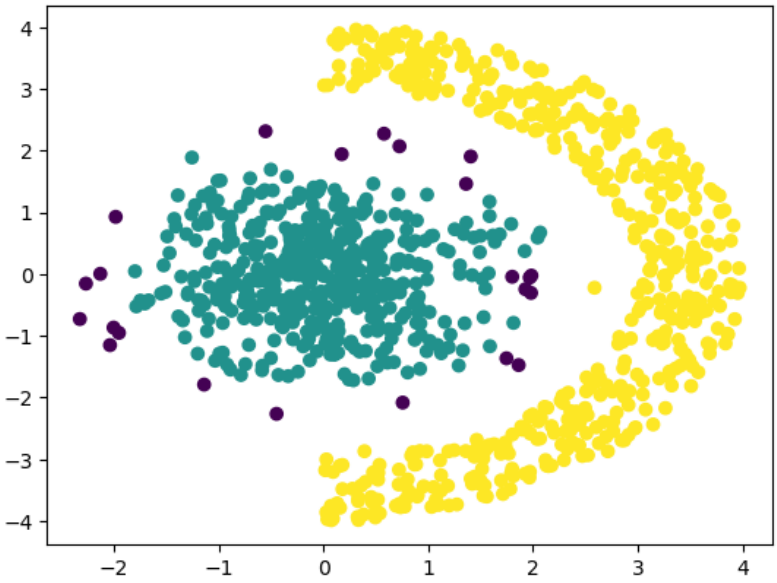
Εφαρμογή DBSCAN

- mydata dataset:

Εκτελώντας την μέθοδο DBSCAN για τις πρώτες δύο διαστάσεις των δεδομένων και με τιμές ε=0,5 και MinPts=15 εμφανίζουμε τα εξής γραφήματα:



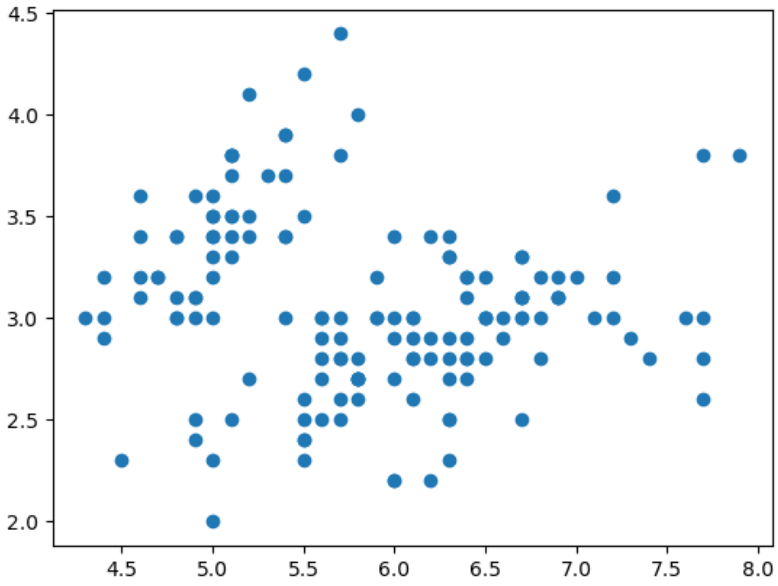
Εικόνα 5 Τα δεδομένα πριν την συσταδοποίηση



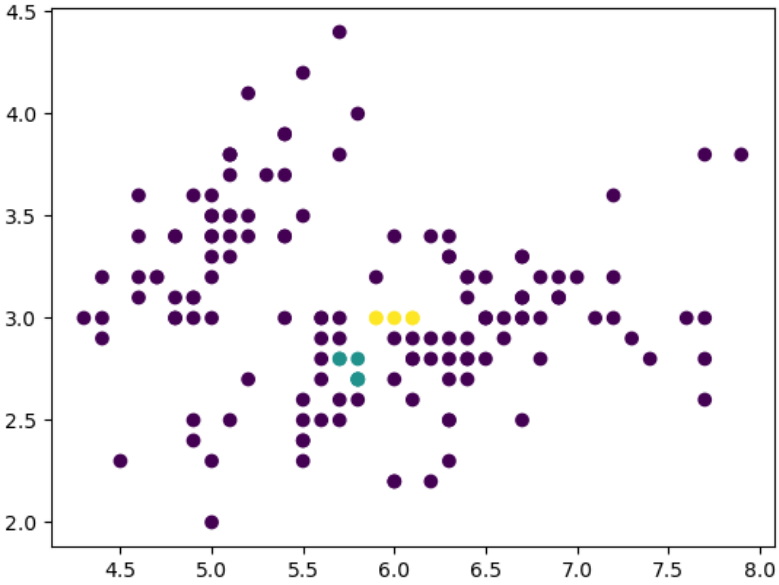
Εικόνα 6 Τα δεδομένα χωρισμένα σε δύο κλάσεις, με μωβ τα δεδομένα που θεωρούνται θόρυβος

- Iris dataset:

Εφαρμόζοντας την μέθοδο DBSCAN στις δύο πρώτες διαστάσεις του dataset με ε=0,1 και MinPts=5 εμφανίζουμε τα εξής γραφήματα:

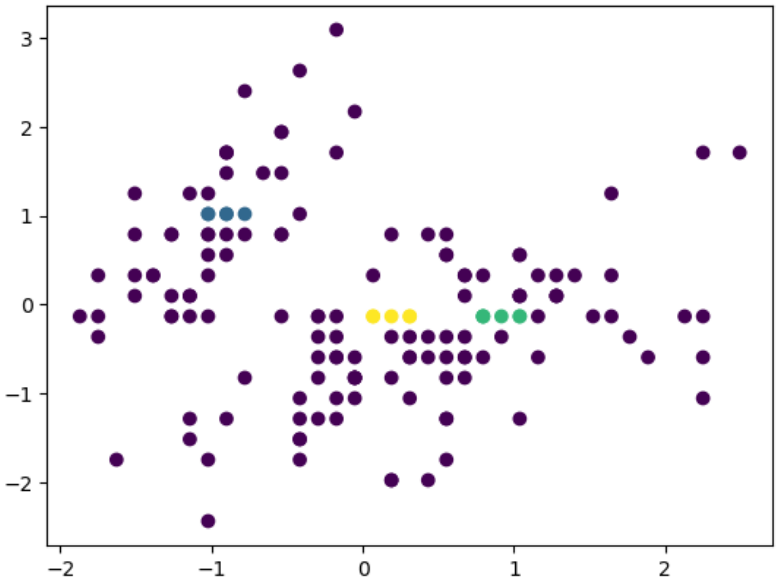


Εικόνα 7 Τα δεδομένα πριν την συσταδοποίηση

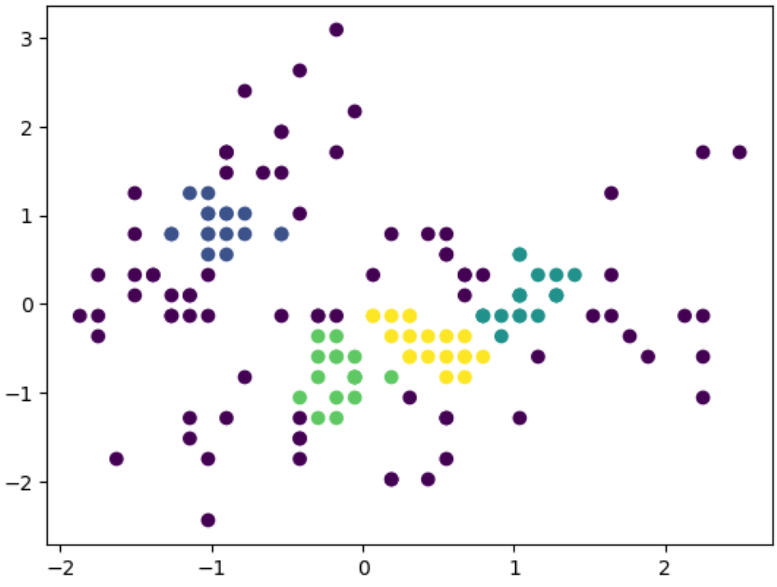


Εικόνα 8 Τα δεδομένα χωρισμένα σε δύο κλάσεις, με μωβ τα δεδομένα που θεωρούνται θόρυβος

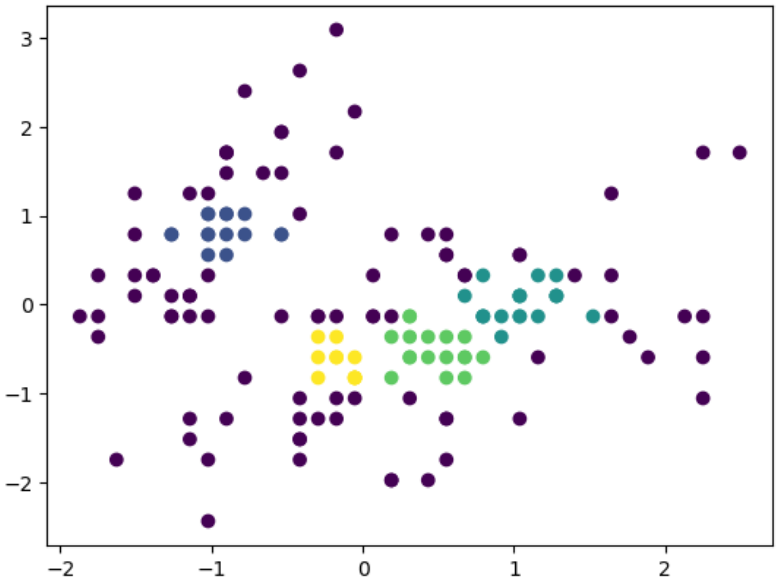
Παρατηρείται ότι κατηγοριοποιεί ελάχιστα δεδομένα σε δυο συστάδες και όλα τα υπόλοιπα τα εμφανίζει σαν θόρυβο. Αν κανονικοποιήσουμε τα δεδομένα με την μέθοδο zscore παρατηρούμε ότι με τις ίδιες παραμέτρους, ο SBSCAN αλγόριθμος δεν φτιάχνει συστάδες. Αν κρατήσουμε το MinPts σταθερό στα 5 σημεία και αυξήσουμε την ακτίνα ε θα παρατηρήσουμε ότι ο αλγόριθμος φτιάχνει συστάδες.



Εικόνα 9 Για ε= 0.2 και MinPts= 5



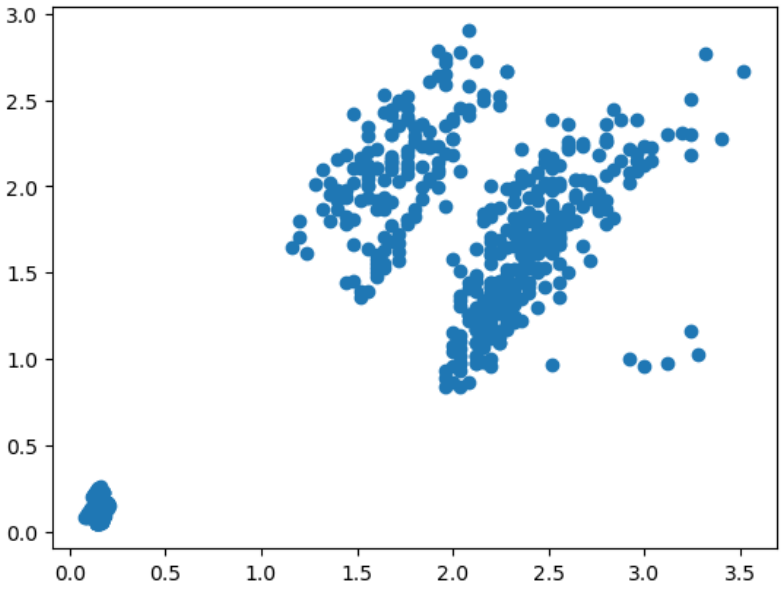
Εικόνα 10 Για ε= 0.3 και MinPts= 10



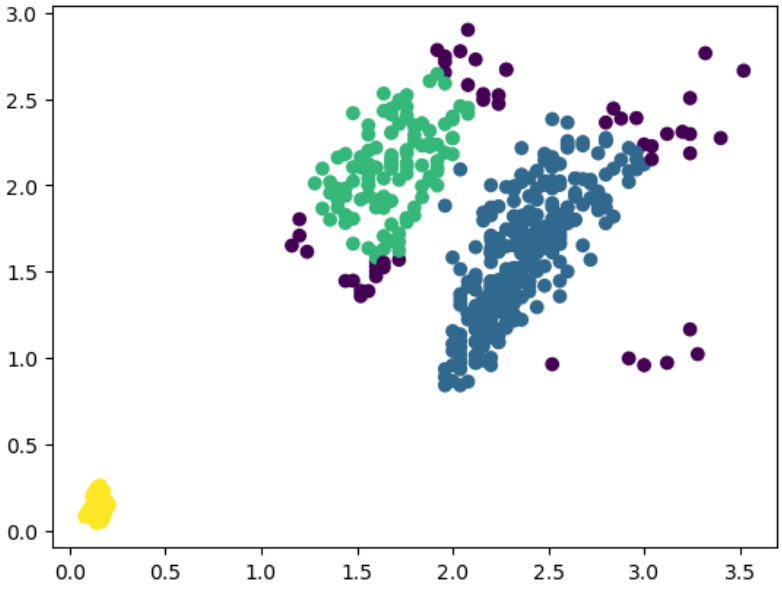
Εικόνα 11 Για ε= 0.4 και MinPts= 15

- xV dataset:

Εφαρμόζοντας την μέθοδο DBSCAN στις δύο πρώτες διαστάσεις του dataset με ε=0,3 και MinPts=50 εμφανίζουμε τα παρακάτω γραφήματα και παρατηρούμε ότι μερικά δεδομένα αρκετά κοντά στην μορφή των συστάδων θεωρούνται ως θόρυβος ενώ με το μάτι θα μπορούσαν να θεωρηθούν δεδομένα των συστάδων:

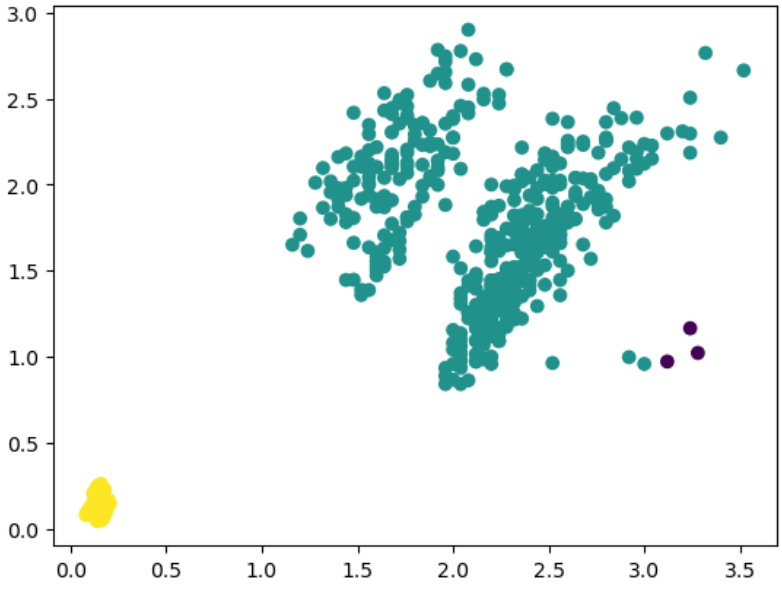


Εικόνα 12 Τα δεδομένα πριν την συσταδοποίηση



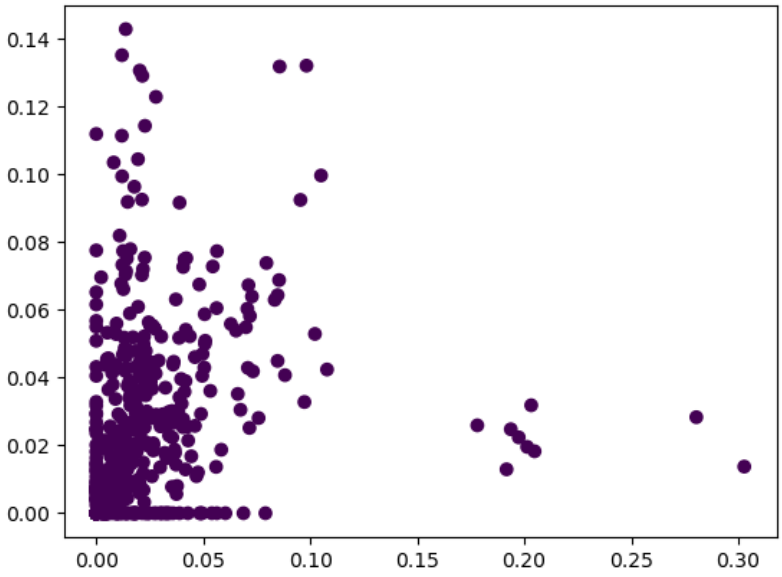
Εικόνα 13 Τα δεδομένα χωρισμένα σε τρεις κλάσεις, με μωβ τα δεδομένα που θεωρούνται θόρυβος

Αυξάνοντας στα συγκεκριμένα δεδομένα το ε=0,6 και μειώνοντας το κατώτατο όριο γειτόνων σε MinPts=43 διαμορφώνει μόνο 2 συστάδες και θέτει ως θόρυβο ελάχιστα δεδομένα

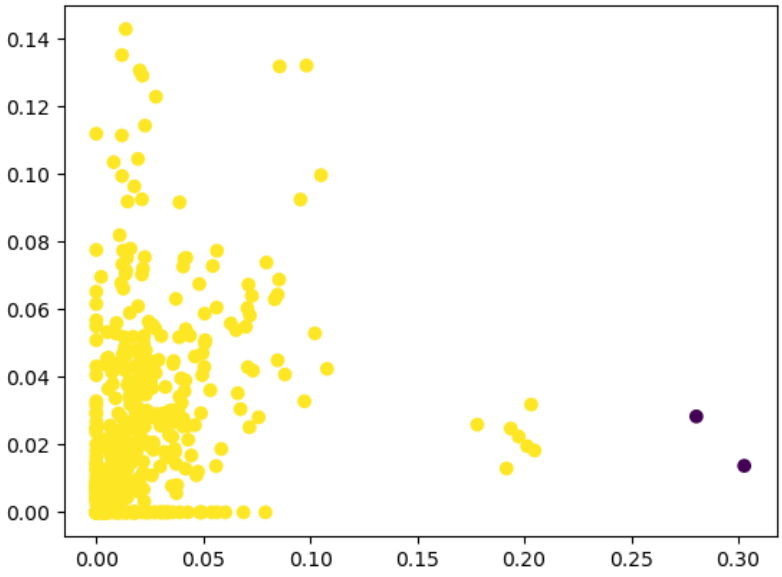


Εικόνα 14 Τα δεδομένα χωρισμένα σε δύο κλάσεις, με μωβ τα δεδομένα που θεωρούνται θόρυβος

Στην εφαρμογή του DBSCAN στις δυο τελευταίες διαστάσεις των δεδομένων παρατηρείται ότι για μεγάλους αριθμούς ε και MinPts δεν δημιουργούνται συστάδες

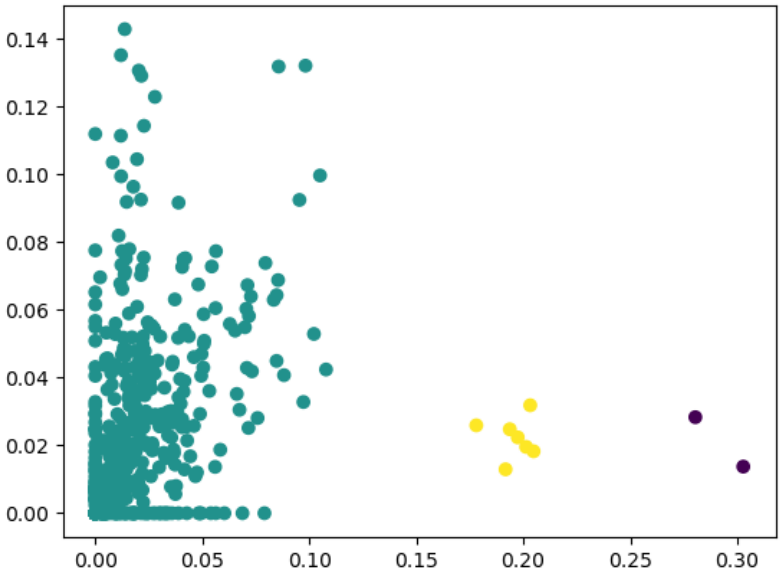


Εικόνα 15 Για ε= 0.5 και MinPts= 5+

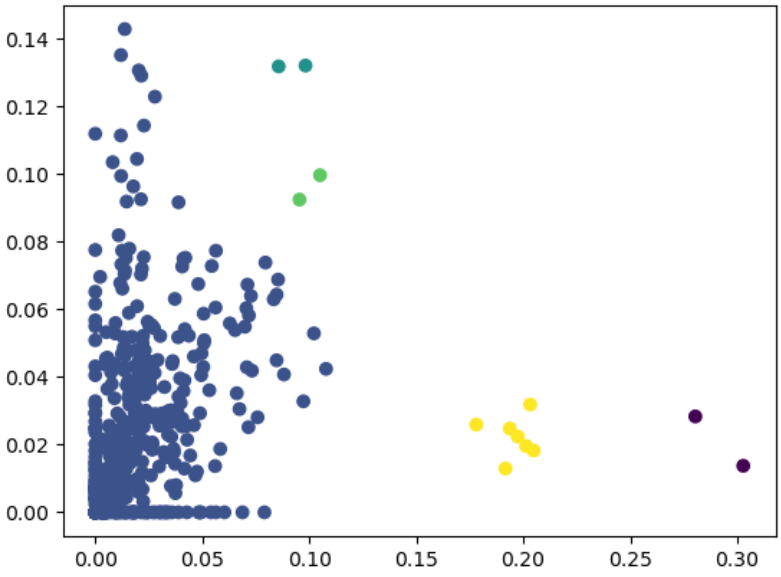


Εικόνα 16 Για ε= 0.9 και MinPts= 10

Όσο αυξάνουμε την ακτίνα την οποία θα εξετάζεται η γειτονιά του κάθε δεδομένου τόσο περισσότερο θα αυξάνεται η πυκνότητα των σημείων που θα θεωρούνται βασικά ή οριακά. Στις συγκεκριμένες διαστάσεις το ε πρέπει να είναι κάτω του 0.1 και ύστερα το πόσες συστάδες θα δημιουργήσει εξαρτάται από το πόσο μικρές τιμές θα έχουν οι παράμετροι ε και MinPts



Εικόνα 17 Για ε= 0.05 και MinPts= 5



Εικόνα 18 Για ε= 0.02 και MinPts= 2