Πανεπιστήμιο Δυτικής Αττικής

Σχολή Μηχανικών

Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

Εξόρυξη Δεδομένων

ΘΩΜΑΣ ΝΙΚΟΛΑΟΣ - ΑΜ: 21390068

ΑΘΗΝΑ

Σάββατο, 11 Ιανουαρίου 2025

Περιεχόμενα

1. K-MEANS

1.1 Εφαρμογή στο σύνολο δεδομένων iris.

1. Φόρτωση του συνόλου δεδομένων Iris από το sklearn.datasets. Χρησιμοποιούνται οι 2 τελευταίες διαστάσεις του πίνακα (μήκος και πλάτος πετάλου).
2. Εκτέλεση του k-means με αριθμό συστάδων k=3.
3. Στο παρακάτω γράφημα απεικονίζονται τα δεδομένα με διαφορετικά χρώματα για κάθε συστάδα και μαρκαρισμένα τα κεντροειδή.

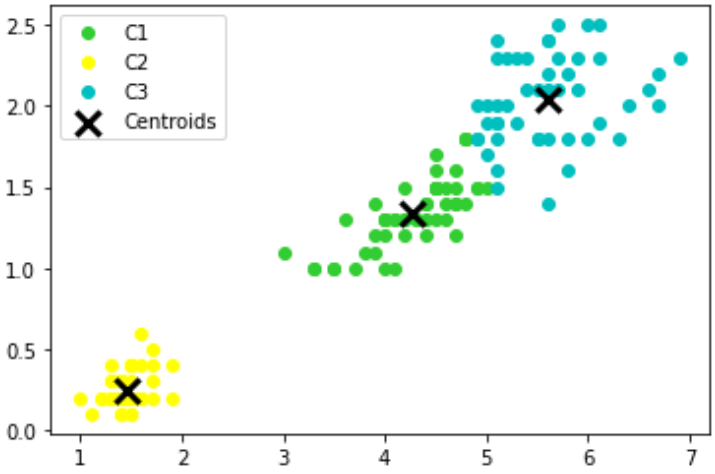
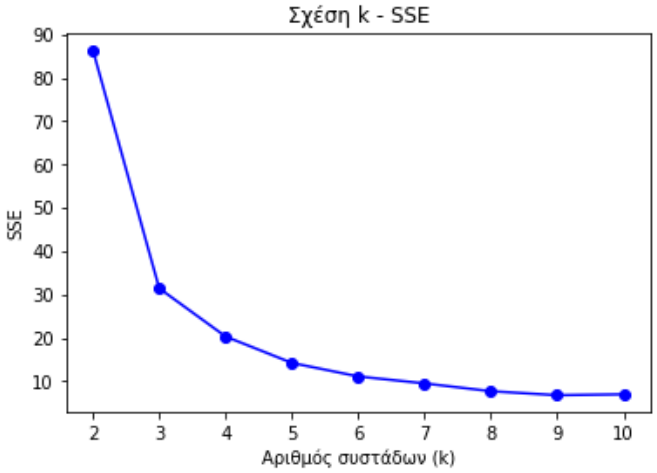


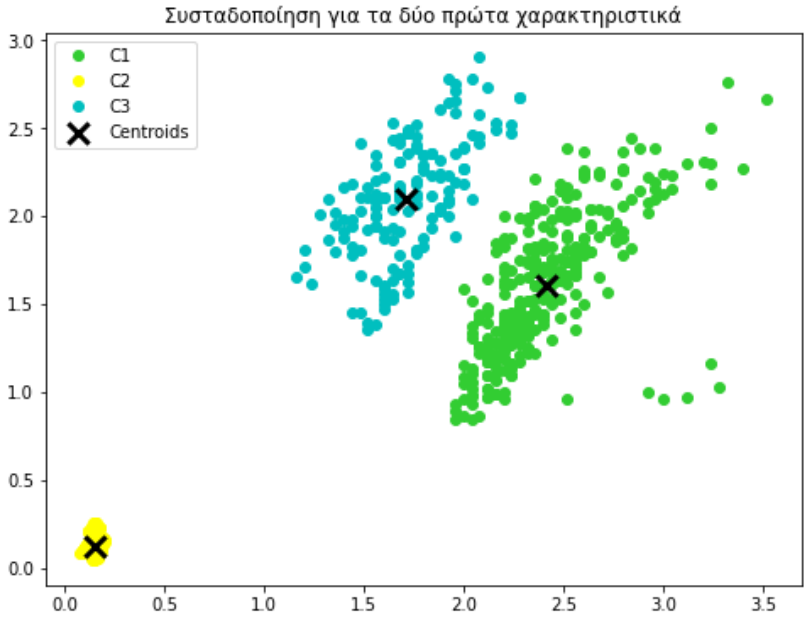
Figure 1: Εκτέλεση του k-means.

4. Μελέτη της επίδρασης του k στη συσταδοποίηση (2 – 10). Υπολογίστηκε το SSE (άθροισμα τετραγωνικών αποστάσεων) και ο συντελεστής Silhouette. Η σχέση k – SSE απεικονίζεται στο παρακάτω γράφημα:

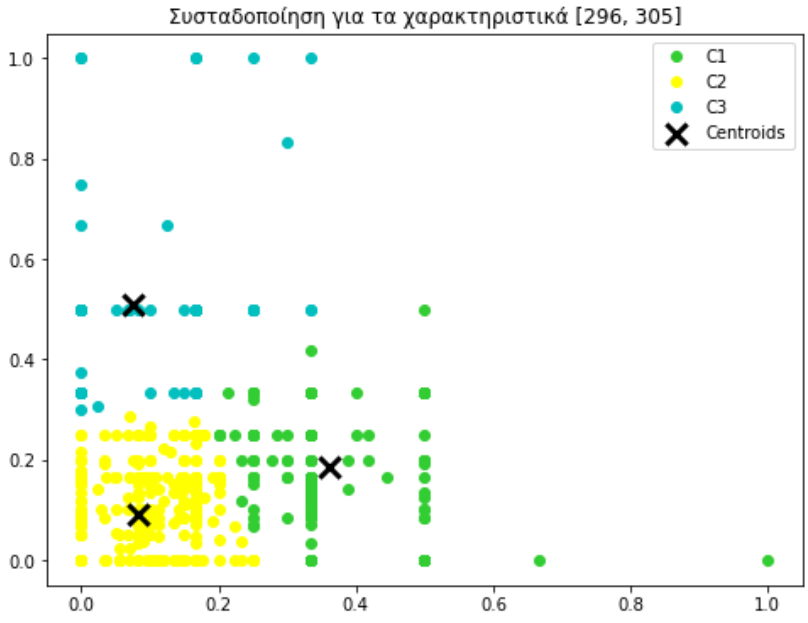


1.2 Εφαρμογή στο σύνολο δεδομένων xV.mat.

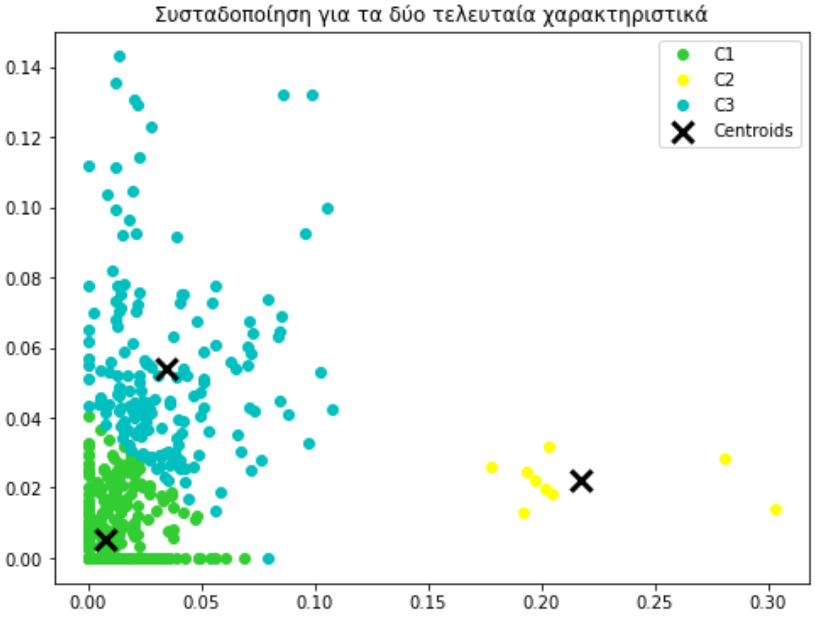
1. Φόρτωση του συνόλου δεδομένων xV.mat με 600 δείγματα και 469 χαρακτηριστικά.
2. Εκτέλεση του k-means για τα δύο πρώτα χαρακτηριστικά [0, 1], απεικονίζεται στο παρακάτω γράφημα:



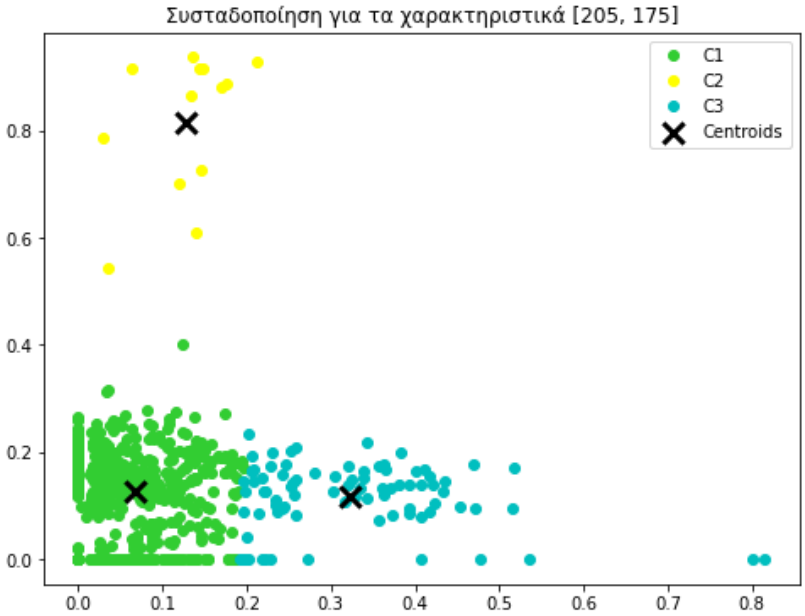
3. Εκτέλεση του k-means για τα χαρακτηριστικά [296, 305], απεικονίζεται στο παρακάτω γράφημα:



4. Εκτέλεση του k-means για τα δύο τελευταία χαρακτηριστικά [467, 468], απεικονίζεται στο παρακάτω γράφημα:



5. Εκτέλεση του k-means για τα χαρακτηριστικά [205, 175], απεικονίζεται στο παρακάτω γράφημα:



6. Σύγκριση SSE:

Βήμα 2 (χαρακτηριστικά [0,1]): 110.00
Βήμα 4 (τελευταία χαρακτηριστικά): 0.33
Βήμα 5 (χαρακτηριστικά [205,175]): 9.04

Βάση των παραπάνω αποτελεσμάτων, για τα δύο πρώτα χαρακτηριστικά το SSE είναι αρκετά υψηλό που σημαίνει ότι οι συστάδες έχουν μεγάλη διασπορά. Για τα τελευταία χαρακτηριστικά το SSE είναι πολύ χαμηλό που μας δείχνει ότι οι συστάδες είναι καλά ορισμένες. Τέλος για τα χαρακτηριστικά [205, 175] έχουμε κάτι στο ενδιάμεσο.

Οπότε συμπεραίνουμε ότι τα τελευταία χαρακτηριστικά [467, 468] προσφέρουν την καλύτερη διάκριση μεταξύ των συστάδων, ενώ τα πρώτα χαρακτηριστικά [0, 1] τη χειρότερη.

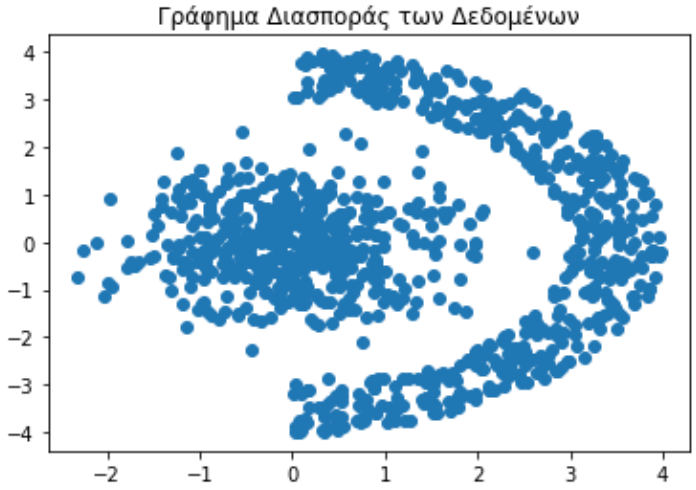
2. DBSCAN

2.1 Εφαρμογή στο σύνολο δεδομένων mydata.

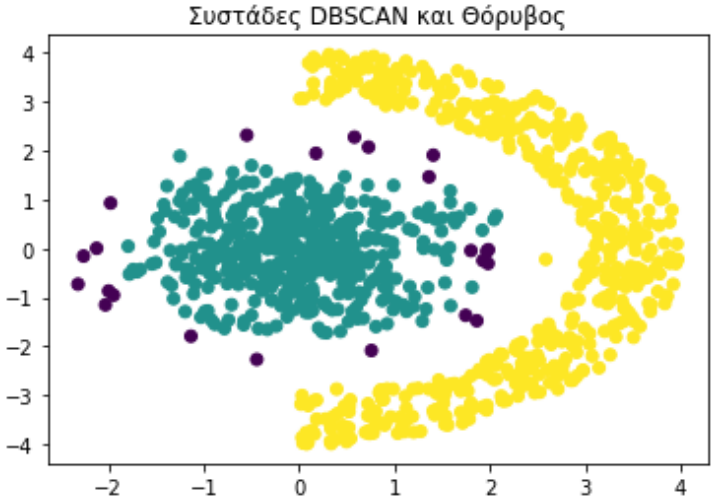
1. Φόρτωση δεδομένων από το mydata.

2. Εφαρμογή DBSCAN στις δύο πρώτες διαστάσεις με παραμέτρους eps=0.5 και MinPts=15.

3. Στο παρακάτω γράφημα διασποράς απεικονίζονται οι τιμές των δύο διαστάσεων του πίνακα X.



4. Στο παρακάτω γράφημα απεικονίζονται οι συστάδες στις οποίες χώρισε τα δεδομένα η μέθοδος DBSCAN, καθώς και τον θόρυβο.

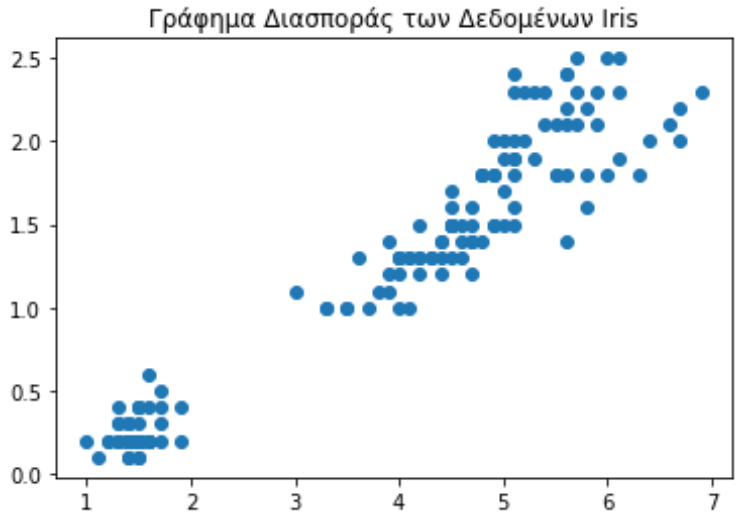


2.2 Εφαρμογή στο σύνολο δεδομένων iris.

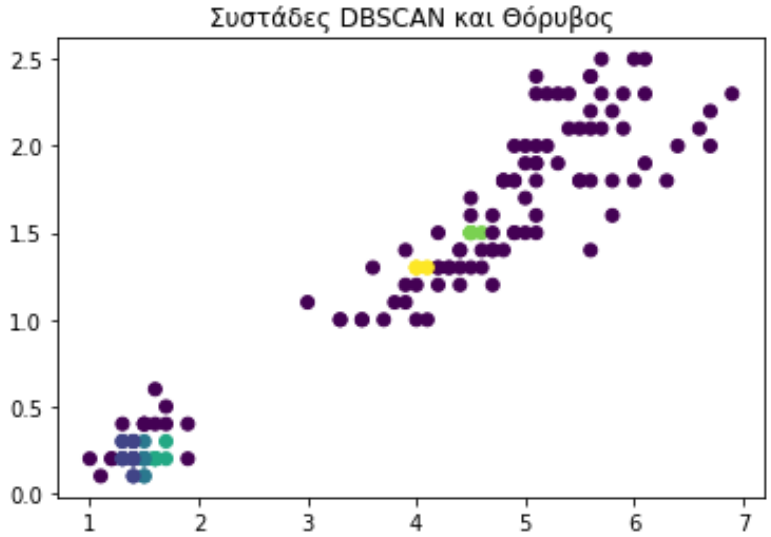
1. Φόρτωση των δεδομένων Iris και επιλογή των διαστάσεων [2,3].

2. Εκτέλεση της DBSCAN με παραμέτρους eps=0.1 και MinPts=5.

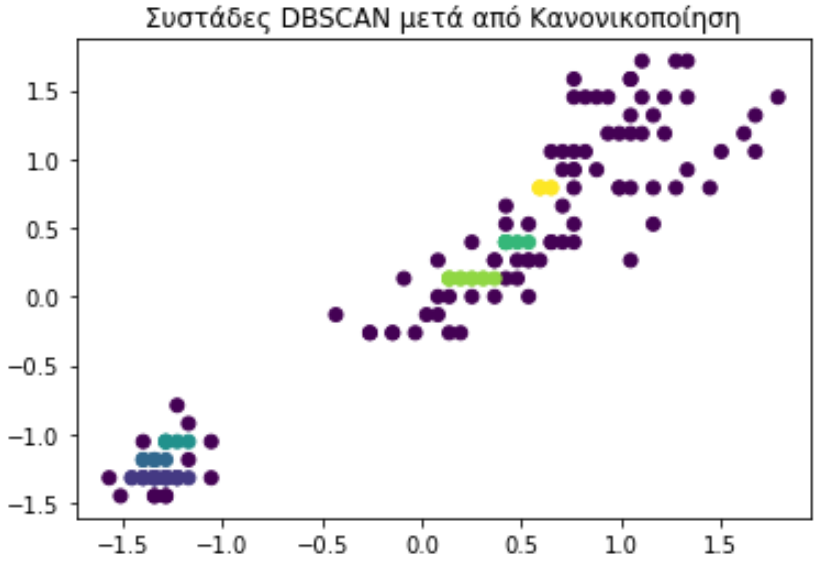
3. Στο παρακάτω γράφημα διασποράς απεικονίζονται οι τιμές των δύο διαστάσεων του X.



4. Στο παρακάτω γράφημα απεικονίζονται οι συστάδες στις οποίες χώρισε τα δεδομένα η μέθοδος DBSCAN, καθώς και τον θόρυβο.



5. Κανονικοποίηση δεδομένων με Z-Score. Στο παρακάτω γράφημα απεικονίζονται οι συστάδες στις οποίες χώρισε τα δεδομένα η μέθοδος DBSCAN, μετά από κανονικοποίηση.



6. Η κανονικοποίηση βελτίωσε τα αποτελέσματα του DBSCAN στο σύνολο Iris, αφού τα χαρακτηριστικά αποκτούν ίδια κλίμακα και κατανομή. Έτσι ο DBSCAN μπορεί να εντοπίσει συστάδες πιο αποτελεσματικά.