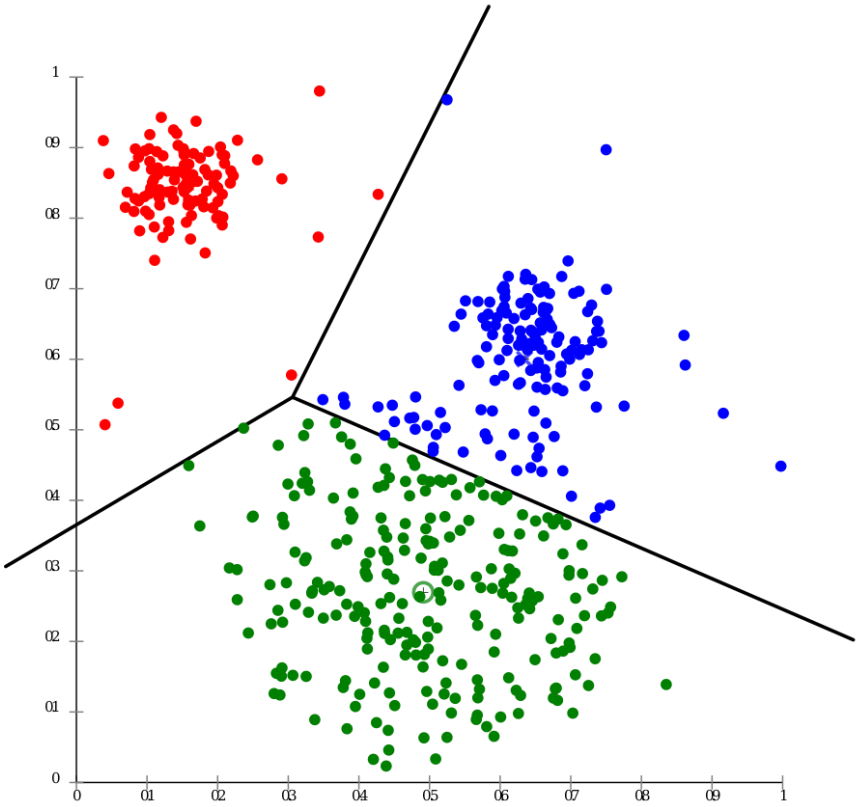
Εργασία 3 στην Αναγνώριση προτύπων



Χατζησάββας Χρήστος

Ακαδημαϊκό έτος 2023-2024

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΤΜΗΜΑ ΗΜ&ΜΥ

Άσκηση 1:

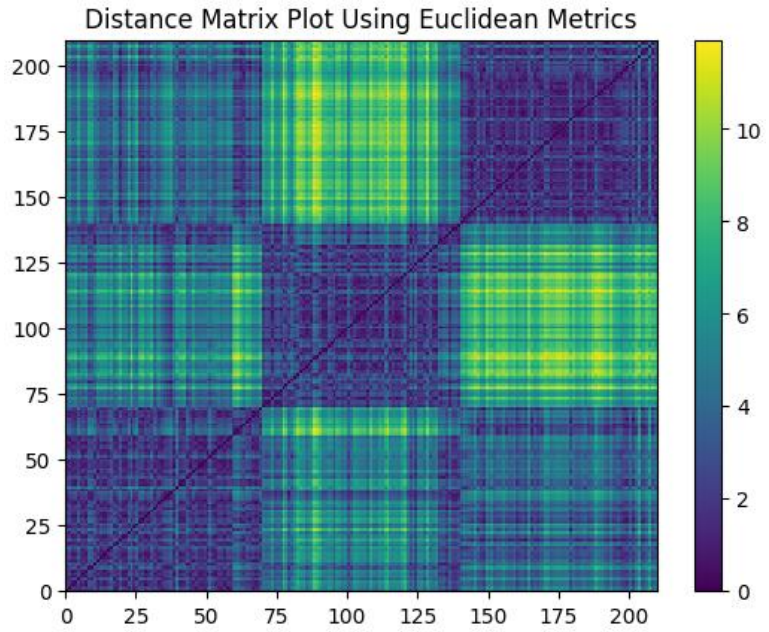
Αρχικά χωρίζουμε τα δεδομένα του dataset σε έναν πίνακα που περιέχει μόνο τα χαρακτηριστικά των δειγμάτων (αφαιρούμε δηλαδή τις κλάσεις). Για την ευκλείδεια απόσταση γνωρίζουμε ότι εάν έχουμε 2 διανύσματα x = (x1,x2..,xn) και y = (y1,y2…,yn) η απόσταση δίνεται από τον τύπο:

$$\text{distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

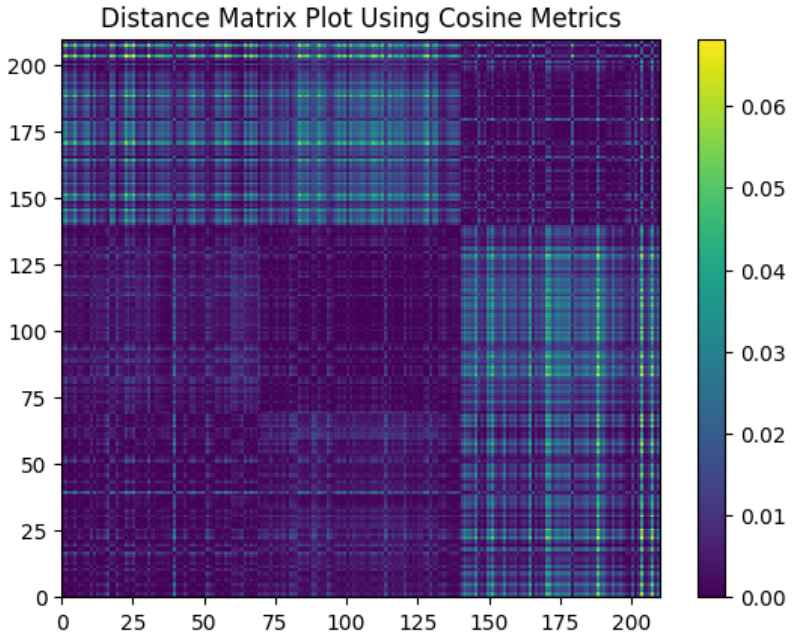
ενώ για την cosine μετρική ο τύπος είναι:

$$\cos(\theta) = \frac{x * y}{\|x\| * \|y\|}$$

A) Στον πίνακα αποστάσεων που δημιουργήθηκε χρησιμοποιώντας Euclidean μετρικές (με την συνάρτηση pairwise_distances) παρατηρούμε πως οι κλάσεις 2 και 3 είναι ευκολότερο να διαχωριστούν καθώς εκεί οι αποστάσεις είναι μεγαλύτερες. Πιο συγκεκριμένα από το plot που ακολουθεί γίνεται εμφανές πως εκεί που έχουμε πιο ανοιχτό χρώμα έχουμε μεγαλύτερη απόσταση μεταξύ των δειγμάτων. Αντιθέτως στις πιο σκούρες περιοχές έχουμε μικρότερες αποστάσεις που δυσκολεύουν τον σωστό διαχωρισμό μεταξύ των κλάσεων.

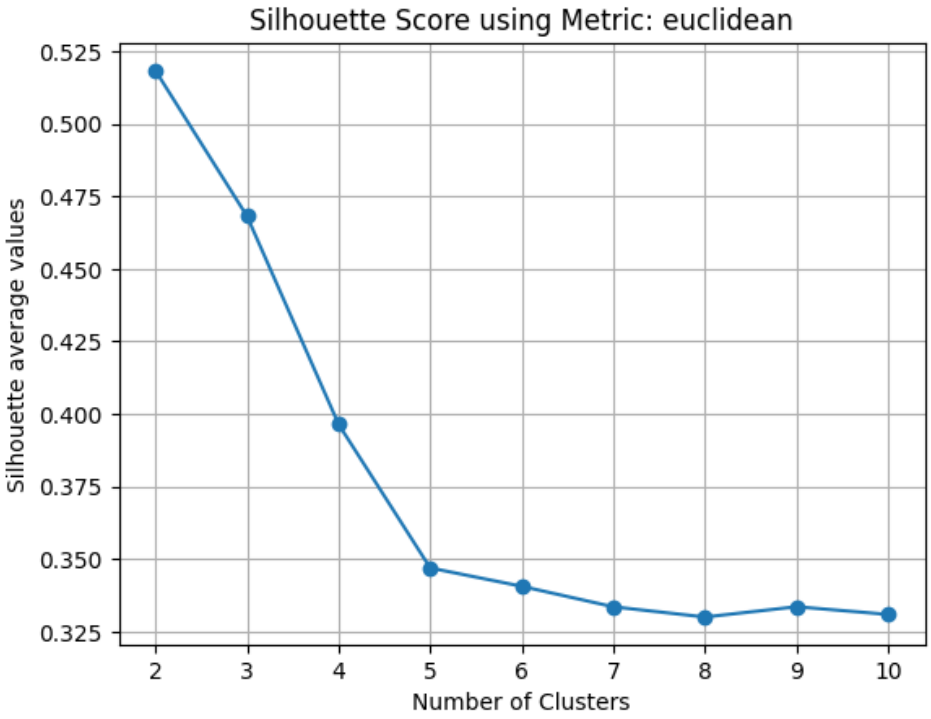


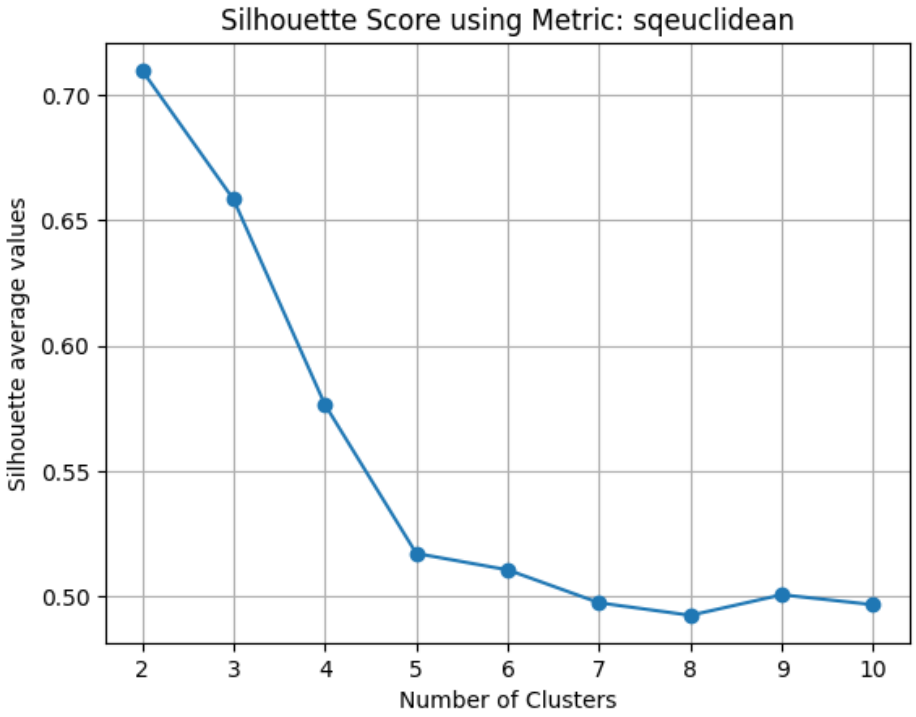
Χρησιμοποιώντας τώρα Cosine μετρικές, λόγω της διαφορετικής προσέγγισης των αποστάσεων σε σχέση με τις Ευκλείδειες παρατηρούμε πως οι κλάσεις 1 και 3 αλλά και οι κλάσεις 2 και 3 μπορούν να διαχωριστούν ευκολότερα λόγω μεγαλύτερων αποστάσεων.



B) Ο αλγόριθμος KMeans είναι ένας αλγόριθμος ομαδοποίησης που χρησιμοποιείται σε μάθηση χωρίς επίβλεψη (unsupervised learning). Πιο συγκεκριμένα, ο αλγόριθμος δίχως να γνωρίζει τις ετικέτες προσπαθεί να δημιουργήσει clusters. Στην περίπτωση του KMeans, χωρίζουμε το σύνολο των δεδομένων σε k clusters όπου κάθε δείγμα θα ανήκει στο cluster με την μικρότερη απόσταση από αυτό. Προκειμένου να αξιολογηθεί ο αλγόριθμος χρησιμοποιείται η μέθοδος silhouette coefficient. Για ένα σημείο, η silhouette coefficient ορίζεται ως: $s = 1 – a / b$ με $a < b$, όπου το a είναι η μέση απόσταση του σημείου από όλα τα υπόλοιπα σημεία του cluster και β είναι η ελάχιστη μέση απόσταση του σημείου από όλα τα υπόλοιπα clusters. Η τιμή του s κυμαίνεται από το 0 στο 1, και όσο πιο κοντά βρίσκεται στο 1 είναι καλύτερα ενταγμένο το συγκεκριμένο σημείο στην εκάστοτε κλάση σε σχέση με τις υπόλοιπες.

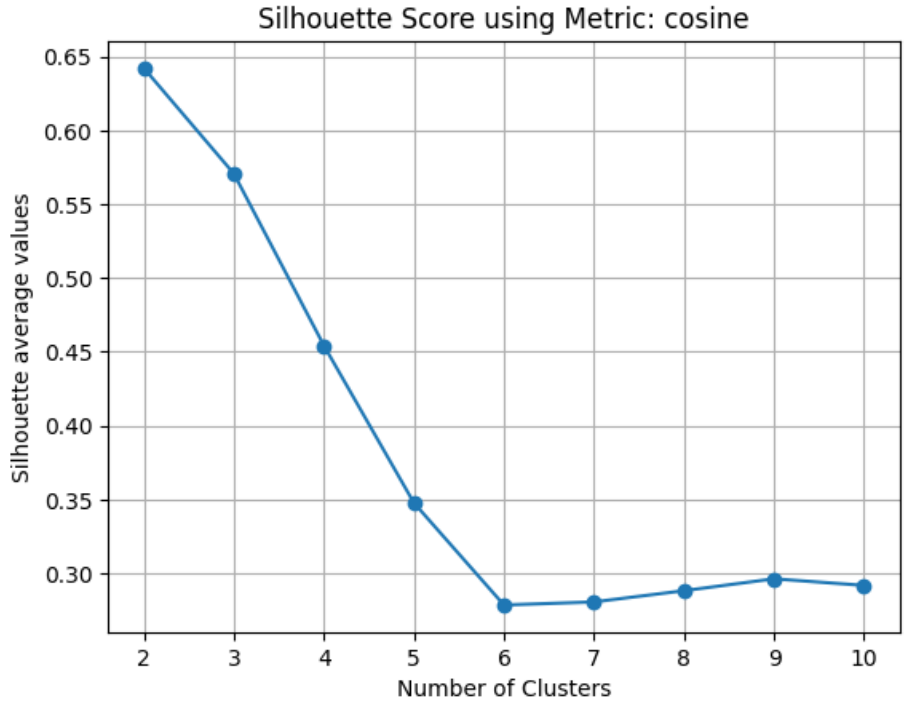
Απεικονίζοντας τα διαγράμματα του Silhouette για Ευκλείδεια και για squared Euclidean μετρική παρατηρούμε πως είναι σχεδόν ίδια και διαφοροποιούνται μόνο στην τιμή που παίρνει η μέση τιμή για τα διαφορετικά k που ορίζουμε. Αυτό οφείλεται κυρίως στο γεγονός πως η κάθε μέθοδος υπολογίζει τις αποστάσεις με διαφορετικό τρόπο. Ο βέλτιστος αριθμός κλάσεων και στις 2 περιπτώσεις είναι οι 2 κλάσεις καθώς εκεί έχουμε την μέγιστη μέση τιμή του Silhouette που προκύπτει από την συνάρτηση silhouette_score().

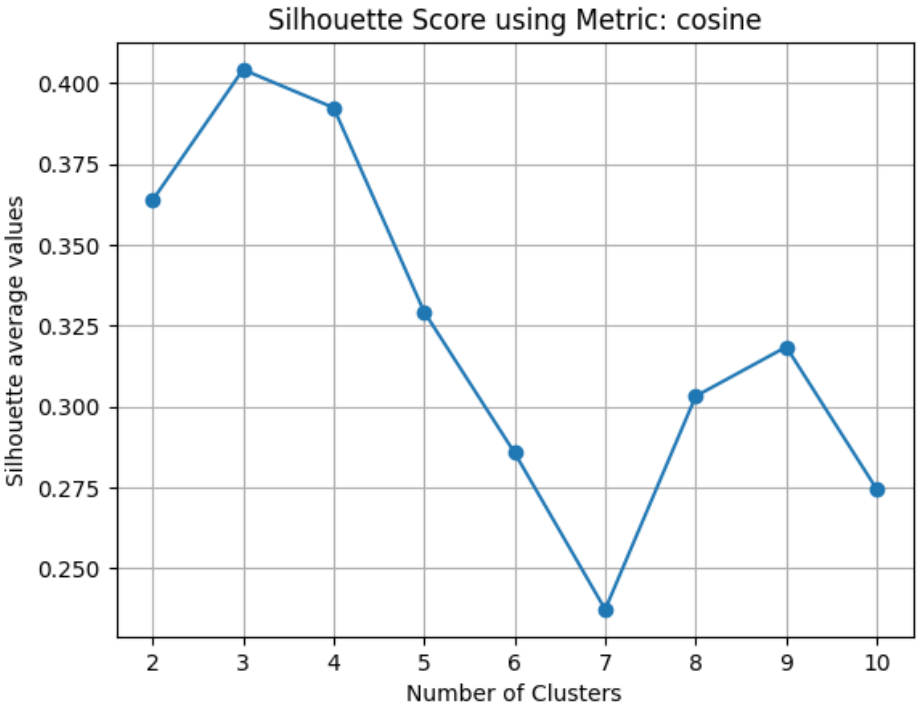




Γ) Χρησιμοποιώντας cosine μετρική με κανονικοποιημένα δεδομένα, δεν παρατηρούμε κάποια αλλαγή στον βέλτιστο αριθμό clusters. Αντιθέτως, αν χρησιμοποιήσουμε τα αρχικά δεδομένα εφαρμόζοντας cosine μετρική παρατηρούμε πως ο ιδανικός αριθμός clusters γίνεται 3. Ένας λόγος που πιθανώς εξηγεί αυτήν την διαφορά είναι ότι τα δεδομένα έπειτα από την κανονικοποίηση είναι πιο εύκολο να διαχωριστούν στον χώρο ώστε να μπορέσει ο αλγόριθμος

KMeans να κάνει αυτήν την κατηγοριοποίηση. Ακολουθεί η γραφική παράσταση του silhouette score με τα κανονικοποιημένα δεδομένα και έπειτα με τα αρχικά δεδομένα.





Δ) Για το συγκεκριμένο ερώτημα με την χρήση του κώδικα που βρίσκεται στο jupyter notebook λαμβάνουμε ως έξοδο για μία εκτέλεση:

```
Rand Index is: 0.8743677375256322
```

Επαναλαμβάνοντας 5 φορές για τυχαία αρχικοποίηση κέντρων (με την χρήση της παραμέτρου init = ‘random’) η μέση τιμή και το variance του Rand Index είναι:

```
Mean Rand Index: 0.8743677375256322
Variance of Rand Index: 0.0
```

E) Χρησιμοποιώντας τώρα cosine μετρική έχουμε:

```
Rand Index is: 0.8616085668717247
```

Και για 5 επαναλήψεις:

```
Mean Rand Index: 0.8616085668717247
Variance of Rand Index: 0.0
```

Αξίζει να σημειωθεί πως για την ομαδοποίηση με KMeans και cosine μετρική, τα δεδομένα αρχικά κανονικοποιήθηκαν ώστε να έχουν μηδενική μέση τιμή και μοναδιαία variance. Στην συνέχεια, επειδή η βιβλιοθήκη για το KMeans δεν υποστηρίζει cosine μετρική, χρησιμοποιήθηκε η προσέγγιση της κανονικοποίησης των διανυσμάτων σε μοναδιαίο μήκος ακολουθούμενο από KMeans με ευκλείδεια μετρική.

Συμπερασματικά, παρατηρούμε με βάση την μέση τιμή του rand index και το variance πως και η ευκλείδεια και η cosine μετρική δίνουν πολύ κοντινά αποτελέσματα. Το rand index αποτελεί μέτρο ομοιότητας μεταξύ δύο ομαδοποιήσεων λαμβάνοντας υπόψη όλα τα ζεύγη δειγμάτων και μετρώντας ζεύγη που έχουν εκχωρηθεί στο ίδιο ή σε διαφορετικό cluster στις πραγματικές και στις προβλεπόμενες συστάδες. Συνεπώς και οι δύο μετρικές επιτυγχάνουν περίπου την ίδια ομαδοποίηση με την Euclidean να είναι λίγο καλύτερη.

Bonus Ερώτημα:

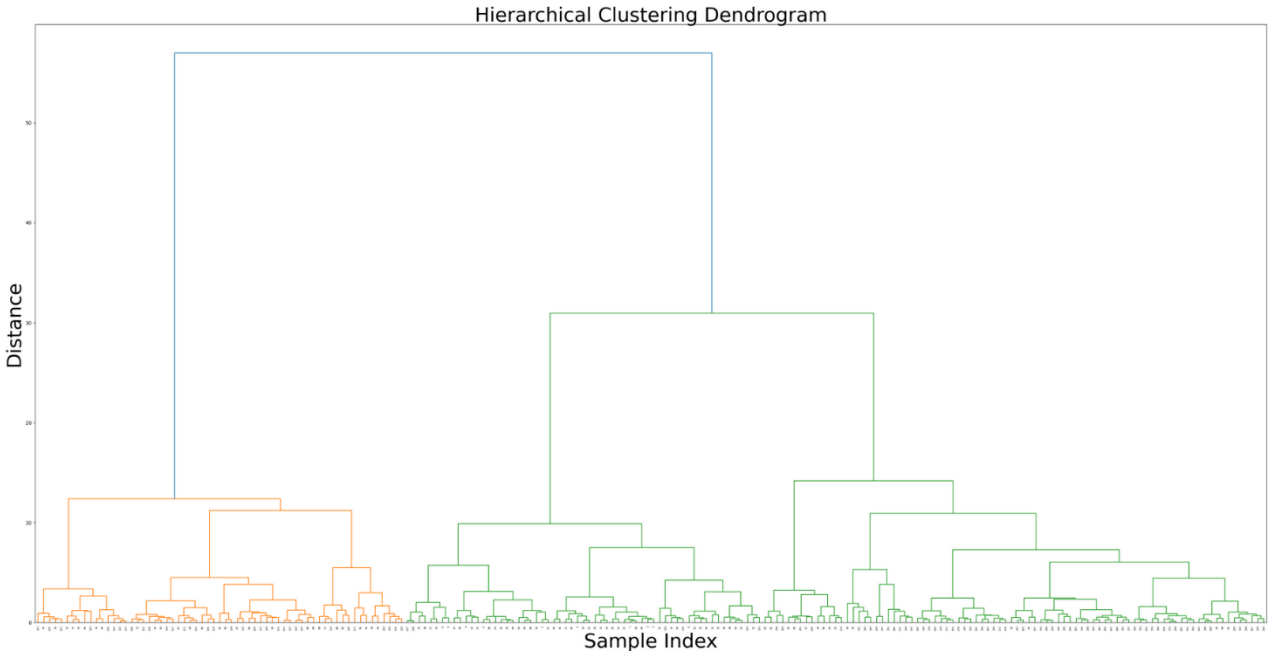
Αρχικά θα πρέπει να γίνει κανονικοποίηση (standarization) των δεδομένων ώστε οι τιμές τους να μην απέχουν πολύ μεταξύ τους (να είναι δηλαδή σε παρόμοια κλίμακα). Στην συνέχεια θα ήταν θεμιτό να επιλεχθεί κατάλληλος αριθμός κ εγγύτερων γειτόνων. Για την επιλογή κέντρων θα επιλέξουμε αλγόριθμο ιεραρχικής ομαδοποίησης ενώ για τις μετρικές των αποστάσεων θα χρησιμοποιηθούν ευκλείδειες ή cosine μετρικές. Τέλος, για την εύρεση των γειτόνων προκειμένου να αποφευχθεί η εύρεση όλων των αποστάσεων θα γίνει χρήση του majority vote μέσα σε κάθε cluster που έχει δημιουργηθεί.

Άσκηση 2:

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης ανήκουν στην κατηγορία των αλγορίθμων clustering και χρησιμοποιούνται για την ομαδοποίηση των δεδομένων σε ιεραρχικές δομές. Για την εκτέλεση του αλγορίθμου αυτού βασική προϋπόθεση αποτελεί ο υπολογισμός του πίνακα απόστασης. Ακολούθως, ανάλογα με την προσέγγιση δημιουργούνται και οι ομάδες. Πιο συγκεκριμένα, στην agglomerative προσέγγιση ή αλλιώς bottom up, κάθε δείγμα θεωρείται ως μια κλάση και στην συνέχεια οι γειτονικές κλάσεις συγχωνεύονται σχηματίζοντας μια νέα κλάση. Αυτό συμβαίνει έως ότου μείνει μόνο μια ομάδα. Για την divisive προσέγγιση, ισχύει ακριβώς το αντίθετο δηλαδή ξεκινά από μία ομάδα και διασπάται έως ότου δημιουργηθούν N. Για την απεικόνιση των ιεραρχικών clusters προτιμώνται τα δενδρογράμματα τα οποία δείχνουν και το μέτρο ομοιότητας μεταξύ των ομάδων.

A) Για την υλοποίηση της άσκησης αυτής επιλέχθηκε η χρήση agglomerative αλγορίθμου ιεραρχικής ομαδοποίησης. Πιο συγκεκριμένα, με την χρήση της βιβλιοθήκης SciPy η συνάρτηση linkage() επιτυγχάνει την παραπάνω ιεραρχική ομαδοποίηση. Προκειμένου όμως να λειτουργήσει σωστά η μέθοδος πρέπει να δοθούν τα κατάλληλα ορίσματα ανάλογα με την δομή των δεδομένων. Επειδή τα δεδομένα μας δεν είναι distance matrix, ως όρισμα για την μετρική θα βάλουμε ‘euclidean’. Ως method επιλέγεται λόγω μικρότερης χρονικής πολυπλοκότητας σύμφωνα με το documentation η ‘ward’, η οποία βασίζει την ομοιότητα των clusters στην αύξηση του τετραγωνικού σφάλματος όταν συγχωνεύονται δύο clusters. Επίσης είναι λιγότερο ευαίσθητη στον θόρυβο και τις ακραίες τιμές ενώ αποτελεί ιεραρχικό ισοδύναμο της μεθόδου KMeans.

B) Το δενδρόγραμμα που προκύπτει από την ομαδοποίηση που αναλύθηκε παραπάνω ακολουθεί.



Στο dataset της συγκεκριμένης άσκησης έχουμε 3 διακριτές κλάσεις. Γενικά στο δενδρόγραμμα, αυτό που δημιουργεί τα νέα clusters είναι η απόσταση μεταξύ των επιμέρους clusters. Όσο πιο μικρή είναι αυτή η απόσταση τόσο μεγαλύτερη είναι η ομοιότητα και έτσι σχηματίζεται ένα νέο μεγαλύτερο cluster. Στο παραπάνω δενδρόγραμμα, γίνεται εύκολα αντιληπτό πως τα δείγματα χωρίζονται σε 3 κλάσεις όταν σε κάποια απόσταση από την βάση φέρουμε μια γραμμή παράλληλη στον οριζόντιο άξονα (που να τέμνει την μπλε και τις 2 πράσινες γραμμές).

Γ) Χρησιμοποιώντας τον πίνακα linkage_matrix από το προηγούμενο ερώτημα και καλώντας την συνάρτηση fcluster βρίσκουμε τις προβλέψεις για τα labels των δεδομένων. Στην συνέχεια, με την χρήση της rand_score συνάρτησης βρίσκουμε το rand index ανάμεσα στα πραγματικά labels και σε αυτά που έγινε πρόβλεψη. Με τον αντίστοιχο κώδικα βρίσκουμε ότι:

```
Rand Index is: 0.8722715880610618
```

Συγκρίνοντας τα αποτελέσματα της ιεραρχικής ομαδοποίησης με αυτά που έδωσε ο αλγόριθμος KMeans στην προηγούμενη άσκηση παρατηρούμε ότι ο αλγόριθμος KMeans αποδίδει καλύτερα με ελάχιστη διαφορά (επιτυγχάνει δηλαδή ομαδοποίηση πιο κοντά στην πραγματική) συγκρίνοντας τις μετρικές Rand Index.

Δ) Οι τεχνικές ιεραρχικής ομαδοποίησης έχουν το πολύ βασικό πλεονέκτημα πως δεν χρειάζεται να προσδιοριστεί ο αριθμός των clusters που θα δημιουργηθούν καθώς προκύπτει από το δενδρόγραμμα. Σε αντίθεση με τον KMeans, οι ιεραρχικές τεχνικές δεν προϋποθέτουν γραμμικά διαχωρίσιμα δεδομένα για να πετύχουν υψηλή απόδοση καθώς μπορούν να αναγνωρίσουν με επιτυχία και κλάσεις με μη γραμμικά διαχωρίσιμα δεδομένα. Αξίζει επίσης να σημειωθεί πως λόγω του ότι οι κλάσεις δημιουργούνται βήμα-βήμα υπάρχει μεγαλύτερη ευαισθησία σε τοπικά πρότυπα που ενδέχεται να υπάρχουν στα δεδομένα. Τέλος, οι ιεραρχικές δομές υποστηρίζουν και την χρήση διαφορετικών μετρικών απόστασης σε σχέση με την τεχνική KMeans που χρησιμοποιεί αποκλειστικά την ευκλείδεια απόσταση.

Άσκηση 3:

Η μείωση των διαστάσεων είναι μια πολύ συνήθης τακτική που χρησιμοποιείται στην ανάλυση δεδομένων ώστε να μειωθεί ο αριθμός των χαρακτηριστικών σε ένα dataset αλλά να διατηρηθεί όσο το δυνατόν περισσότερη πληροφορία. Στο πλαίσιο της άσκησης αυτής θα εστιάσουμε στην μέθοδο PCA (Principal Component Analysis) και στην LDA (Linear Discriminant Analysis). Ξεκινώντας από την PCA, είναι μια μέθοδος που έχει ως στόχο την εύρεση της κατεύθυνσης της μέγιστης τυπικής απόκλισης στο dataset. Πιο συγκεκριμένα, κεντράρει τα δεδομένα γύρω από το μέσο και έπειτα βρίσκει τα ιδιοδιανύσματα και τις ιδιοτιμές του πίνακα συνδιασποράς. Τα ιδιοδιανύσματα αναπαριστούν την κατεύθυνση της μέγιστης τυπικής απόκλισης ενώ οι ιδιοτιμές το μέγεθος της τυπικής απόκλισης που περιγράφει το κάθε ιδιοδιάνυσμα. Ο αριθμός

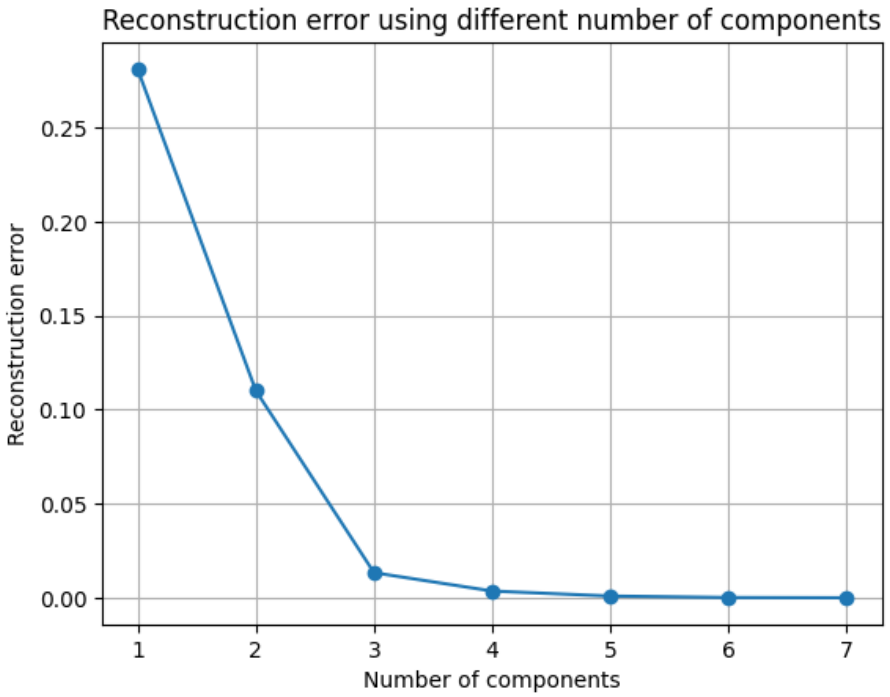
των principal components που κρατάμε εξαρτάται από την τιμή της τυπικής απόκλισης που θέλουμε να διατηρήσουμε.

Από την άλλη πλευρά, η μέθοδος LDA αποσκοπεί στην την εύρεση ενός γραμμικού συνδυασμού χαρακτηριστικών ώστε να διαχωρίσει καλύτερα τις κλάσεις που βρίσκονται στο σύνολο δεδομένων. Αρχικά, υπολογίζει την μέση τιμή και τον πίνακα συνδιασποράς για κάθε κλάση. Έπειτα, υπολογίζει τον scatter matrix μεταξύ των κλάσεων και τον scatter matrix μέσα στην κλάση. Ο σκοπός είναι όπως προαναφέρθηκε να βρεθεί μια προβολή που μεγιστοποιεί την αναλογία μεταξύ του scatter matrix μεταξύ των κλάσεων με τον scatter matrix μέσα στην κλάση.

A) Για τον υπολογισμό των ελάχιστων κύριων συνιστωσών που πρέπει να διατηρηθούν ώστε να εξηγείται τουλάχιστον το 90% και το 99% της τυπικής απόκλισης του αρχικού dataset αρχικά κανονικοποιούμε τα δεδομένα. Στην συνέχεια, κάνουμε fit και μετασχηματίζουμε τα κανονικοποιημένα δεδομένα στον νέο χώρο και υπολογίζουμε το explained_variance_ratio. Έπειτα, με την συνάρτηση cumsum βρίσκουμε το cumulative explained variance όπου συγκρίνοντας για τα επιθυμητά variance που ζητούνται προσδιορίζουμε πόσα είναι τα ζητούμενα components. Με την εκτέλεση του κατάλληλου κώδικα προκύπτει πως:

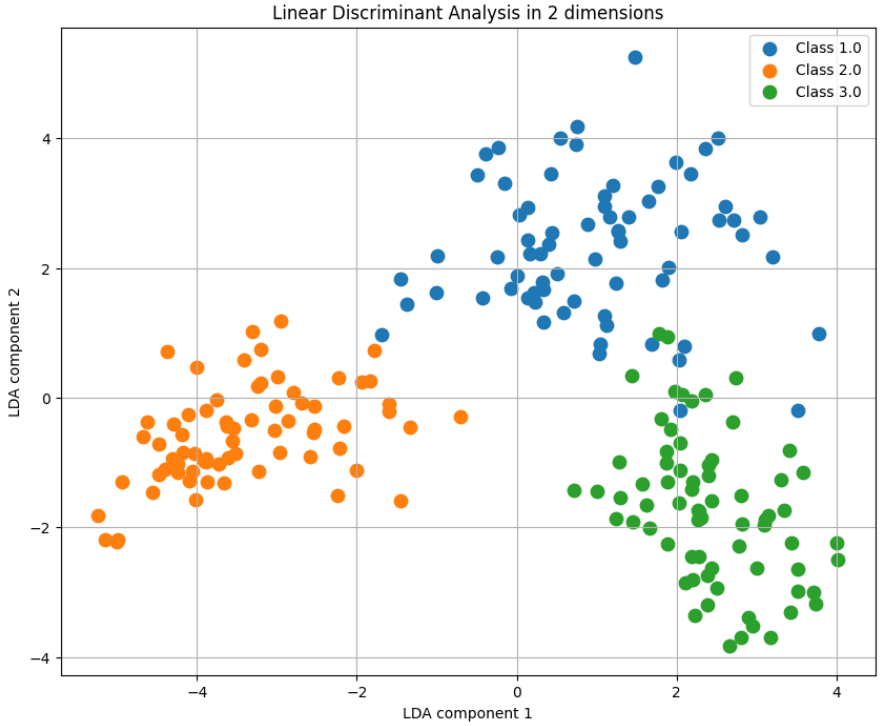
```
Number of components to retain 90.0% of variance is: 3
Number of components to retain 99.0% of variance is: 4
```

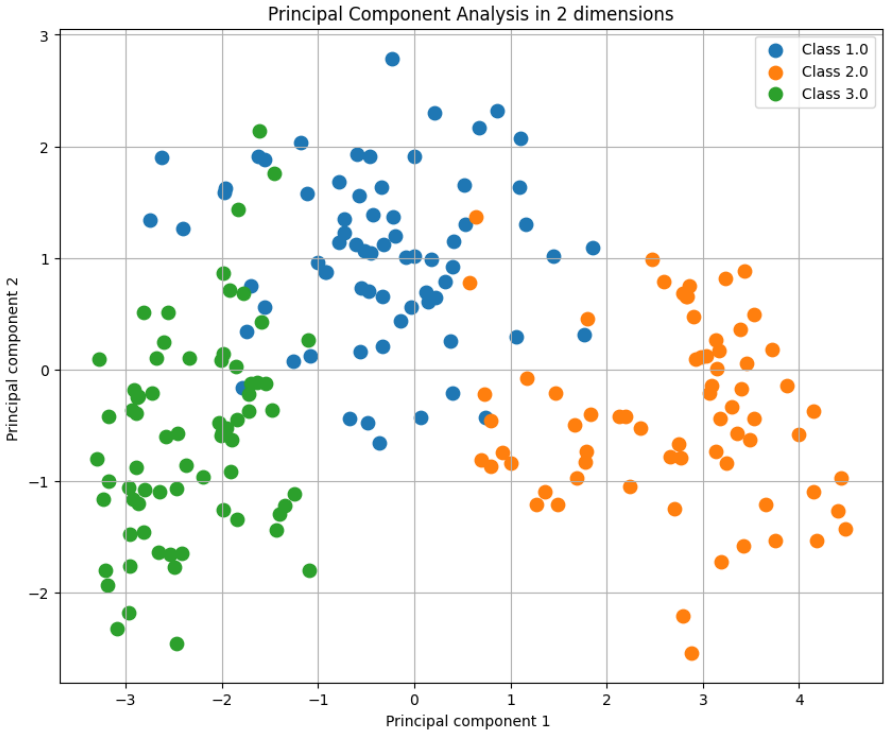
B) Το ζητούμενο διάγραμμα ακολουθεί:



Για τον υπολογισμό του σφάλματος χρησιμοποιήθηκε το MSE (Μέσο τετραγωνικό σφάλμα) ενώ τα δεδομένα ανακατασκευής προκύπτουν από την συνάρτηση pca.inverse_transform της βιβλιοθήκης sklearn.

Γ) Για την απεικόνιση του dataset σε 2 διαστάσεις με την LDA, αξιοποιώντας τα κανονικοποιημένα δεδομένα εφαρμόζουμε την συνάρτηση pca.fit_transform() ώστε να γίνει fit και να μετασχηματιστούν τα δεδομένα. Έπειτα, κάνοντας plot τα δεδομένα που προκύπτουν στον νέο χώρο έχουμε:



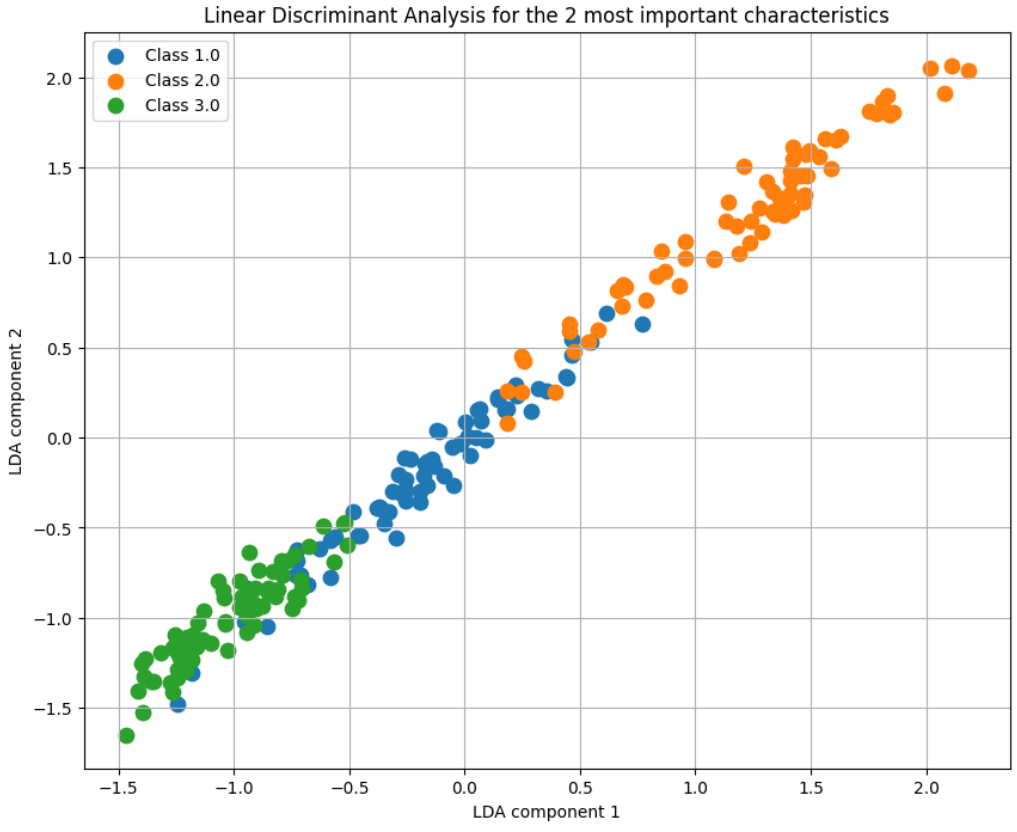


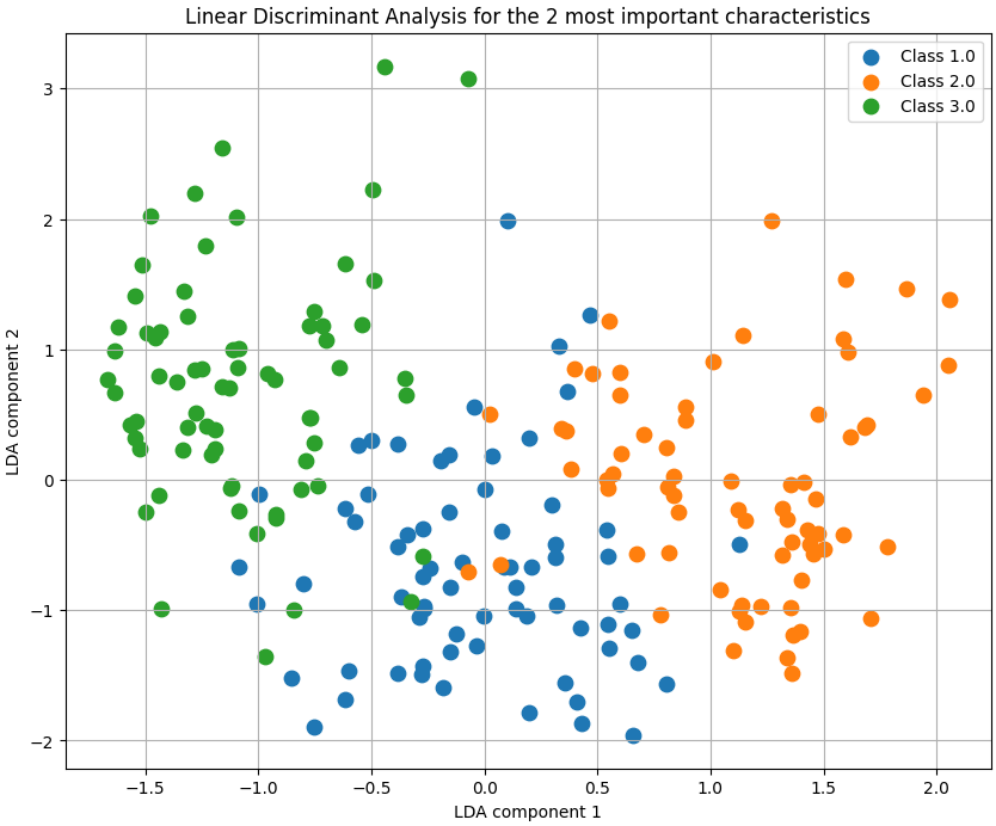
Όπως αναφέρθηκε και προηγουμένως, η μέθοδος PCA προσπαθεί να διατηρήσει όσο το δυνατόν περισσότερη διακύμανση. Αυτό φαίνεται χαρακτηριστικά και στο διάγραμμα από το πως είναι διατεταγμένα τα σημεία. Αντιθέτως η LDA προσπαθεί να διατηρεί την διακύμανση μεταξύ των κλάσεων και παρατηρούμε πως υπάρχει μια ικανοποιητική διαχωρισιμότητα μεταξύ των δεδομένων. Αξίζει να σημειωθεί πως η LDA αξιοποιεί τα labels των δεδομένων ενώ η PCA δεν τα λαμβάνει υπόψη.

Δ) Προκειμένου να βρεθούν τα δύο χαρακτηριστικά που συνεισφέρουν περισσότερο στον διαχωρισμό των κλάσεων και τα δύο χαρακτηριστικά που συνεισφέρουν λιγότερο αρκεί να βρούμε αρχικά τον πίνακα προβολής με την εντολή lda.coef_. Έπειτα, εφόσον υπάρχουν αρνητικές τιμές βρίσκουμε την απόλυτη τιμή κάθε στοιχείου και υπολογίζουμε τον μέσο όρο των στηλών. Αυτό συμβαίνει διότι έχουμε 3 κλάσεις και 7 χαρακτηριστικά. Έτσι τα 2 χαρακτηριστικά που έχουν μεγαλύτερη συνεισφορά είναι τα 2 στοιχεία με τις μεγαλύτερες τιμές. Αντίστοιχα, αυτά με την μικρότερη συνεισφορά είναι αυτά με τις μικρότερες τιμές. Από την εκτέλεση του κώδικα προκύπτει:

```
The most important characteristics located at 0 and 1
The less important characteristics located at 4 and 5
```

Χρησιμοποιώντας τα χαρακτηριστικά που βρέθηκαν παραπάνω τα plots που προκύπτουν ακολουθούν. Αυτό που παρατηρούμε είναι πως με τα δύο καλύτερα χαρακτηριστικά τα δεδομένα βρίσκονται πάνω στην ίδια γραμμή και είναι πιο εύκολα διαχωρίσιμα. Εν αντιθέσει, με τα δύο χαρακτηριστικά που συνεισφέρουν λιγότερο βλέπουμε πως τα δεδομένα είναι «απλωμένα» στον χώρο και πιο δύσκολα διαχωρίσιμα.

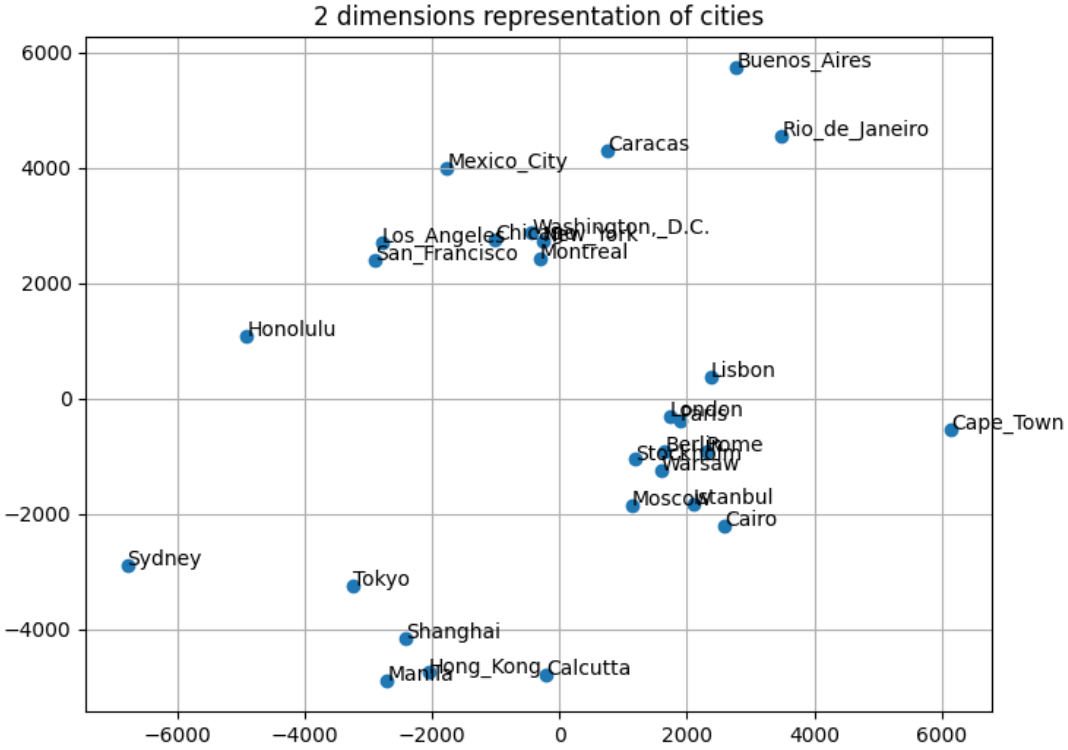


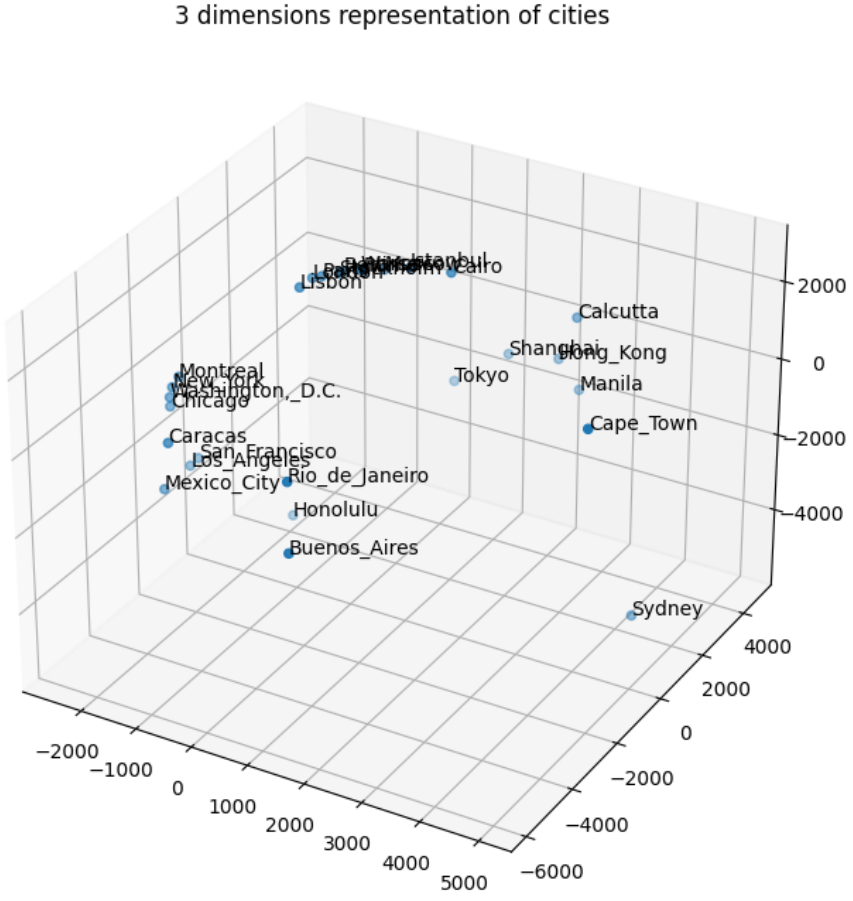


Άσκηση 4:

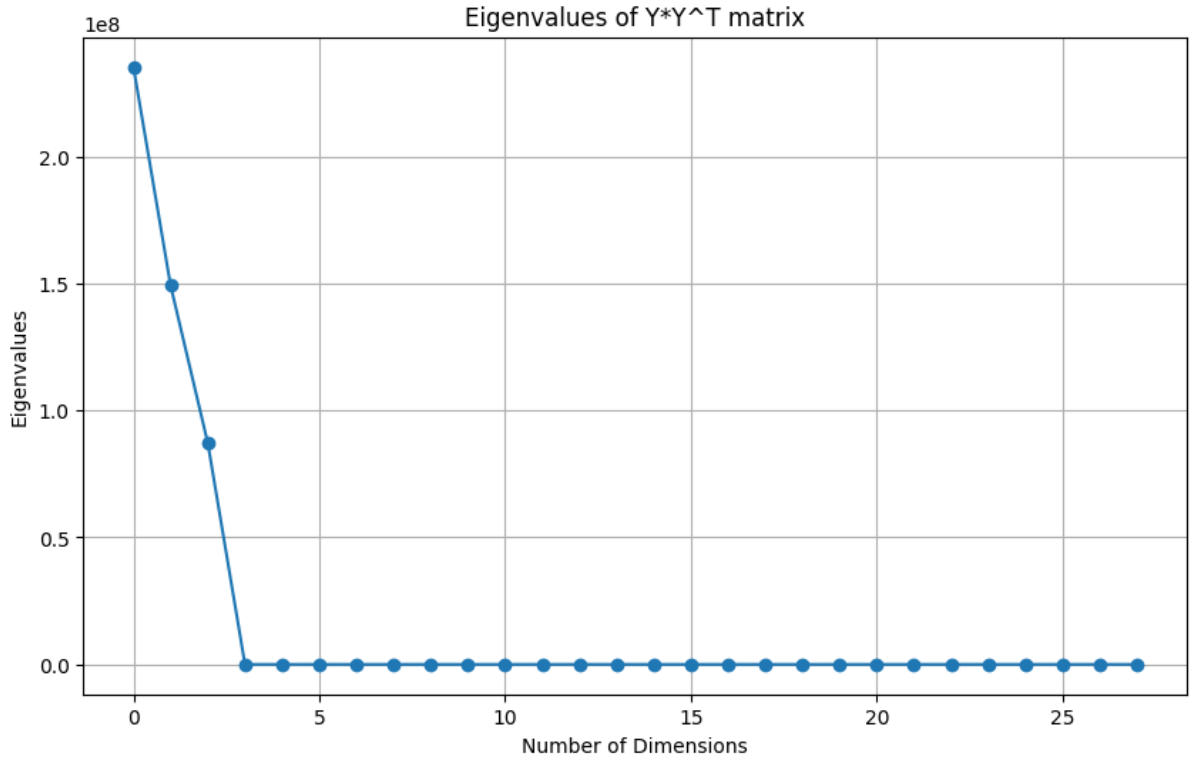
Το classical Multidimensional Scaling (MDS) είναι μια τεχνική μείωσης των διαστάσεων των δεδομένων ενώ παράλληλα διατηρεί τις αποστάσεις μεταξύ των δειγμάτων. Αρχικά, με βάση τον πίνακα που δίνεται ως είσοδος, υπολογίζονται οι ευκλείδιες αποστάσεις ώστε να δημιουργηθεί ο double centered matrix και να βρεθούν τα ιδιοδιανύσματα και οι ιδιοτιμές του. Τα ιδιοδιανύσματα αυτά εμπεριέχουν τις νέες διαστάσεις των δειγμάτων. Επιλέγονται οι πρώτες N ιδιοτιμές και ιδιοδιανύσματα ώστε να γίνει απεικόνιση σε νέο χώρο χαμηλότερων διαστάσεων. Σε άλλη περίπτωση, η επιλογή πολλών ιδιοτιμών και ιδιοδιανυσμάτων μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting). Για την υλοποίηση της άσκησης αυτής, τις απαραίτητες συναρτήσεις τις παίρνουμε από την βιβλιοθήκη sklearn.

Α) Παρατηρούμε στα διαγράμματα που ακολουθούν ότι στις 3 διαστάσεις υπάρχει μεγαλύτερη λεπτομέρεια στην απεικόνιση των πόλεων κάτι που δεν είναι άμεσα αντιληπτό στις 2 διαστάσεις. Συνεπώς, οδηγούμαστε στο συμπέρασμα πως στις 3 διαστάσεις εμπεριέχετε περισσότερη πληροφορία σχετικά με τις αποστάσεις των πόλεων.


2 dimensions representation of cities
Buenos_Aires
Rio_de_Janeiro
Caracas
Mexico_City
Los_Angeles
San_Francisco
Montreal
Honolulu
Lisbon
Cape_Town
Warsaw
Cairo
Sydney
Tokyo
Shanghai
Hong_Kong
Calcutta
Manila
6000
4000
2000
0
−2000
−4000
−6000
−4000
−2000
0
2000
4000
6000


3 dimensions representation of cities
Lisbon
Calcutta
Shanghai
Hong_Kong
Tokyo
Manila
Montreal
Chicago
Caracas
Cape_Town
San_Francisco
Los_Angeles
Mexico_City
Rio_de_Janeiro
Honolulu
Buenos_Aires
Sydney
2000
0
−2000
−4000
4000
2000
0
−2000
−4000
−6000
−2000
−1000
0
1000
2000
3000
4000
5000

B) Παρατηρούμε πως στο διάγραμμα με τις ιδιοτιμές συναρτήσει του αριθμού των διαστάσεων που ακολουθεί ότι μετά την χρήση τριών διαστάσεων οι ιδιοτιμές είναι πολύ μικρές (αρκετές τάξεις μεγέθους κάτω από τις πρώτες). Συνεπώς το να καταφύγουμε σε διαστάσεις μεγαλύτερες των τριών δεν έχει αξία. Άρα ο βέλτιστος αριθμός διαστάσεων για τα δεδομένα του συγκεκριμένου προβλήματος είναι οι 3 διαστάσεις.



Bonus Ερώτημα:

Οι μη μηδενικές ιδιοτιμές μετά την $3^{η}$ διάσταση πιθανώς να οφείλονται στην πολυπλοκότητα των χωρικών σχέσεων μεταξύ των πόλεων που δεν μπορούν να αποτυπωθούν με τρεις διαστάσεις. Πιο συγκεκριμένα, οι εναέριες αποστάσεις δεν είναι πάντα απόλυτα γραμμικές λόγω της καμπυλότητας της γης. Τα δεδομένα μπορεί να εξαρτώνται από γεωγραφικά χαρακτηριστικά της επιφάνειας της γης και έτσι να χρειάζονται περισσότερες διαστάσεις για να αποτυπωθούν. Συγκρίνοντας τα δύο διαγράμματα ιδιοτιμών συναρτήσει του αριθμού των διαστάσεων παρατηρούμε πως για τις πόλεις των Ηνωμένων Πολιτειών χρειαζόμαστε μόνο 2 διαστάσεις καθώς μετά από εκεί οι ιδιοτιμές είναι πολύ κοντά στο 0. Αυτό οφείλεται στο γεγονός πως οι πόλεις ανά τον κόσμο έχουν μεταξύ τους μεγαλύτερες αποστάσεις και άρα τα τόξα που βαίνουν στην ευθεία της απόστασης είναι μεγαλύτερα. Έτσι, δεδομένης αυτής της διαφοράς, οι πόλεις που έχουν κοντινές αποστάσεις (στο ερώτημα αυτό, οι πόλεις των ΗΠΑ) με την μέθοδο MDS χρειάζονται λιγότερες διαστάσεις λόγω της φύσης και της πολυπλοκότητας των δεδομένων.

Ακολουθεί η δισδιάστατη αναπαράσταση των πόλεων και το διάγραμμα των ιδιοτιμών.