ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Στατιστική Μάθηση-Υπολογιστική Νοημοσύνη

Spectral Embedding & Clustering

Σουράνης Παναγιώτης

ΑΕΜ:17

Το παρακάτω κείμενο αποτελεί μια σύντομη αναφορά της εφαρμογής των αλγορίθμων μείωσης διάστασης των δεδομένων μέσω φασματικής ανάλυσης όπως (Isomap,LLE,MDS,tSNE) και ακολούθως την ομαδοποιήση των δεδομένων μέσω αλγορίθμων ομαδοποίησης – ταξινόμησης όπως για παράδειγμα (KMeans , DBScan , etc.)

Το σύνολο δεδομένων στο οποίο εργαστήκαμε ήταν εικόνες από γεωμετρικά σχήματα.Αναλυτικότερα στα δεδομένα μας είχαμε 4 κλάσεις (Τρίγωνο ,Κύκλος,Αστέρι και Τετράγωνο) τα οποία όμως σε αντίθεση με την handwritten digits Mnist είχαν την ιδιαιτερότητα να εμφανίζουν τα σχήματα μας σε περιστροφή.

Στόχος μας είναι να μπορέσουμε να ομαδοποιήσουμε όσο καλύτερα τα δεδομένα μας και να επιτύχουμε μείωση διάσταση των δεδομένων διατηρώντας ταυτόχρονα την γεωμετρία τους.

Το dataset πάρθηκε από την ιστοσελίδα:

https://www.kaggle.com/smeschke/four-shapes/home

Σε μορφή εικόνων οι οποίες στην συνέχεια μετατράπηκαν σε μορφή ανάλογη της Mnist.

Η διάσταση τους πρέπει να αναφέρουμε ότι είναι 64x64.

Ας δούμε λοιπόν πώς δείχνει το σύνολο δεδομένων μας.

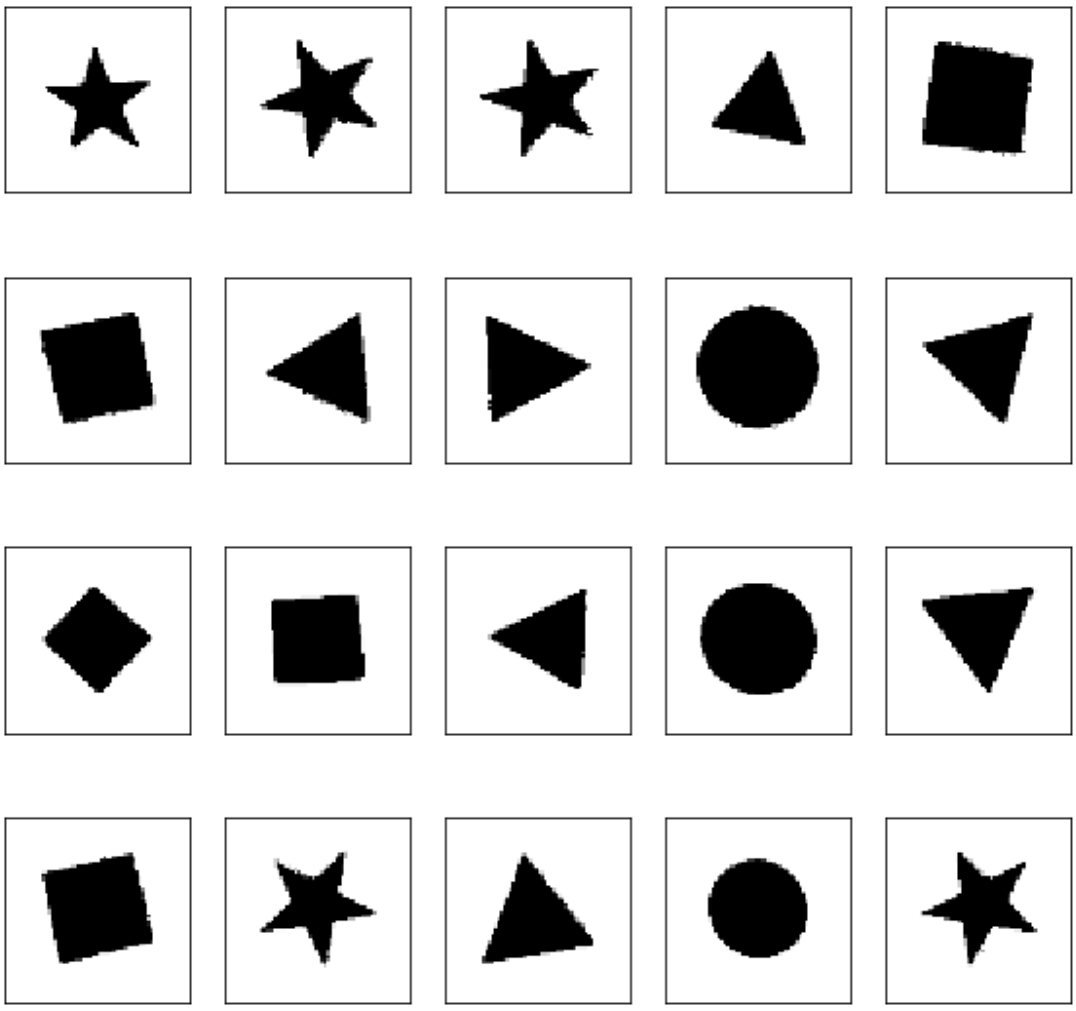
Ανάλυση Dataset:

Τα πρώτα 5 μας δεδομένα έχουν την παρακάτω μορφή.

	0	1	2	3	4	5	6	7	8	9	...	4086	4087	4088	4089	4090	4091	4092	4093	4094	4095
0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	...	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0
1	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	...	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0
2	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	...	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0
3	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	...	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0
4	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	...	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0	255.0

Όμως ας δόυμε και μια εικόνα τους για να έχουμε μια καλύτερη άποψη ιδέα

Σχήμα 1.1



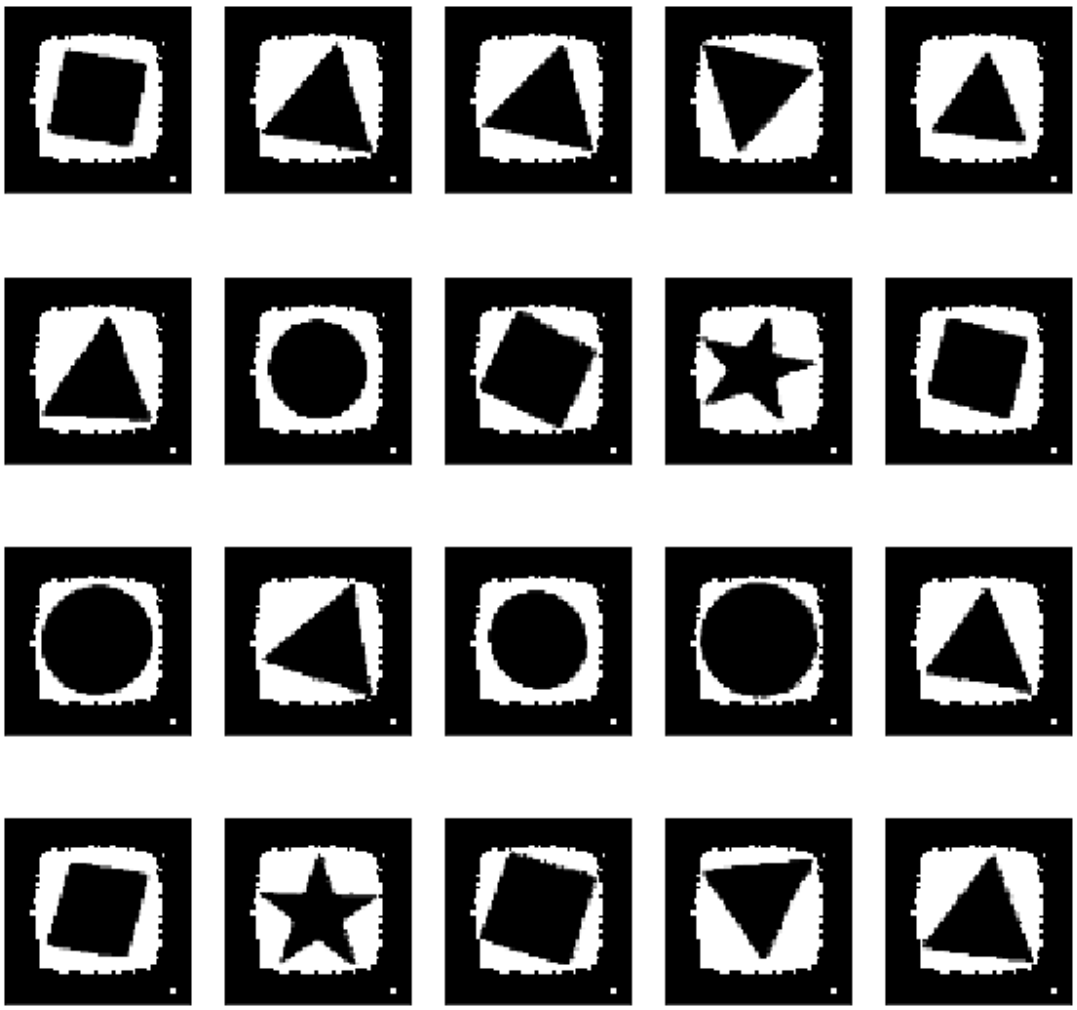
Πρωτόυ προχωρήσουμε παρακάτω να αναφέρουμε ότι οι ετικέτες των αντικειμένων μετατράπηκαν από την αρχική μορφή που είχαν [**'circle'**, **'square'**, **'star'**, **'triangle'**] στις κλάσεις [**0** , **1** , **2** , **3**] αντίστοιχα.

Ακόμη επειδη το σύνολο των δεδομένων μας ήτανε μεγάλο για τους συγκεκριμένους αλγορίθμους τους οποίους καλούμαστε να εφαρμόσουμε το περιορίσαμε σε ένα μικρότερο σύνολο της τάξης των 3000 δειγμάτων (Αρχικό πλήθος δειγμάτων ήταν 15000).Αυτο οφείλεται κυρίως στο ότι οι αλγόριθμοι μας έχουν σαν βάση τους τον υπολογισμό των αποστάσεων των δειγμάτων μεταξύ τους σε όλες τις διαστάσεις προκειμένου να μπορέσουν να δημιουργήσουν τον πίνακα ομοιοτήτων-ανομοιοτήτων (Dissimilarity η αλλιώς Affinity Matrix) οπότε αν είχαμε για παράδειγμα αρχικα 15000 δείγματα θα έπρεπε οι αλγόριθμοι μας να υπολογίσουν 15000^2 αποστάσεις μόνο για να βρούν τον πίνακα ομοιοτήτων κατι που βέβαια απαιτεί αρκετή υπολογιστική ισχύ και χρόνο. Βέβαια υπάρχουν τεχνικές οι οποίες μπορούν να μείωσουν την περιπλοκότητα του αλγορίθμου (complexity) για τις οποίες όμως θα μιλήσουμε στην συνέχεια.

Προκειμένου επίσης να είναι ευκολότεροι οι υπολογισμόι για τους αλγορίθμους μας θα πραγματοποιήσουμε MinMaxScaling έτσι ώστε να φέρουμε το εύρος τιμών τους στο [0,1]

Η εικόνα που έχουν μετά το scaling είναι η παρακάτω.

Σχήμα 1.2



Ας περάσουμε λοιπόν στο κυρίως θέμα της αναφοράς μας το οποίο είναι οι αλγόριθμοι μείωσης διάστασης των δεδομένων μέσω φασματικής ανάλυσης.

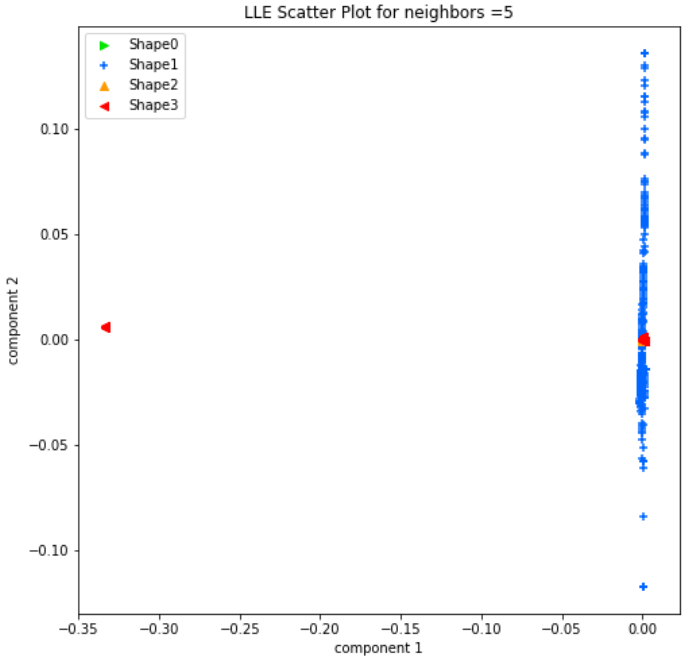
Οι αλγόριθμοι που θα χρησιμοποιηθούν είναι οι:

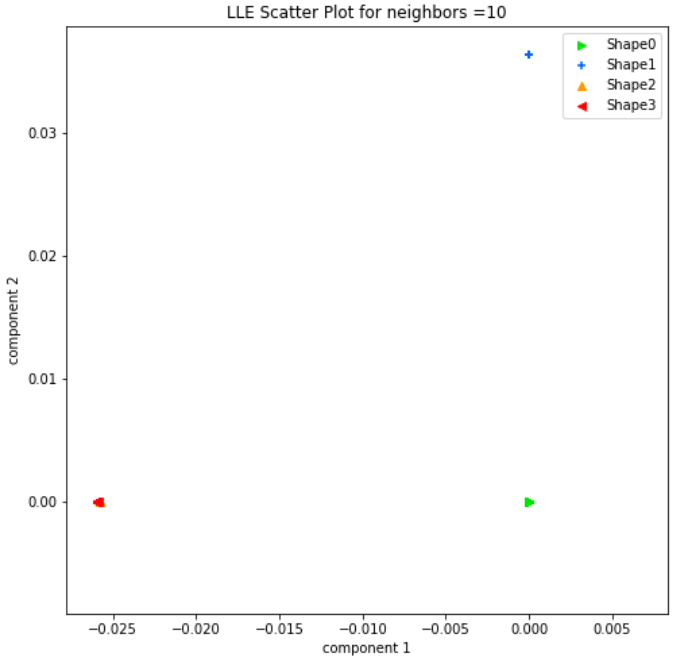
(Isomap , LLE , MDS ,tSNE) οι οποίοι ανοίκουν όλοι στην οικογένεια των Spectral Embedding αλγορίθμων και επιλέχθηκαν αυτοί προκειμένου να παρουσιάσουμε περισσότερο τις διαφορές που υπάρχουν.

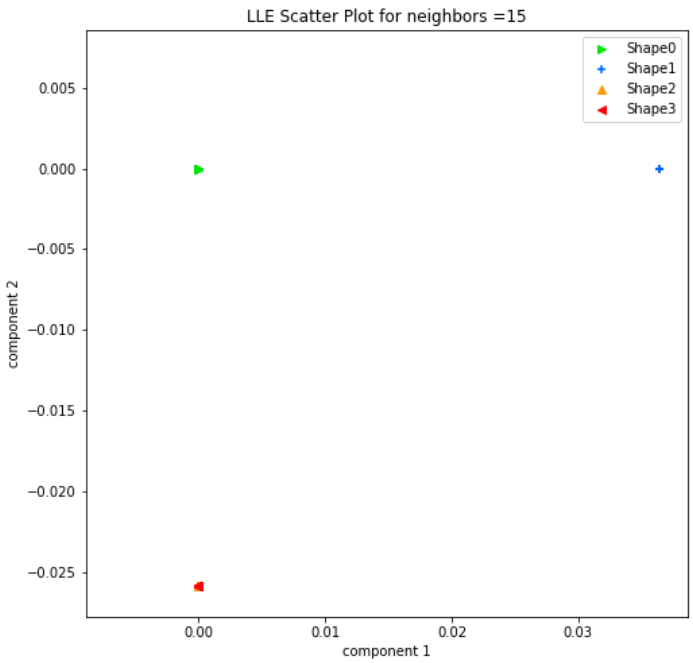
Ας ξεκινήσουμε πρώτα με τον αλγόριθμο LLE

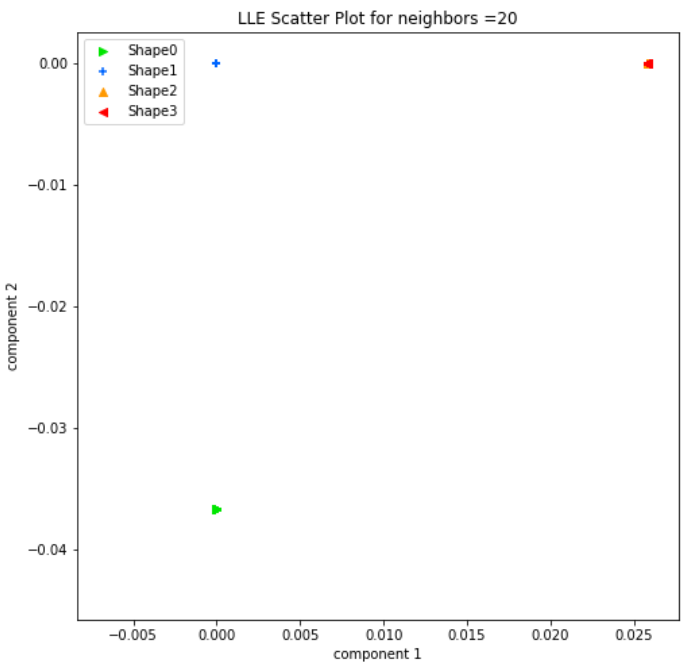
LLE

Τα αποτελέσματα που προέκυψαν μετα την εφαρμογή του αλγορίθμου LLE ήταν τα παρακάτω:









Ακόμη οι χρόνοι που χρειάστηκαν σε **seconds** ήταν:

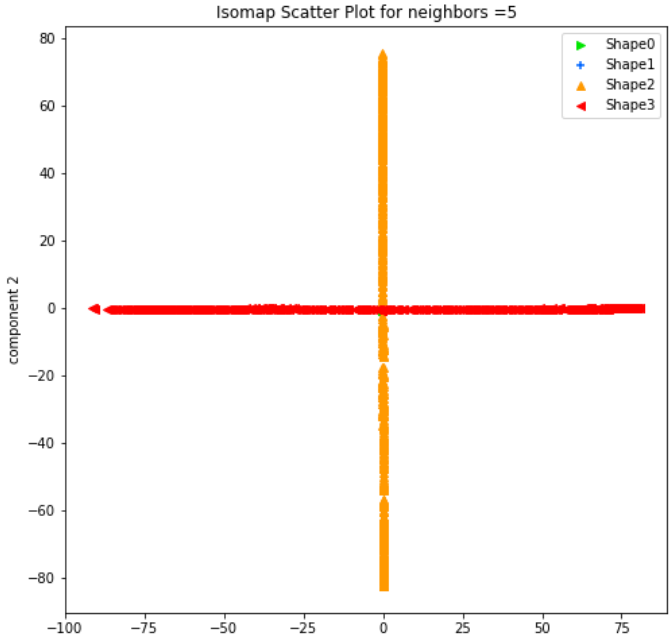
	1	2	3	4
Time needed to complete	54.720471	54.651132	54.563388	55.032771

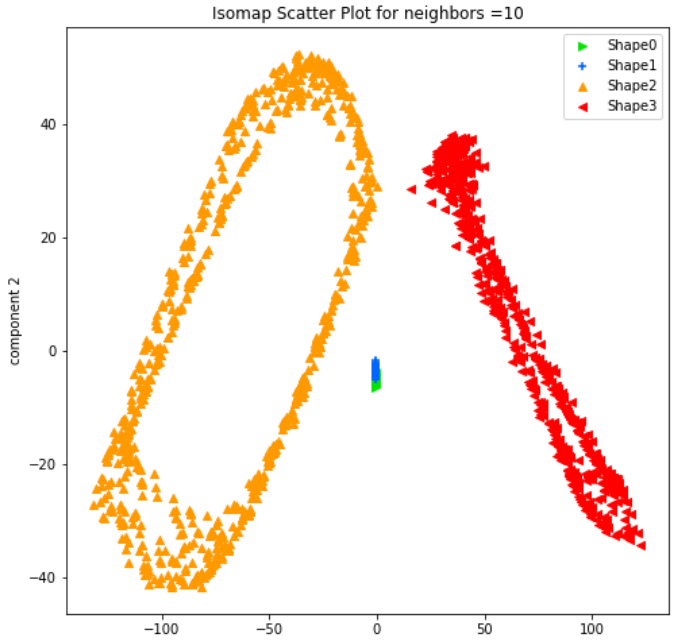
Παρατηρούμε ότι για πλήθος γειτόνων από 10 και πάνω τα στοιχεία των κλάσεων έχουν πέσει ακριβώς το ένα πάνω στο άλλο.Αυτο οφείλεται στο Collapse problem που σχολιάζεται περαιτέρω στους ‘αλγόριθμους’.

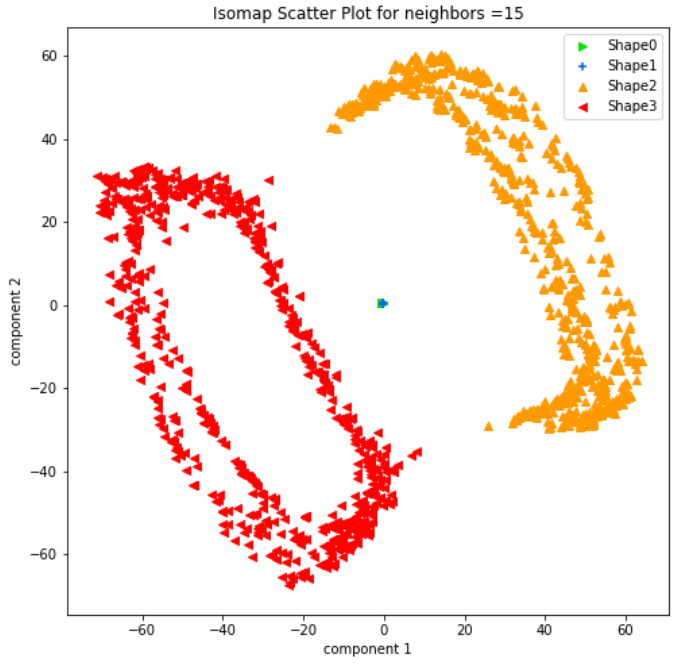
Ακολουθεί ο αλγόριθμος Isomap.

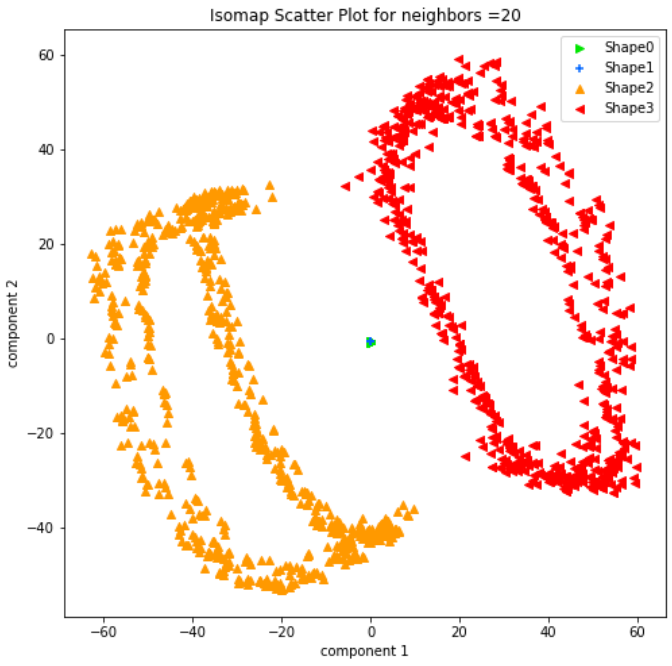
Isomap

Τα αποτελέσματα που προέκυψαν ήταν τα παρακάτω:







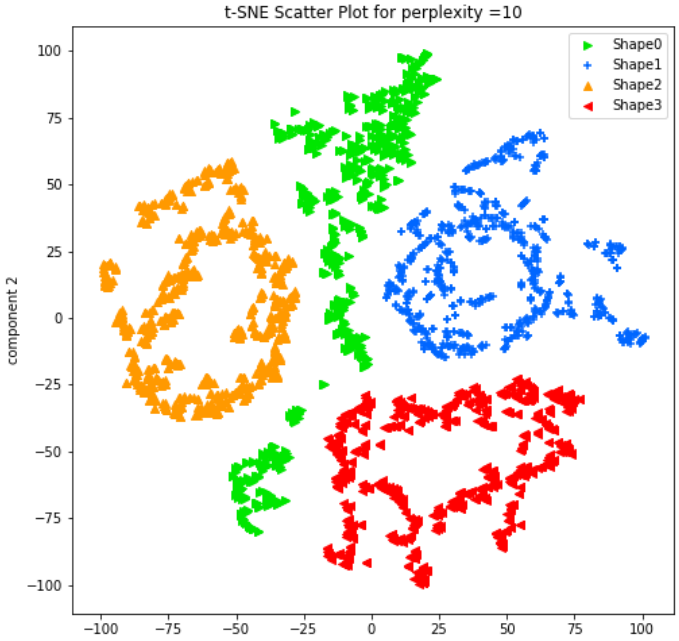


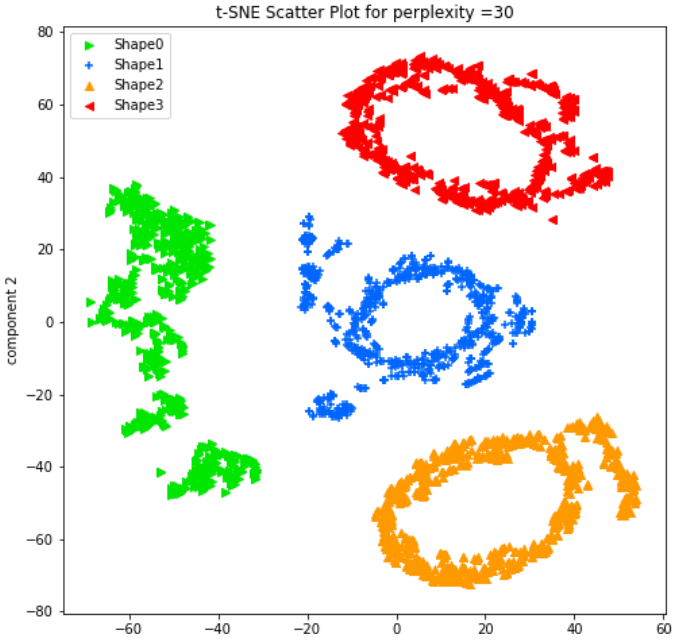
Οι χρόνοι που χρειάστηκαν αυτή την φορά ήταν:

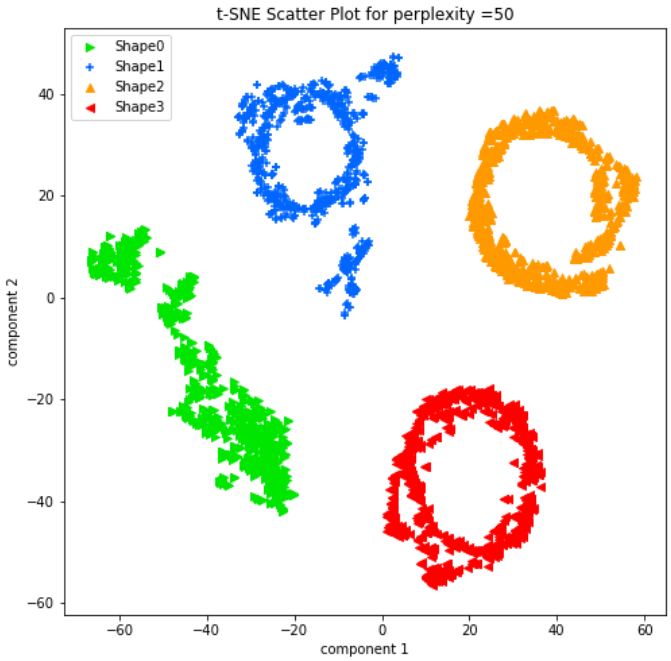
	1	2	3	4
Time needed to complete	53.453003	54.431137	53.587795	53.729394

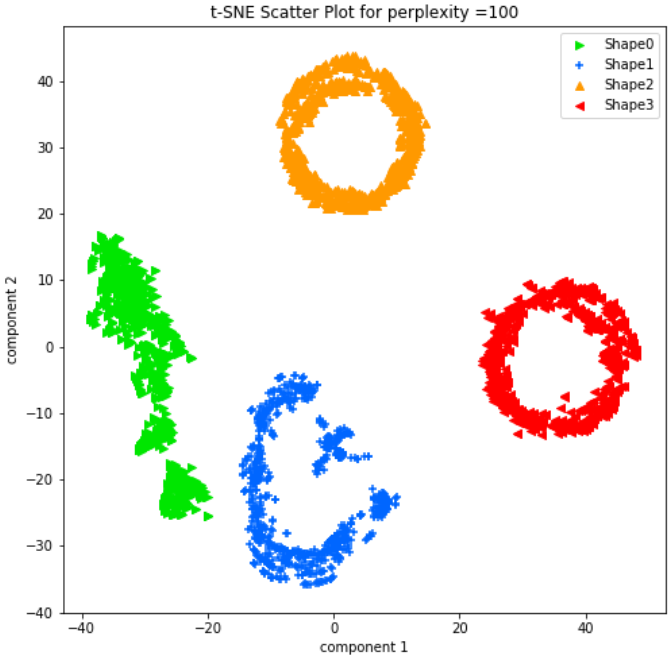
tSNE (t-Stochastic Neighbor Embedding)

Τα αποτελέσματα που προέκυψαν ήταν τα εξής:









Οι χρόνοι που χρειάστηκαν αυτή την φορά ήταν:

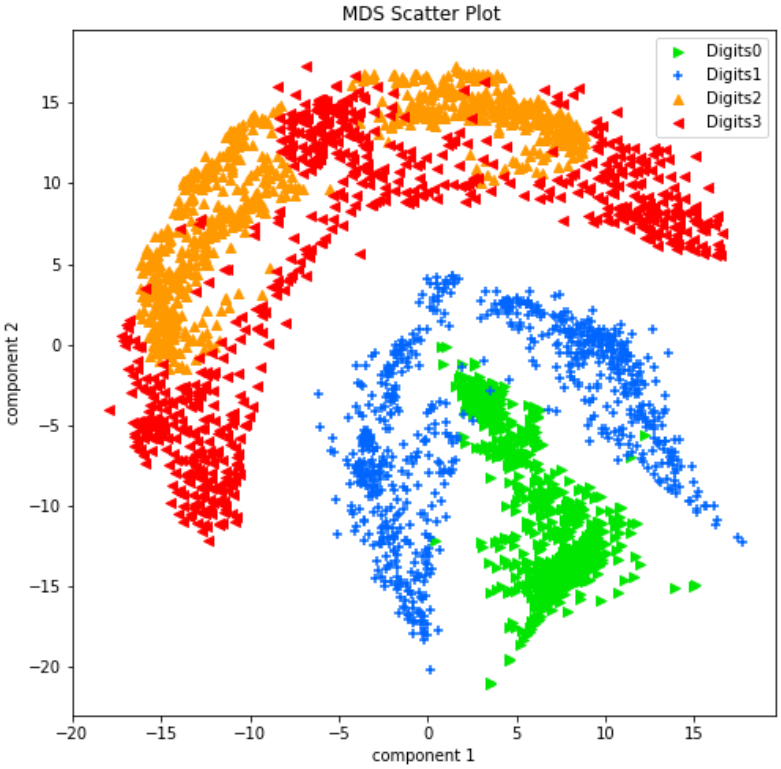
	1	2	3	4
Time needed to complete	92.759602	112.449648	128.420372	174.779208

Note!

Να αναφέρουμε ότι επειδή ο αλγόριθμος tSNE είναι ευριστικός δεν υπάρχει καμία εγγύηση ότι αν ξανατρέξουμε τον αλγόριθμο με τις ίδιες παραμέτρους θα πάρουμε τα ίδια αποτελέσματα.Επίσης παρατηρούμε ότι όσο αυξάνεται η παράμετρος perplexity τα clusters που σχηματίζονται τείνουν να μαζεύονται πιο κοντά το οποίο εξηγείται στο κείμενο «Αλγόριθμοι».

MDS

Τα αποτελέσματα που πήραμε από την εφαρμογή του MDS ήταν τα παρακάτω:



Ο χρόνος που χρειάστηκε αυτή την φορά ήταν:

Time needed to complete : 375.1160955429077 seconds

Κατι που περιμέναμε καθώς ο MDS είναι ο πιο αργός από όλους τους άλλους αλγορίθμους.

Ας περάσουμε τωρα στο δεύτερο κομμάτι της εργασίας μας το οποίο είναι η εφαρμογή clustering αλγορίθμων.Στο πείραμα μας θα χρησιμοποιήσουμε τον αλγόριθμο KMeans ο οποίος είναι απλοικότερος στην υλοποίηση απ'ότι αλλοι αλγόριθμοι (πχ DBScan)

Στην συνέχεια θα αξιολογήσουμε το μοντέλο μας μέσω μετρικών οι οποίοι περιγράφονται στο αρχείο «Αξιολόγηση Μοντέλων».

Ας αρχίσουμε από τον αλγόριθμο LLE

<u>LLE</u> (Clusters number = 4)

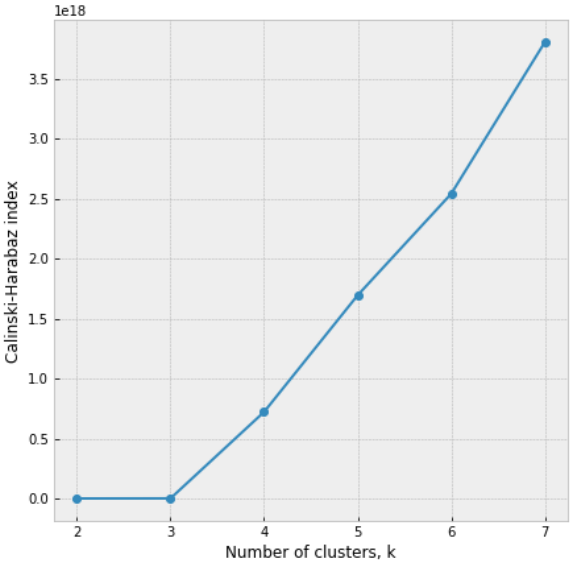
```
--Homogenity of clusters is 1.000 --

--Silhouette measure of clusters is 0.994--

--Completness of clusters is 1.000 --

--V measure score of clusters is 1.000--

--Adjusted Mutual Information score of clusters is 1.000--

--Calinski Harabaz Index score is 62499751916764252l6.000--

--Purity Score is 1.000--
```

The higher the better

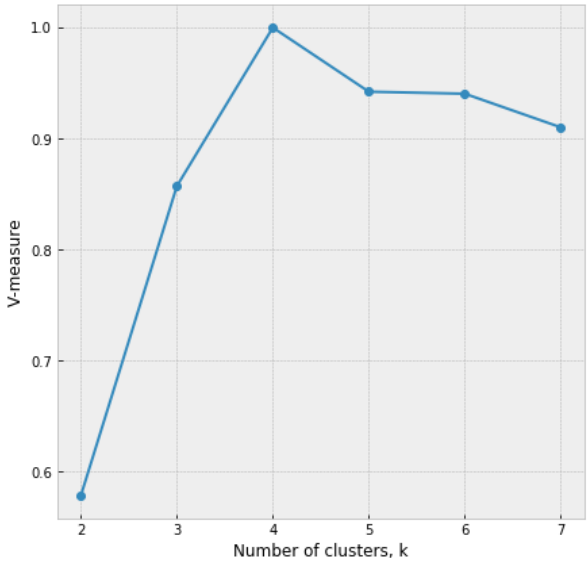
Βλέπουμε ότι ο αλγόριθμος ομαδοποιήσης για 4 κλάσεις έχει πετύχει τέλεια αποτελέσματα.

Ας δούμε όμως πως μεταβάλονται οι μετρικές ανάλογα με τις κλάσεις:

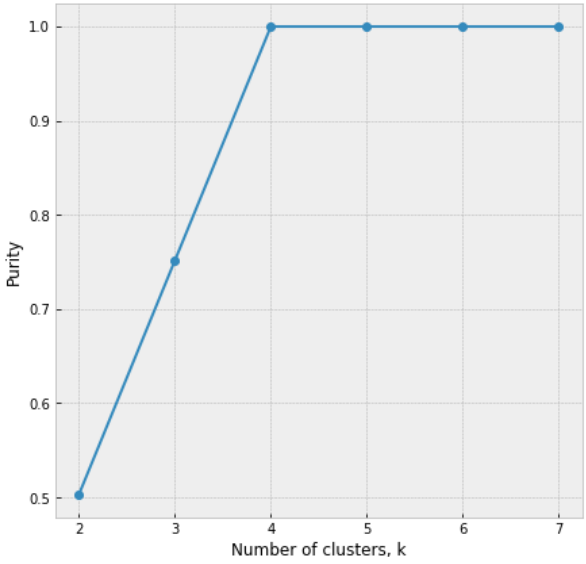
Calinski-Harabaz index



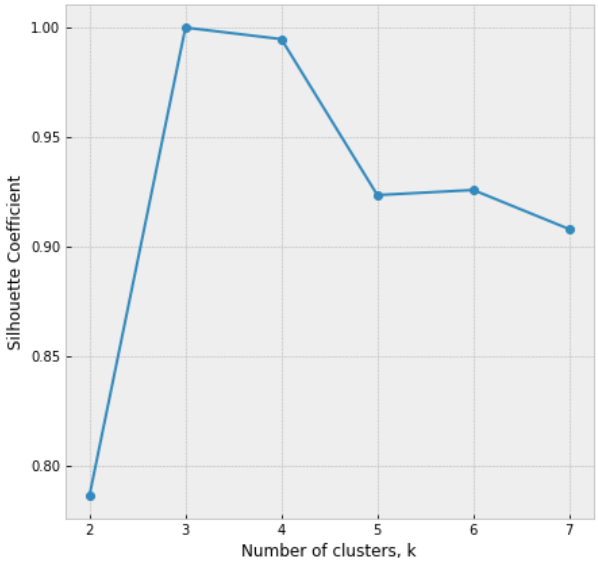
V-measure



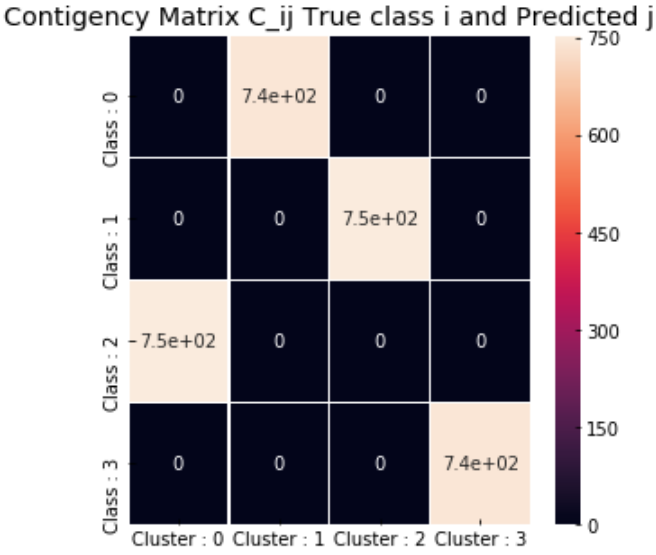
Purity



Silhouette Coefficient



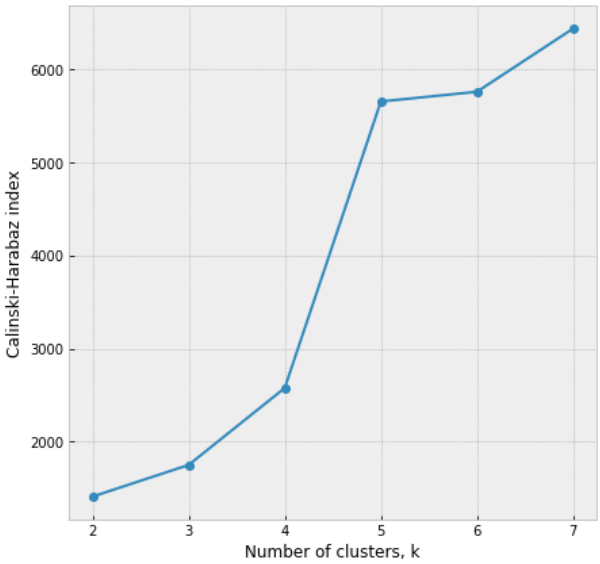
Contigency Matrix



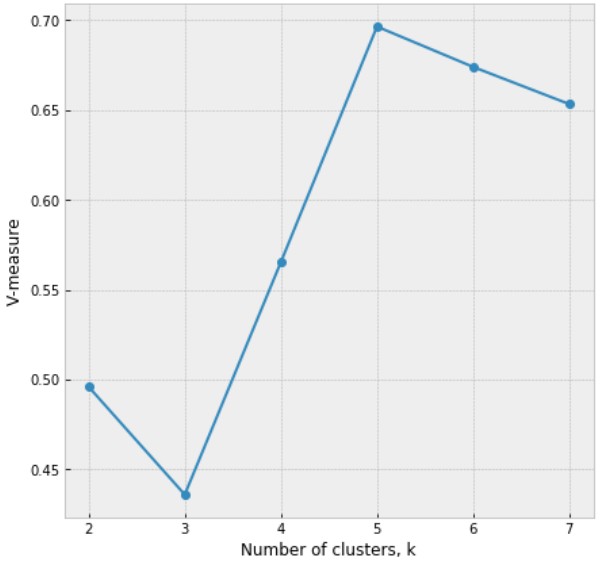
ISOMAP *(Clusters number = 4)*

```
--Number of clusters is 4 --

--Homogenity of clusters is 0.503 --

--Silhouette measure of clusters is 0.651--

--Completness of clusters is 0.646 --

--V measure score of clusters is 0.566--

--Adjusted Mutual Information score of clusters is 0.503--

--Calinski Harabaz Index score is 2567.061--

--Purity Score is 0.632--
```

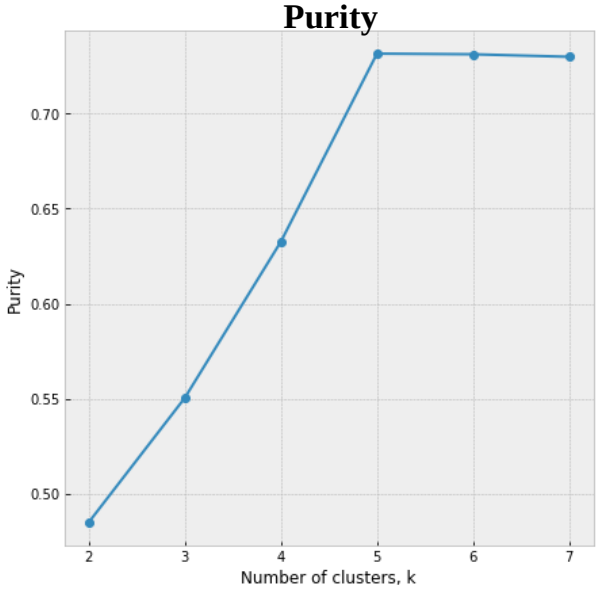
Calinski Harabaz-Index



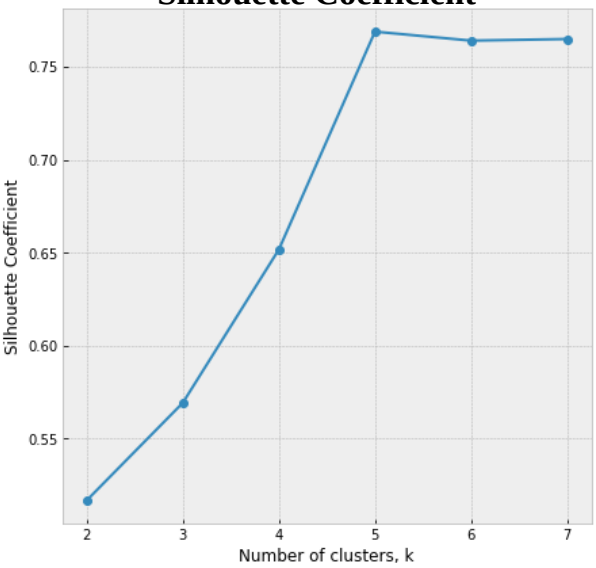
V-Measure



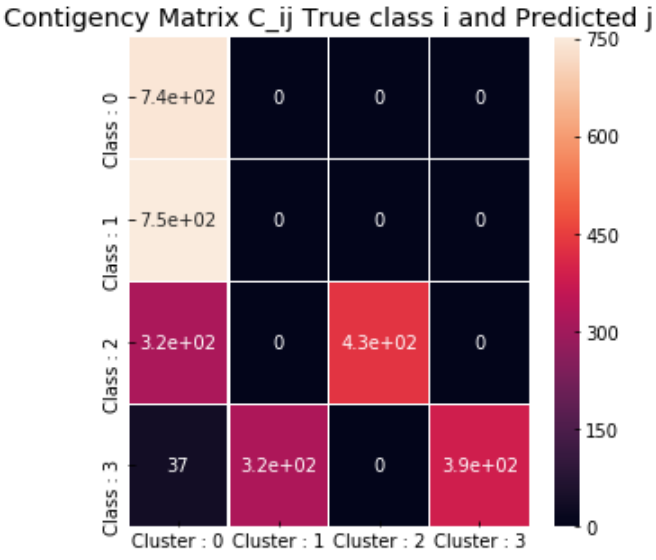
Purity



Silhouette Coefficient



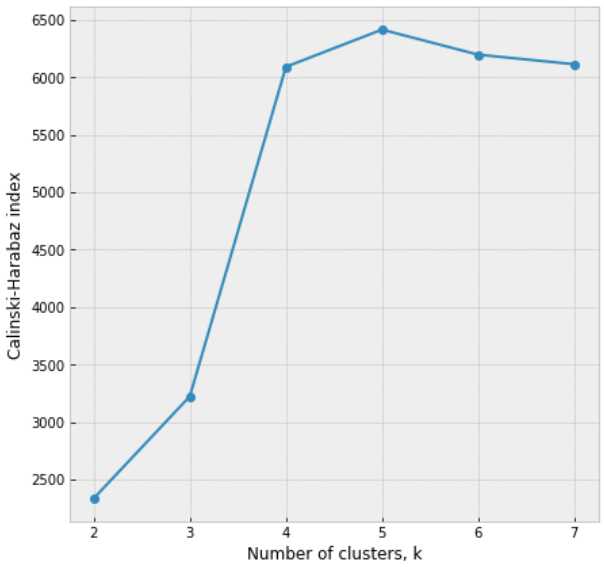
Contigency Matrix



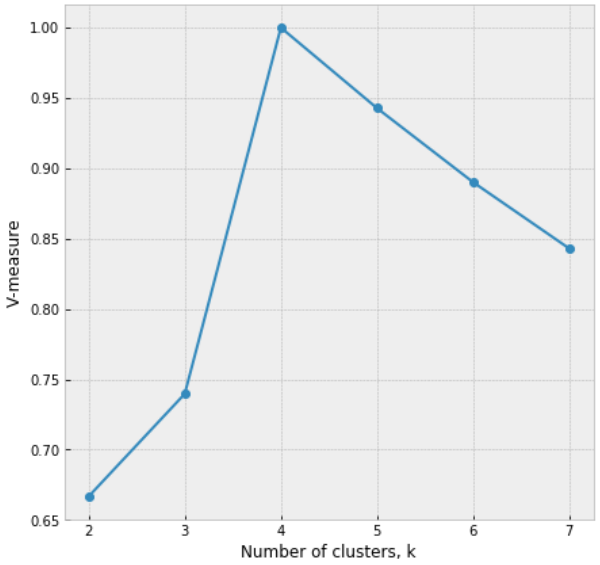
***tSNE* (Clusters number = 4)**

```
--Number of clusters is 4 --

--Homogenity of clusters is 1.000 --

--Silhouette measure of clusters is 0.605--

--Completness of clusters is 1.000 --

--V measure score of clusters is 1.000--

--Adjusted Mutual Information score of clusters is 1.000--

--Calinski Harabaz Index score is 6091.401--

--Purity Score is 1.000--
```

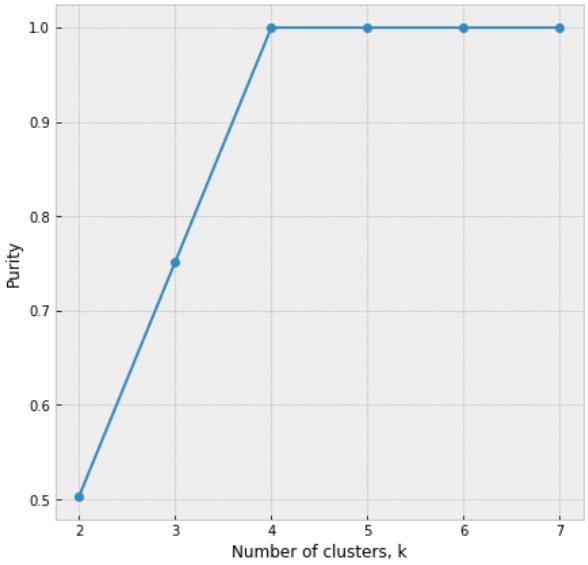
Calinski-Harabaz Index



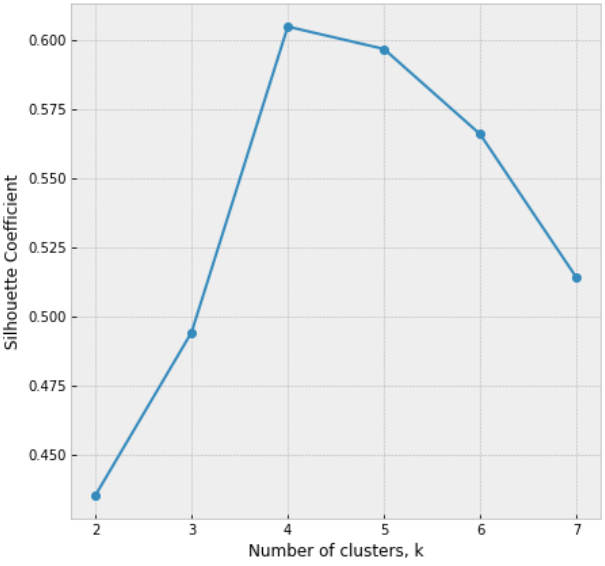
V-measure



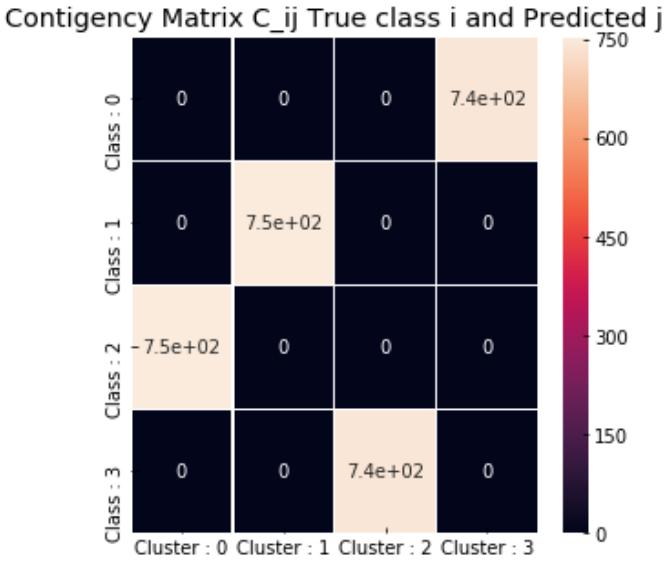
Purity



Silhouette Coefficient



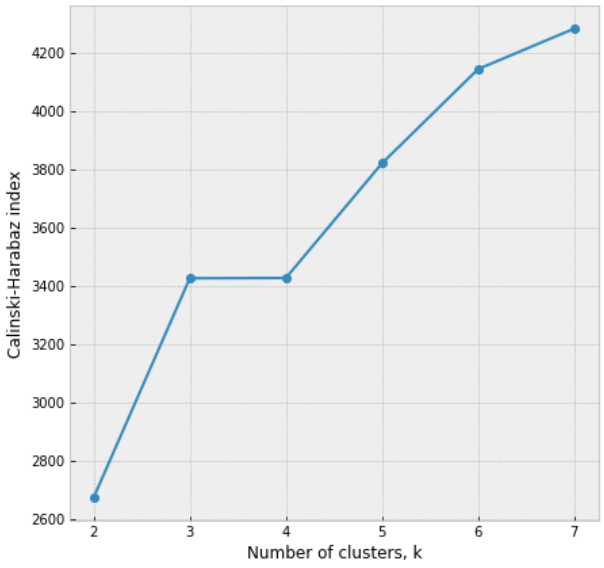
Contigency Matrix



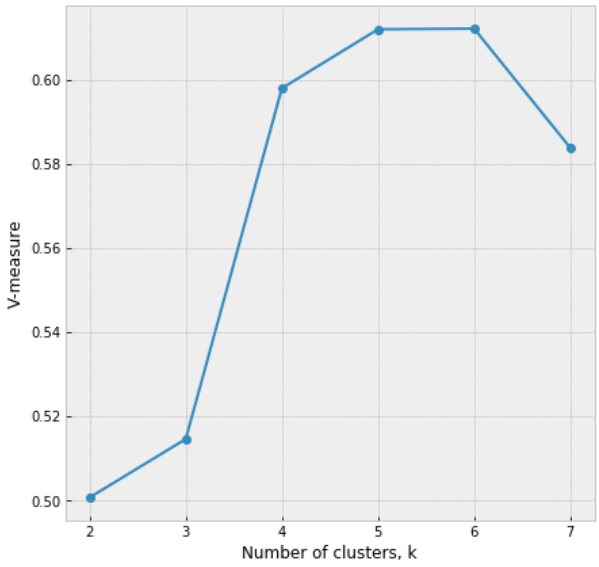
MDS *(Clusters number = 4)*

```
--Number of clusters is 4 --

--Homogenity of clusters is 0.584 --

--Silhouette measure of clusters is 0.447--

--Completness of clusters is 0.613 --

--V measure score of clusters is 0.598--

--Adjusted Mutual Information score of clusters is 0.584--

--Calinski Harabaz Index score is 3426.192--

--Purity Score is 0.695--
```

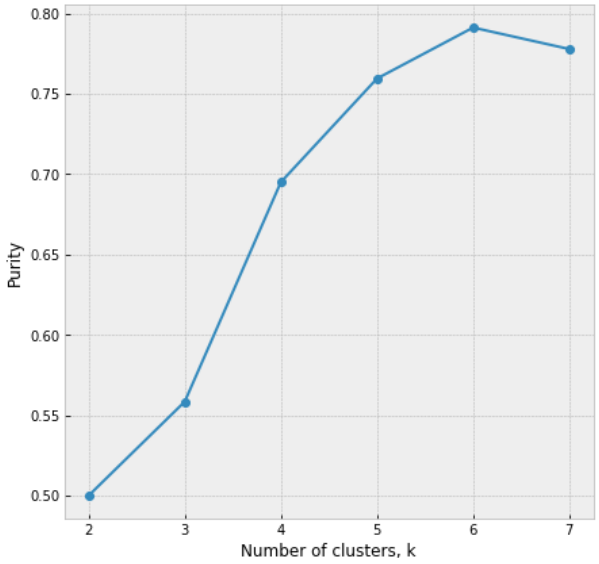
Calinski-Harabaz Index



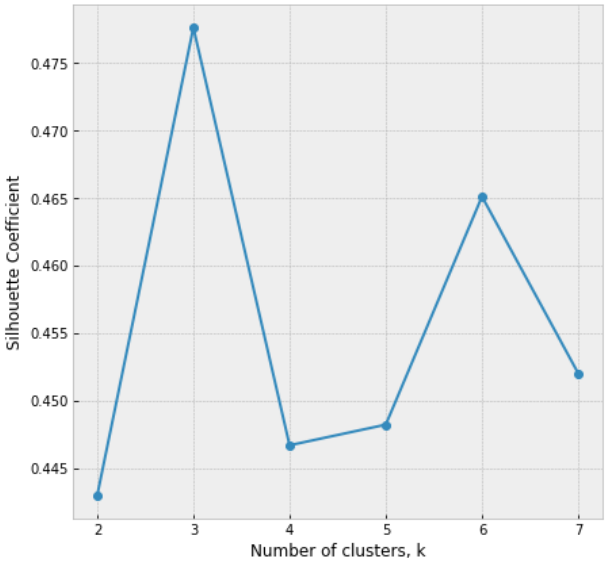
V-measure



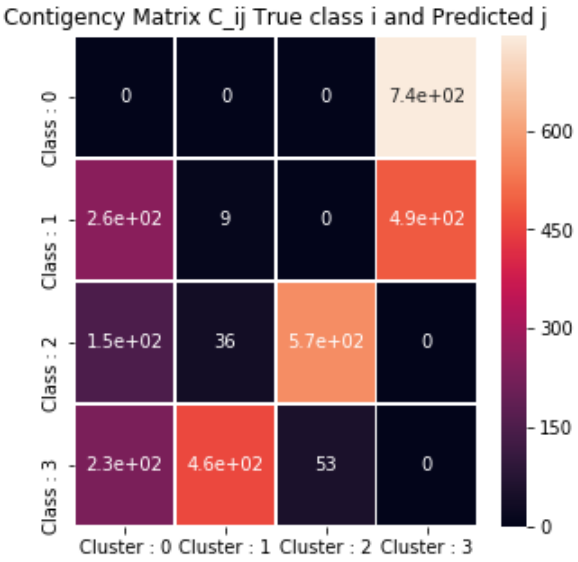
Purity



Silhouette coefficient



Contigency Matrix

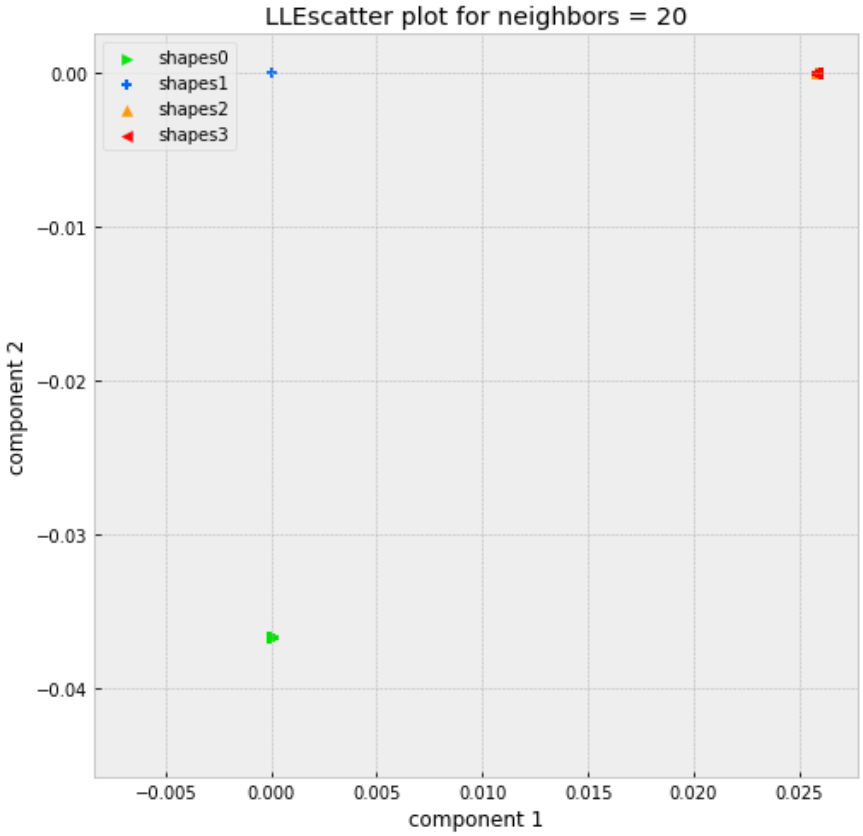


Αφού βγάλαμε τα αποτελέσματα μας λοιπόν για τα δεδομένα εκπαίδευσης ας δούμε πως μπορούμε να τα εφαρμόσουμε στα δεδομένα ελέγχου.

Για τους αλγόριθμους **tSNE** και **MDS** επειδή όπως εξηγούμε στο κείμενο ‘αλγόριθμοι’ δεν έχουν την ιδιότητα outsampling αυτό που θα κάνουμε είναι να εκπαιδεύσουμε έναν μοντέλο που θα πραγματοποιεί Multioutput Regression προκειμένου να μπορέσουμε να προβάλουμε τα καινούρια μας δεδομένα.

***LLE* (*Clusters number = 4*)**

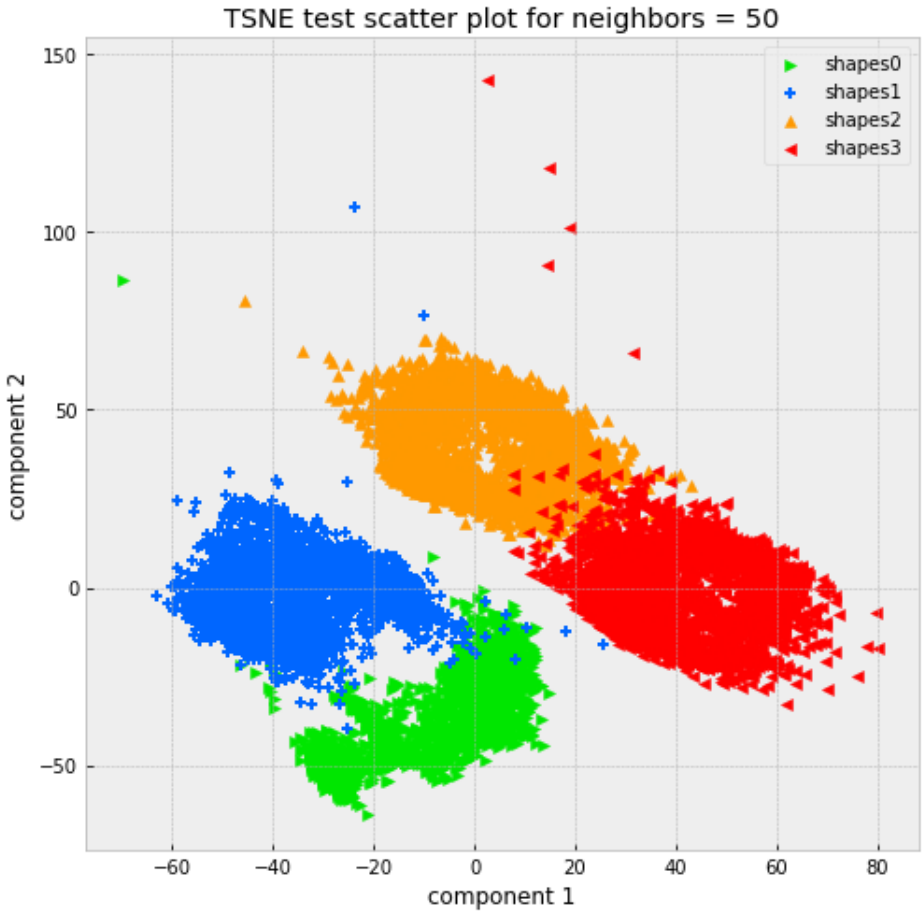
Τα αποτελέσματα που προέκυψαν ήταν τα παρακάτω



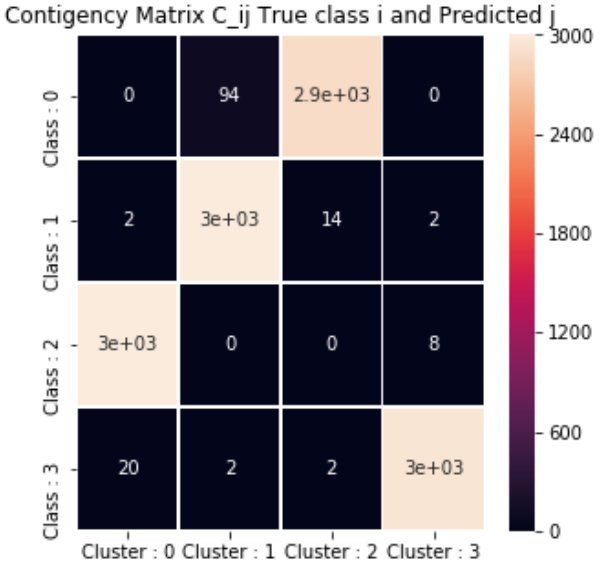

```
--Number of clusters is 4 --

--Homogenity of clusters is 1.000 --

--Silhouette measure of clusters is 0.994--

--Completness of clusters is 1.000 --

--V measure score of clusters is 1.000--

--Adjusted Mutual Information score of clusters is 1.000--

--Calinski Harabaz Index score is 17132109718775896064.000--

--Purity Score is 1.000--
```

***tSNE* (Number of clusters = 4)**

Τα αποτελέσματα που προέκυψαν ήταν τα παρακάτω:

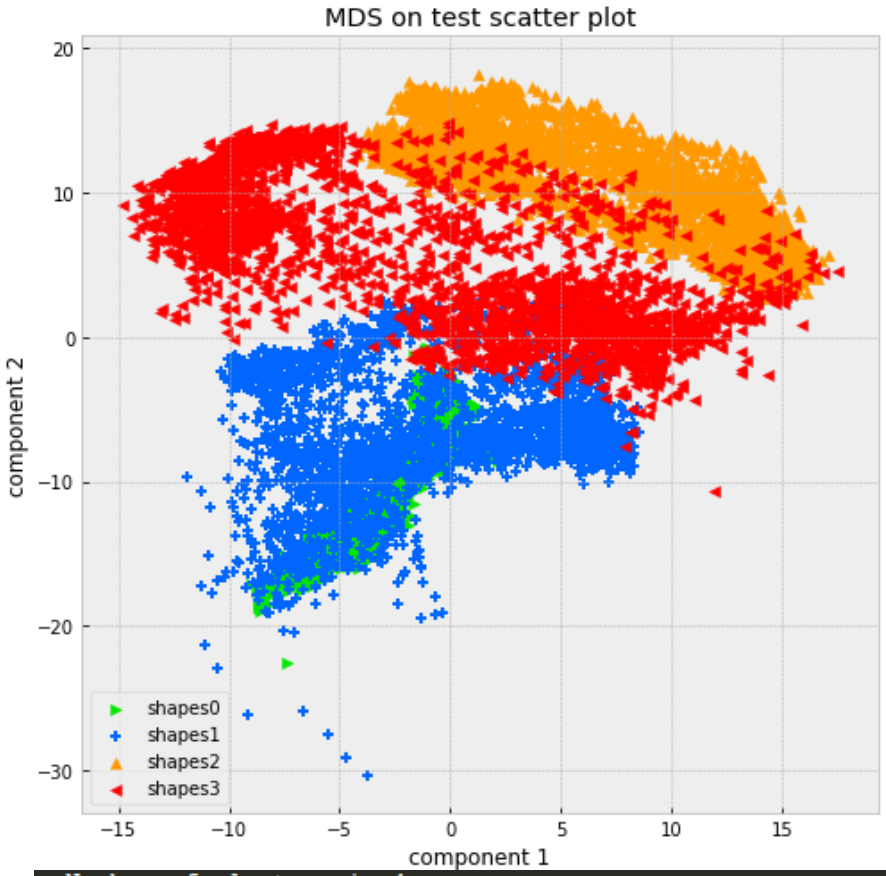



```
--Number of clusters is 4 --

--Homogenity of clusters is 0.955 --

--Silhouette measure of clusters is 0.576--

--Completness of clusters is 0.955 --

--V measure score of clusters is 0.955--

--Adjusted Mutual Information score of clusters is 0.955--

--Calinski Harabaz Index score is 22461.003--

--Purity Score is 0.988--
```

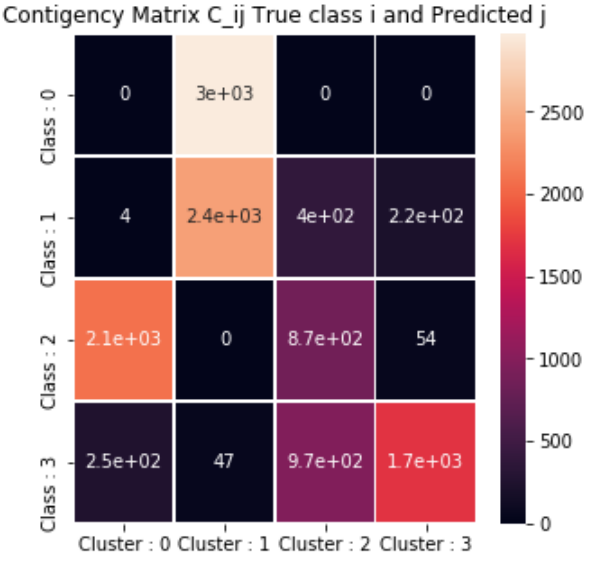


MDS *(Number of Clusters = 4)*

Τα αποτελέσματα που προέκυψαν ήταν τα παρακάτω:

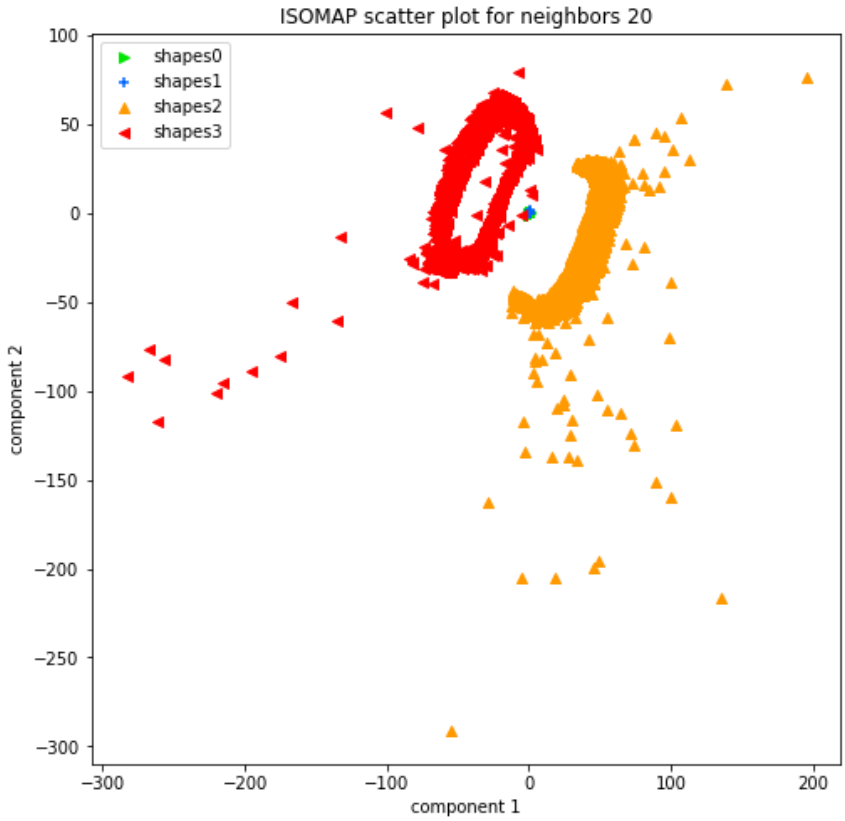



```
--Number of clusters is 4 --

--Homogenity of clusters is 0.516 --

--Silhouette measure of clusters is 0.477--

--Completness of clusters is 0.555 --

--V measure score of clusters is 0.535--

--Adjusted Mutual Information score of clusters is 0.516--

--Calinski Harabaz Index score is 15215.582--

--Purity Score is 0.646--
```

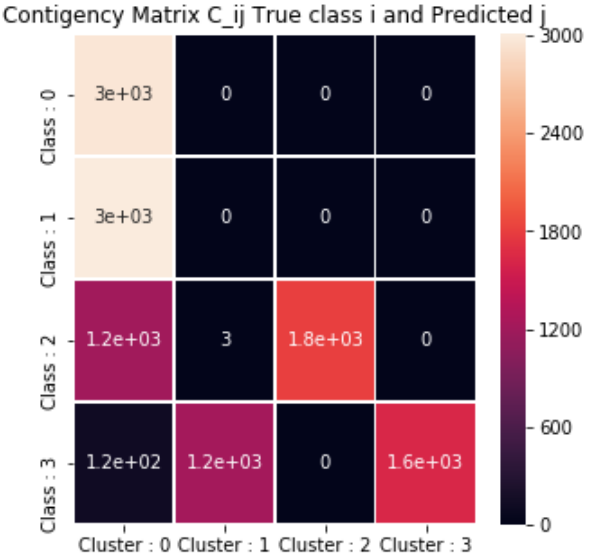


ISOMAP *(Number of Clusters = 4)*

Τα αποτελέσματα που προέκυψαν ήταν τα παρακάτω




```
--Number of clusters is 4 --

--Homogenity of clusters is 0.516 --

--Silhouette measure of clusters is 0.657--

--Completness of clusters is 0.655 --

--V measure score of clusters is 0.577--

--Adjusted Mutual Information score of clusters is 0.516--

--Calinski Harabaz Index score is 9336.813--

--Purity Score is 0.641--
```



Συμπεράσματα:Απ’οσο είδαμε οι αλγόριθμοι που μας έδωσαν τις καλύτερες επιδόσεις ήτανε αυτοί του LLE και tSNE, παρ’όλα αυτά και στους δύο εχουμε να αντιμετπωπίσουμε αρκετά ζητήματα,στον LLE όπως είδαμε είναι κυρίαρχο το Collapse Problem και στον tSNE το πρόβλημα είναι ότι δεν έχουμε την δυνατότητα outsampling και βασιζόμαστε σε άλλες τεχνικές για να μπορέσουμε να κατηγοριοποιήσουμε τα δεδομένα ελέγχου.