ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Εργασία 2

KPCA+LDA

Υπολογιστική Νοημοσύνη-Στατιστική Μάθηση

Σουράνης Παναγιώτης

ΑΕΜ:17

Το παρακάτω κείμενο είναι μια σύντομη αναφορά της εφαρμογής των αλγορίθμων KPCA plus LDA στο dataset **RNA Gene Sequence** το οποίο πάρθηκε από την σελίδα

https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

Περιγραφή Dataset:

Στο παρακάτω σύνολο δεδομένων μας δίνονται οι πληροφορίες από αλληλουχίες γονιδίων RNA οι οποίες αφορούν 5 συγκεκριμένα γονίδια τα οποία είναι τα (BRCA, COAD,KIRC,LUAD,PRAD).Στόχος μας είναι λοιπόν να κατατάξουμε τις αλληλουχίες RNA στις παραπάνω κλάσεις

Ας ξεκινήσουμε με την περιγραφή του συνόλου δεδομένων:

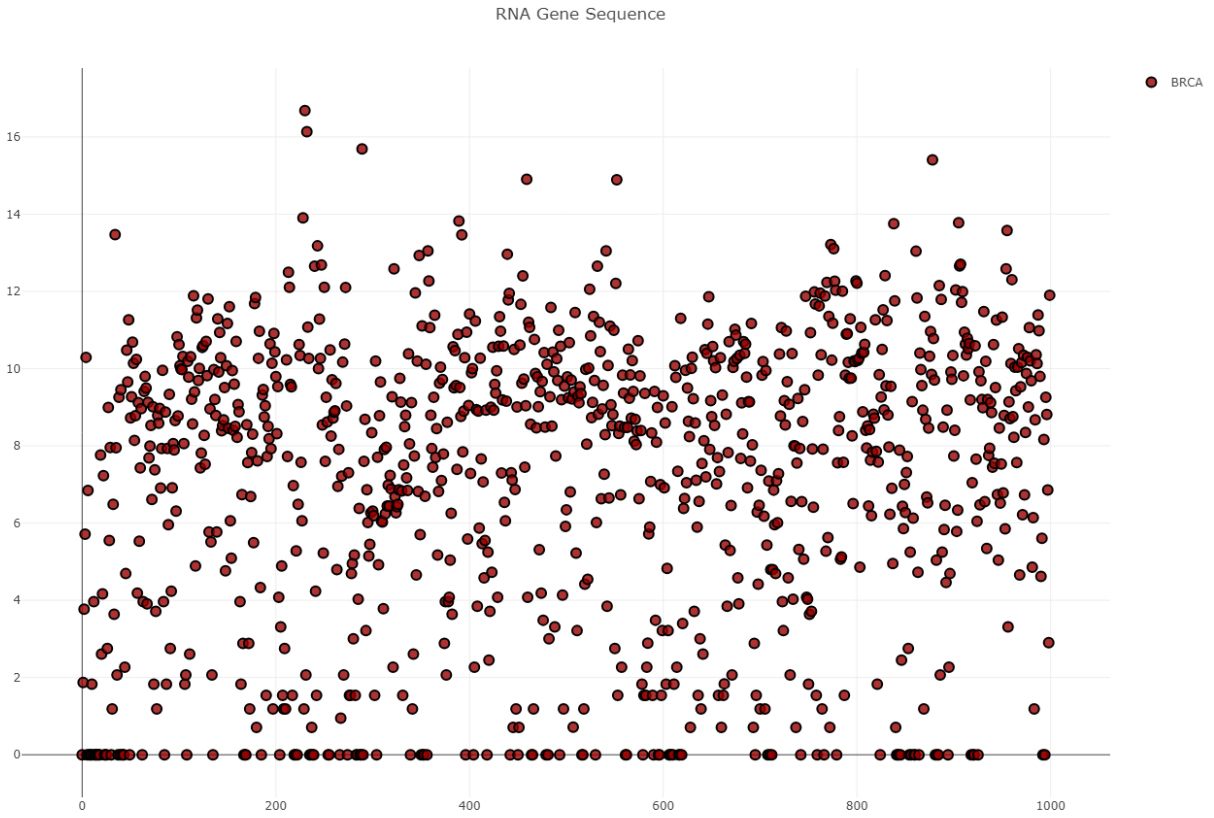
	Unnamed: 0	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	...	gene_20521	gene_20522	gene_20523	gene_20524
0	sample_0	0.0	2.017209	3.265527	5.478487	10.431999	0.0	7.175175	0.591871	0.0	...	4.926711	8.210257	9.723516	7.220030
1	sample_1	0.0	0.592732	1.588421	7.586157	9.623011	0.0	6.816049	0.000000	0.0	...	4.593372	7.323865	9.740931	6.256586
2	sample_2	0.0	3.511759	4.327199	6.881787	9.870730	0.0	6.972130	0.452595	0.0	...	5.125213	8.127123	10.908640	5.401607
3	sample_3	0.0	3.663618	4.507649	6.659068	10.196184	0.0	7.843375	0.434882	0.0	...	6.076566	8.792959	10.141520	8.942805
4	sample_4	0.0	2.655741	2.821547	6.539454	9.738265	0.0	6.566967	0.360982	0.0	...	5.996032	8.891425	10.373790	7.181162

5 rows × 20532 columns

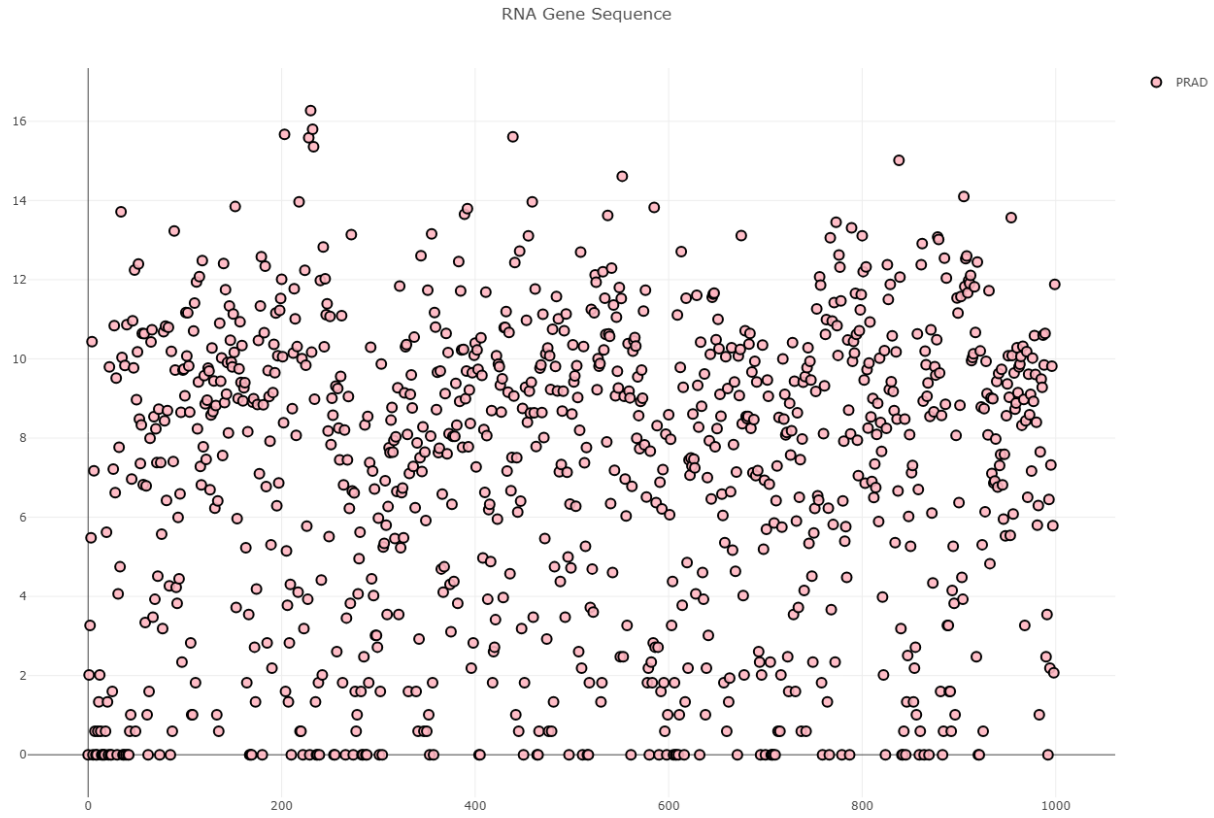
Παρατηρούμε ότι έχουμε δεδομένα τα οποία δεν ανήκουν σε καμία κατηγορία και δεν μπορούμε να τα συμπεριλάβουμε στην εκπαίδευση του αλγορίθμου και για αυτό πρέπει να απομακρυνθούν.

Ας ρίξουμε μια ματία λοιπόν σε κάποια τυχαία δεδομένα

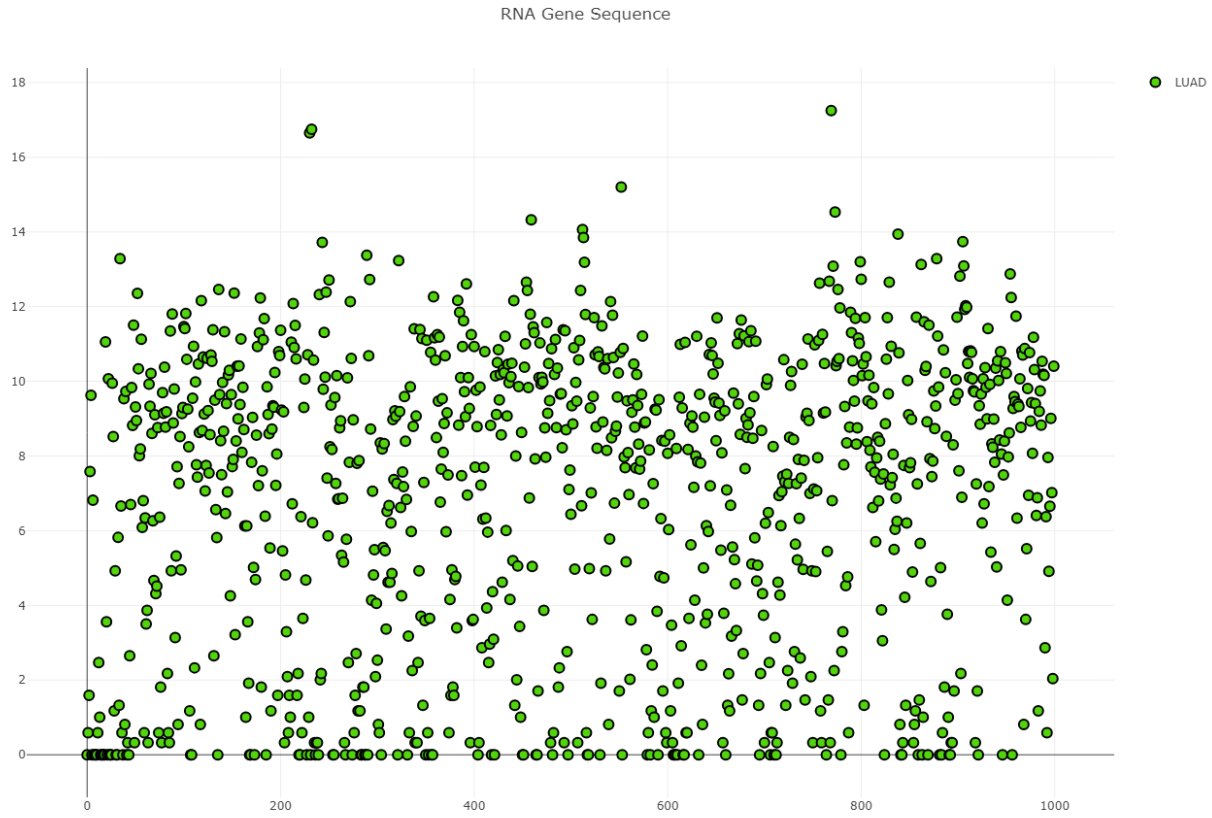
Σχήμα 1.1 (Γονίδιο BRCA)



Σχήμα 1.2 (Γονίδιο PRAD)

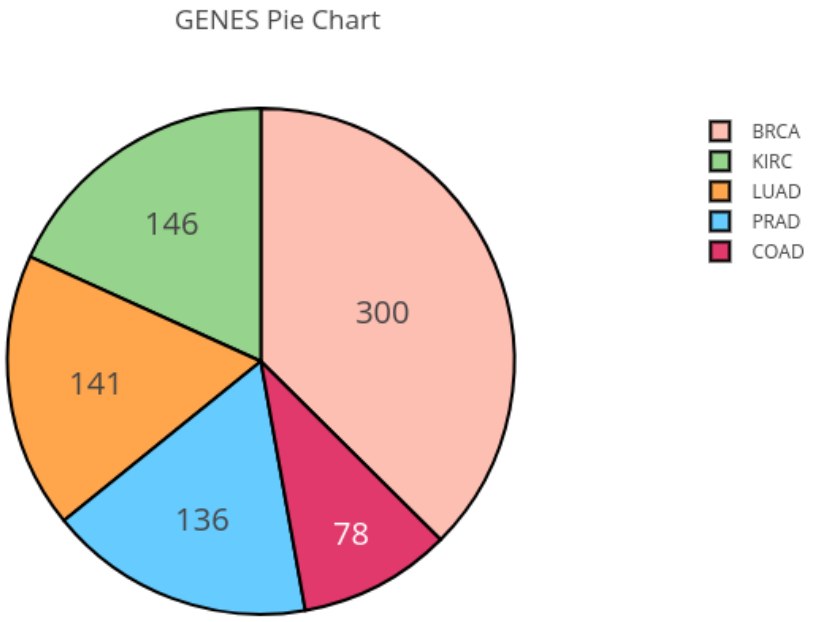


Σχήμα 1.3 (Γονίδιο LUAD)



Έπειτα θα ήταν χρήσιμο να χρησιμοποιήσουμε ένα διάγραμμα πίτας για να μπορέσουμε να καταλάβουμε την κατανομή αυτων των γονιδίων στο συγκεκριμένο dataset έτσι ώστε να καταλάβουμε αν οι κλάσεις του συνόλου μας δεν είναι ισορροπημένες.

Σχήμα 1.4 (Genes Pie Chart)

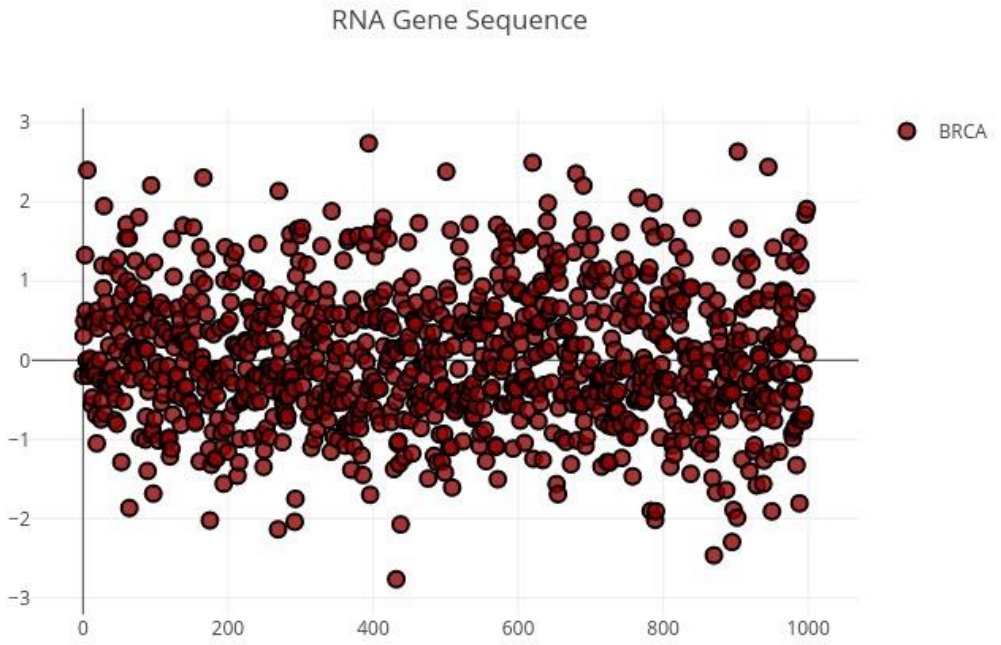


Φαίνεται λοιπόν να υπερέχει μεταξύ των άλλων η κλάση με το γονίδιο BRCA,μένει να δούμε λοιπόν κατά πόσο αυτό θα επηρεάσει το τελικό αποτέλεσμα.

Θα εφαρμόσουμε λοιπόν στην συνέχεια κανονικοποίηση στα δεδομένα μας (Standard Scaling) διότι είναι ευκολότερο για τον αλγόριθμο μας να επεξεργαστεί τα δεδομένα όταν οι τιμές τους είναι «σχετικά» μικρότερες απ'ότι ήταν στο αρχικό σύνολο.

Η εικόνα που προκύπτει μετά την εφαρμογή scaling είναι η παρακάτω.

Σχήμα 1.5 (Γονίδιο BRCA)



Βλέπουμε ότι όλες οι τιμές έχουν συρρικνωθεί στο διάστημα [-3 , 3] απ'ότι ήταν στην αρχή στο [0,16].

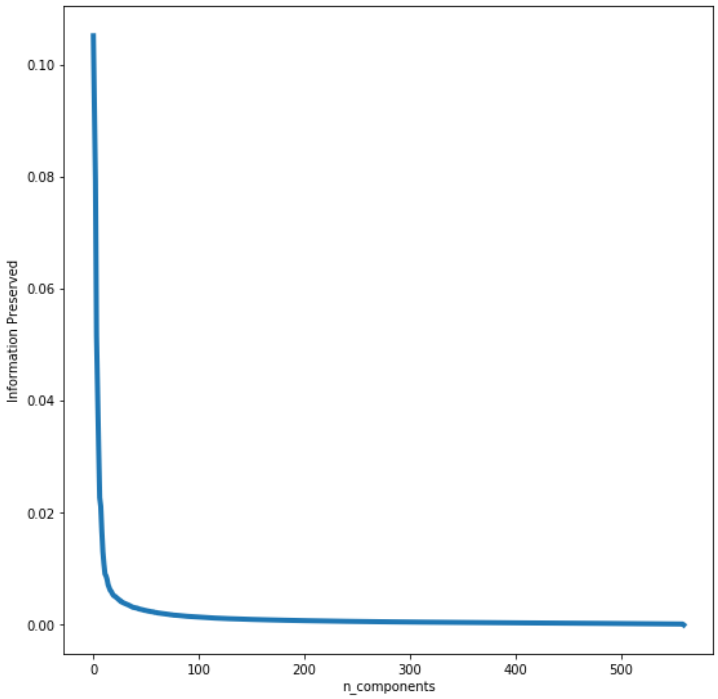
Εφόσον λοιπόν πραγματοποιήσαμε και το scaling ήρθε η ώρα να προχωρήσουμε στην εφαρμογή του Kernel PCA με τους 3 γραμμικούς πυρήνες (Linear,Polynomial,Radial Basis Function)

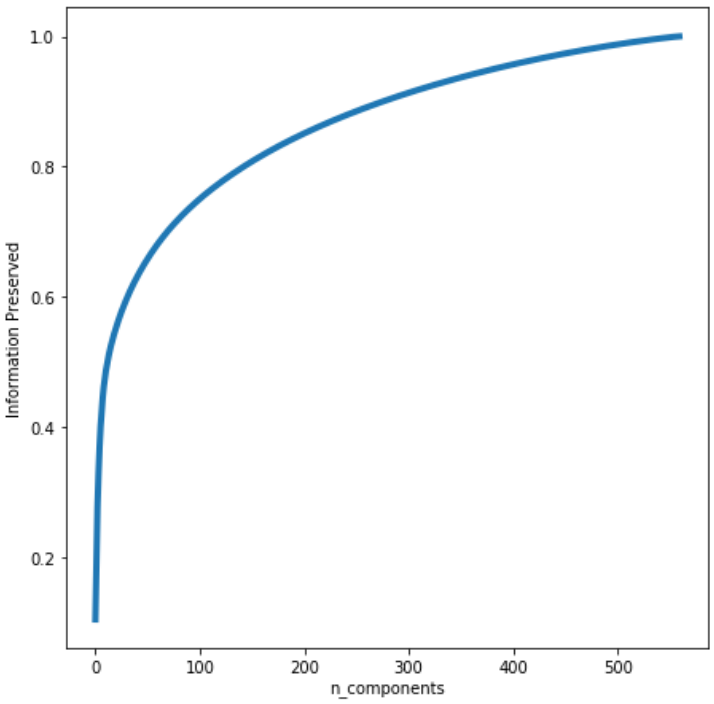
- **Linear Kernel PCA**

Στον κώδικα ζητήθηκε να κρατήσει τουλάχιστον το 93% της συνολικής πληροφορίας οπότε το αποτέλεσμα που μας δώθηκε ήταν ότι θα έπρεπε να κρατήσουμε τουλάχιστον 350 χαρακτηριστικά τουλάχιστον, η αρχική διάσταση ξεπερνούσε τις 20000

Η διακύμανση της πληροφορίας φαίνεται στο παρακάτω σχήμα

Σχήμα 1.6α και 1.6β

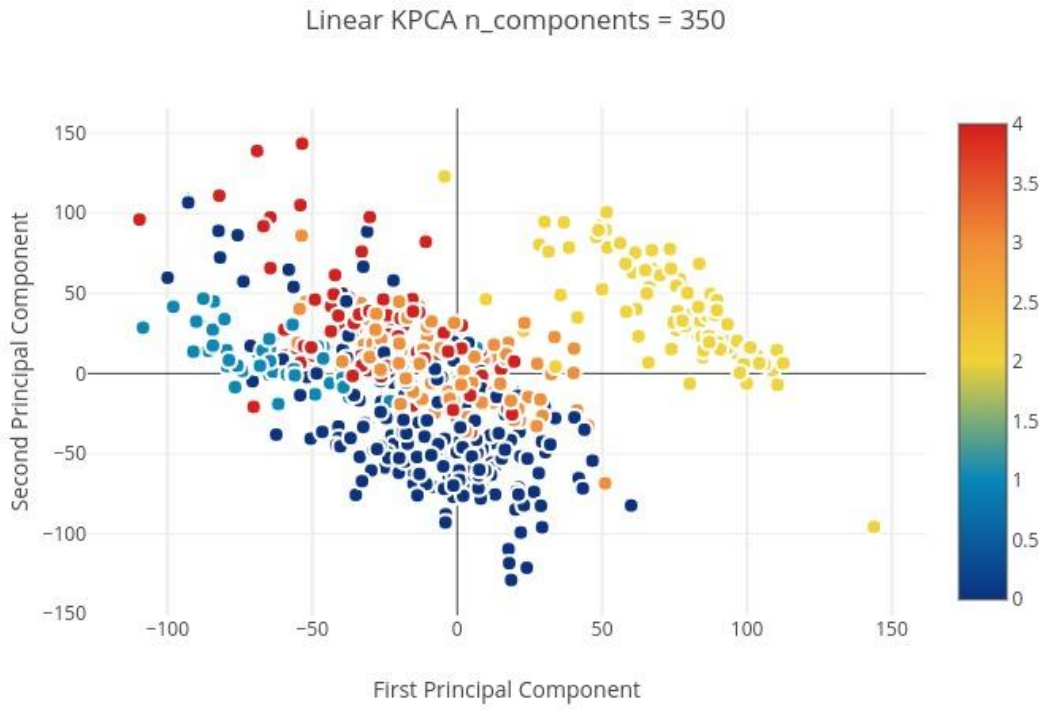




Η δεύτερη εικόνα αφορά το cumulative percentage δηλαδή το συνολικό ποσοστό της πληροφορίας που διατηρείται όσο αυξάνουμε το πλήθος των χαρακτηριστικών.

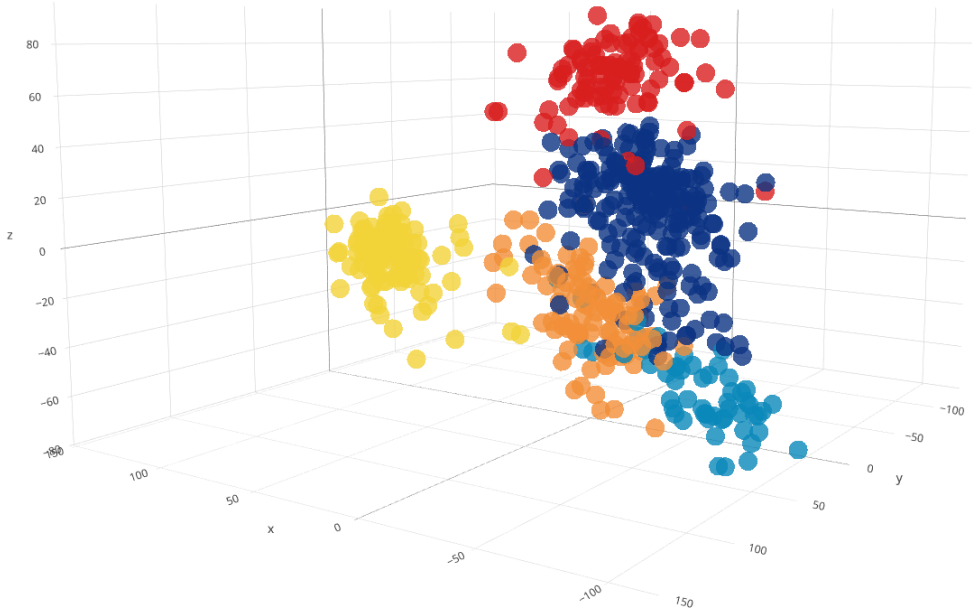
Η προβολή στις 2 διαστάσεις των δεδομένων μας φαίνεται στο παρακάτω σχήμα

Σχήμα 1.7



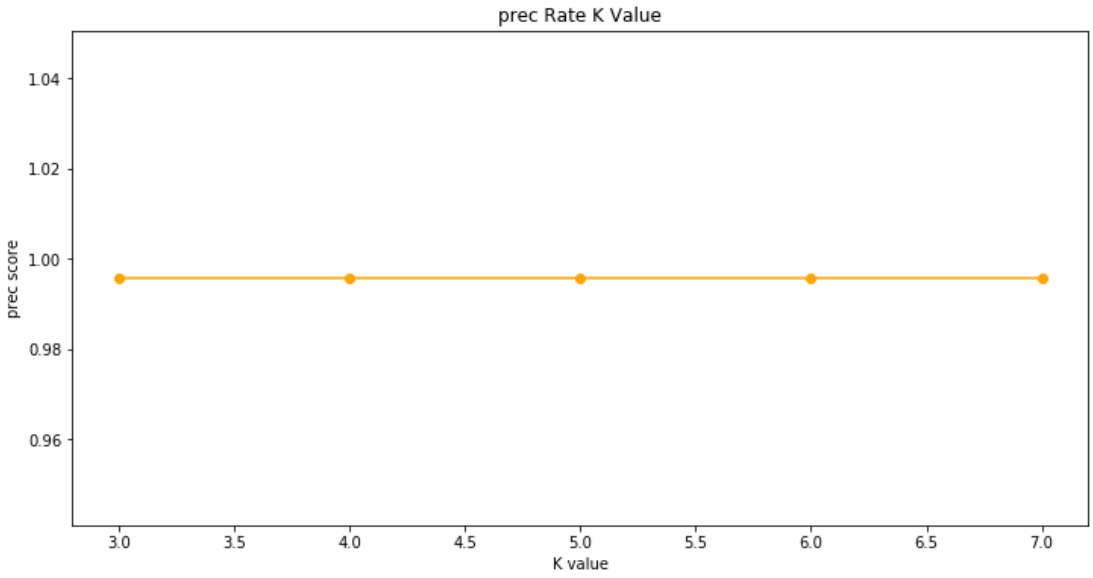
Προβολή στις 3 διαστάσεις

Σχήμα 1.8

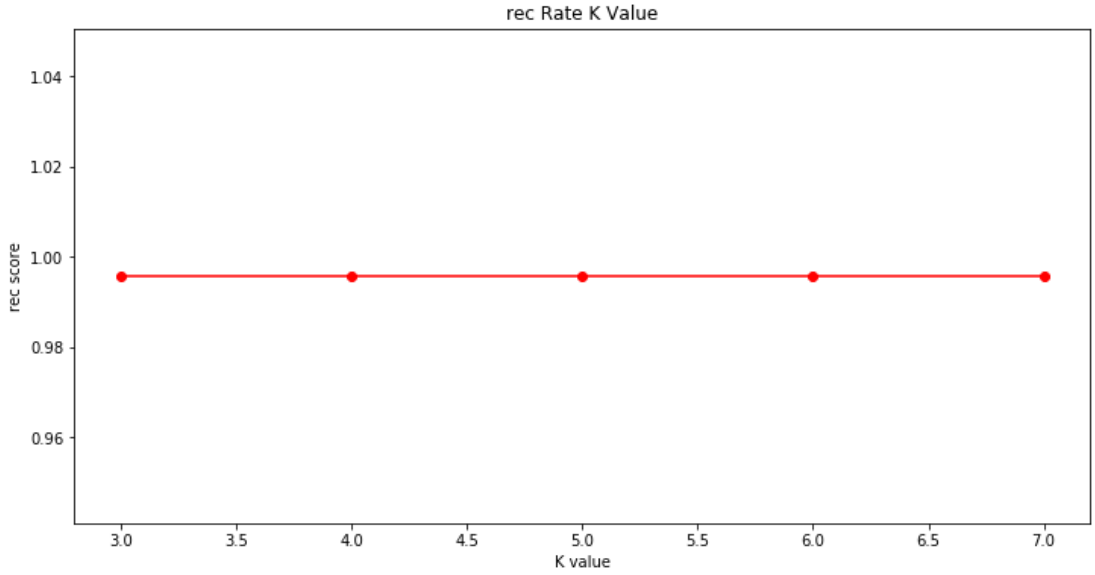


Στην συνέχεια εφαρμόστηκαν 2 αλγόριθμοι για την κατηγοριοποίηση (K-Nearest Neighbors και LogisticRegression) των οποίων τα αποτελέσματα εμφανίζονται παρακάτω.

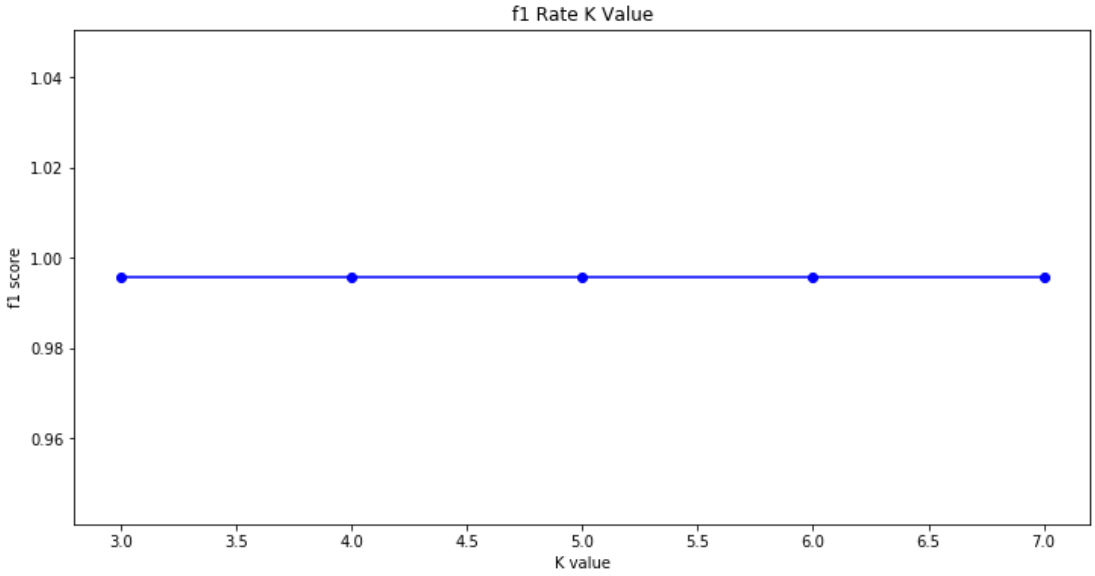
Precision (Nearest Neighbors)



Recall (KNN)



F1 Score (KNN)



Classification Report

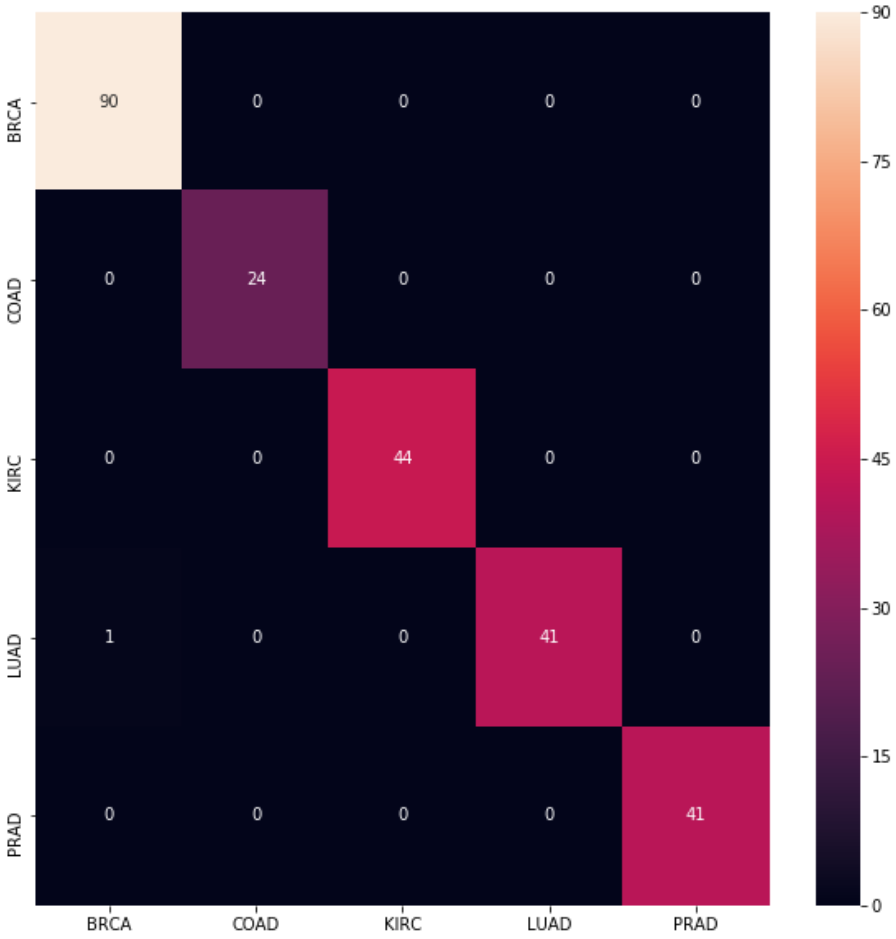
Test

	precision	recall	f1-score	support
0	1.00	0.99	0.99	91
1	1.00	1.00	1.00	24
2	0.98	1.00	0.99	43
3	0.98	1.00	0.99	41
4	1.00	0.98	0.99	42
avg / total	0.99	0.99	0.99	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Confussion Matrix



Classification Report (Logistic Regression)

	precision	recall	f1-score	support
0	0.99	1.00	0.99	90
1	1.00	1.00	1.00	24
2	1.00	1.00	1.00	44
3	1.00	0.98	0.99	42
4	1.00	1.00	1.00	41
avg / total	1.00	1.00	1.00	241

Test

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Train

- **Radial Basis Function KernelPCA**

Για την περίπτωση του RBF PCA έγινε έλεγχος στο gamma για τις παρακάτω τιμές

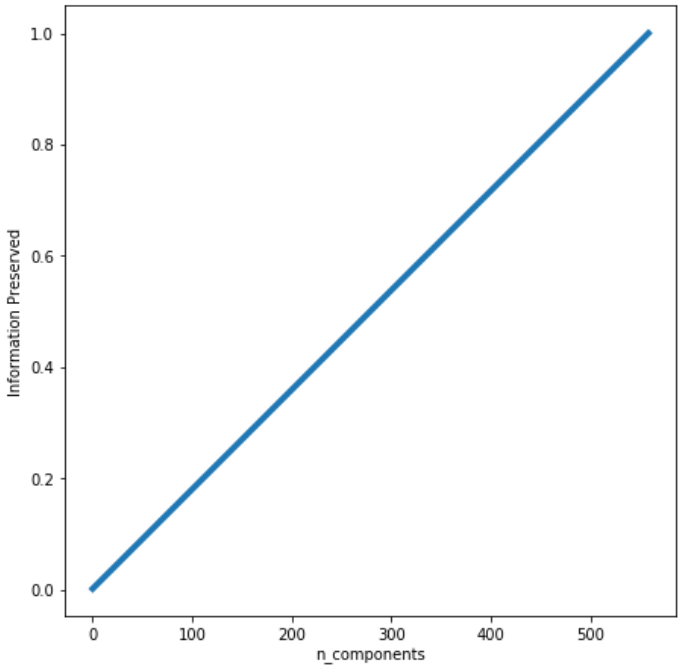
	1	2	3
values for gamma	0.5	0.05	0.005

- **RBF gamma = 0.5**

Στην περίπτωση του RBF PCA χρειάστηκαν όπως είναι λογικό περισσότερα components για να διατηρήσει το 93% της πληροφορίας της τάξης των 530.

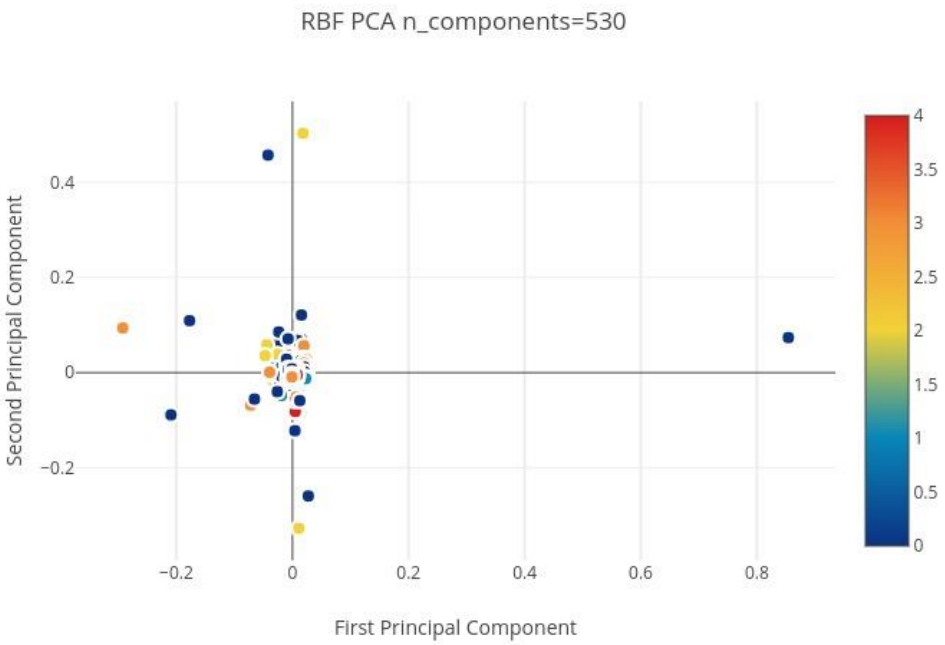
Στην συγκεκριμένη περίπτωση σε κάθε component αντιστοιχούσε 0.00178891 της συνολικής πληροφορίας.

Παρακάτω το αθροιστικό διάγραμμα για το μέγεθος της πληροφορίας



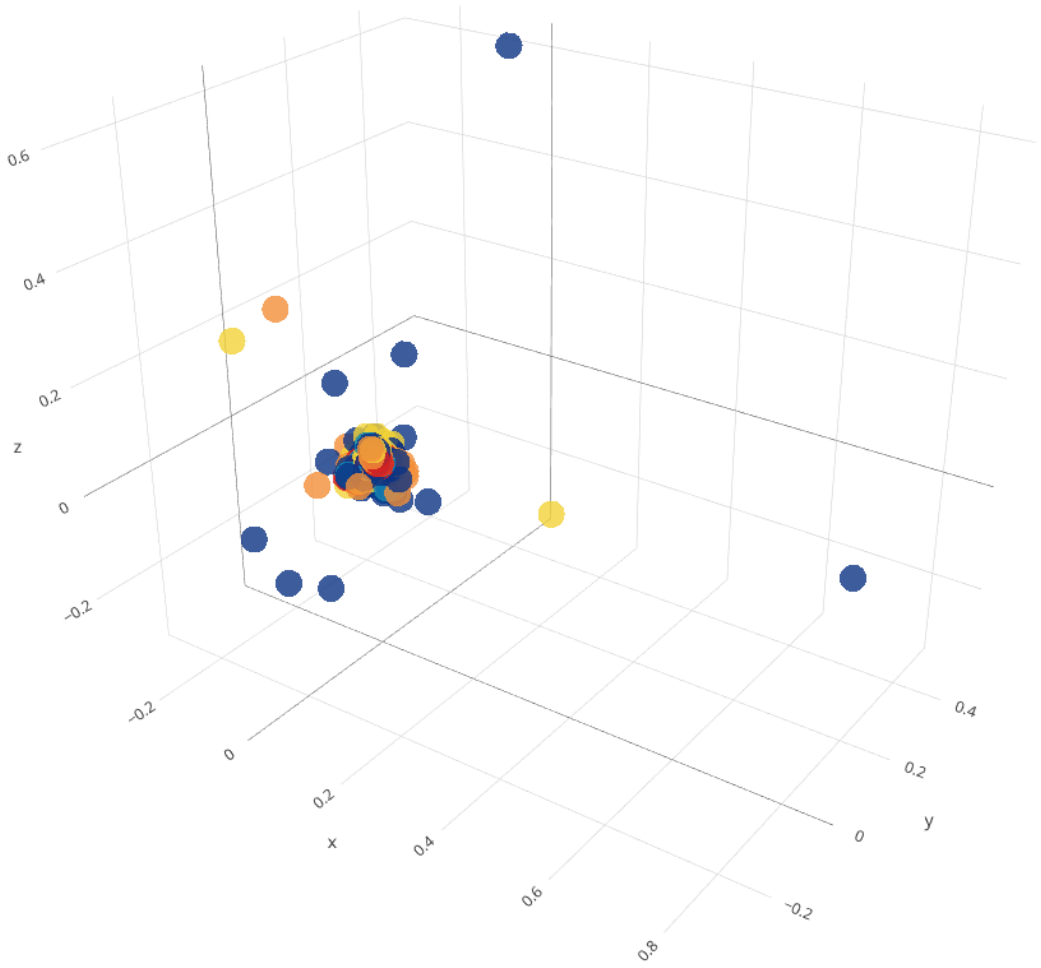
Ας δούμε στη συνέχεια την προβολή των στοιχείων μας στις 2 και ακολούθως στις 3 διαστάσεις.

2D Projection of points



Όπως βλέπουμε στη συγκεκριμένη περίπτωση τα δείγματα έχουν πέσει το ένα πάνω στο άλλο κατι που δεν είναι επιθυμητό καθώς στόχος μας είναι να τα διαχωρίσουμε όσο καλύτερα μπορούμε.

3D Projection of points



Note!

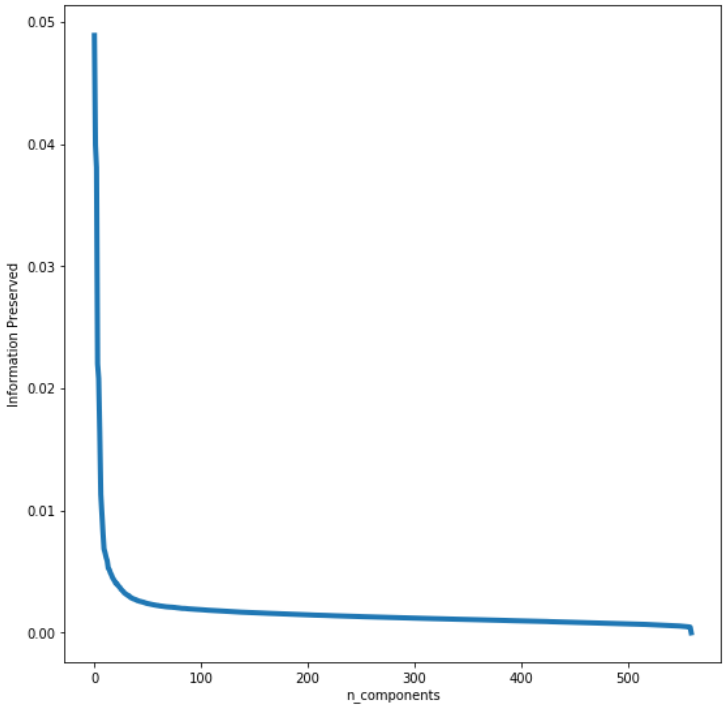
Τα αποτελέσματα του overfitting είναι περισσότερο ευδιάκριτα στις 3 διαστάσεις και η εφαρμογή του μετασχηματισμού στο test σύνολο δεδομένων μας επέστρεψε πίνακα του οποίου όλες οι τιμές ήταν μηδενικές και για αυτό δεν προχωρήσαμε στην εφαρμογή του classifier.

Τα ίδια αποτελέσματα προέκυψαν και για τις υπόλοιπες τιμές του gamma για αυτό προχωρήσαμε σε επανακαθορισμό των αρχικών τιμών,οι νέες τιμες είναι

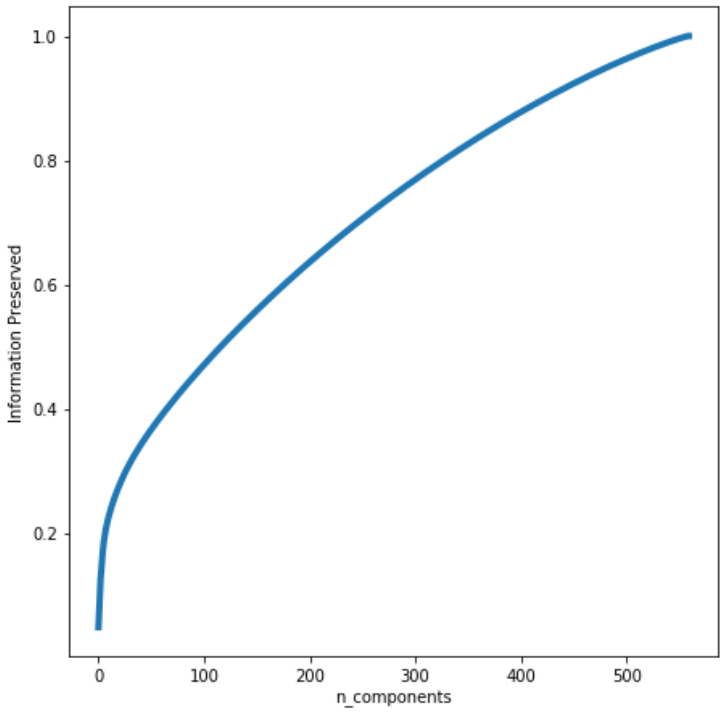
Gamma values : **0.00005** , **0.000005**

- **RBF (gamma = 0.00005)**

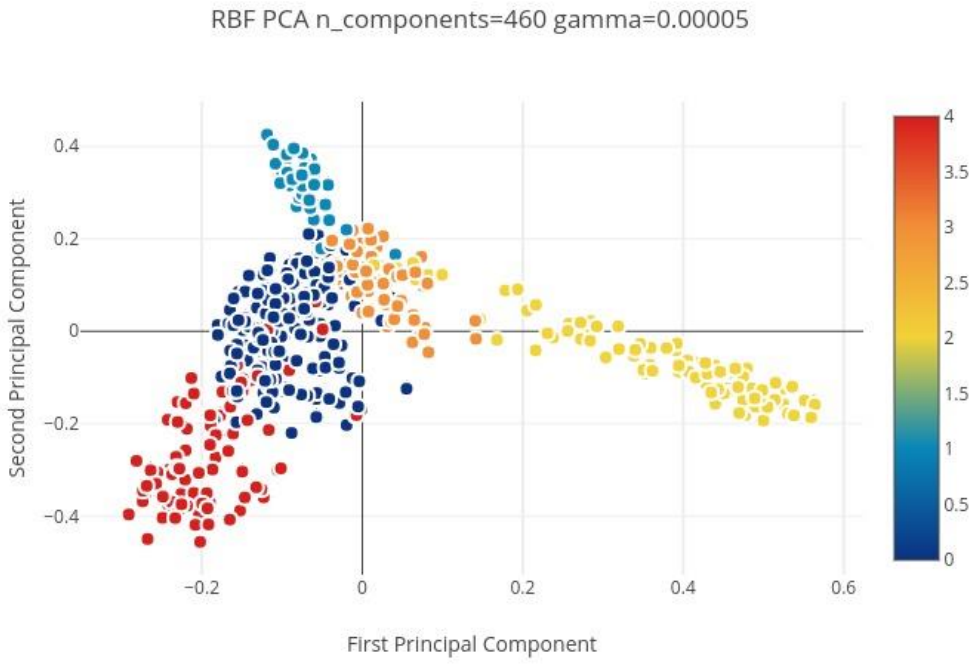
Πληροφορία που διατηρεί το κάθε χαρακτηριστικό



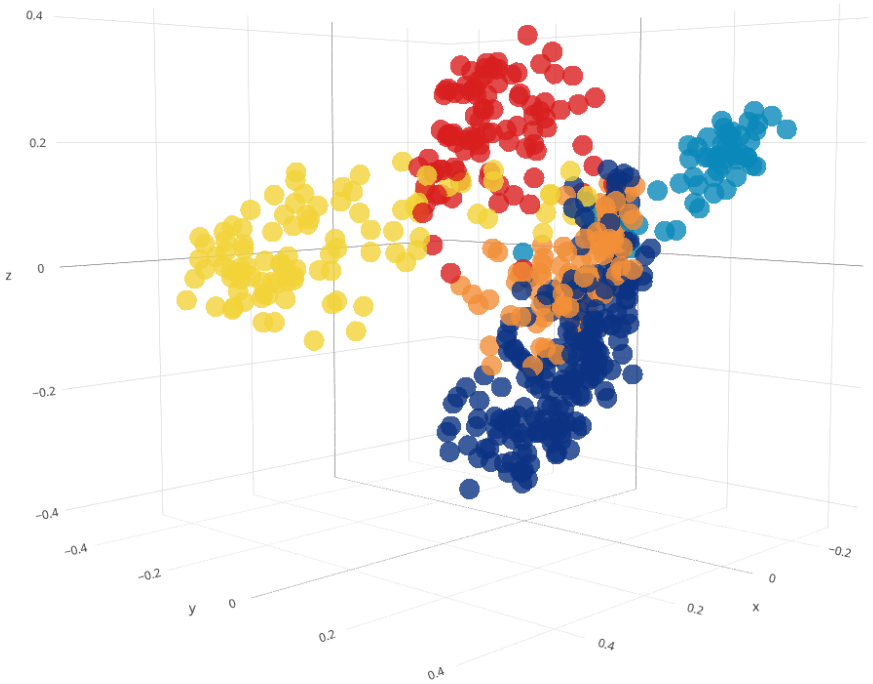
Αθροιστικό διάγραμμα της πληροφορίας



2D Projection of points

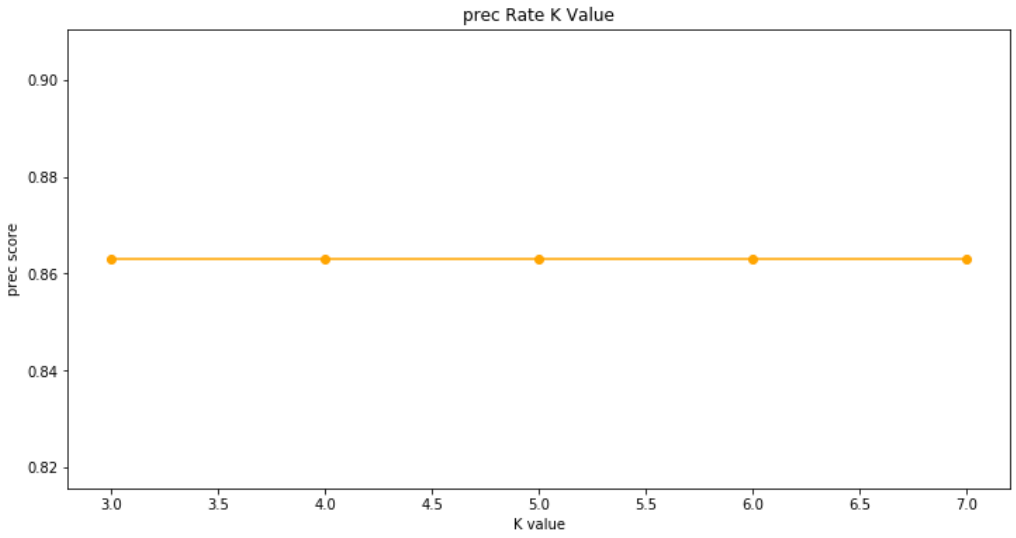


3D Projection of points

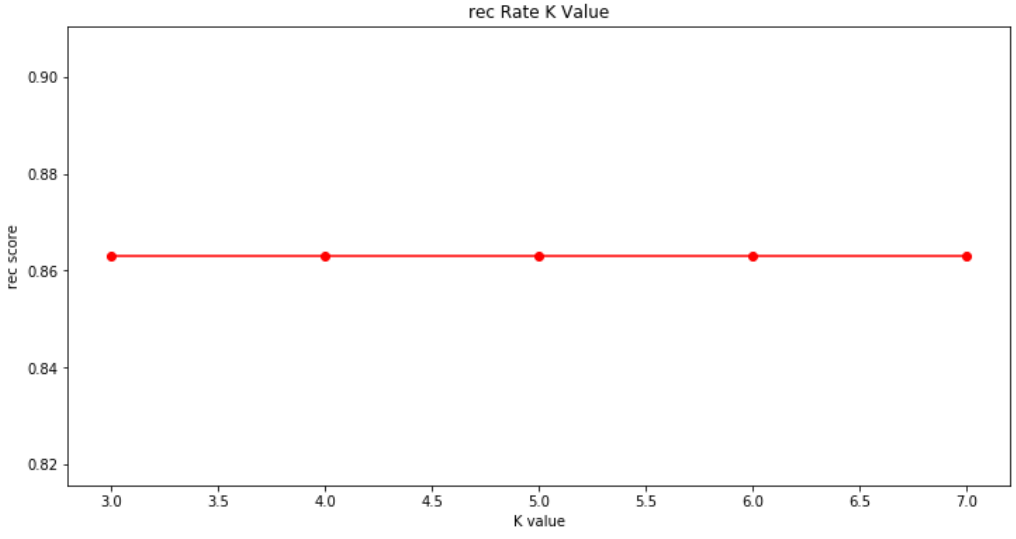


Ας προχωρήσουμε λοιπόν στην εφαρμογή του classifier και στα αποτελέσματα του.

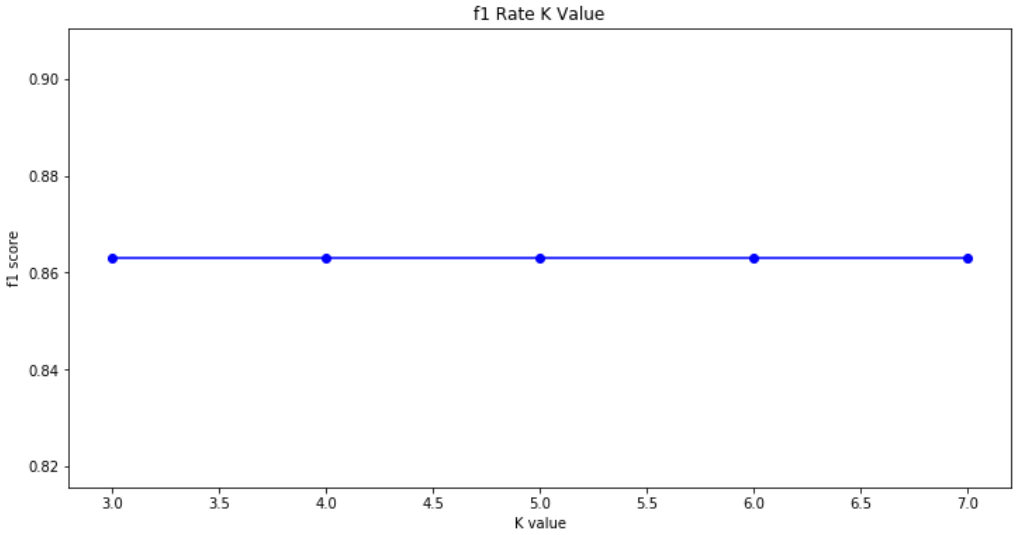
Precision



Recall



F1 Score



Classification Report

Test

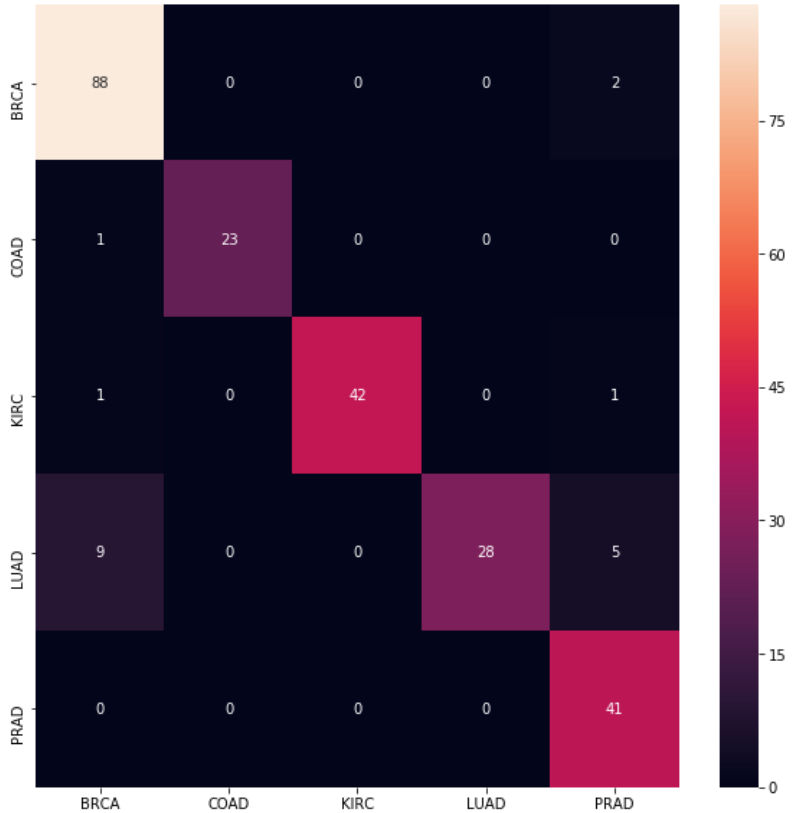
	precision	recall	f1-score	support
0	0.98	0.89	0.93	99
1	0.96	1.00	0.98	23
2	0.95	1.00	0.98	42
3	0.67	1.00	0.80	28
4	1.00	0.84	0.91	49
avg / total	0.94	0.92	0.92	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

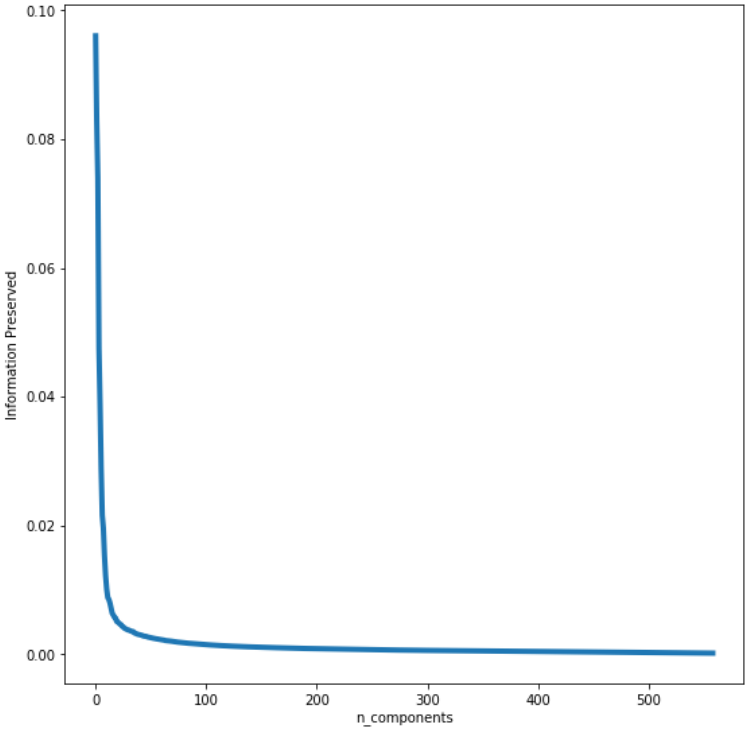
Όπως είναι φανερό έχει γίνει overfitting και για αυτό η απόδοση του αλγορίθμου στα test δεδομένα είναι αρκετά μικρότερη από την αποδόση στα train η οποία βρέθηκε να είναι άριστη.

Confussion Matrix

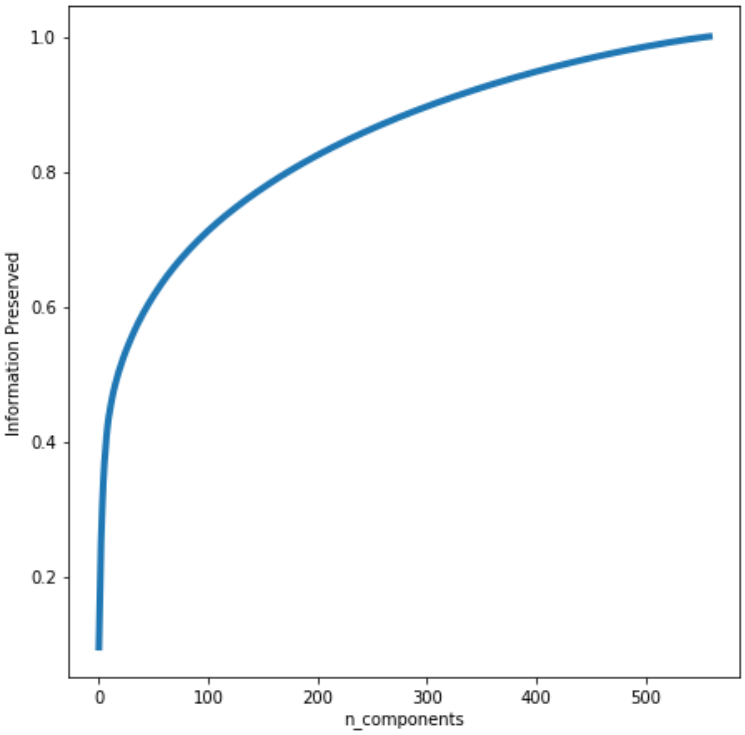


- **RBF (gamma = 0.000005)**

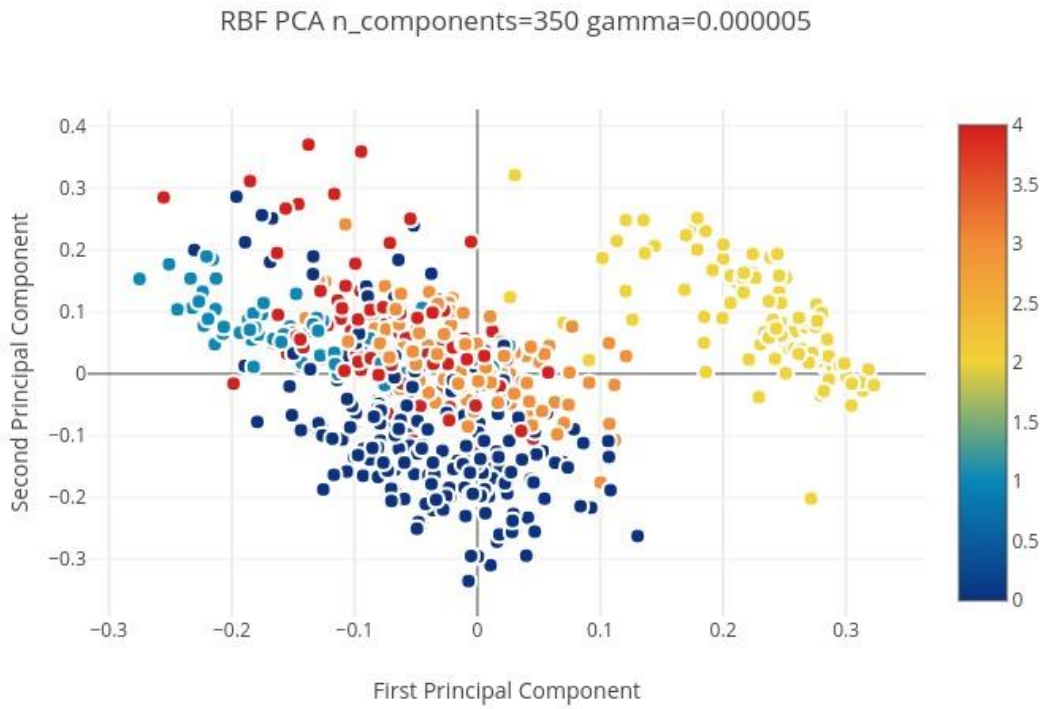
Πληροφορία που διατηρεί το κάθε χαρακτηριστικό



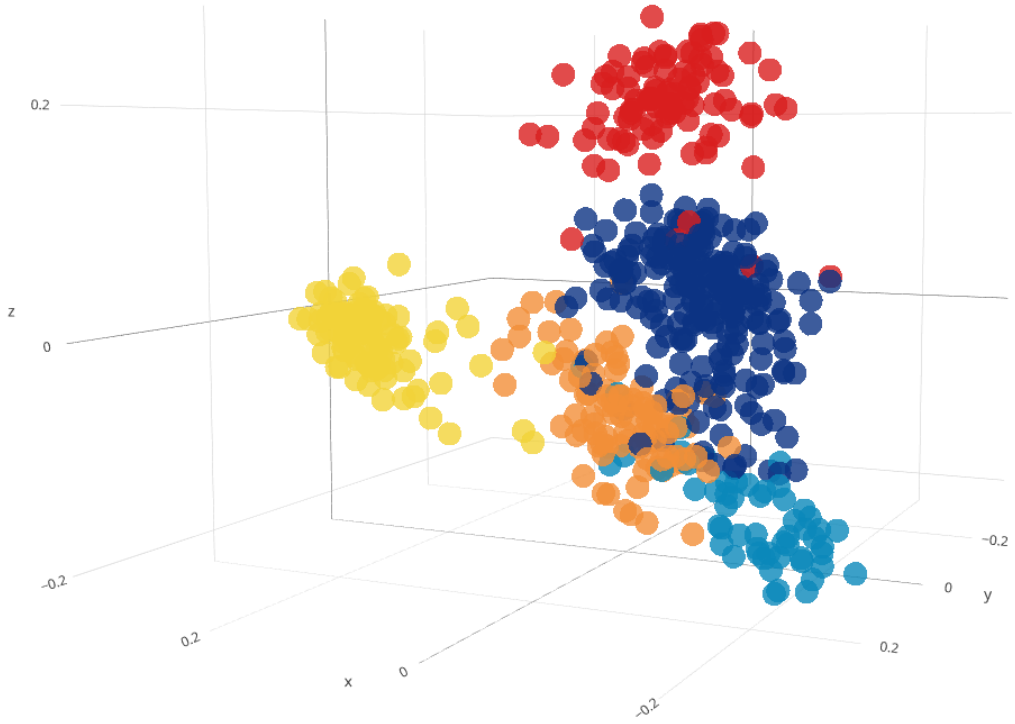
Αθροιστικό διάγραμμα της πληροφορίας



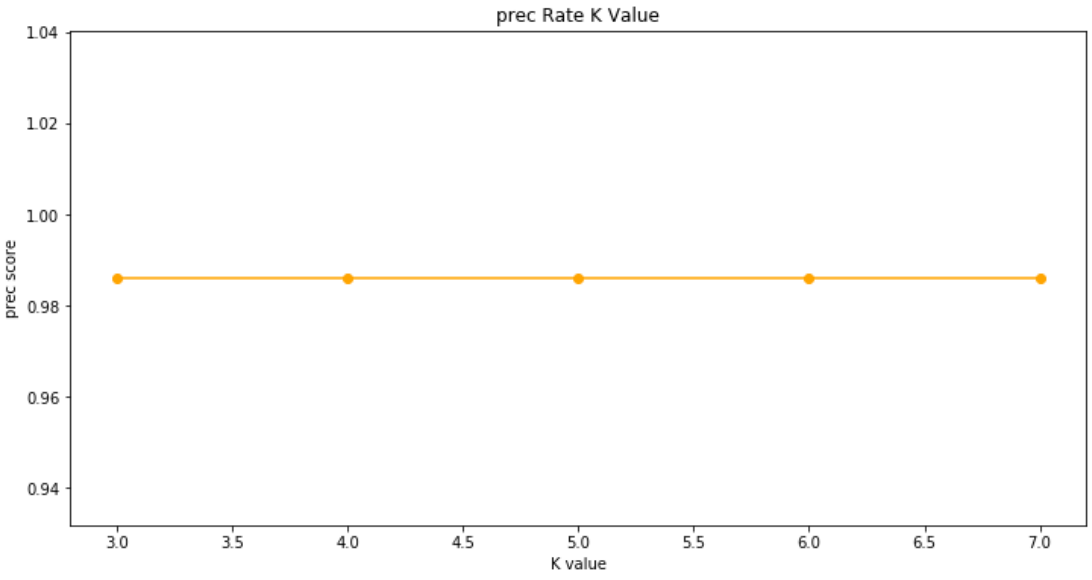
2D Projection of points



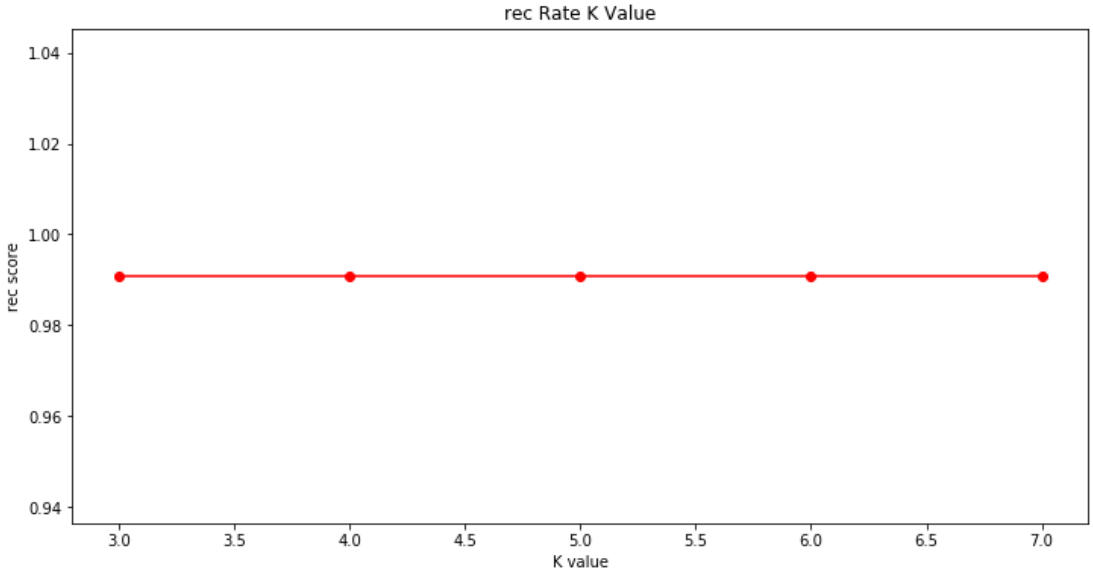
3D Projection of points



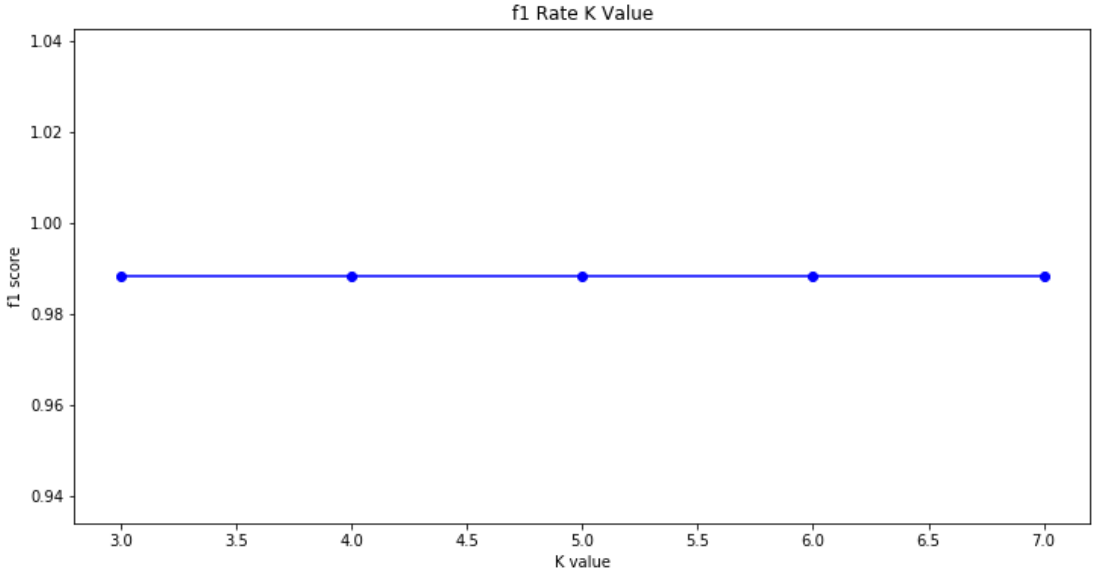
Precision



Recall



F1 Score



Classification Report

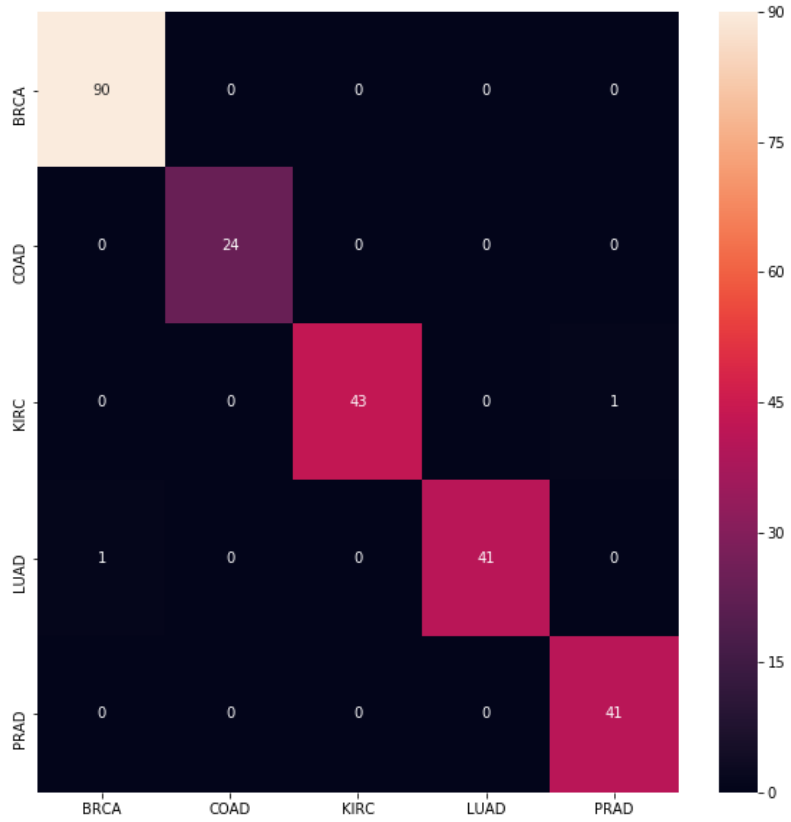
Test

	precision	recall	f1-score	support
0	1.00	0.99	0.99	91
1	1.00	1.00	1.00	24
2	0.98	1.00	0.99	43
3	0.98	1.00	0.99	41
4	1.00	0.98	0.99	42
avg / total	0.99	0.99	0.99	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Confussion Matrix



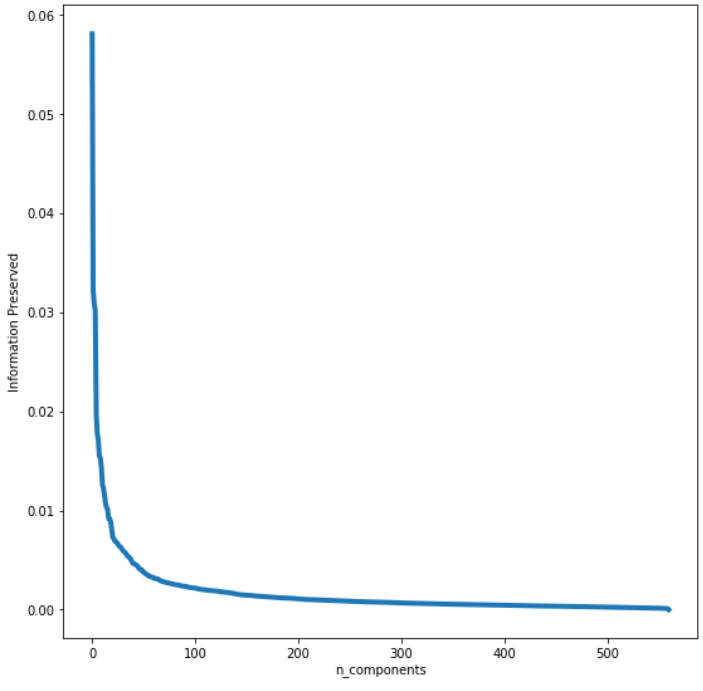
- **Polynomial Kernel PCA**

Στην συνέχεια θα εφαρμόσουμε τον πολυωνυμικό πυρήνα για βαθμούς πολυωνύμου 3 και 4

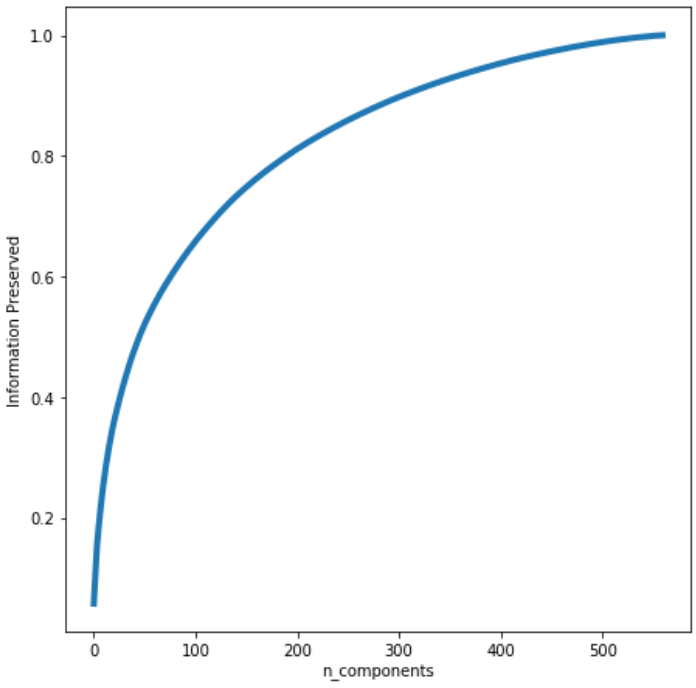
- **Polynomial Kernel PCA degree = 3**

Διατηρήσαμε 350 components στη συγκεκριμένη περίπτωση

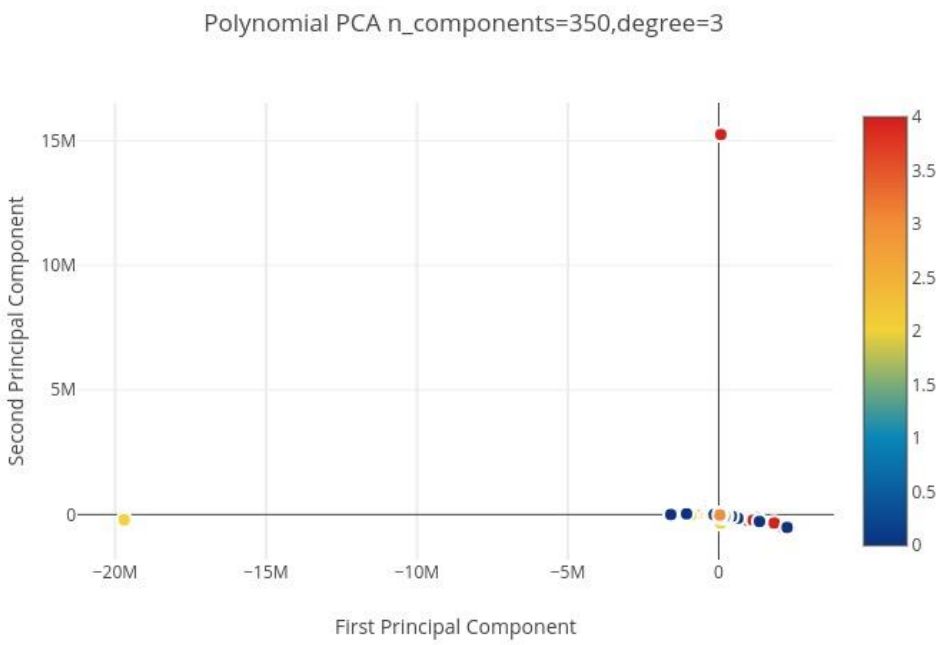
Πληροφορία που διατηρεί το κάθε χαρακτηριστικό



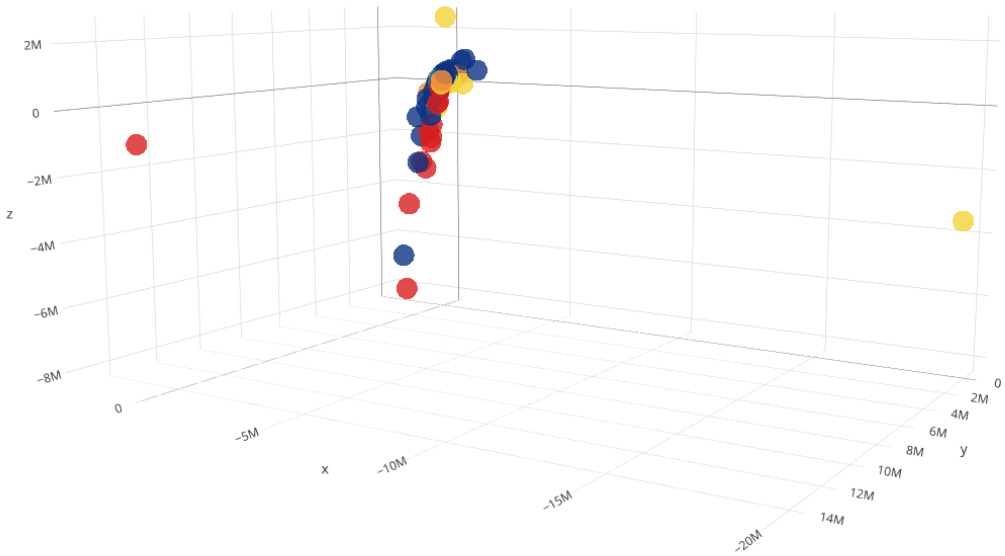
Αθροιστικό διάγραμμα της πληροφορίας



2D projection of points

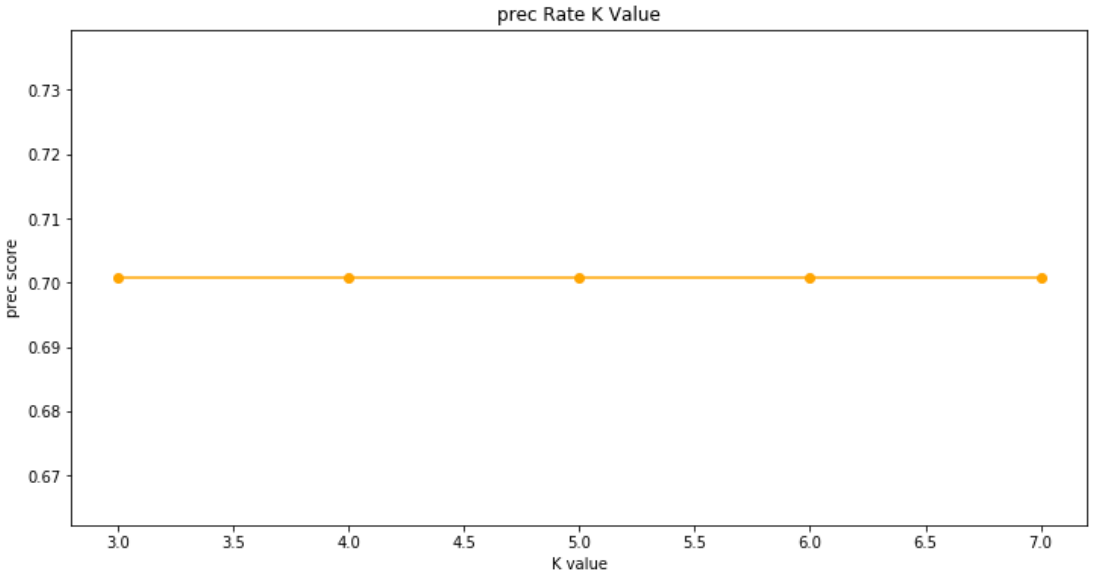


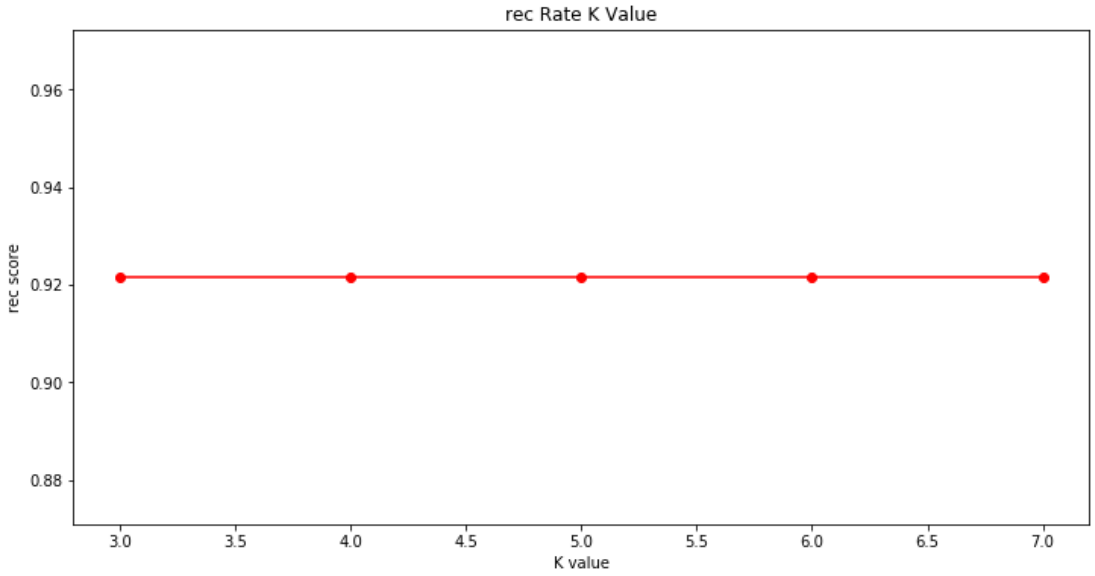
3D projection of points



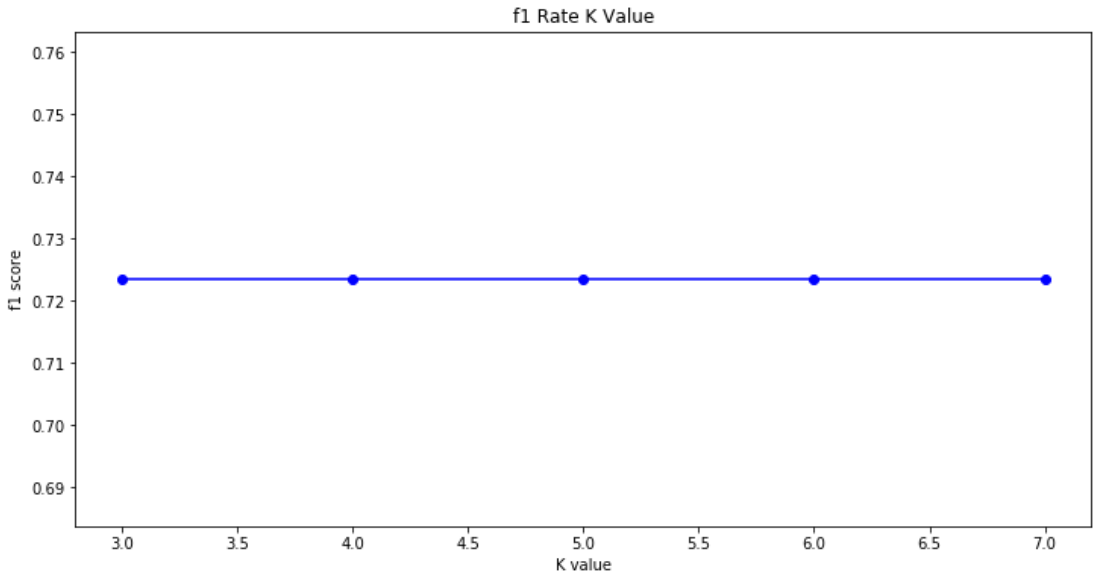
Όπως παρατηρούμε και σε αυτήν την περίπτωση έχει γίνει overfitting στα δεδομένα μας.

Precision





F1 Score



Classification Report

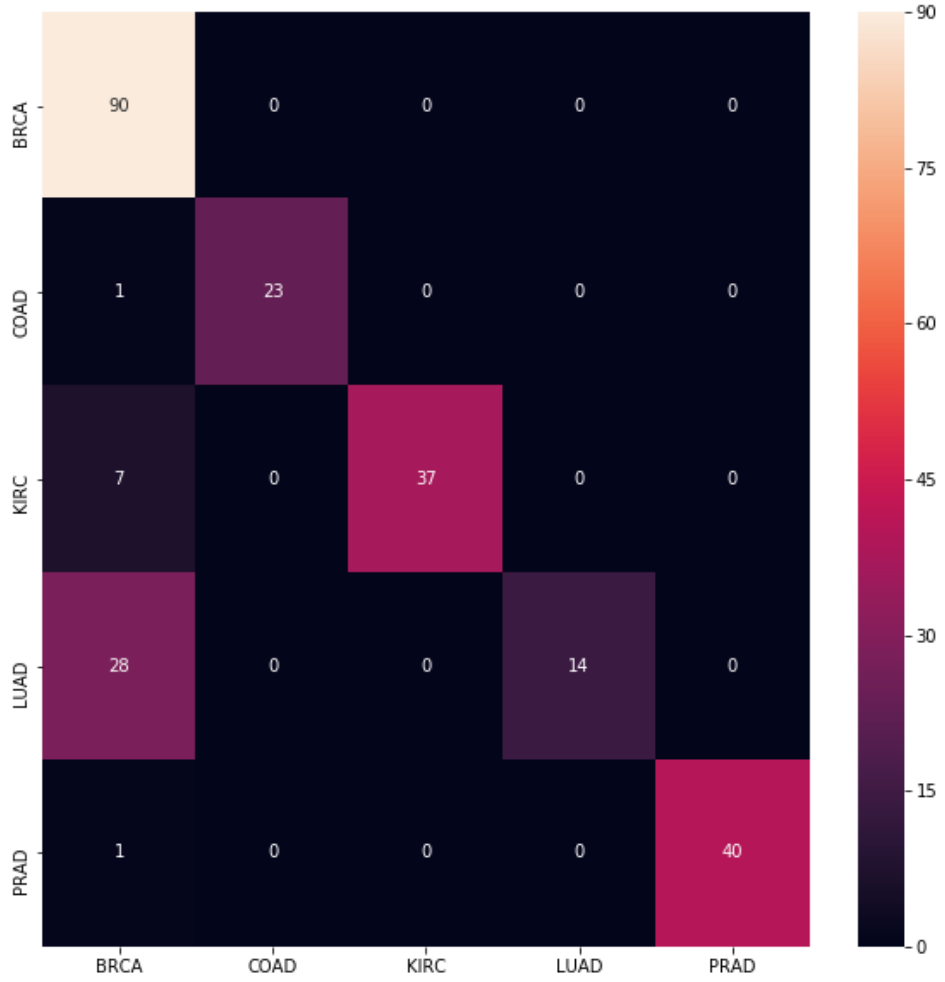
Test

	precision	recall	f1-score	support
0	1.00	0.71	0.83	127
1	0.96	1.00	0.98	23
2	0.84	1.00	0.91	37
3	0.33	1.00	0.50	14
4	0.98	1.00	0.99	40
avg / total	0.93	0.85	0.86	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Confussion Matrix



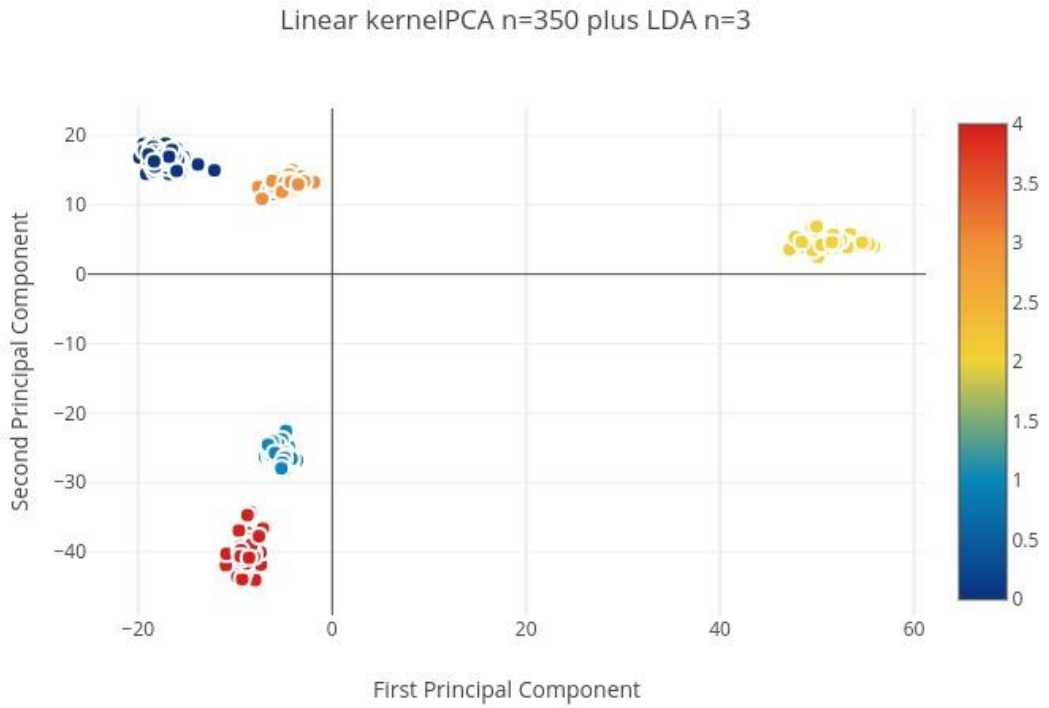
Note!

Όπως σε προηγούμενες περιπτώσεις έτσι και σε αυτή είχαμε overfitting στην εφαρμογή του αλγορίθμου που προκύπτει ξεκάθαρα από το classification report βλέποντας πόσο χαμηλό είναι το precision για την κλάση 3,για αυτό τον λόγο θα παραλλείψουμε να εφαρμόσουμε τον Polynomial KernelPCA για βαθμο πολυωνύμου

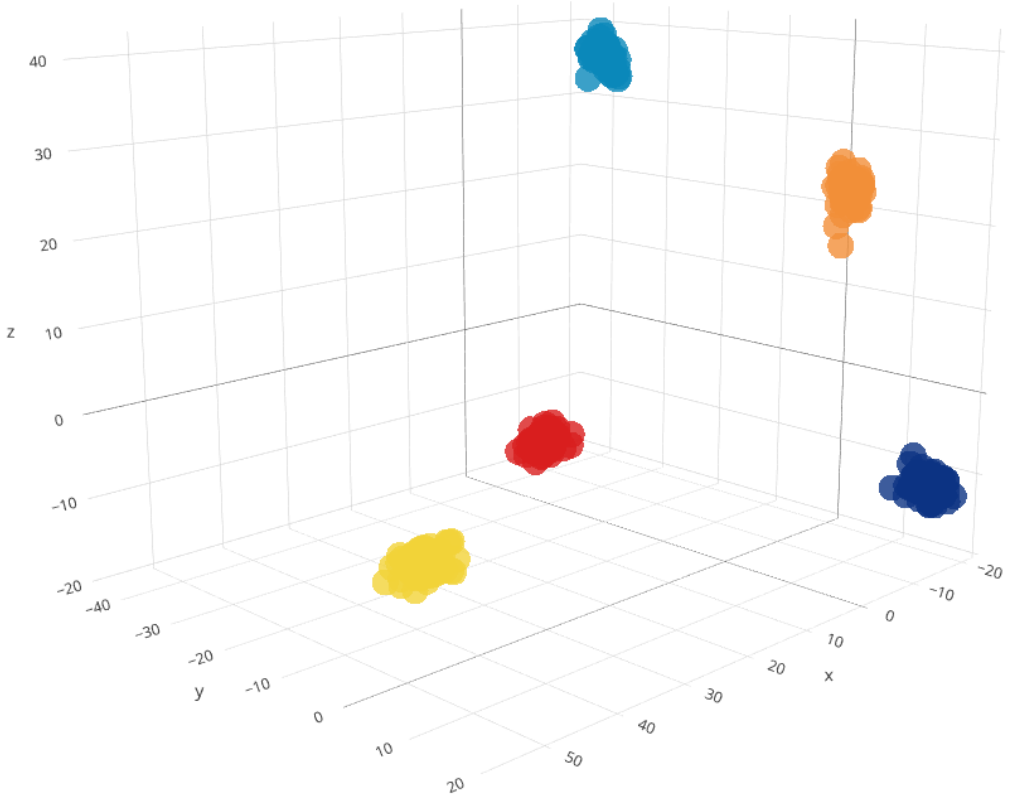
μεγαλύτερο του 3.Θα περάσουμε στην συνέχεια στην εφαρμογή του KernelPCA plus LDA διαδοχικά.

- **Linear KernelPCA (n_components=350) plus LDA (n=3)**

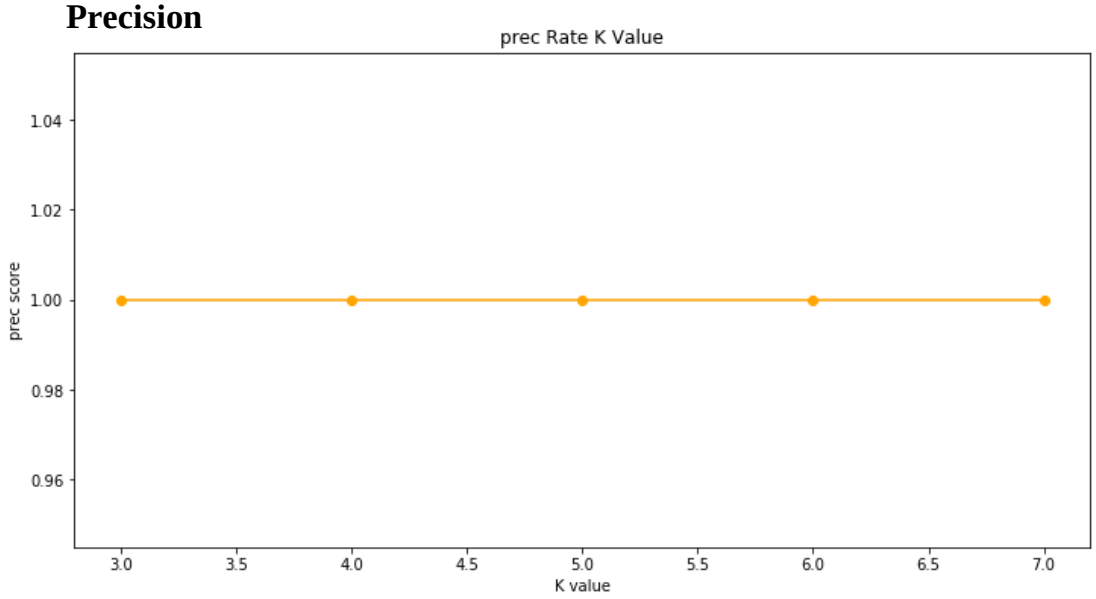
2D projection of points



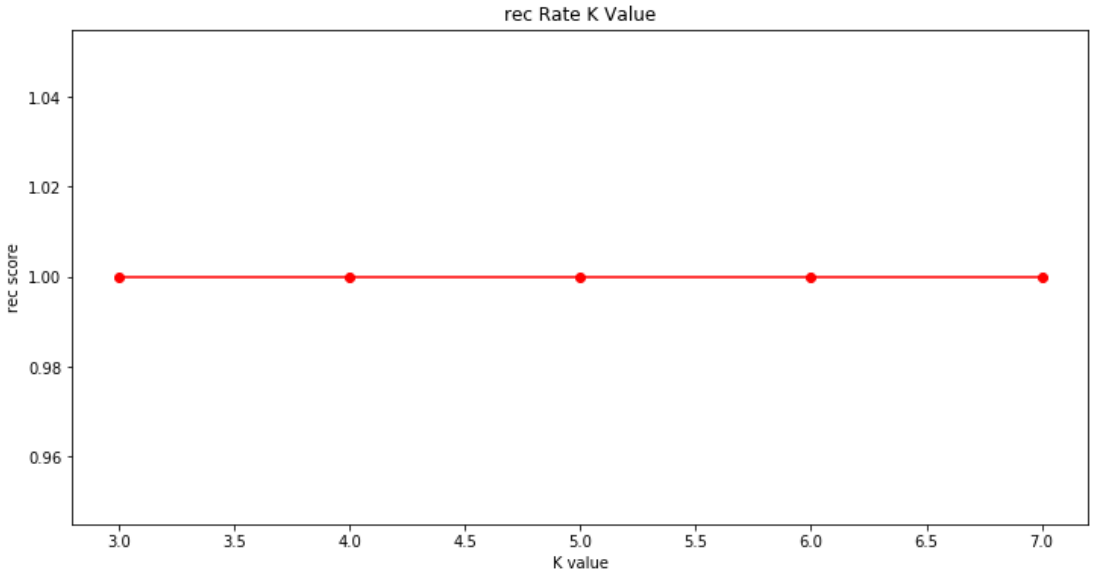
3D projection of points



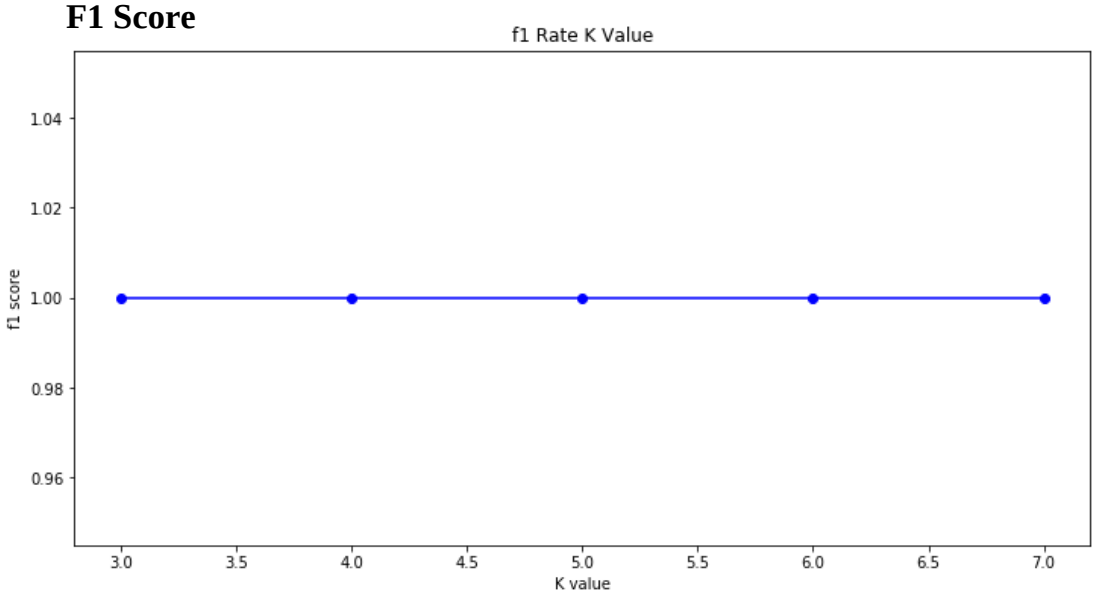
Precision



Recall



F1 Score



Classification Report

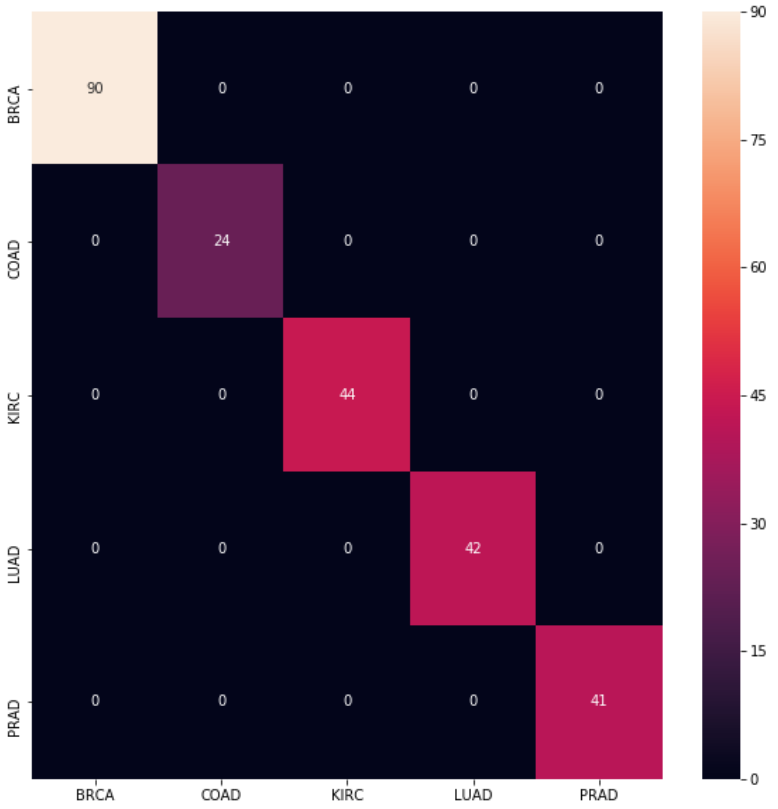
Test

	precision	recall	f1-score	support
0	1.00	1.00	1.00	90
1	1.00	1.00	1.00	24
2	1.00	1.00	1.00	44
3	1.00	1.00	1.00	42
4	1.00	1.00	1.00	41
avg / total	1.00	1.00	1.00	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Confussion Matrix

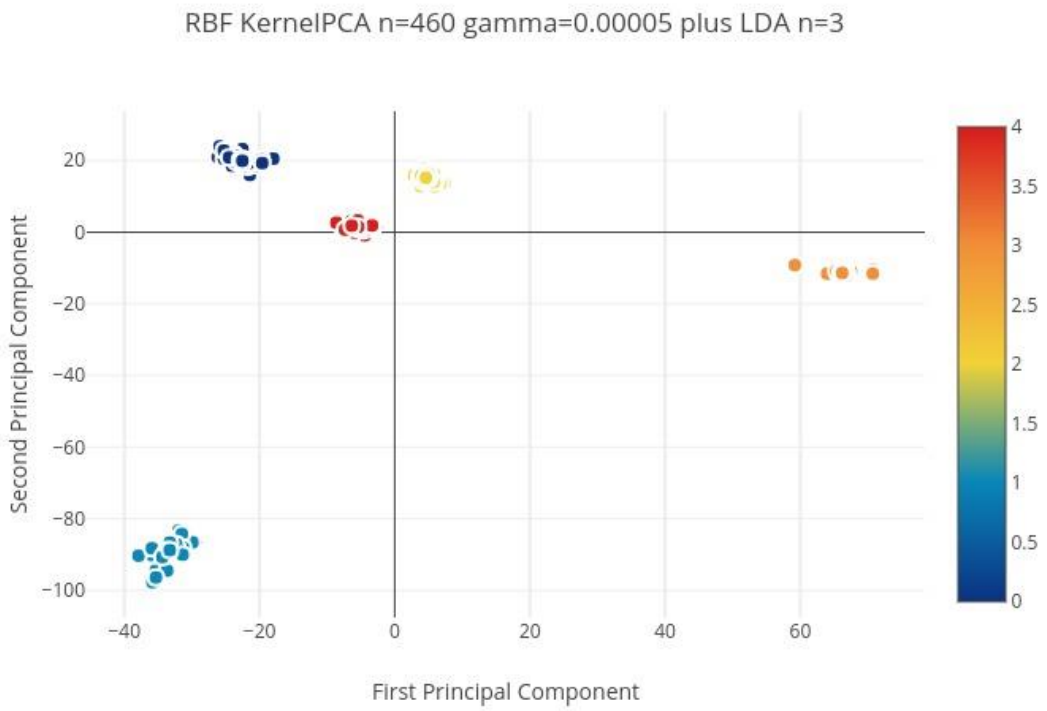


- **Linear KernelPCA (n_components=350) plus LDA (n=4)**

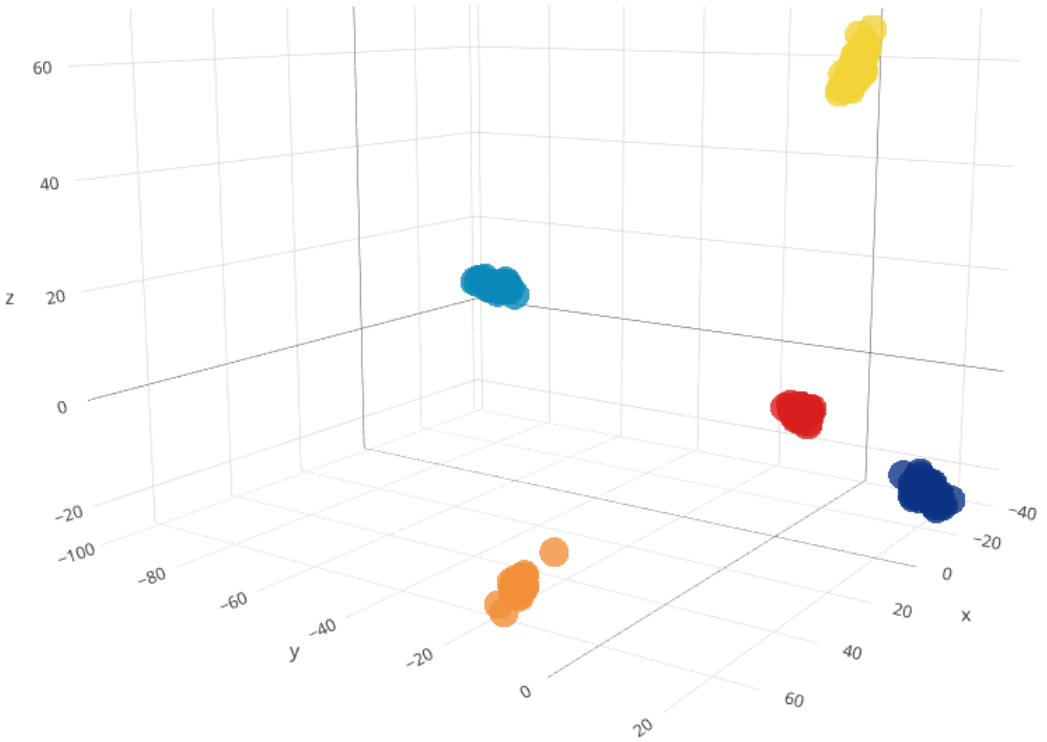
Η προβολή στις 2 και στις 3 διαστάσεις καθώς και τα αποτελέσματα των score προέκυψαν ακριβώς τα ίδια με την περίπτωση όπου n=3 και για αυτό παραλλείπονται οι εικόνες.

- **RBF Kernel PCA (0.00005, 0.000005) plus LDA (n=3)**
- **Gamma value 0.00005**

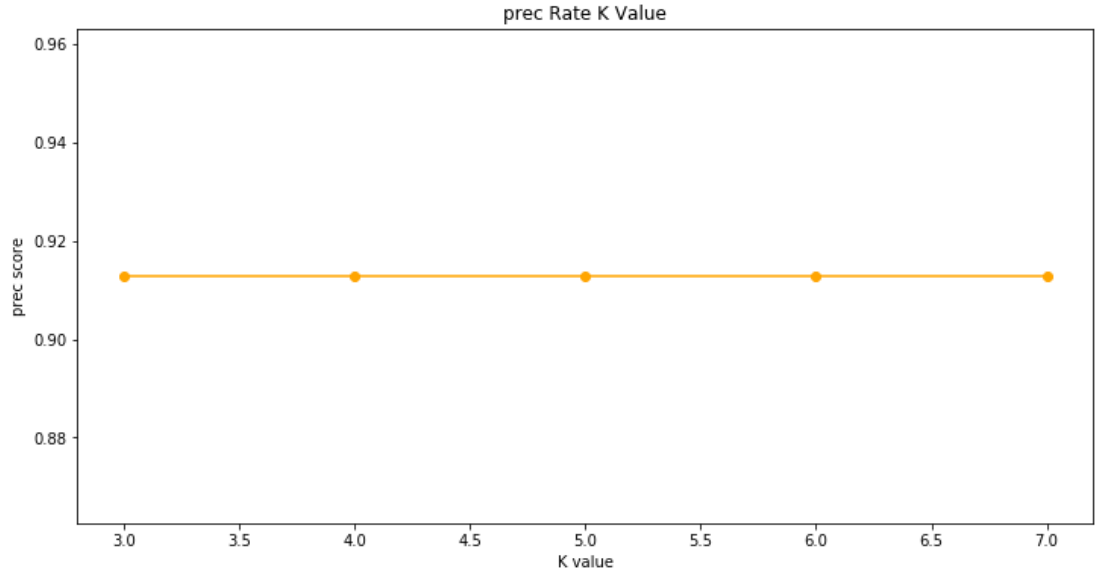
2D projection of points



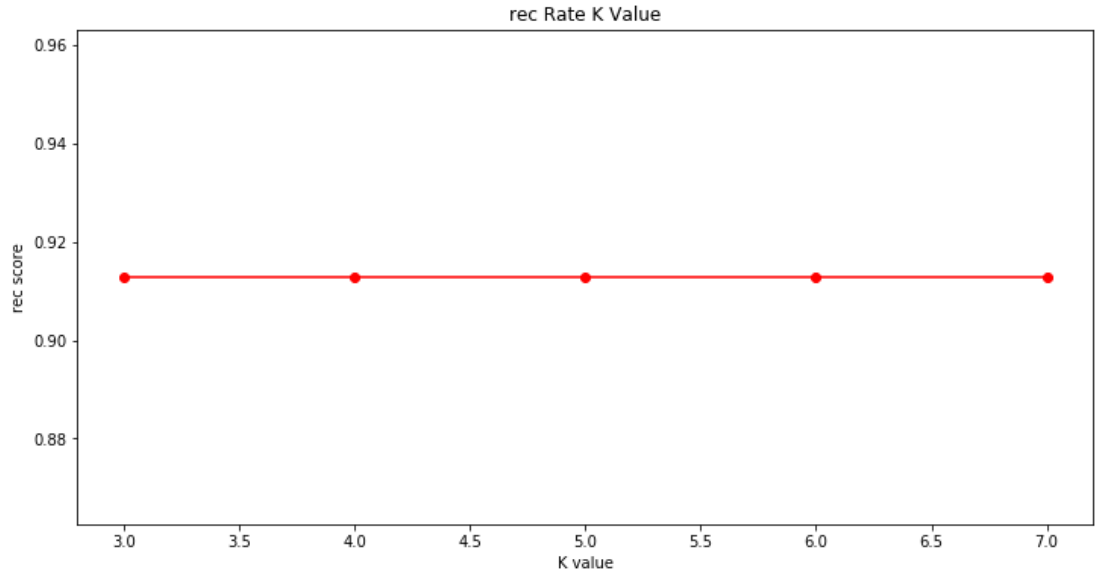
3D projection of points



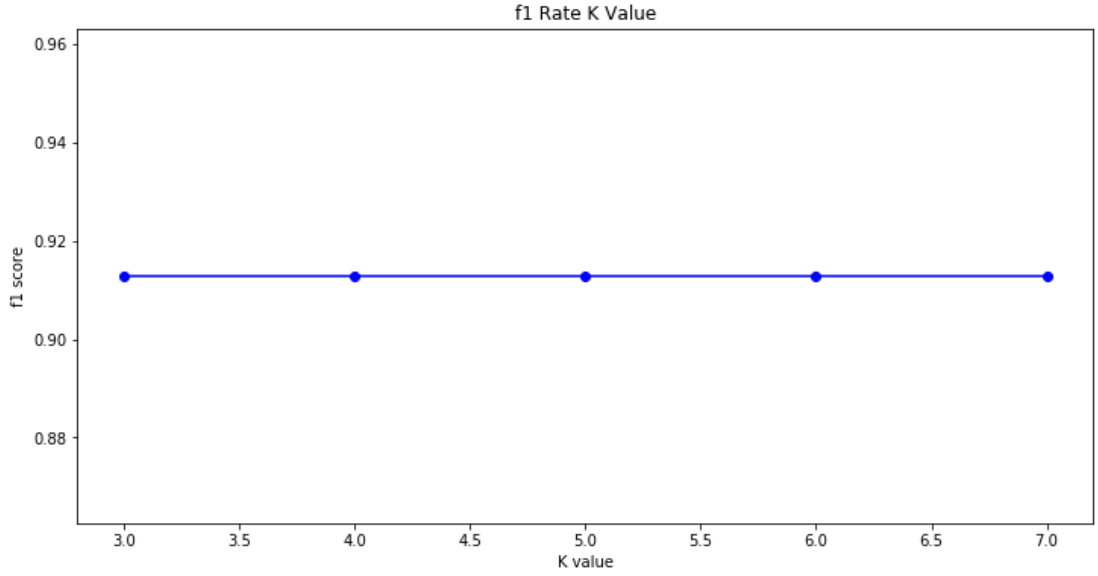
Precision



Recall



F1 Score



Classification report

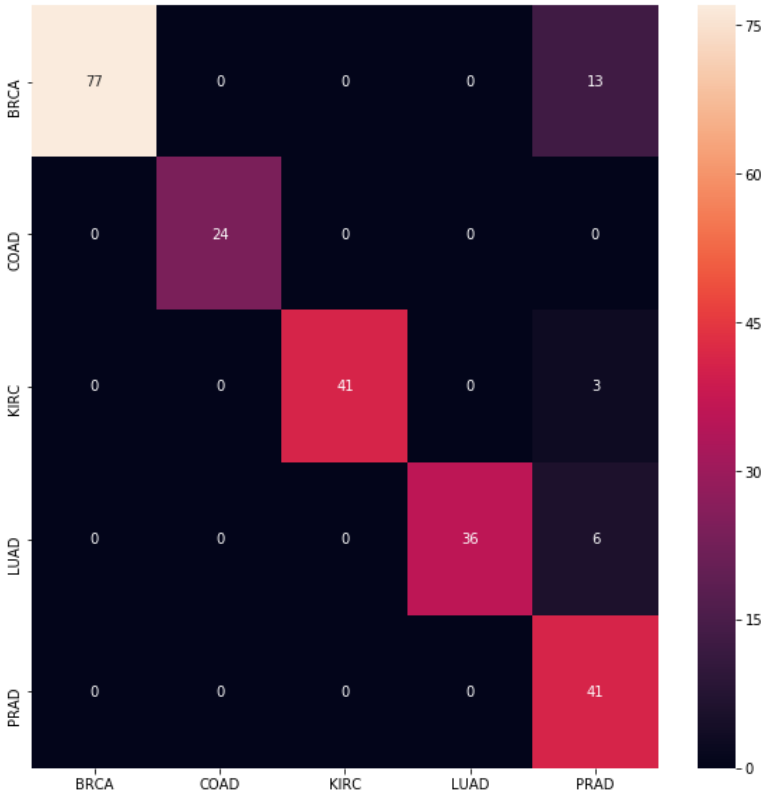
Test

	precision	recall	f1-score	support
0	0.86	1.00	0.92	77
1	1.00	1.00	1.00	24
2	0.93	1.00	0.96	41
3	0.86	1.00	0.92	36
4	1.00	0.65	0.79	63
avg / total	0.92	0.91	0.90	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Confussion Matrix



- **RBF Kernel PCA (0.00005) plus LDA (n=4)**

Classification Report

Test

	precision	recall	f1-score	support
0	1.00	0.97	0.98	93
1	1.00	1.00	1.00	24
2	0.95	1.00	0.98	42
3	0.98	1.00	0.99	41
4	1.00	1.00	1.00	41
avg / total	0.99	0.99	0.99	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

- **RBF Kernel PCA (0.000005) n=350 plus LDA (n=3)**

Test

	precision	recall	f1-score	support
0	1.00	1.00	1.00	90
1	1.00	1.00	1.00	24
2	0.98	1.00	0.99	43
3	1.00	0.98	0.99	43
4	1.00	1.00	1.00	41
avg / total	1.00	1.00	1.00	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

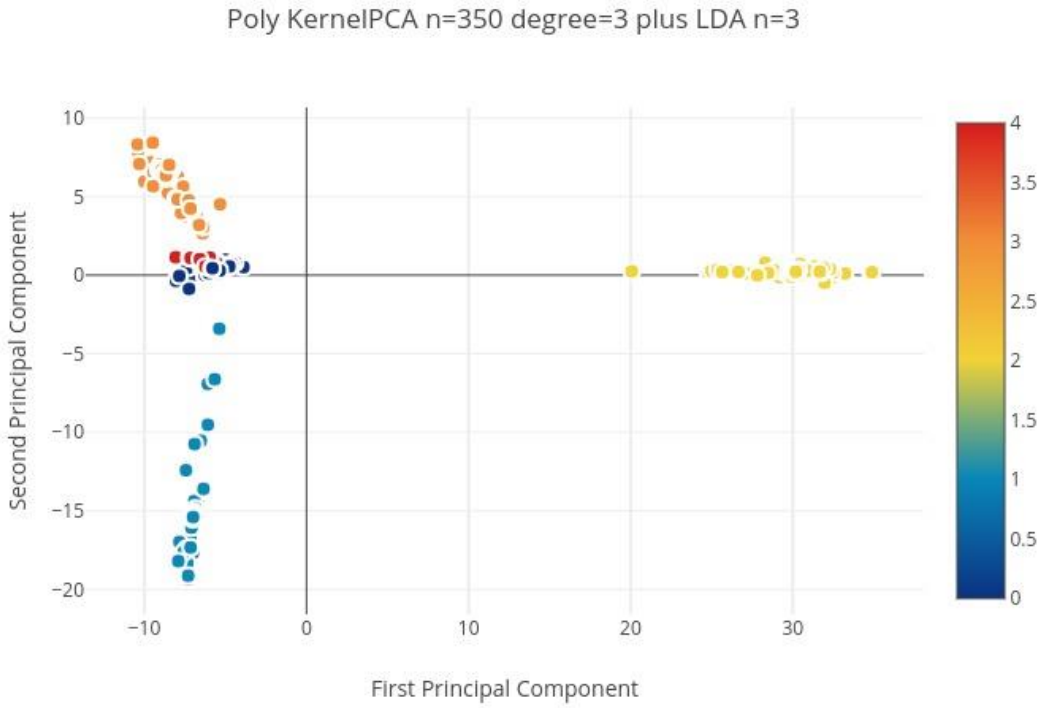
- **RBF Kernel PCA (0.000005) n=350 plus LDA (n=4)**

	precision	recall	f1-score	support
0	1.00	1.00	1.00	90
1	1.00	1.00	1.00	24
2	1.00	1.00	1.00	44
3	1.00	1.00	1.00	42
4	1.00	1.00	1.00	41
avg / total	1.00	1.00	1.00	241

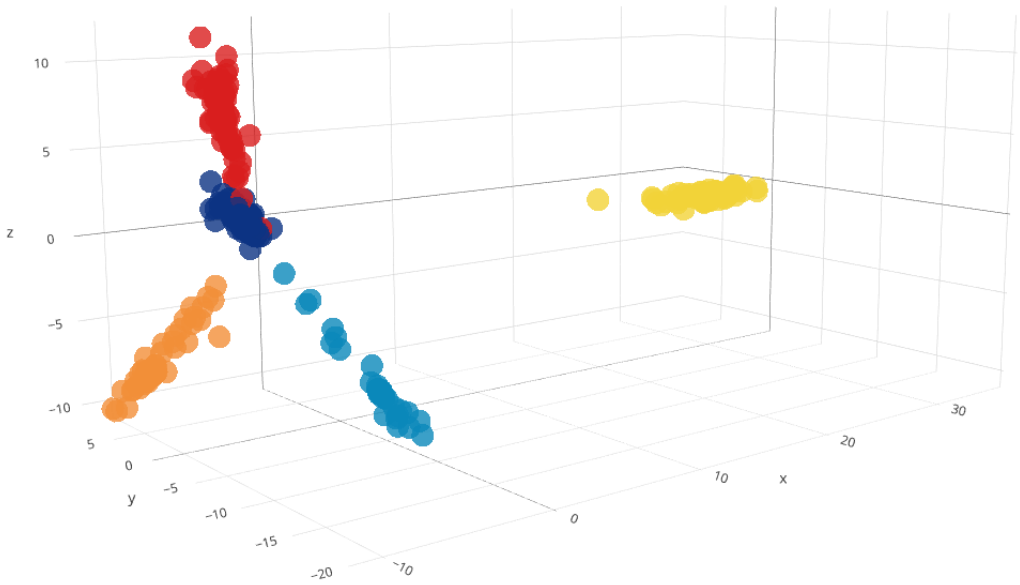
	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

- **Polynomial KernelPCA (degree = 3) plus LDA (n=3)**

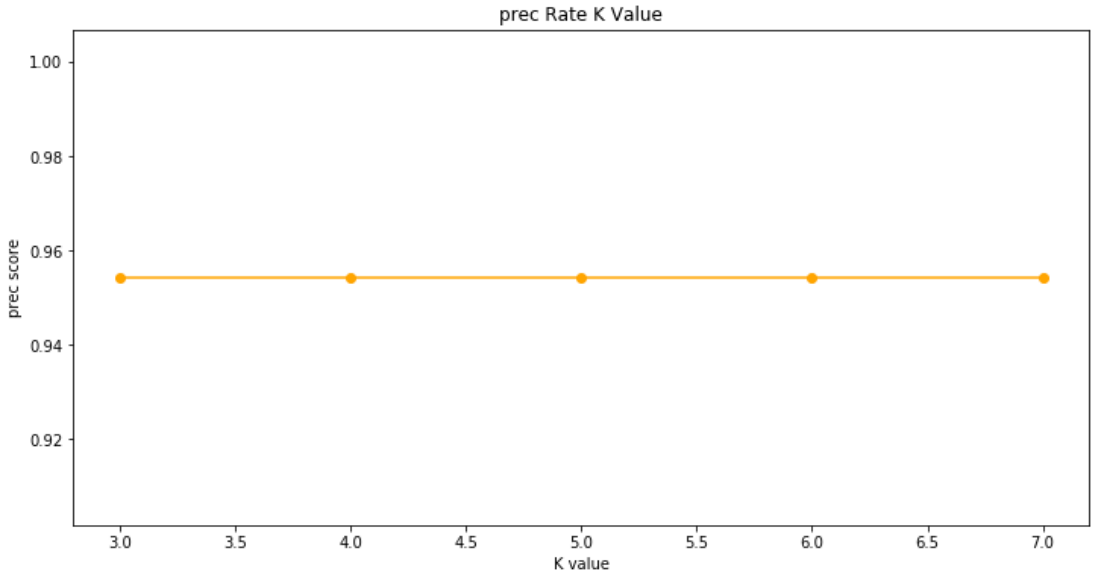
2D projection of points



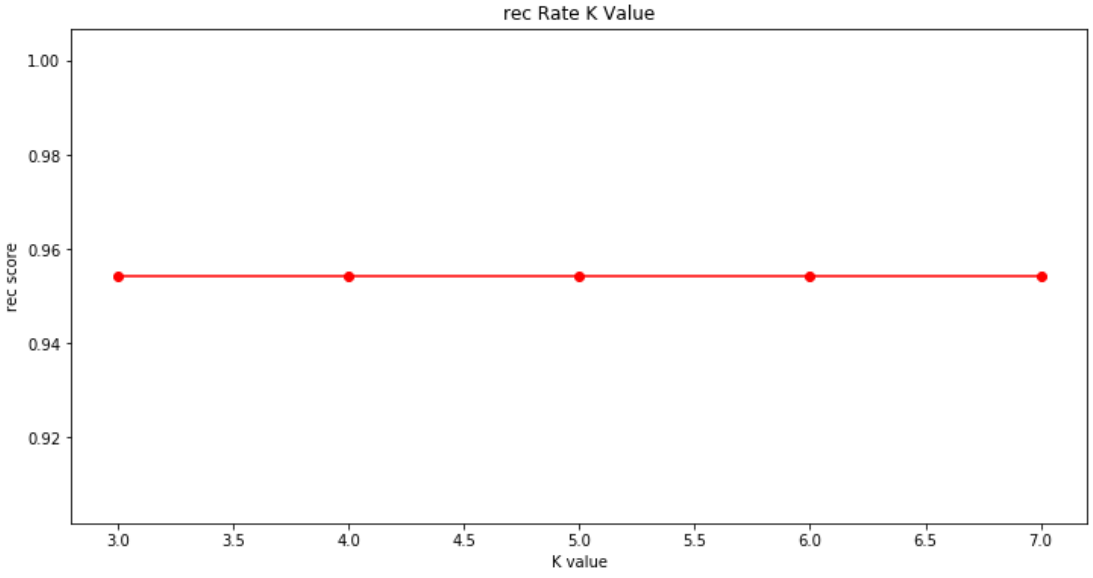
3D projection of points



Precision



Recall



Classification Report

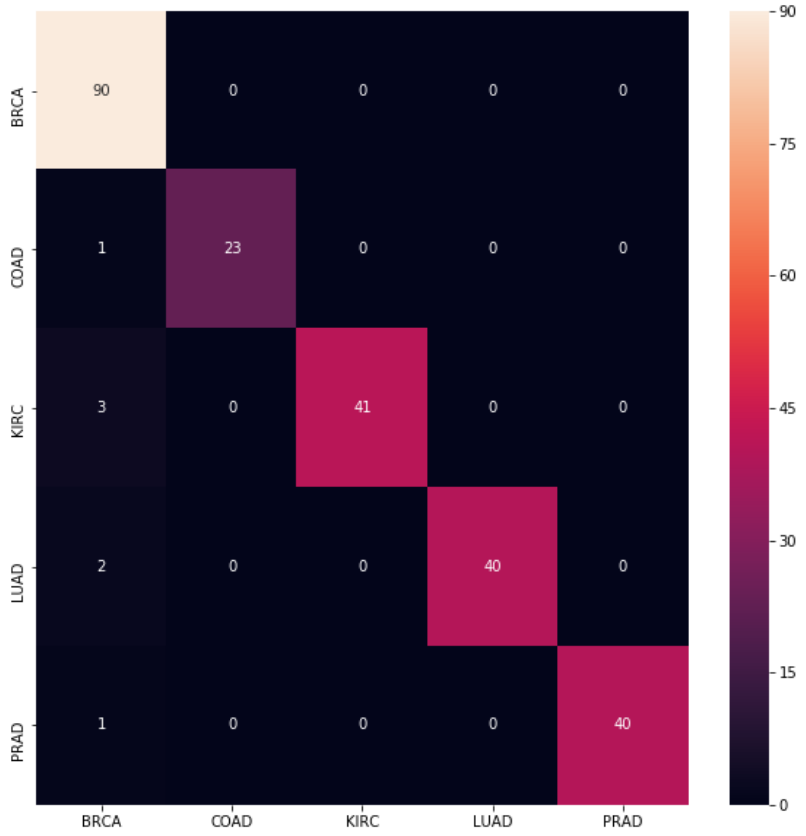
Test

	precision	recall	f1-score	support
0	1.00	0.93	0.96	97
1	0.96	1.00	0.98	23
2	0.93	1.00	0.96	41
3	0.95	1.00	0.98	40
4	0.98	1.00	0.99	40
avg / total	0.97	0.97	0.97	241

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Confussion Matrix



Στην συνέχεια θα προχωρήσουμε στην εφαρμογή του **SVM** αλγορίθμου προκειμένου να συγκρίνουμε τα αποτελέσματα με τον αλγόριθμο **KNearestNeighbors**

Έγινε εφαρμογή μετα από Linear KPCA και στην συνέχεια Grid Search για εύρεση καλύτερων παραμέτρων.

Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	210
1	1.00	1.00	1.00	54
2	1.00	1.00	1.00	102
3	1.00	1.00	1.00	99
4	1.00	1.00	1.00	95
avg / total	1.00	1.00	1.00	560

Test

	precision	recall	f1-score	support
0	0.99	1.00	0.99	90
1	1.00	1.00	1.00	24
2	1.00	1.00	1.00	44
3	1.00	0.98	0.99	42
4	1.00	1.00	1.00	41
avg / total	1.00	1.00	1.00	241

Τα καλύτερα λοιπόν αποτελέσματα που προέκυψαν συγκεντρωτικά ήταν τα παρακάτω

- **Linear KPCA (n_components=350) plus LDA (n=3)**
  Precision = 100%
  Recall = 100%
  F1-Score = 100%

- **RBF Kernel PCA (gamma =0.000005) n=350 plus LDA (n=3)**
  Precision = 100%
  Recall = 100%
  F1-Score = 100%

- **Polynomial KPCA (degree=3 n=350) plus LDA(n=3)**
  Precision = 100%
  Recall = 100%
  F1-Score = 100%

- **Linear KPCA (n_components=350) plus SVM (Polynomial , degree = 3 ,C=1)**
  Precision = 100%
  Recall = 100%
  F1-Score = 100%