# *Μηχανική Μάθηση*

**Εργασία 2: Clustering problems**

Ομάδα:
Φλάμος Κάρολος dai19200
Βασιλειάδης Παντελεήμων dai19216

15/01/2023

Ομαδοποιητής Πουλιών

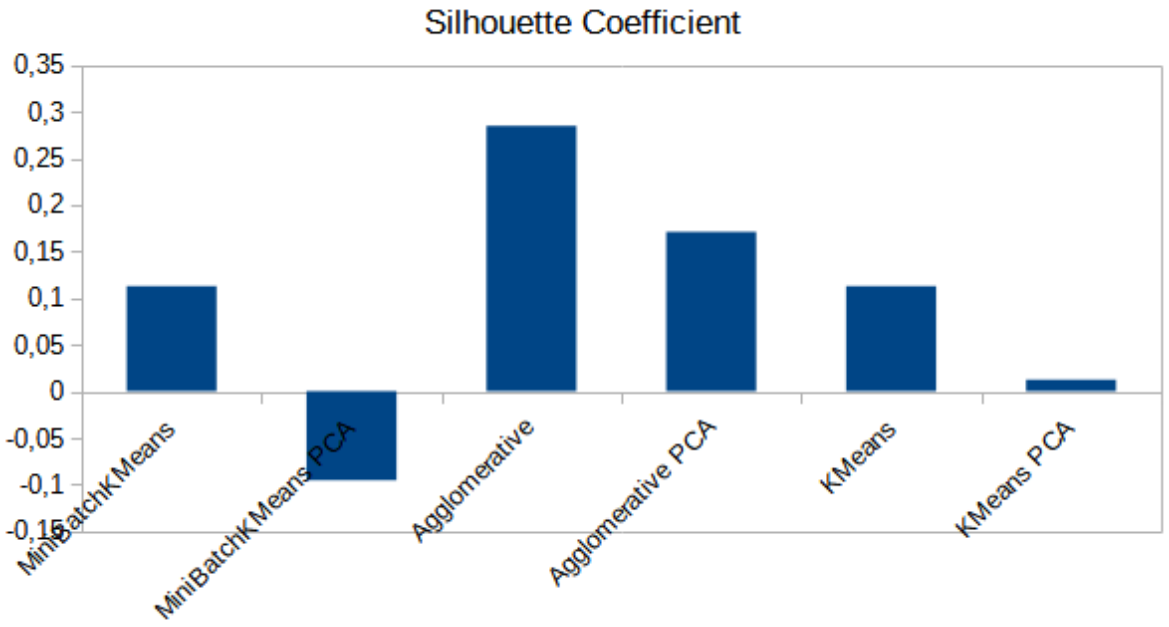
Περιεχόμενα

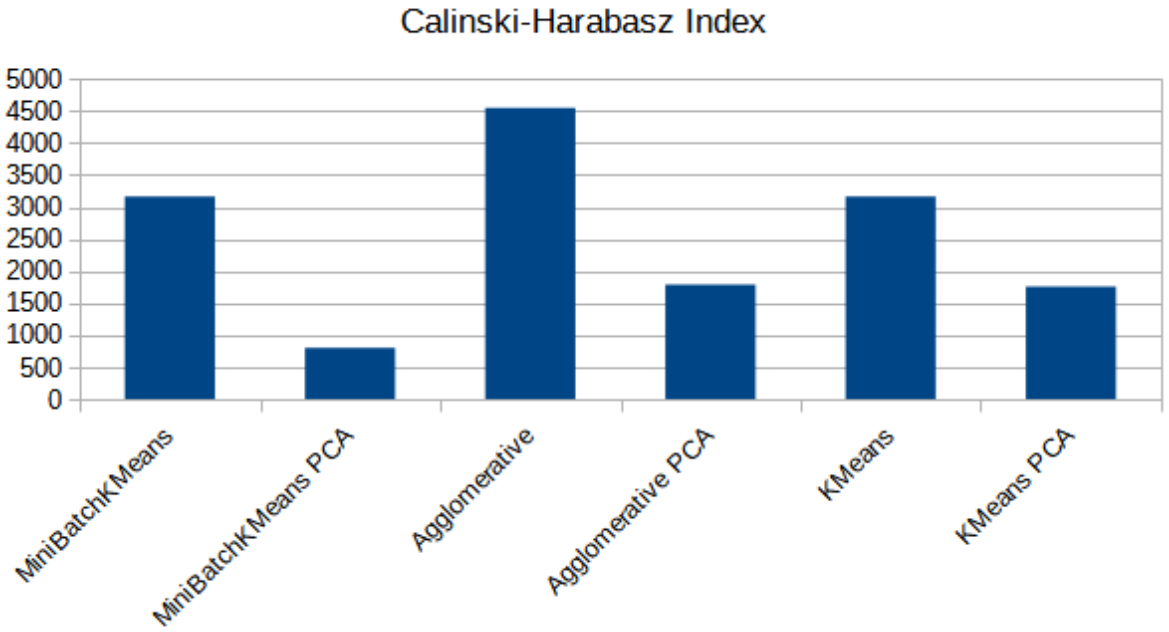
# Εισαγωγή

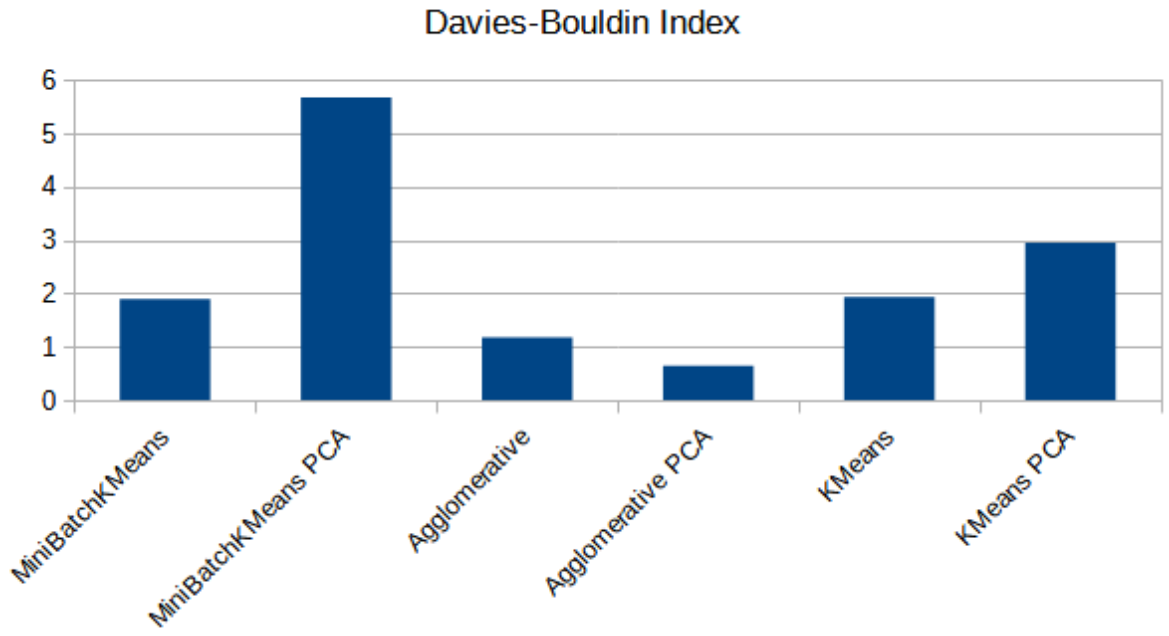
Οι βασικοί στόχοι της εργασίας είναι η ομαδοποίηση(clustering) ενός συνόλου εγγραφών απο το dataset fashion-mnist. Αρχικά πρέπει να βρούμε έναν τρόπο να μεταφέρουμε τα δεδομένα, στον υπολογιστή του collab. Για να δουλέψουμε σωστά τα δεδομένα και να εκπαιδεύσουμε σωστά τα μοντέλα μας, τα χωρίζουμε σε training, validation and test set.
Το dataset αυτο περιέχει δεδομένα υψηλότερων διαστάσεων, και γι'αυτο θα προσπαθήσουμε να τις μειώσουμε με την χρήση του Principal component analysis(PCA). Με το PCA θα αναλυθούν οι διαστάσεις σε συνιστώσες, και σε όσα μεγέθη(στήλες) υπάρχει μεγάλη συσχέτιση, θα συγχωνευτούν. Στόχος είναι να μειώσουμε τις διαστάσεις με τρόπο τέτοιο ώστε να μην χάνεται πάνω από 15% της πληροφορίας. Παράλληλα θα πρέπει η συσταδοποίηση να γίνεται επαρκώς καλά. Για την μέτρηση της επάρκειας του PCA καλούμαστε να χρησιμοποιήσουμε κάποιες μετρικές απόδοσης των αλγοριθμων/μοντέλων συσταδοποίησης.
Συγκρίνοντας τις μετρικές πρίν και μετά την μείωση διαστάσεων, θα ξέρουμε εάν έχει γίνει σωστά η διαδικασία.

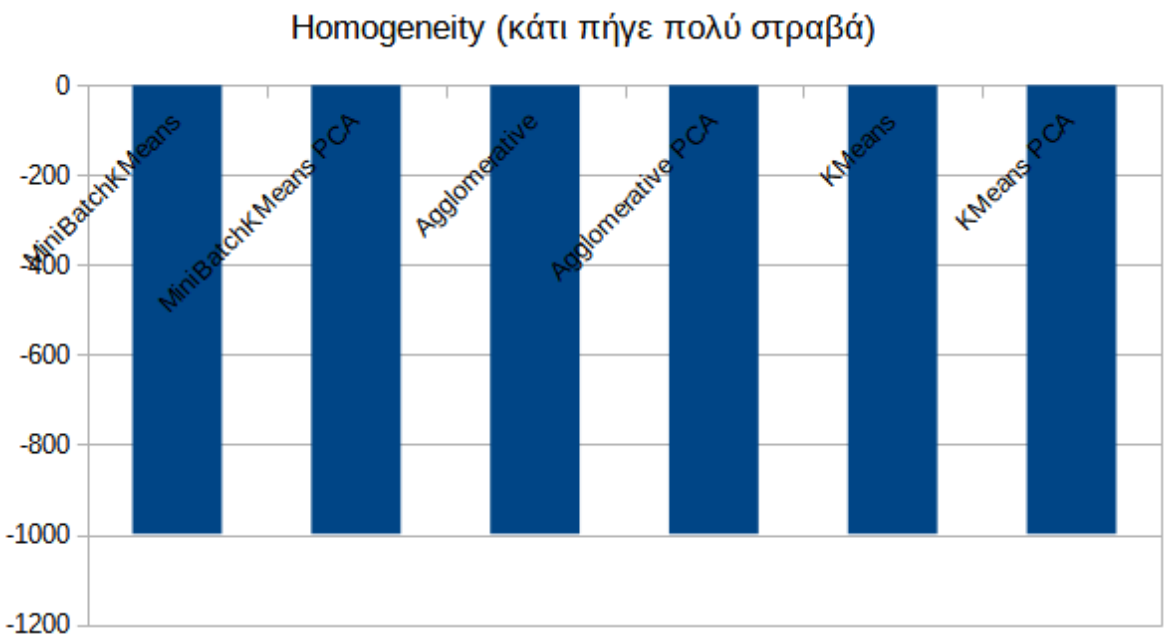
# Μέθοδοι που εφαρμόστηκαν

Εισάγαμε το dataset fashion-mnist στον σερβερ μας για να μπορούμε να χειριστούμε τα δεδομένα. Τα χωρίσαμε σε training, validation και test sets. Κρατήσαμε και τις πραγματικές κλασεις των εγγραφών για supervised learning.
Οι διαστάσεις του σετ είναι σε πολύ υψηλό βαθμο, για να επεξεργαστούμε δεδομένα πανω σε αυτον. Θα μειώσουμε τις διαστάσεις με Principal component Analysis. Αναλύουμε σε συνιστώσες τα μεγέθη, και έτσι μπορούμε να δούμε αν καποια απο αυτα είναι θετικά συσχετισμένα. Εάν δούμε οτι συνδιακυμαίνονται, μπορούμε να συγχωνεύσουμε τα μεγέθη σε ένα. Υλοποιήσαμε τον μετασχηματισμό με τις βιβλιοθήκες της python κρατώντας το 85% της πληροφορίας. Παίρνουμε δειγματολειπτικά μερικές ενδείξεις απο απο τις εγγραφές του training set και όλα φαινεται να βαίνουν καλώς. Ο μετασχηματισμός εχει γίνει, και απο την ποιότητα των εικόνων φαίνεται να έχει χαθεί λιγη ποιότητα, αλλα είναι ακομη ευδιακριτο το ποια αντικείμενα απεικονίζονται. Τον ίδιο μετασχηματισμό έχουμε κάνει και στο validation set. Απο το δείγμα που πήραμε δίνεται η διαίσθηση οτι όλα βαίνουν καλώς. Έχουμε εφαρμόσει και την αντίστροφη διαδικασία/συνάρτηση στα τροποποιημένα δεδομένα, ωστε να τα επαναφέρουμε στον αρχικό χωρο. Η επιτυχής επαναφορά μας δειχνει οτι δεν είχε χαθεί σημαντικο κομματι της πληροφορίας.Οι εκτυπώσεις των εικόνων το αποδεικνύουν και με το ματι.
Στη συνέχεια εφαρμώσαμε τον ιδιο μετασχηματισμό και στο τεστ σετ. Κάνοντας οπτικοποίηση φαίνεται πως η μετατροπή εχει γίνει και εδω, αλλα με λιγο χειρότερα αποτελέσματα απο πριν.
Μετά φτιάξαμε τα μοντέλα ομαδοποίησης (clustering). Επιλέξαμε τα: kmeans, min batch kmeans και hierarchical clustering, με υλοποιηση απο την sklearn. Εκπαιδεύσαμε το μοντέλο πάνω στο traning set.
Συγκεκριμενα:
min batch Kmeans: χρησιμοποιήσαμε 3 αρχικά clusters(κ) για την εκπαίδευση του μοντέλου.
Kmeans : και πάλι το ίδιο(κ=3)
hierarchical : του επιτρέψαμε να φτάσει στο μέγιστο επίπεδο διαχωρισμού
Στην συνέχεια προσπαθήσαμε να επιλέξουμε τις κατάλληλες παραμέτρους για την εκκίνηση των αλγορίθμων. Η επιλογή αυτή εγινε με εκπαίδευση πάνω στο validation set. Πήραμε διαδοχικές τιμές εκκινησης, και δοκιμάσαμε μέχρι να δωθεί το καλύτερο σκορ(τα σκορ αναφαίρονται αναλυτικά στο τέλος) που μπορουσαμε να βρούμε. Τελος εκπαίδευσης του Μοντέλου. Τέλος δοκιμάσαμε τα μοντέλα πάνω στο τεστ σετ. Τα αποτελέσματα που λάβαμε μετρήθηκαν ως προς την ποιότητα με τις μετρικές που φαίνονται παρακάτω:
-Silhouette Coefficient: Δείχνει αν τα σημεία είναι κοντά με τα στοιχεία του cluster τους και ταυτόχρονα μακριά απο τα αλλα
`-Calinski-Harabasz Index` : Δίχνει το πόσο καλά εχουν διααχωριστεί τα clusters με βάση τις ιδιότητες των συνλικών εγγραφών. Δδεδομένου οτι χρησιμοποιήσαμε unsupervised learning, θεωρουμε οτι είναι πολύ βοηθητικό κριτήριο
`-Davies-Bouldin Index`: άλλη μία μέτρηση βασισμένη σε εσωτερικά χαρακτηριστικά. Ιδανική για Unsupervised learning.
Παρατίθενται οι εικόνες των γραφημάτων για τις αποδόσεις των μετρικών.


Silhouette Coefficient
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
-0,05
-0,1
-0,15
MiniBatchKMeans
MiniBatchKMeans PCA
Agglomerative
Agglomerative PCA
KMeans
KMeans PCA


Calinski-Harabasz Index
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
MiniBatchKMeans
MiniBatchKMeans PCA
Agglomerative
Agglomerative PCA
KMeans
KMeans PCA


Davies-Bouldin Index
6
5
4
3
2
1
0
MiniBatchKMeans
MiniBatchKMeans PCA
Agglomerative
Agglomerative PCA
KMeans
KMeans PCA


Homogeneity (κάτι πήγε πολύ στραβά)
0
-200
-400
-600
-800
-1000
-1200
MiniBatchKMeans
MiniBatchKMeans PCA
Agglomerative
Agglomerative PCA
KMeans
KMeans PCA

# Συμπεράσματα-αποτιμήσεις

Ως προς το Silhouette Coefficient, τα αποτελέσματα είναι μακράν καλύτερα πρίν την μετατροπή PCA. Το ιεραρχικο φαίνεται να αντιστέκεται περισσότερο στην πτώση αποδόσεων μετα το PCA μετατροπη. Γενικά τα silhouette scores είναι πολύ κακά, με εξαίρεση τον agglomerative.
Το `Calinski-Harabasz Index` επιβεβαιώνει αυτα τα αποτελέσματα. Είναι φανερό οτι τα αποτελέσματα είναι πεσμένα μετά το PCA και αυτό σημαίνει πεσμένη αποδωση.
Στο `Davies-Bouldin` τα νούμερα είναι πιο ανεβασμένα μετά το PCA , ωστόσο για την συγκεκριμένη μετρική αυτο δεν συνεπάγεται καλύτερες αποδόσεις.
Ο δείκτης που χρησιμοποιήσαμε μονοι μας ηταν της ομοιογένειας, αλλά δεν πήγε καλά, οποτε τον αποκλείουμε απο την εξαγωγή συμπερασμάτων ούτε μιλάμε για αυτον.

Ο καλύτερος αλγόριθμος είναι ξεκαθαρα του μοντέλου agglomerative clustering. Δεν παίρνουμε τις ίδιες μετρικές. Θα λεγαμε οτι οι αποδόσεις είναι πεσμένες.

Ο καλύτερος συνδιασμός είναι agglomerative με ανεπεξέργαστα δεδομένα.
Οι μετρικές του είναι:

```
.. Silhouette Coefficient score is 0.28
.. Calinski-Harabasz Index score is 4552.63
.. Davies-Bouldin Index score is 1.18
```

Τα αποτελέσματα θεωρούνται αποδεκτά, αφού δηλώνουν οτι τα cluster είναι διαχωρισμένα αρκετά. Δεν είναι καλά διαχωρισμένα αλλα είναι αποδεκτά. Μπορούμε να είμαστε αρκετά σίγουροι οτι το καθε σημείο ανήκει (στατιστικά) στο cluster του. Επίσης δεν επικαλύπτωνται τα clusters. Θα μπορούσαμε ενδεχομένως να κρατήσουμε μεγαλύτερο ποσοστο πληροφορίας κατα τον PCA. Έτσι ισως να ηταν αποδεκτά τα αποτελέσματα και στους άλλους αλγοριθμους.