Π.Μ.Σ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΠΤΥΞΗ ΛΟΓΙΣΜΙΚΟΥ & ΝΕΦΟΣ

Μέθοδοι και Εργαλεία Τεχνητής Νοημοσύνης (SD0203)

Εργασία 3 – Μάθηση χωρίς επίβλεψη – Συσταδοποίηση

Ονοματεπώνυμο : Χρυσοχοΐδης Αναστάσιος
Αριθμός Μητρώου : mai25067
email : mai25067@uom.edu.gr
Ακαδημαϊκό έτος 2024-2025
Ιούνιος 2025

Περιεχόμενα

Πίνακας Διαγραμμάτων

1. Εισαγωγή

Το παρόν έγγραφο αποτελεί αναφορά για την εργασία πάνω στην Μάθηση χωρίς Επίβλεψη (Unsupervised learning) και την Συσταδοποίηση (Clusterings) στα πλαίσια του μαθήματος Μέθοδοι και Εργαλεία Τεχνητής Νοημοσύνης. Σκοπός της εργασίας είναι να συγκριθούν 2 τεχνικές clustering ενός dataset, σε διαφορετικές περιπτώσεις μορφής δεδομένων. Συγκεκριμένα θα τεθούν σε σύγκριση οι τεχνικές Minibatch K means και ο Aggromerative Clustering στις περιπτώσεις όπου τα δεδομένα είναι ακατέργαστα (raw data) και όπου στα δεδομένα χρησιμοποιούνται σύνθετα περιγραφικά χαρακτηριστικά, αποτέλεσμα τεχνικών Dimensionality Reduction. Στην παρούσα μελέτη θα χρησιμοποιηθούν 2 τεχνικές dimensionality reduction.

2. Μεθοδολογία

Τα δεδομένα που χρησιμοποιήθηκαν αφορά στο fashion-mnist dataset, ένα σύνολο εικόνων 28x28 grayscale εικόνων που απεικονίζουν διάφορα είδη μόδας. Οι τεχνικές clustering που θα συγκριθούν είναι :

1. Minibatch kMeans
2. Aggromerative clustering

Για την ζητούμενη σύγκριση, εφαρμόστηκαν οι μέθοδοι clustering όταν το σύνολο των δεδομένων είναι ακατέργαστα (raw data) , που σημαίνει ότι οι εικόνες θα έχουν τις πραγματικές τιμές pixel και όλα τα χαρακτηριστικά τους , και όταν τα δεδομένα θα έχουν υποστεί Dimensionality reduction, που σημαίνει ότι οι εικόνες θα έχουν σύνθετα περιγραφικά χαρακτηριστικά. Για dimensionality reduction χρησιμποιήθηκαν οι μέθοδοι Principal Component Analysis (PCA) και Stacked Autoencoder (SAE).

Συνεπώς, έχουμε τις εξής περιπτώσεις :

1. Εφαρμογή Minibatch kMeans στα δεδομένα που έχουν υποστεί dimensionality reduction με PCA
2. Εφαρμογή Agglomerative Clustering στα δεδομένα που έχουν υποστεί dimensionality reduction με PCA
3. Εφαρμογή Minibatch kMeans στα δεδομένα που έχουν υποστεί dimensionality reduction με SAE
4. Εφαρμογή Agglomerative clustering στα δεδομένα που έχουν υποστεί dimensionality reduction με SAE
5. Εφαρμογή Minibatch kMeans σε raw data
6. Εφαρμογή Agglomerative clustering σε raw data

Προκειμένου να κάνουμε τις παραπάνω δοκιμές, πραγματοποιήθηκε κανονικοποίηση στο σύνολο του dataset , και ύστερα χωρίστηκε σε train, validation και test set. Για την εκπαίδευση των μοντέλων dimensionality reduction χρησιμοποιήθηκε το train set και για την δοκιμή των μεθόδων clustering το test set, χρησιμοποιώντας το ίδιο μοντέλο που εκπαιδεύτηκε. Για κάθε τεχνική clustering κρατήσαμε τις εξής μετρικές :

- Silhouette score
- Calinski – Harabasz index

- Davies – Bouldin index

Για να αποφασίσουμε με ποιον αριθμό clusters θα γίνουν οι δοκιμές στο test set, τρέξαμε τις τεχνικές clustering στο validation set. Οι δοκιμές έγιναν για 5-10 clusters και κρατήσαμε τον αριθμό cluster που σημείωσε την καλύτερη επίδοση μα βάση το Silhouette score, χρησιμοποιώντας πάντα το μοντέλο που εκπαιδεύτηκε.

Ως τελικός καλύτερος συνδιασμός τεχνικών για την συγκεκριμένη περίπτωση θεωρήθηκε αυτός που σημείωσε το καλύτερο Silhouette Score.

3. Πειραματικά Αποτελέσματα

3.1 Περιγραφή Dataset

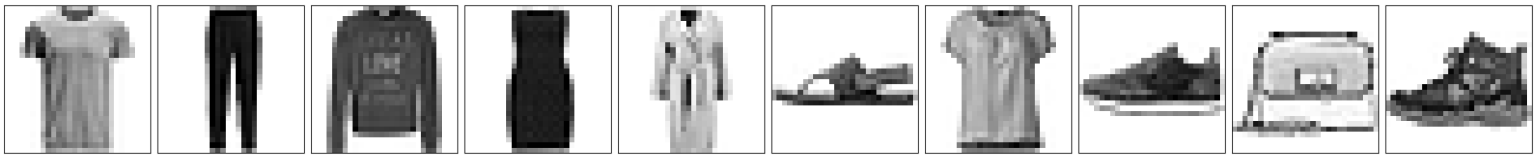
Το dataset αποτελείται από ένα σύνολο 28x28 grayscale εικόνων που απεικονίζουν είδη μόδα, και σύμφωνα με το [1] είναι ήδη χωρισμένο σε train και test set με το train να αποτελείται από 60000 δείγματα και το test από 10000 δείγματα.

Ως πρώτο βήμα της μεθοδολογίας, με την φόρτωση των δεδομένων χρειάστηκε να κόψουμε ένα κομμάτι από το train set για να φτιάξουμε το validation set, που θα μας βοηθήσει στην συνέχεια να επιλέξουμε τον κατάλληλο αριθμό clusters για την περίπτωσή μας.

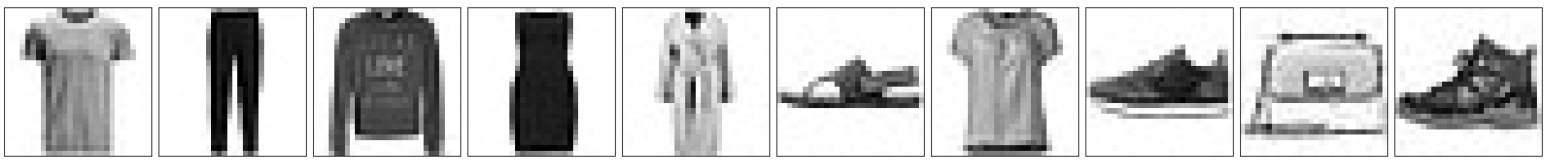
Ύστερα, το σύνολο των δεδομένων κανονικοποιήθηκαν για να μπορούμε να έχουμε καλύτερα αποτελέσματα στις δοκιμές μας. Πολλές φορές κατά την διάρκεια των δοκιμών χρειάστηκε τα δεδομένα μας να τα αλλάξουμε μορφή προκειμένου να μπορούν να αξιοποιηθούν από μεθόδους όπως εκπαίδευση μοντέλων ή εκτύπωση τυχαίων εικόνων.

3.2 Dimensionality Reduction με PCA

Αρχικά εφαρμόσθηκε η τεχνική PCA στο train set. Η διάρκεια της εκπαίδευσης διήρκησε 4.11 seconds. Ακολουθεί ένα set εικόνων πριν την εφαρμογή του PCA και ένα σετ με τις ίδιες εικόνες μετά την εφαρμογή του PCA. Τα σετ έχουν μία εικόνα ανά κλάση και η επιλογή έγινε με τυχαίο τρόπο.

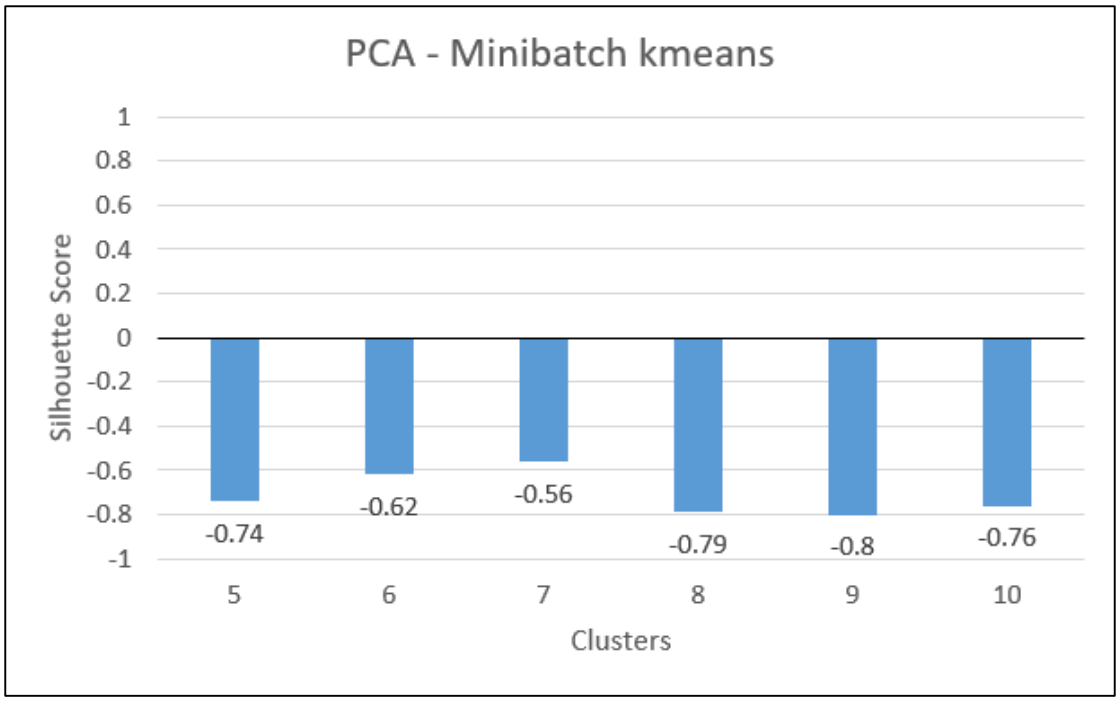


Διάγραμμα 1 - Εικόνες ΠΡΙΝ την εφαρμογή PCA

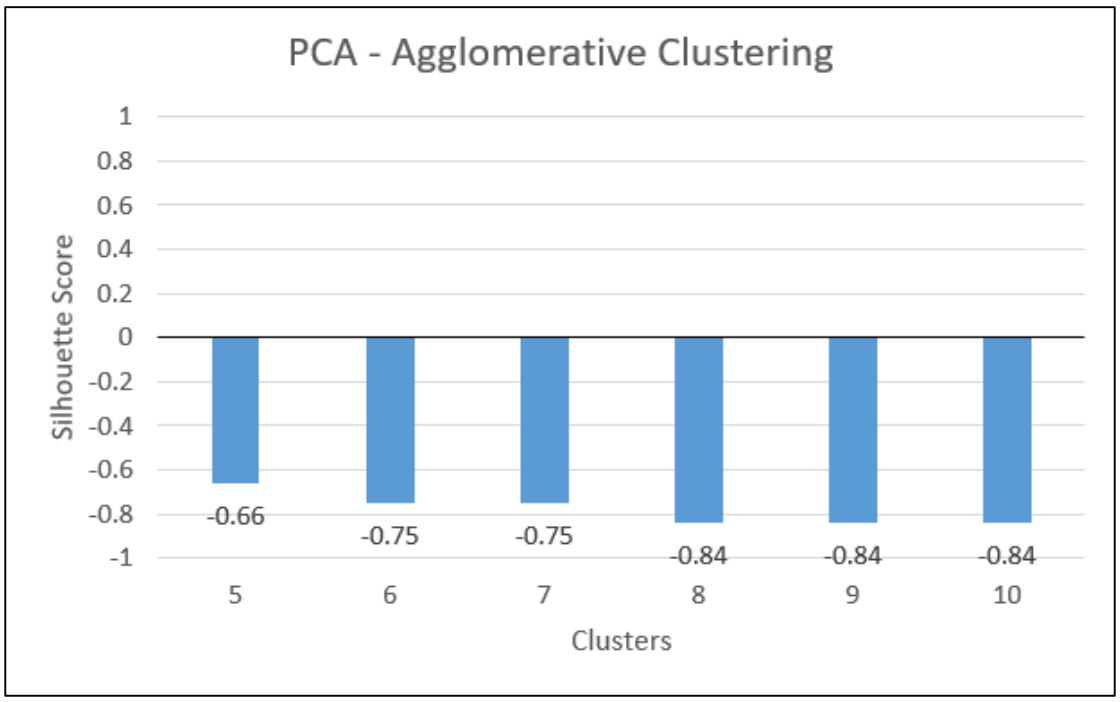


Διάγραμμα 2 - Εικόνες ΜΕΤΑ την εφαρμογή PCA

Στη συνέχεια, χρησιμοποιώντας το ίδιο μοντέλο, έγιναν δοκιμές για την εύρεση του ιδανικότερου αριθμού cluster με την χρήση του validation set και με τις δύο τεχνικές clustering. Δοκιμάστηκαν 5 έως 10 clusters και θεωρήσαμε την καλύτερη επίδοση αυτήν με το υψηλότερο Silhouette score. Οι επιδόσεις φαίνονται στα διαγράμματα που ακολουθούν :



Διάγραμμα 3 - Silhouette Score με Minibatch kmeans μετά από PCA



Διάγραμμα 4 - Silhouette score με Agglomerative clustering μετά από PCA

Υπενθυμίζεται ότι το Silhouette Score κυμαίνεται στο διάστημα [-1, 1], με το υψηλότερο score να δηλώνει καλή συσταδοποίηση. Σύμφωνα με τα Διαγράμματα 3 και 4 επιλέχθηκε η χρήση **7 clusters** για την τεχνική minibatch kmeans και **5 clusters** για την τεχνική Agglomerative clustering. Επομένως, για αυτά τα k τρέξαμε τις δύο τεχνικές clustering στο test set και οι επιδόσεις τους φαίνονται στα διαγράμματα της Ενότητας 4.

3.3 Dimensionality Reduction με SAE

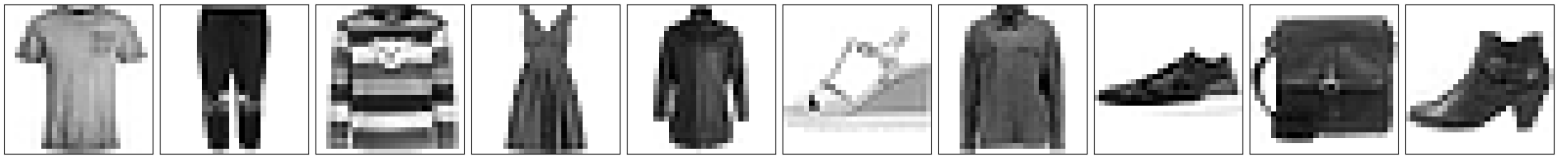
Στη συνέχεια, εφαρμόστηκε Dimensionality reduction με SAE. Για την υλοποίηση του, εφαρμόστηκε η παρακάτω αρχιτεκτονική :

- **Είσοδος (input layer):** Έχει μέγεθος ίσο με τις διαστάσεις των χαρακτηριστικών των δεδομένων εισόδου, δηλαδή του μετασχηματισμένου train set.
- **Κωδικοποιητές (encoding layers):** Δύο επίπεδα με 512 και 256 νευρώνες αντίστοιχα, χρησιμοποιώντας τη συνάρτηση ενεργοποίησης ReLU. Αυτά τα επίπεδα κάνουν προοδευτικά dimensionality reduction, κρατώντας τις σημαντικότερες πληροφορίες.
- **Σημείο συμφόρησης (bottleneck layer):** Ένα επίπεδο με 64 νευρώνες, το οποίο αποτελεί την "συμπιεσμένη" αναπαράσταση των δεδομένων. Ο αριθμός 64 επιλέχθηκε ως συμβιβασμός μεταξύ συμπίεσης και διατήρησης πληροφορίας.
- **Αποκωδικοποιητές (decoding layers):** Αντίστροφη διαδικασία με δύο επίπεδα 256 και 512 νευρώνων που αποσυμπιέζουν τα δεδομένα πίσω στις αρχικές τους διαστάσεις.
- **Έξοδος (output layer):** Ένα επίπεδο με μέγεθος ίσο με το train set, και συνάρτηση ενεργοποίησης **sigmoid** για να παράγει τιμές στο διάστημα [0,1], κατάλληλο για κανονικοποιημένα δεδομένα εικόνας ή χαρακτηριστικών. Το autoencoder εκπαιδεύεται ώστε να ανακατασκευάζει τα αρχικά δεδομένα δηλαδή το train set, από την είσοδο στην έξοδο, ελαχιστοποιώντας το σφάλμα μέσης τετραγωνικής απόκλισης (mean squared error) με τη βοήθεια του optimizer Adam.

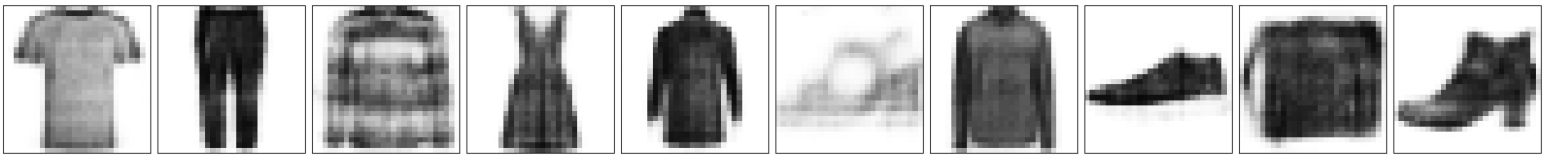
Για το παραπάνω μοντέλο επιλέχθηκαν 20 εποχές, προκειμένου να επιτευχθεί καλή σύγκλιση χωρίς να overfitting, επιτρέψαμε την ανάμειξη των δεδομένων ανά εποχή για καλύτερη γενίκευση, και χρησιμοποιήσαμε και το validation set για παρακολούθηση της απόδοσης του μοντέλου.

Η διάρκεια της εκπαίδευσης διήρκησε 190 seconds.

Ακολουθεί ένα set εικόνων πριν την εφαρμογή του SAE και ένα σετ με τις ίδιες εικόνες μετά την εφαρμογή του SAE. Τα σετ έχουν μία εικόνα ανά κλάση και η επιλογή έγινε με τυχαίο τρόπο.

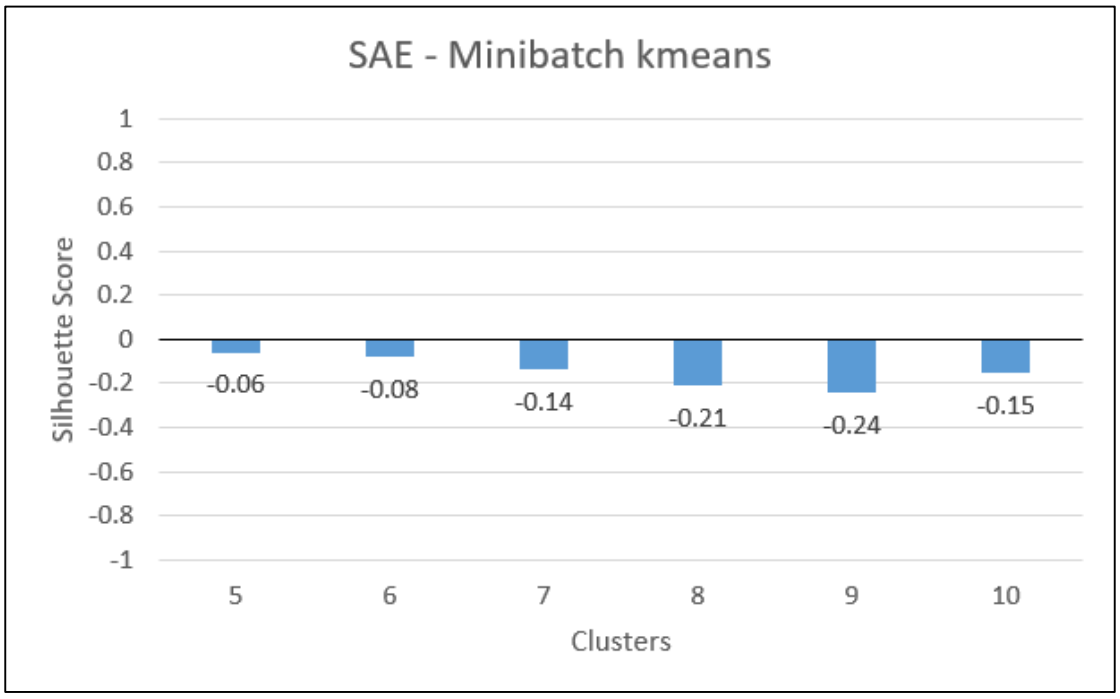


Διάγραμμα 5 - Εικόνες ΠΡΙΝ την εφαρμογή SAE

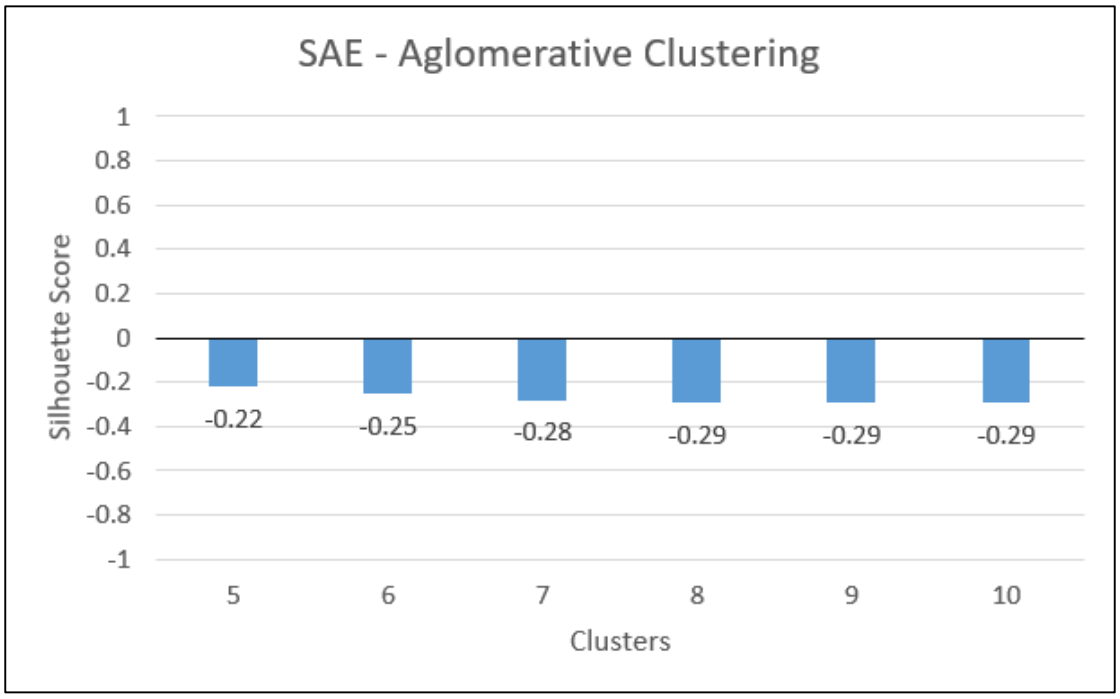


Διάγραμμα 6 - Εικόνες ΜΕΤΑ την εφαρμογή SAE

Στη συνέχεια, ακολουθήσαμε την ίδια διαδικασία όπως και με τον PCA, για την εύρεση του ιδανικότερου αριθμού cluster για κάθε τεχνική clustering με την χρήση του validation set. Όπως και προηγουμένως, δοκιμάστηκαν 5 έως 10 clusters και θεωρήσαμε την καλύτερη επίδοση αυτήν με το υψηλότερο Silhouette score. Οι επιδόσεις φαίνονται στα διαγράμματα που ακολουθούν :



Διάγραμμα 7 - Silhouette score με Minibatch kmeans μετά από SAE

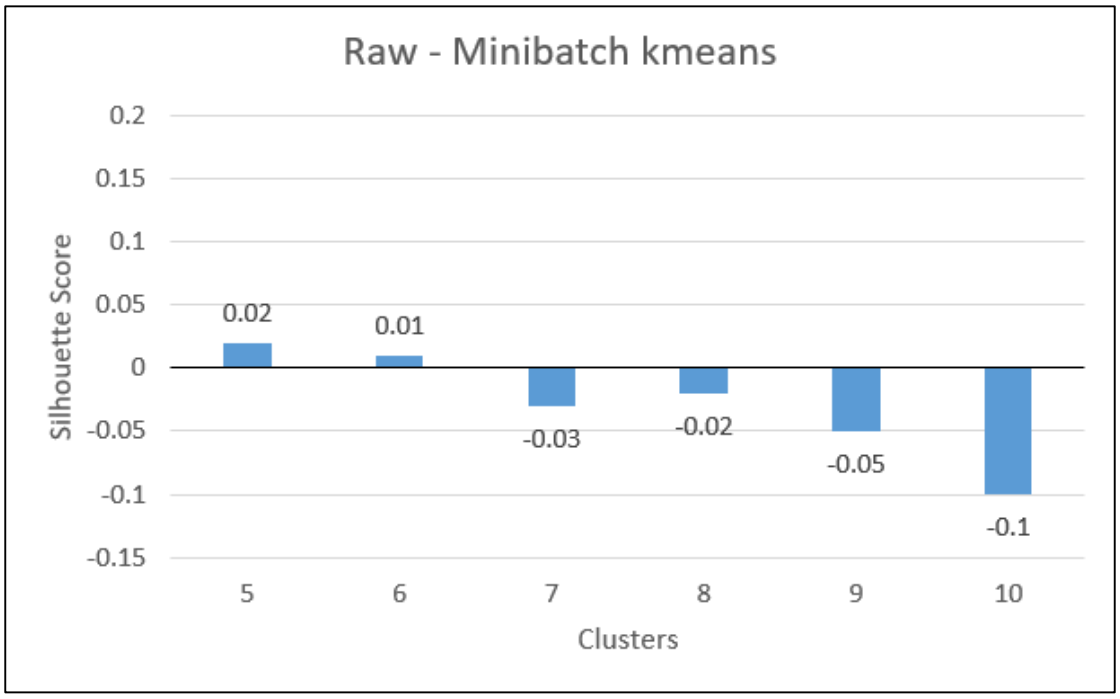


Διάγραμμα 8 - Silhouette score με Agglomerative clustering μετά από SAE

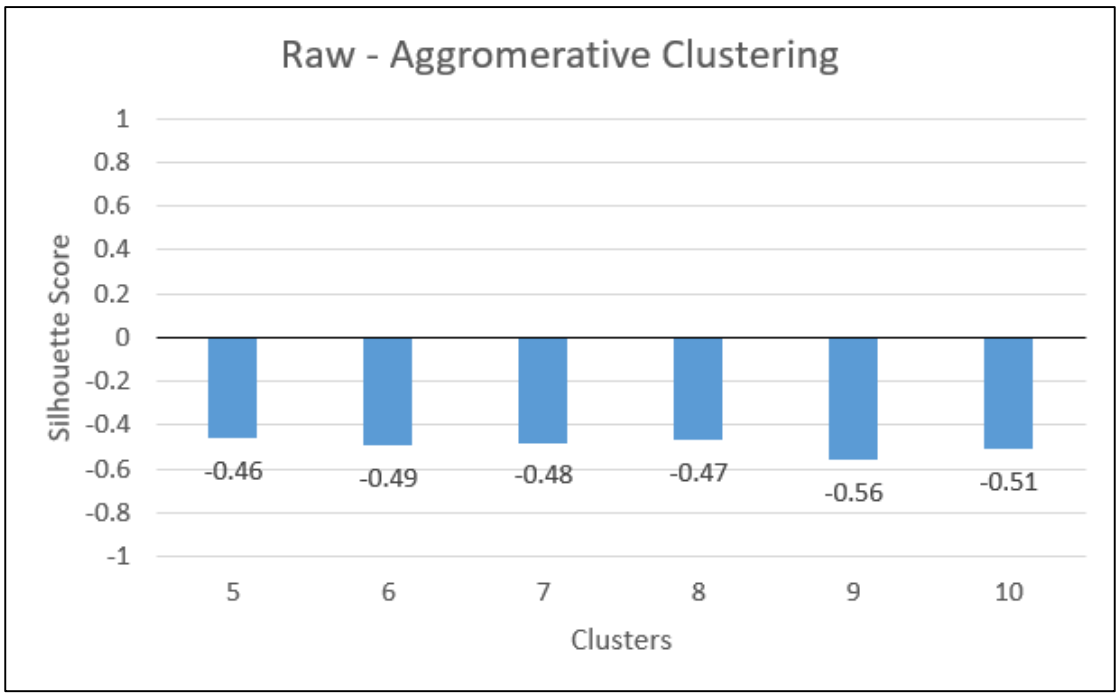
Σύμφωνα με τα Διαγράμματα 7 και 8 επιλέχθηκε η χρήση **5 clusters** για την τεχνική minibatch kmeans και **5 clusters** για την τεχνική Agglomerative clustering. Επομένως, για αυτά τα k τρέξαμε τις δύο τεχνικές clustering στο test set και οι επιδόσεις τους φαίνονται στα διαγράμματα της Ενότητας 4.

3.4 Χρήση Raw data

Σε αυτήν την περίπτωση δεν απαιτείται κάποια εκπαίδευση ενός μοντέλου καθώς τα δεδομένα χρησιμοποιούνται δίχως κάποια αλλαγή, πέραν της αρχικής κανονικοποίησης. Ωστόσο, όπως και στις προηγούμενες περιπτώσεις, για να δοκιμάσουμε τις τεχνικές clustering πρέπει να βρεθεί ο ιδανικότερος αριθμός cluster με την χρήση του validation set, όπως φαίνεται στα διαγράμματα που ακολουθούν :



Διάγραμμα 9 - Silhouette score με Minibatch kmeans στα raw data

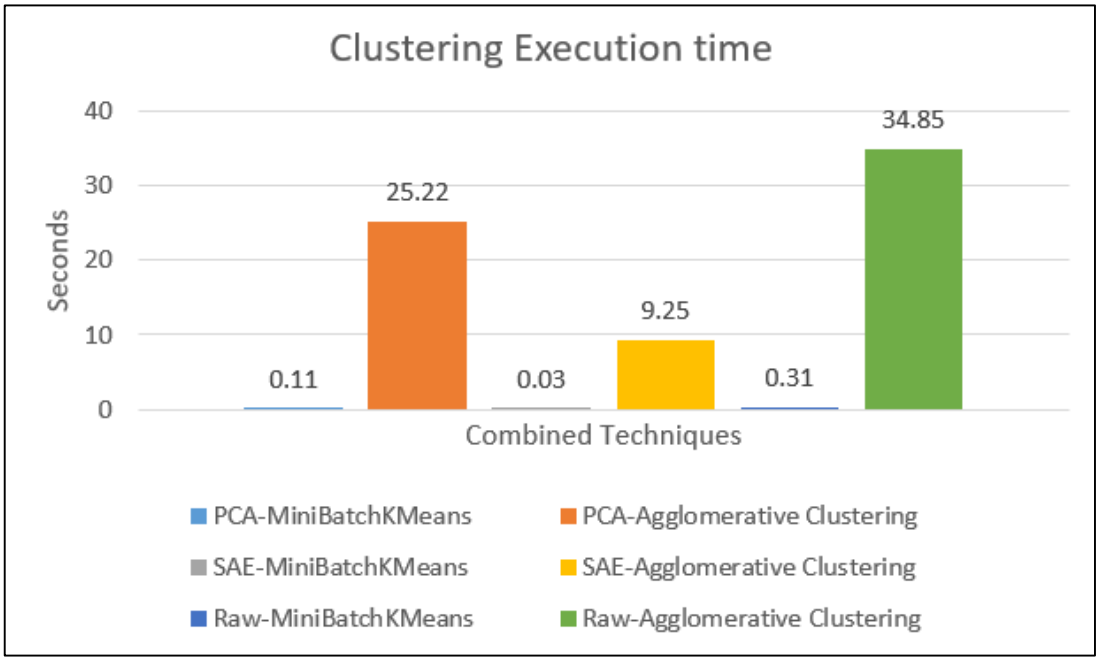


Διάγραμμα 10 - Silhouette score με Agglomerative clustering στα raw data

Σύμφωνα με τα Διαγράμματα 9 και 10, επιλέχθηκε η χρήση **5 clusters** για την τεχνική minibatch kmeans και **5 clusters** για την τεχνική Agglomerative clustering. Επομένως, για αυτά τα k τρέξαμε τις δύο τεχνικές clustering στο test set και οι επιδόσεις τους φαίνονται στα διαγράμματα της Ενότητας 4.

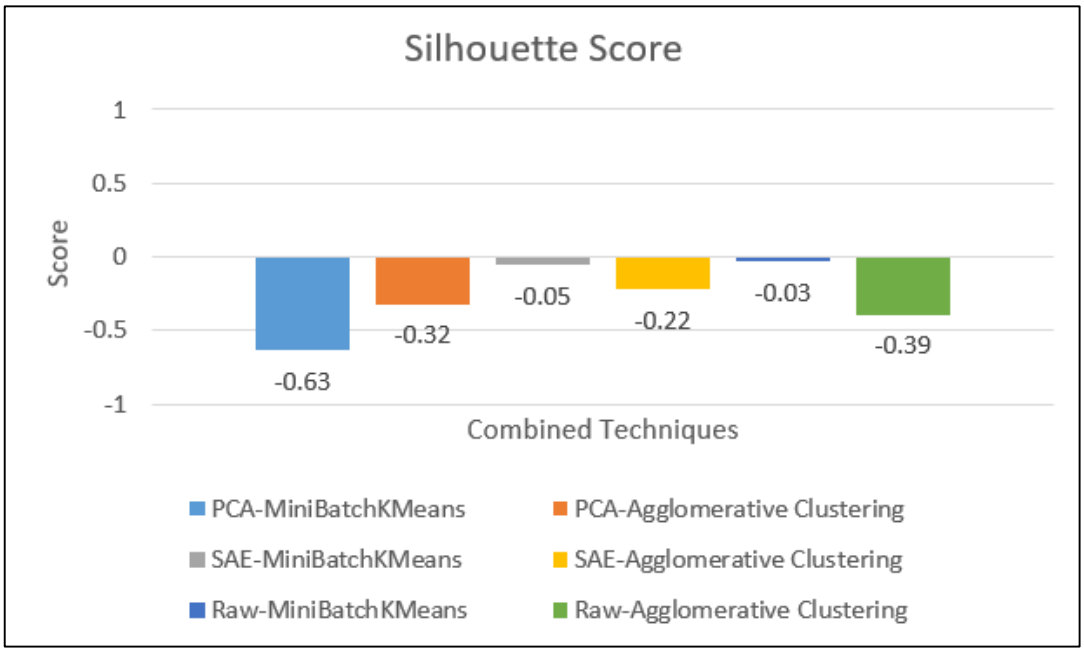
4. Συμπεράσματα

Οι συνδυασμοί των τεχνικών έτρεξαν με αριθμό clusters τον ιδανικότερο βάσει της προηγούμενης ανάλυσης των validation set. Στα διαγράμματα που ακολουθούν, παρουσιάζονται τα αποτελέσματα των συνδυασμών τεχνικών που τρέξαμε. Σημειώνεται ότι είναι διαφορετικά τα διαγράμματα, για την καλύτερη απεικόνιση των ευρημάτων, διότι κάθε δείκτης έχει την δική του κλίμακα.



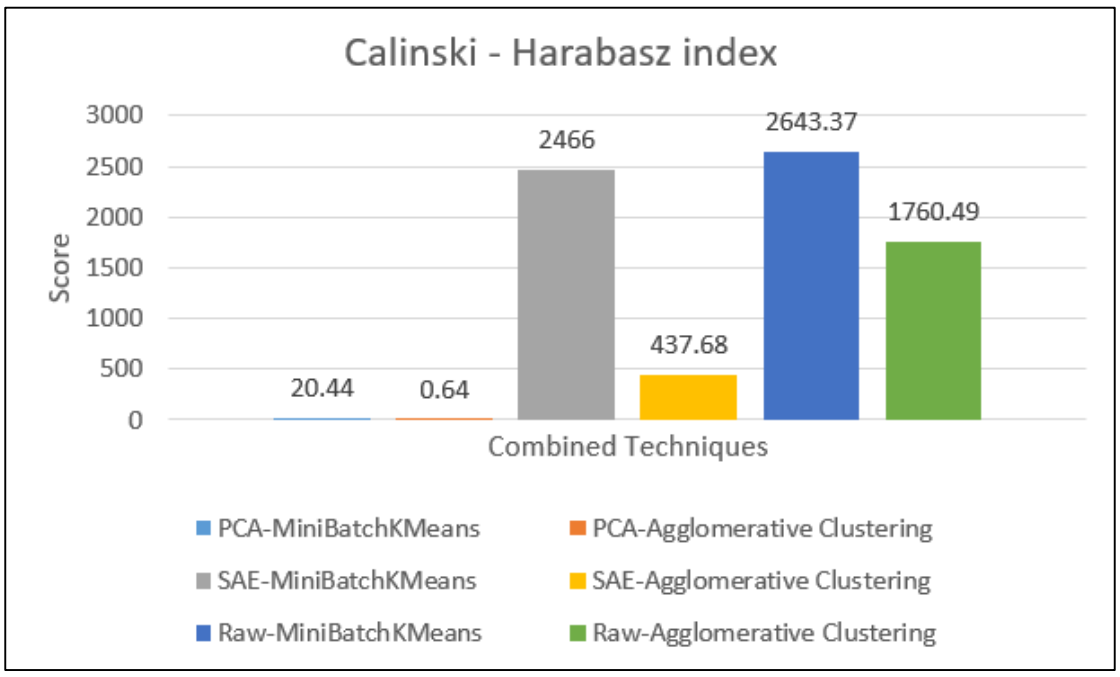
Διάγραμμα 11 – Χρόνοι εκτέλεσης Τεχνικών Clustering

Παρατηρείται ότι σε όλες τις περιπτώσεις ο Minibatch kMeans έκανε το clustering σε σημαντικά λιγότερο χρόνο έναντι του Agglomerative clustering. Τον λιγότερο χρόνο τον σημείωσε ο **Minibatch kMeans μετά από μείωση διαστάσεων με SAE**, στα 0.03 seconds. Σημειώνεται ότι η τεχνική έτρεξε με 5 clusters. Δεύτερη καλύτερη επίδοση σημείωσε ο Minibatch kMeans μετά από εφαρμογή PCA, στα 0.11 seconds, με 7 clusters.



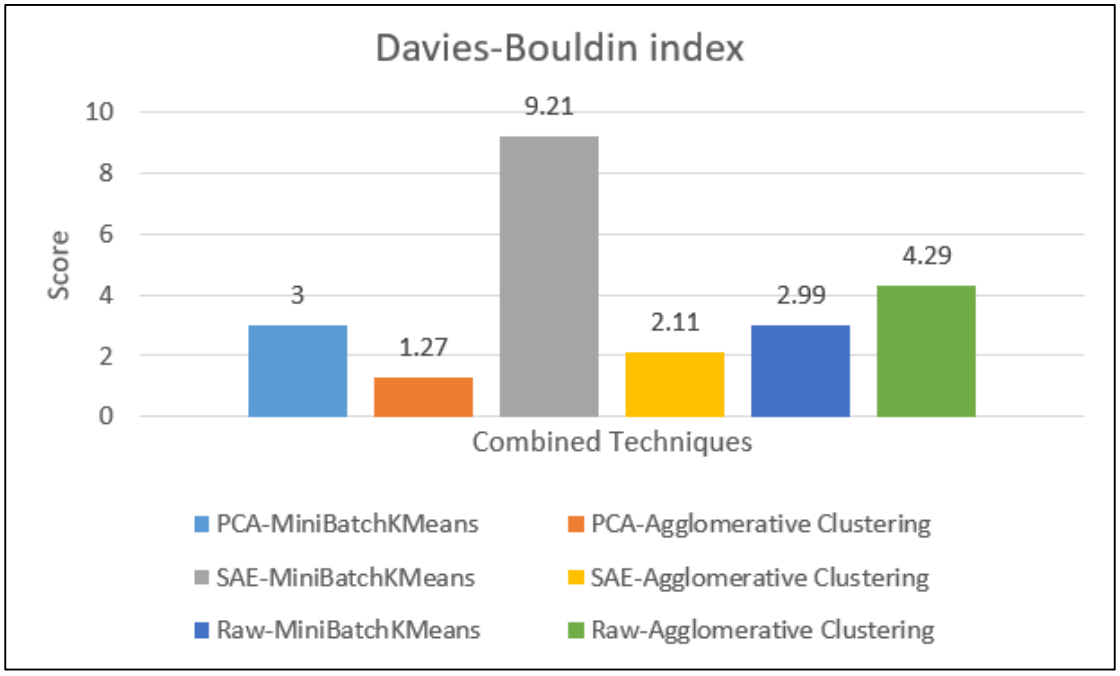
Διάγραμμα 12 - Δείκτης Silhouette Score Τεχνικών Clustering

Ο δείκτης Silhouette score μετρά πόσο καλά ταιριάζει ένα σημείο στον cluster του, συγκρίνοντας το με τα γειτονικά του σημεία. Κανένα από τα μοντέλα δεν σημείωσε αρκετά ικανοποιητικό Silhouette score, μιας και το σκορ κυμαίνεται στο διάστημα [-1,1]. Την καλύτερη επίδοση σημείωσε ο **Minibatch kMeans χρησιμοποιώντας τα raw data**, με Silhouette score -0.03 και 5 clusters. Σημειώνεται ότι η τεχνική έτρεξε με 5 clusters. Δεύτερη καλύτερη επίδοση σημείωσε ο Minibatch kMeans μετά από SAE, με 5 clusters και score -0.05.



Διάγραμμα 13 - Δείκτης Calinski - Harabasz Τεχνικών Clustering

Ο δείκτης Calinski – Harabasz αξιολογεί την αναλογία της διασποράς μεταξύ των clusters και της διασποράς εντός των clusters, με το υψηλό score να είναι το επιθυμητό. Στις δοκιμές μας το υψηλότερο score σημείωσε πάλι ο **Minibatch kMeans χρησιμοποιώντας τα raw data,** με score 2643.37 . Σημειώνεται ότι η τεχνική έτρεξε με 5 clusters. Δεύτερη καλύτερη επίδοση σημείωσε ο Minibatch kMeans μετά από SAE, με 5 clusters και score 2466.



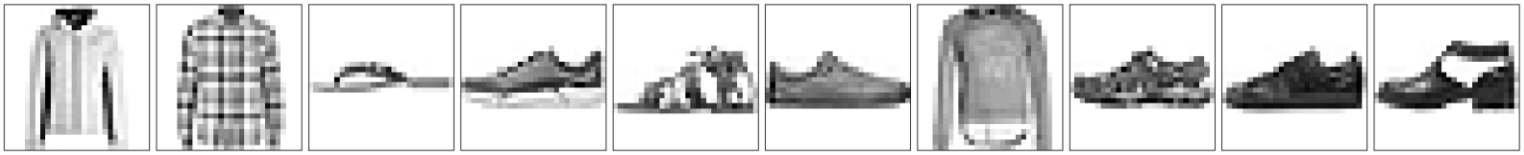
Διάγραμμα 14 - Δείκτης Davies - Bouldin Τεχνικών Clustering

Ο δείκτης Davies - Bouldin δηλώνει την ομοιογένεια, δηλαδή το πόσο κοντά είναι τα σημεία εντός ενός cluster, και το κατά πόσο καλά είναι διαχωρισμένοι οι clusters μεταξύ τους, δηλαδή πόσο μακριά απέχουν οι clusters μεταξύ τους. Επιθυμητή είναι η μικρότερη τιμή του δείκτη. Την καλύτερη επίδοση την σημείωσε ο Agglomerative μετά την εφαρμογή του PCA, με 5 clusters και δείκτη 1.27 .

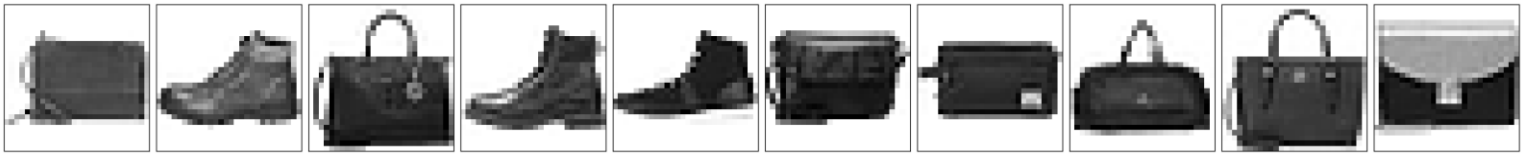
Ως καλύτερο συνδυασμός τεχνικών λήφθηκε υπόψιν το Silhouette Score. Όπως φαίνεται στο Διάγραμμα 12, τη καλύτερη επίδοση σημείωσε η χρήστη του **Minibatch kMeans με την χρήση raw data.** Αυτός ο συνδυασμός σημείωσε επίσης 3^{η} καλύτερη επίδοση σε χρόνο εκτέλεσης, την καλύτερη επίδοση στον δείκτη Calinski – Harabazs και την 3^{η} καλύτερη επίδοση στον δείκτη Davies –Bouldin.

Ακολούθησε με μικρές διαφορές ο συνδυασμός : χρήση του Minibatch kMeans μετά από dimensionality reduction με χρήση του SAE.

Ακολουθούν 10 τυχαίες εικόνες από 2 τυχαίες clusters που επιλέχθηκαν μετά την εφαρμογή του Minibatch kMeans σε raw data .



Διάγραμμα 15 - Εικόνες τυχαίου cluster καλύτερης μεθόδου



Διάγραμμα 16 - Εικόνες τυχαίου cluster καλύτερης μεθόδου

Παρατηρείται ότι υπάρχει μια τάση του cluster να έχει label με χαρακτηρισμό Παπούτσι και Τσάντα αντίστοιχα, ωστόσο φαίνεται να μην είναι πλήρως ομοιογενείς οι clusters.

Εν μέρει αυτό ίσως οφείλεται στο γεγονός ότι στην καλύτερη μέθοδο καλύτερο score επίδειξε όταν έτρεξε σε 5 clusters, ενώ στην πραγματικότητα οι κλάσεις είναι 10. Παράδοξο είναι πως όταν έτρεξαν όλοι οι αλγόριθμοι σε 10 clusters, καμιά δεν σημείωσε το καλύτερο score.

Βάσει των αποτελεσμάτων, οι διάφοροι συνδυασμοί των τεχνικών δεν απέδωσαν πολύ ικανοποιητικά βάσει του Silhouette Score, καθώς δεν θεωρείται αρκετά καλή η συσταδοποίηση με score κοντά στο 0 ή και < 0. Επίσης, φάνηκε πως με τον PCA γενικώς υπήρχαν ικανοποιητικά αποτελέσματα σε σύγκριση με την χρήση του SAE.

5. Βιβλιογραφία

[1]. https://keras.io/api/datasets/fashion_mnist/

[2]. Διαφάνειες διαλέξεων μαθήματος «Μέθοδοι και Εργαλεία Τεχνητής Νοημοσύνης»

[3]. Παραδείγματα κώδικα πάνω σε προβλήματα Clustering και Dimensionality Reduction του μαθήματος «Μέθοδοι και Εργαλεία Τεχνητής Νοημοσύνης»