
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

# Τμήμα Εφαρμοσμένης Πληροφορικής Επιστήμης και Τεχνολογίας Υπολογιστών

## ΜΑΘΗΜΑ Μηχανική Μάθησης

### Εργασία 2η – Clustering problems

**Φοιτητές**

**Θωμασιάδης Κωσταντίνος** ics21058

**Κανίδου Ελισάβετ-Περσεφόνη** ics21095

**Κοτούλα Αριστέα** ics21

**Τσαβαλιάς Βασίλειος Εφραίμ** ics21083

# Υλικό Επίλυσης Εργασίας

https://drive.google.com/drive/folders/1CILAiOTMNsQuo7bGJXfDkpT4w9cj99Fc?usp=sharing

Το παραπάνω link για το drive εμπεριέχει όλο το υλικό επίλυσης της εργασίας. Πιο συγκεκριμένα υπάρχει ο κώδικας της εργασίας στο περιβάλλον του Google Colab, το dataframe ,  καθώς και ο κατάλογος όλων των γραφημάτων.

# ΠΕΡΙΕΧΟΜΕΝΑ:

# Περίληψη:

Αυτή η μελέτη διερευνά τη ομαδοποίηση στην ανάλυση δεδομένων εικόνας, εστιάζοντας συγκεκριμένα στο σύνολο δεδομένων Fashion MNIST, το οποίο αποτελείται από 70.000 εικόνες σε κλίμακα του γκρι από διάφορα είδη μόδας, το καθένα με ανάλυση 28x28 pixel. Η έρευνα αντιμετωπίζει την υψηλών διαστάσεων φύση της ταξινόμησης εικόνων χρησιμοποιώντας μια συστηματική προσέγγιση για τη μείωση διαστάσεων, χρησιμοποιώντας την Ανάλυση Κύριων Στοιχείων (PCA), τον Αυτοκωδικοποιητή Στοίβαξης (SAE) και την Ενσωμάτωση Στοχαστικού Γείτονα t-Distributed (t-SNE). Το PCA χρησιμοποιείται για τη μετατροπή των δεδομένων σε ένα χώρο με λιγότερες διαστάσεις διατηρώντας παράλληλα τη μεγαλύτερη διακύμανση, το SAE για την εκμάθηση μιας συμπιεσμένης αναπαράστασης χωρίς επίβλεψη και το t-SNE για την ικανότητά του να διατηρεί τοπικές δομές σε χώρο μειωμένων διαστάσεων. Στη συνέχεια, η μελέτη εξετάζει την ομαδοποίηση μέσω αλγορίθμων όπως το MiniBatch KMeans, το DBSCAN και το Agglomerative Clustering. Το MiniBatch KMeans επιλέγεται για την αποτελεσματικότητά του με μεγάλα σύνολα δεδομένων, το DBSCAN για την ικανότητά του να ανιχνεύει συστάδες ποικίλων σχημάτων και το Agglomerative Clustering για την προσέγγισή του στην ιεραρχική ομαδοποίηση. Η αποτελεσματικότητα αυτών των μεθόδων αξιολογείται διεξοδικά χρησιμοποιώντας διάφορες μετρήσεις: ο δείκτης Calinski-Harabasz για την εγκυρότητα συμπλέγματος με βάση τη διακύμανση εντός του συμπλέγματος, ο δείκτης Davies-Bouldin για τη μέση ομοιότητα μεταξύ κάθε συστάδας και του πλησιέστερου αντίστοιχου, ο δείκτης Silhouette για τη μέτρηση της εγγύτητας των σημείων δεδομένων μέσα στο σύμπλεγμα τους σε σύγκριση με άλλες συστάδες και τον προσαρμοσμένο δείκτη Rand για τον ποσοτικό προσδιορισμό της ακρίβειας ομαδοποίησης σε σχέση με ένα γνωστό σύνολο ετικετών αληθείας βάσης. Τα ευρήματα από αυτή τη λεπτομερή ανάλυση αποκαλύπτουν ποικίλες επιδόσεις στη ομαδοποίηση, φωτίζοντας την περίπλοκη ισορροπία και την αλληλεπίδραση μεταξύ τεχνικών μείωσης διαστάσεων και αλγορίθμων ομαδοποίησης. Αυτή η έρευνα όχι μόνο ενισχύει την κατανόηση της δυναμικής ομαδοποίησης στα δεδομένα εικόνας, αλλά ανοίγει επίσης το δρόμο για μελλοντικές έρευνες στη μηχανική μάθηση και την ανάλυση δεδομένων, εστιάζοντας στη βελτίωση και τη βελτιστοποίηση αυτών των μεθοδολογιών για πιο αποτελεσματική εφαρμογή.

# Abstract:

This study investigates clustering in image data analysis, focusing specifically on the Fashion MNIST dataset, which consists of 70,000 grayscale images of various fashion items, each with a resolution of 28x28 pixels. The research addresses the high-dimensional nature of image classification by employing a systematic approach to dimensionality reduction, utilizing Principal Component Analysis (PCA), Stacked Autoencoder (SAE), and t-Distributed Stochastic Neighbor Embedding (t-SNE). PCA is used to transform the data into a space with fewer dimensions while retaining most variance, SAE for learning a compressed representation in an unsupervised manner, and t-SNE for its ability to preserve local structures in a reduced dimension space. The study then examines clustering through algorithms such as MiniBatch KMeans, DBSCAN, and Agglomerative Clustering. MiniBatch KMeans is chosen for its efficiency with large datasets, DBSCAN for its capability to detect clusters of varied shapes, and Agglomerative Clustering for its approach to hierarchical clustering. The efficacy of these methods is thoroughly evaluated using several metrics: the Calinski-Harabasz index for cluster validity based on within-cluster variance, the Davies-Bouldin index for average similarity between each cluster and its closest counterpart, the Silhouette index for measuring the closeness of data points within their cluster compared to other clusters, and the adjusted Rand index for quantifying clustering accuracy against a known set of ground truth labels. The findings from this detailed analysis reveal varying performances in clustering, illuminating the intricate balance and interplay between dimensionality reduction techniques and clustering algorithms. This research not only enhances understanding of clustering dynamics in image data but also paves the way for future investigations in machine learning and data analysis, focusing on refining and optimizing these methodologies for more effective application.

# 1. Θεωρητικό Υπόβαθρο:

Αυτή η έρευνα βασίζεται στην ανάλυση δεδομένων εικόνας υψηλών διαστάσεων, που εφαρμόζονται ειδικά στο σύνολο δεδομένων Fashion MNIST. Επικεντρώνεται στις προκλήσεις της υψηλής διάστασης στα δεδομένα εικόνας, όπου κάθε εικόνα είναι ένα πολυδιάστατο σημείο σε έναν μεγάλο χώρο χαρακτηριστικών.

Για να αντιμετωπιστεί αυτό, η μελέτη χρησιμοποιεί τεχνικές μείωσης διαστάσεων. Το Principal Component Analysis (PCA) χρησιμοποιείται για γραμμικούς μετασχηματισμούς για τη μείωση δεδομένων υψηλών διαστάσεων σε κύρια στοιχεία που καταγράφουν σημαντική διακύμανση. Stacked Autoencoders (SAE), μια μέθοδος που βασίζεται σε νευρωνικά δίκτυα, κωδικοποιεί δεδομένα σε μια συμπιεσμένη αναπαράσταση. Επιπλέον, το t-Distributed Stochastic Neighbor Embedding (t-SNE) εφαρμόζεται για την ικανότητά του να διατηρεί τοπικές δομές δεδομένων, βοηθώντας στην οπτικοποίηση συμπλεγμάτων δεδομένων υψηλών διαστάσεων.

Στη συνέχεια, η έρευνα εξετάζει αλγόριθμους ομαδοποίησης για τα επεξεργασμένα δεδομένα εικόνας. Το MiniBatch KMeans επιλέγεται για την αποτελεσματικότητά του στο χειρισμό μεγάλων συνόλων δεδομένων όπως το Fashion MNIST. Το DBSCAN χρησιμοποιείται για την ανίχνευση συστάδων διαφόρων σχημάτων και μεγεθών, χαρακτηριστικό των δεδομένων εικόνας. Η αθροιστική ομαδοποίηση, μια τεχνική ιεραρχικής ομαδοποίησης, χρησιμοποιείται επίσης για την παροχή πληροφοριών σχετικά με τη δομή δεδομένων.

Η αποτελεσματικότητα αυτών των μεθόδων ομαδοποίησης αξιολογείται χρησιμοποιώντας μετρήσεις όπως ο δείκτης Calinski-Harabasz για τη διασπορά συστάδων, ο δείκτης Davies-Bouldin για ομοιότητα μεταξύ συστάδων, ο δείκτης Silhouette για τη συνοχή συμπλέγματος και ο προσαρμοσμένος δείκτης Rand για ακρίβεια σε σχέση με ένα γνωστό έδαφος αλήθεια.

Αυτό το πλαίσιο συνδυάζει κλασικές και σύγχρονες τεχνικές μηχανικής μάθησης για την ανάλυση και την εξαγωγή μοτίβων από το σύνολο δεδομένων Fashion MNIST.

# 2. Εισαγωγή:

Αυτή η έρευνα επικεντρώνεται στο σύνολο δεδομένων Fashion MNIST, ένα πρότυπο στη μηχανική μάθηση με 70.000 εικόνες σε κλίμακα του γκρι σε 10 κατηγορίες μόδας. Κάθε εικόνα, ένα πλέγμα εικονοστοιχείων 28x28, είναι ένα σημείο δεδομένων υψηλών διαστάσεων με πολύπλοκα μοτίβα, που παρουσιάζουν προκλήσεις στις εργασίες ταξινόμησης. Ο κύριος στόχος είναι η αποτελεσματική ομαδοποίηση αυτών των δεδομένων εικόνας υψηλών διαστάσεων. Η ομαδοποίηση, ένα βασικό συστατικό της μάθησης χωρίς επίβλεψη, ομαδοποιεί τα σημεία δεδομένων έτσι ώστε αυτά που βρίσκονται στο ίδιο σύμπλεγμα να μοιάζουν περισσότερο μεταξύ τους παρά με εκείνα σε διαφορετικά συμπλέγματα. Αυτό είναι σημαντικό για την κατανόηση των χαρακτηριστικών δεδομένων και την υποβοήθηση της ταξινόμησης εικόνων.

Η προσέγγιση περιλαμβάνει την εφαρμογή και τη σύγκριση διαφόρων τεχνικών μείωσης διαστάσεων. Το Principal Component Analysis (PCA) χρησιμοποιείται για τη μείωση των δεδομένων στα πιο ενημερωτικά στοιχεία του. Ο Stacked Autoencoder (SAE) μαθαίνει μια αναπαράσταση δεδομένων χαμηλότερης διάστασης και η t-Distributed Stochastic Neighbor Embedding (t-SNE) χρησιμοποιείται για την ικανότητά του να διατηρεί τοπικές δομές σε μειωμένες διαστάσεις. Αυτές οι τεχνικές απλοποιούν την πολυπλοκότητα των δεδομένων διατηρώντας ταυτόχρονα βασικές πληροφορίες.

Παράλληλα με αυτές τις μεθόδους, διερευνώνται αρκετοί αλγόριθμοι ομαδοποίησης. Το MiniBatch KMeans επιλέγεται για την επεκτασιμότητα και την αποτελεσματικότητά του. Το DBSCAN επιλέγεται για την ικανότητά του να αναγνωρίζει συστάδες διαφορετικών σχημάτων και η αθροιστική ομαδοποίηση χρησιμοποιείται για ιεραρχική ανάλυση συμπλέγματος. Αυτός ο συνδυασμός αλγορίθμων μείωσης διαστάσεων και ομαδοποίησης είναι κεντρικός στην ανάλυση, αναζητώντας αποτελεσματικές στρατηγικές για τη ομαδοποίηση δεδομένων εικόνας υψηλών διαστάσεων. Τα ευρήματα αναμένεται να συμβάλουν στη μηχανική μάθηση και στην επιστήμη δεδομένων, ιδιαίτερα στη διαχείριση πολύπλοκων συνόλων δεδομένων εικόνων.

# 3. Μεθοδολογία:

## 3.1 Περιγραφή και προ επεξεργασία Δεδομένων:

### 3.1.1 Φόρτωση Δεδομένων:

Το πρώτο βήμα σε αυτήν την έρευνα ήταν η φόρτωση δεδομένων, χρησιμοποιώντας το TensorFlow και το Keras για την εισαγωγή του συνόλου δεδομένων Fashion MNIST, το οποίο περιλαμβάνει ασπρόμαυρες εικόνες ρούχων.

Η μεθοδολογία ξεκίνησε με την προεπεξεργασία του συνόλου δεδομένων Fashion MNIST, ένα ουσιαστικό βήμα για την ανάλυση δεδομένων. Αυτό το σύνολο δεδομένων περιλαμβάνει 60.000 εικόνες εκπαίδευσης και 10.000 δοκιμαστικές εικόνες, η καθεμία μια εικόνα σε κλίμακα του γκρι 28x28 διαφόρων ειδών μόδας. Αυτές οι εικόνες αντιπροσωπεύουν έναν χώρο υψηλών διαστάσεων όταν μετατρέπονται σε μια γραμμική διάταξη 784 χαρακτηριστικών ανά εικόνα.

### 3.1.2 Κανονικοποίηση Δεδομένων:

Η προεπεξεργασία ξεκίνησε με την κανονικοποίηση για την κλίμακα των τιμών των εικονοστοιχείων κάθε εικόνας μεταξύ 0 και 1. Αυτό το βήμα είναι ζωτικής σημασίας στη μηχανική εκμάθηση για να διασφαλιστεί μια συνεπής κλίμακα εισόδου, αποφεύγοντας την προκατάληψη προς μεγαλύτερες αριθμητικές τιμές εικονοστοιχείων και βοηθώντας στην ταχύτερη, πιο σταθερή εκμάθηση. Συγκεκριμένα, τα pixel στην εκπαίδευση (train_images) και οι δοκιμαστικές εικόνες (test_images) διαιρούνται με το 255,0, καθώς οι τιμές των pixel στο σύνολο δεδομένων κυμαίνονται από 0 (μαύρο) έως 255 (λευκό). Η κανονικοποίηση με διαίρεση με το 255 κλιμακώνει τις τιμές μεταξύ 0 και 1, ευθυγραμμίζοντας με το τυπικό εύρος για εισόδους μοντέλου νευρωνικού δικτύου.

### 3.1.3 Χωρισμός Δεδομένων:

Επιπλέον, ο αριθμός των τάξεων (αριθμός_κλάσεων) και τα ονόματα των τάξεων από το σύνολο δεδομένων Fashion MNIST καταγράφηκαν σε μια λίστα. Δημιουργήθηκε μια συνάρτηση (display_sample_images) για την εμφάνιση μιας επιλογής εικόνων, λήψης εικόνων και ετικετών ως εισόδου.

Στη συνέχεια, το σύνολο δεδομένων χωρίστηκε σε δεδομένα εκπαίδευσης (train_images και train_labels) και σε ένα υποσύνολο επικύρωσης (val_images, val_labels), με το 20% των δεδομένων να διατίθεται για επικύρωση (test_size=0,2).

### 3.1.4 Ανάπτυξη Αλγορίθμων Μείωσης Διαστάσεων:

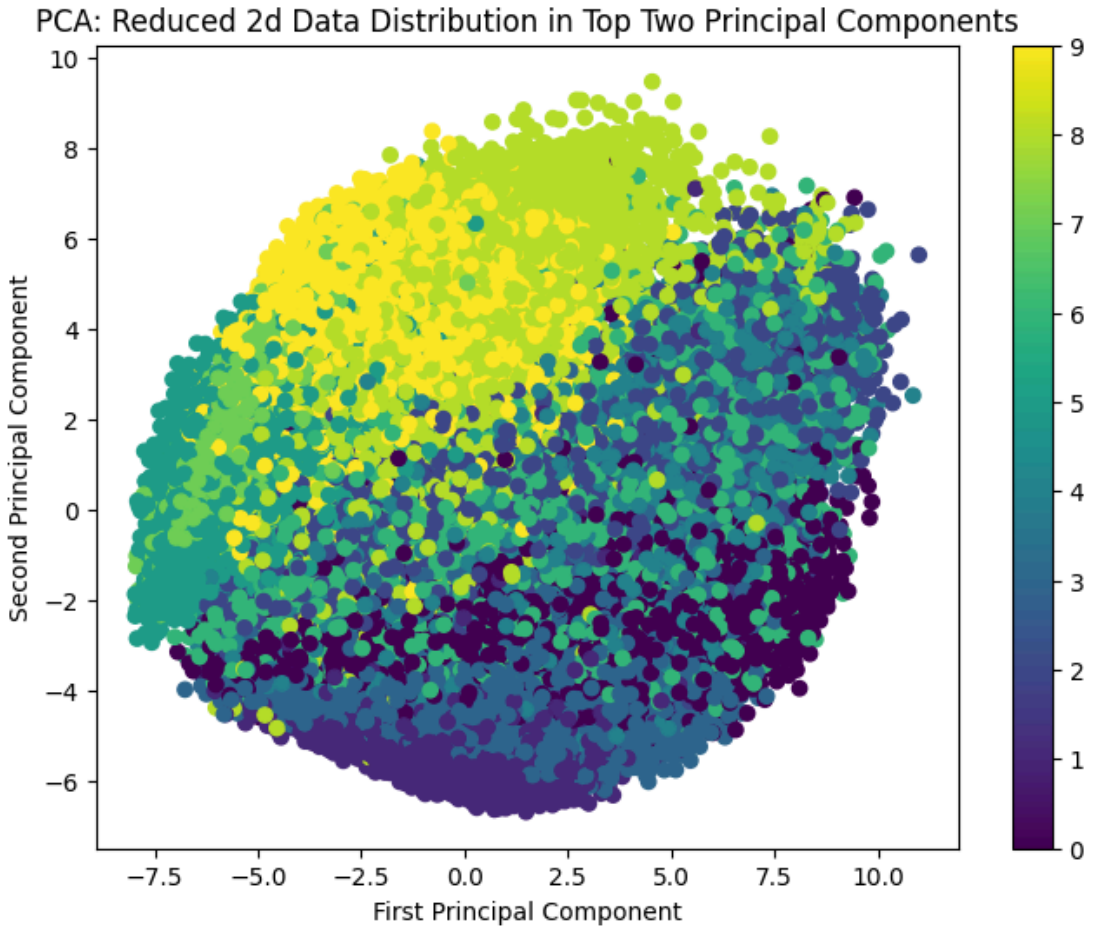
Για να μειωθούν αποτελεσματικά οι διαστάσεις διατηρώντας ταυτόχρονα τις βασικές πληροφορίες, χρησιμοποιήθηκαν τρεις τεχνικές, καθεμία από τις οποίες προσφέρει ξεχωριστά οφέλη. Αρχικά, χρησιμοποιήθηκε η ανάλυση κύριου στοιχείου (PCA) για τη μετατροπή του συνόλου δεδομένων σε ένα σύνολο γραμμικά ασυσχέτιστων στοιχείων. Αυτή η μέθοδος εστιάζει σε στοιχεία που ευθύνονται για τη μεγαλύτερη διακύμανση, μειώνοντας έτσι τις διαστάσεις διατηρώντας παράλληλα τις μέγιστες πληροφορίες. Στη συνέχεια, ένας Stacked Autoencoder (SAE), μια παραλλαγή νευρωνικού δικτύου, χρησιμοποιήθηκε για τη συμπίεση δεδομένων σε ένα χώρο χαμηλότερης διάστασης και στη συνέχεια την ανακατασκευή τους. Αυτή η τεχνική εκμάθησης χωρίς επίβλεψη μειώνει τις διαστάσεις και καταγράφει μη γραμμικές σχέσεις μέσα στα δεδομένα. Τέλος, εφαρμόστηκε η t-Distributed Stochastic Neighbor Embedding (t-SNE), μια τεχνική που σημειώθηκε για την αποτελεσματικότητά της στην οπτικοποίηση δεδομένων υψηλών διαστάσεων. Το t-SNE είναι έμπειρο στη διατήρηση της τοπικής δομής των δεδομένων, καθιστώντας το χρήσιμο για την αναγνώριση συστάδων εντός του συνόλου δεδομένων. Αυτές οι μέθοδοι μείωσης διαστάσεων στόχευαν στην απλοποίηση των σύνθετων δεδομένων εικόνας για πιο αποτελεσματική ανάλυση ομαδοποίησης στα επόμενα στάδια της μελέτης.

Στον αλγόριθμο SAE, η κλάση EarlyStopping από την Keras ενσωματώθηκε ως συνάρτηση επανάκλησης. Η πρόωρη διακοπή είναι ευεργετική για την πρόληψη της υπερβολικής προπόνησης και τη μείωση του χρόνου εκτέλεσης, καθώς διακόπτει την προπόνηση όταν δεν παρατηρείται περαιτέρω βελτίωση.

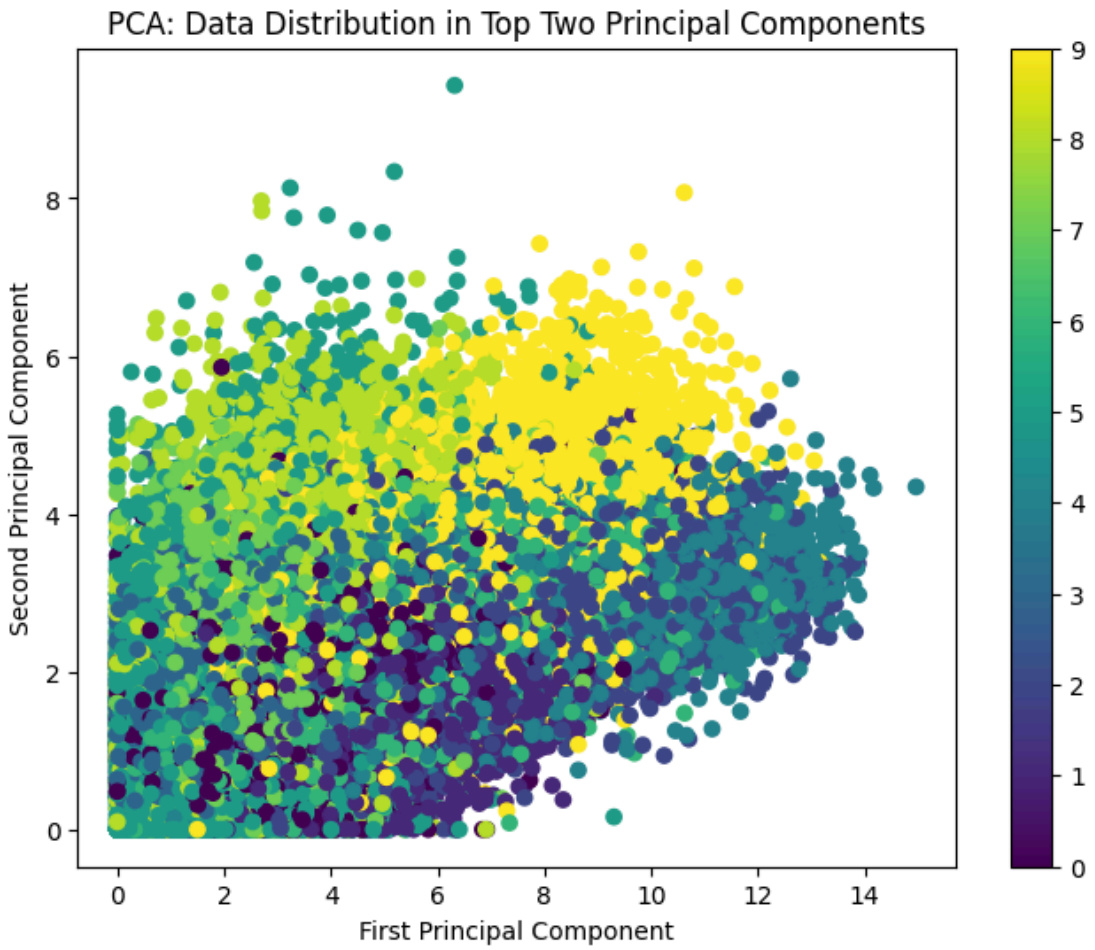
## 3.2 Οπτικοποίηση Αλγορίθμων Μείωσης Διαστάσεων:

### 3.2.1 Οπτικοποιήστε τα αποτελέσματα PCA ή SAE

Η συνάρτηση visualize_pca_sae_2D έχει σχεδιαστεί για να σχεδιάζει τα δεδομένα μειωμένα σε δύο διαστάσεις χρησιμοποιώντας είτε την ανάλυση κύριου στοιχείου (PCA) είτε έναν αυτόματο κωδικοποιητή στοίβαξης (SAE). Η γραφική παράσταση παρουσιάζει ένα δισδιάστατο γράφημα διασποράς όπου κάθε σημείο αντιπροσωπεύει ένα παράδειγμα δεδομένων. Ο άξονας x και ο άξονας y αντιστοιχούν στην πρώτη και δεύτερη κύρια συνιστώσα, αντίστοιχα, στην περίπτωση της PCA. Αυτά τα στοιχεία είναι οι κατευθύνσεις στις οποίες τα δεδομένα ποικίλλουν περισσότερο και είναι γραμμικοί συνδυασμοί των αρχικών χαρακτηριστικών. Για το SAE, αυτοί οι άξονες αντιπροσωπεύουν τις δύο κύριες διαστάσεις στον μειωμένο χώρο χαρακτηριστικών, καταγράφοντας σημαντικά μοτίβα ή χαρακτηριστικά από τα δεδομένα υψηλών διαστάσεων. Παρακάτω φαίνεται η υλοποίηση για PCA:




***Εικόνα 1: δεδομένα μειωμένα σε δύο διαστάσεις με PCA***

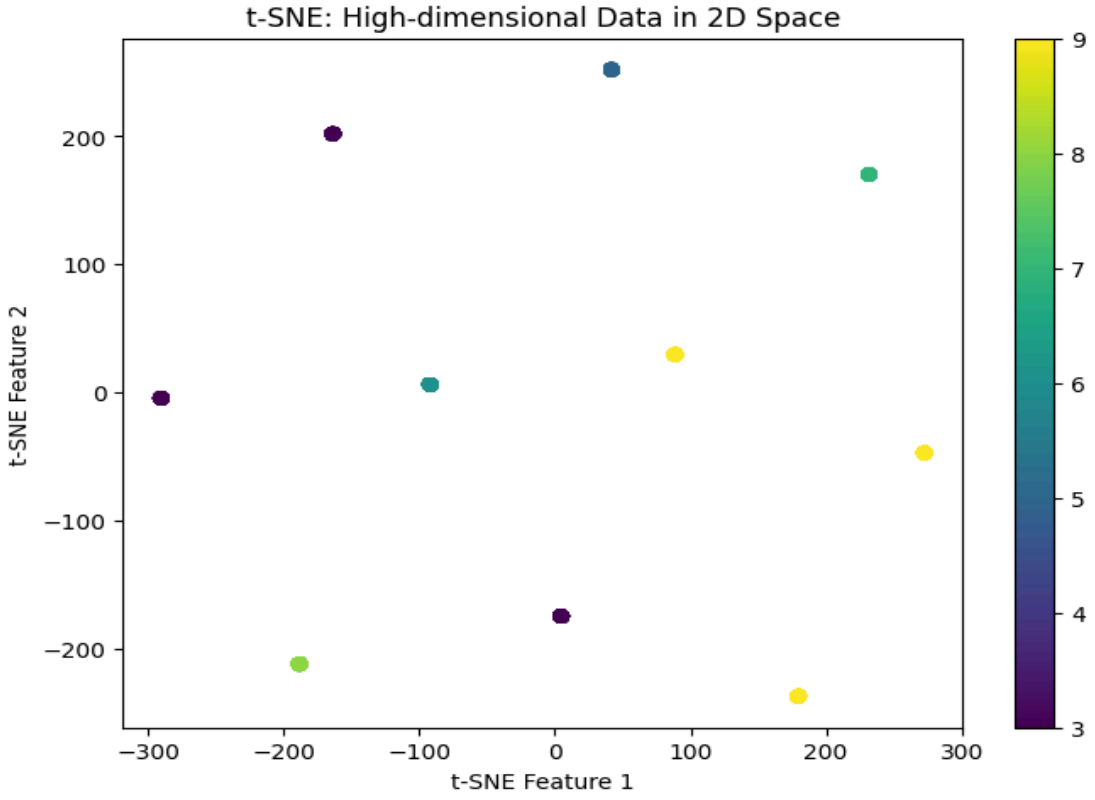



***Εικόνα 2: δεδομένα μειωμένα σε δύο διαστάσεις με SAE***

Αυτή η εικόνα χρησιμοποιεί χρωματική κωδικοποίηση για να αναπαραστήσει διαφορετικές κλάσεις ή κατηγορίες στο σύνολο δεδομένων, όπως ορίζεται από τις ετικέτες. Η χρήση χρωμάτων βοηθά στην αξιολόγηση της αποτελεσματικότητας της τεχνικής μείωσης διαστάσεων στον διαχωρισμό διαφορετικών κατηγοριών. Στην ιδανική περίπτωση, τα σημεία που ανήκουν στην ίδια κατηγορία θα πρέπει να σχηματίζουν συστάδες, υποδηλώνοντας ότι η τεχνική έχει καταγράψει με επιτυχία την υποκείμενη δομή των δεδομένων. Αυτή η οπτικοποίηση βοηθά στην κατανόηση των εγγενών προτύπων των δεδομένων και για την αξιολόγηση της απόδοσης του αλγορίμου SAE όσον αφορά τον διαχωρισμό κλάσεων.

### 3.2.2 Οπτικοποιήστε τα αποτελέσματα του t-SNE

Η συνάρτηση visualize_tsne_2D δημιουργεί μια οπτικοποίηση δεδομένων υψηλών διαστάσεων που μετατρέπονται σε δισδιάστατο χώρο χρησιμοποιώντας t-Distributed Stochastic Neighbor Embedding (t-SNE). Το t-SNE είναι ιδιαίτερα αποτελεσματικό στη διατήρηση της τοπικής δομής των δεδομένων, καθιστώντας το ένα ισχυρό εργαλείο για την οπτικοποίηση συμπλεγμάτων ή ομαδοποιήσεων. Σε αυτό το διάγραμμα διασποράς, κάθε σημείο αντιπροσωπεύει ένα μεμονωμένο στιγμιότυπο δεδομένων, με τους άξονες γραφικής παράστασης (T-SNE Feature 1 και t-SNE Feature 2) να αντιπροσωπεύουν τις δύο διαστάσεις που λαμβάνονται μετά τη μείωση t-SNE.




***Εικόνα 3: δεδομένα μειωμένα σε δύο διαστάσεις με t-sne***

Το χρώμα κάθε σημείου αντιστοιχεί στην ετικέτα κλάσης του, παρέχοντας μια οπτική ένδειξη του τρόπου με τον οποίο τα σημεία δεδομένων από την ίδια κατηγορία ομαδοποιούνται στον μειωμένο χώρο. Τα διαγράμματα t-SNE χρησιμοποιούνται συχνά για την οπτική αξιολόγηση της παρουσίας συστάδων ή μοτίβων σε δεδομένα υψηλών διαστάσεων, καθιστώντας ευκολότερο τον εντοπισμό εγγενών ομαδοποιήσεων ή δομών. Είναι ιδιαίτερα χρήσιμο για διερευνητική ανάλυση δεδομένων και για την κατανόηση των σχέσεων και των ομοιοτήτων μεταξύ σημείων δεδομένων σε ένα σύνολο δεδομένων.

### 3.2.4. Εμφάνιση πρωτότυπων και ανακατασκευασμένων εικόνων

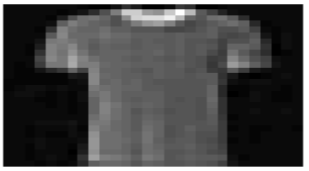
Η συνάρτηση display_images έχει σχεδιαστεί για σύγκριση πρωτότυπων και ανακατασκευασμένων εικόνων, ιδιαίτερα στην αξιολόγηση μοντέλων αυτόματου κωδικοποιητή όπως το SAE. Αυτή η συνάρτηση εμφανίζει ζεύγη εικόνων για κάθε κλάση στο σύνολο δεδομένων, με την αρχική εικόνα στα αριστερά και την ανακατασκευασμένη εκδοχή της στα δεξιά, μετά την επεξεργασία μέσω ενός μοντέλου μείωσης διαστάσεων όπως ένας αυτόματος κωδικοποιητής.

Αυτή η συγκριτική εικόνα είναι χρήσιμη για την οπτική αξιολόγηση της ποιότητας ανακατασκευής που επιτυγχάνεται από το μοντέλο. Μια επιτυχημένη ανακατασκευή υποδηλώνει ότι το μοντέλο έχει συλλάβει αποτελεσματικά τα κύρια χαρακτηριστικά και τα πρότυπα των δεδομένων. Για κάθε κλάση, η συνάρτηση αξιολογεί πόσο καλά το μοντέλο διατηρεί τα χαρακτηριστικά, κάτι που είναι απαραίτητο σε εφαρμογές όπως η αποθορυβοποίηση εικόνας, η ανίχνευση ανωμαλιών ή η εξαγωγή χαρακτηριστικών σε εργασίες κατάντη. Η σύγκριση πρωτότυπων και ανακατασκευασμένων εικόνων παρέχει ένα άμεσο οπτικό μέσο για τη μέτρηση της απόδοσης του μοντέλου στην εκμάθηση μιας συμπιεσμένης αλλά ακριβούς αναπαράστασης των δεδομένων.

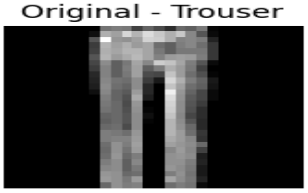
***Εικόνα 4:*** **σύγκριση πρωτότυπων και ανακατασκευασμένων εικόνων**

Original - T-shirt/top

Reconstructed - T-shirt/top



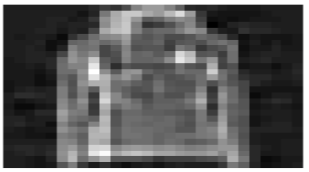
Original - Trouser



Reconstructed - Trouser

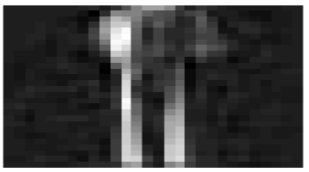
Original - Pullover

Reconstructed - Pullover



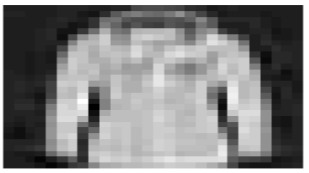
Original - Dress

Reconstructed - Dress



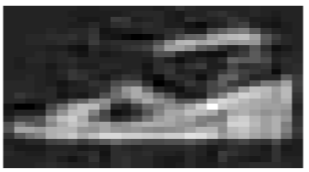
Original - Coat

Reconstructed - Coat



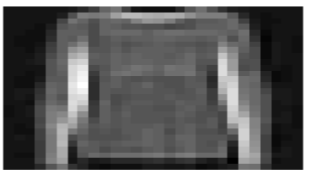
Original - Sandal

Reconstructed - Sandal



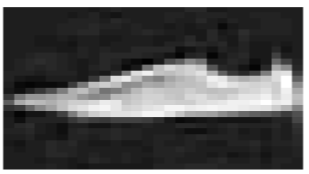
Original - Shirt

Reconstructed - Shirt



Original - Sneaker

Reconstructed - Sneaker

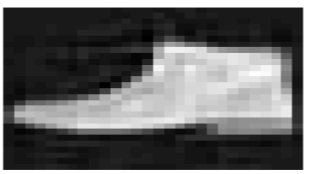


Original - Bag

Reconstructed - Bag

Original - Ankle boot

Reconstructed - Ankle boot



## 3.3 Ανάλυση Μοντέλων Ομαδοποίησης:

Μετά την προεπεξεργασία και τη μείωση των διαστάσεων του συνόλου δεδομένων Fashion MNIST, η μελέτη ανέλυσε την κρίσιμη φάση της ανάπτυξης και ανάπτυξης μοντέλων ομαδοποίησης. Επιλέχθηκαν τρεις αλγόριθμοι ομαδοποίησης για τα διακριτά χαρακτηριστικά και τα πλεονεκτήματά τους, για να αξιολογηθεί η απόδοσή τους στα μετασχηματισμένα δεδομένα.

Το MiniBatch KMeans, που επιλέχθηκε για την αποτελεσματικότητά του, είναι μια παραλλαγή του κλασικού αλγόριθμου KMeans, κατάλληλο για μεγάλα σύνολα δεδομένων λόγω του μειωμένου υπολογιστικού του κόστους. Αυτή η δυνατότητα είναι ευεργετική για το χειρισμό του σημαντικού αριθμού εικόνων στο σύνολο δεδομένων Fashion MNIST. Η ικανότητα του αλγορίθμου να ομαδοποιεί δεδομένα γρήγορα, η επεκτασιμότητα και η ταχύτητά του ήταν βασικοί τομείς εστίασης, μαζί με την απόδοσή του με διάφορες εισόδους χαμηλών διαστάσεων.

Ο δεύτερος αλγόριθμος, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), έρχεται σε αντίθεση με το MiniBatch KMeans καθώς δεν απαιτεί προκαθορισμένο αριθμό συστάδων. Σχηματίζει συμπλέγματα με βάση την πυκνότητα σημείων δεδομένων, επιτρέποντάς του να αναγνωρίζει συστάδες διαφορετικών σχημάτων και πυκνοτήτων, τυπικά σε σύνθετα σύνολα δεδομένων εικόνας. Η χρησιμότητα του DBSCAN ήταν στη δυνατότητά του να διακρίνει πιο λεπτές, ακανόνιστου σχήματος συμπλέγματα σε δεδομένα εικόνας, ειδικά με δεδομένα που μετασχηματίζονται με μεθόδους όπως το t-SNE που διατηρούν τις τοπικές δομές.

Η συγκεντρωτική ομαδοποίηση(Hierarchical Clustering), ο τρίτος αλγόριθμος, είναι μια ιεραρχική μέθοδος ομαδοποίησης. Διαφέρει από τους προηγούμενους αλγόριθμους δημιουργώντας μια ιεραρχία συμπλέγματος, που συχνά απεικονίζεται ως δενδρογράφημα. Αυτή η μέθοδος επιλέχθηκε για την ικανότητά της να αποκαλύπτει πολύπλοκες σχέσεις δεδομένων, προσφέροντας μια μοναδική προοπτική στην ομαδοποίηση δεδομένων εικόνας. Η εφαρμογή της Συσσωρευτικής Ομαδοποίησης(Agglomerative Clustering) σε δεδομένα χαμηλών διαστάσεων είχε ως στόχο να διερευνήσει πόσο καλά οι ιεραρχικές μέθοδοι θα μπορούσαν να χειριστούν την πολυπλοκότητα των δεδομένων εικόνας υψηλών διαστάσεων.

Η χρήση αυτών των αλγορίθμων σε δεδομένα που υποβλήθηκαν σε επεξεργασία μέσω PCA, SAE και t-SNE επέτρεψε τη διερεύνηση της αλληλεπίδρασης μεταξύ διαφορετικών τεχνικών μείωσης διαστάσεων και μεθόδων ομαδοποίησης. Αυτή η φάση ήταν κρίσιμη για την κατανόηση των συνεργειών και των συμβιβασμών κατά τον συνδυασμό αυτών των τεχνικών, προσφέροντας μια ολοκληρωμένη άποψη για το πώς κάθε συνδυασμός επηρεάζει τα αποτελέσματα της ομαδοποίησης. Ο στόχος ήταν να καθοριστούν οι βέλτιστες στρατηγικές για ομαδοποίηση εντός της σφαίρας δεδομένων εικόνας υψηλών διαστάσεων.

## 3.4 Αξιολόγηση Μετρικών Ομαδοποίησης:

Η φάση αξιολόγησης των μοντέλων ομαδοποίησης ήταν κρίσιμη, περιλαμβάνοντας μια εκτενή αξιολόγηση της αποτελεσματικότητας κάθε συνδυασμού τεχνικής μείωσης διαστάσεων και αλγορίθμου ομαδοποίησης. Χρησιμοποιήσαμε ένα σύνολο μετρήσεων για να προσφέρουμε ποικίλες πληροφορίες σχετικά με την απόδοση της ομαδοποίησης. Ο δείκτης Calinski-Harabasz ήταν ένα βασικό μέτρο, που αξιολογούσε τη διασπορά εντός των συστάδων σε σχέση με τη διασπορά μεταξύ των συστάδων. Υπολογίζει τον λόγο της διακύμανσης μεταξύ συστάδων προς διακύμανση εντός συστάδας για όλα τα συμπλέγματα, με υψηλότερες τιμές που υποδεικνύουν καλά διαχωρισμένα και πυκνά συσσωρευμένα συμπλέγματα. Αυτή η μέτρηση παρείχε πληροφορίες για τη συμπαγή και τη διακριτικότητα των συμπλεγμάτων.

Ο δείκτης Davies-Bouldin ήταν μια άλλη σημαντική μέτρηση, που αξιολογούσε τη μέση «ομοιότητα» μεταξύ κάθε συστάδας και του πιο παρόμοιου αντίστοιχου. Εδώ, η ομοιότητα μετρά την αναλογία της απόστασης μεταξύ των συστάδων προς το μέγεθος των συστάδων. Οι χαμηλότερες τιμές αυτού του δείκτη υποδηλώνουν καλύτερη ποιότητα ομαδοποίησης, με πιο απομακρυσμένες και λιγότερο διασκορπισμένες συστάδες.

Υπολογίσαμε επίσης τη βαθμολογία Silhouette για κάθε ρύθμιση ομαδοποίησης. Αυτή η βαθμολογία μετράει πόσο παρόμοιο είναι ένα αντικείμενο με το δικό του σύμπλεγμα (συνοχή) έναντι άλλων συστάδων (διαχωρισμός), που κυμαίνεται από -1 έως 1. Μια υψηλή βαθμολογία υποδηλώνει καλή αντιστοίχιση σε ένα σύμπλεγμα και κακή αντιστοίχιση με γειτονικά συμπλέγματα, βοηθώντας στην αξιολόγηση του καταλληλότητα ανάθεσης συμπλέγματος.

Επιπλέον, χρησιμοποιήθηκε ο Προσαρμοσμένος Δείκτης Rand (ARI), ιδιαίτερα χρήσιμος όταν είναι διαθέσιμη η βασική αλήθεια. Το ARI μετρά την ομοιότητα μεταξύ δύο αναθέσεων, λαμβάνοντας υπόψη την τυχαία κανονικοποίηση και αγνοώντας τις μεταθέσεις. Παρείχε ένα μετρήσιμο μέτρο για την αξιολόγηση του πόσο στενά τα αποτελέσματα της ομαδοποίησης αντικατοπτρίζουν την πραγματική κατανομή δεδομένων, δεδομένου ενός γνωστού συνόλου πραγματικών ετικετών.

Αυτές οι μετρήσεις προσέφεραν συλλογικά μια ολοκληρωμένη εικόνα της απόδοσης της ομαδοποίησης, τονίζοντας τα δυνατά σημεία και τους περιορισμούς κάθε συνδυασμού μείωσης διαστάσεων και τεχνικής ομαδοποίησης. Αυτή η προσέγγιση επέτρεψε τον εντοπισμό των πιο αποτελεσματικών μεθόδων για τη ομαδοποίηση δεδομένων υψηλών διαστάσεων.

## 3.5 Τελική Απόφαση και δημιουργία Dataframe:

Κατά την ανάλυση της απόδοσης ομαδοποίησης στο μειωμένο διαστάσεων σύνολο δεδομένων Fashion MNIST, προέκυψαν αρκετά βασικά ευρήματα. Κάθε αλγόριθμος ομαδοποίησης εμφάνιζε μοναδικά χαρακτηριστικά και αποτελεσματικότητες όταν εφαρμόστηκε στα διαφορετικά επεξεργασμένα δεδομένα.

Το MiniBatch KMeans, γνωστό για τη γρήγορη επεξεργασία του, δοκιμάστηκε για πρώτη φορά με δεδομένα μειωμένα με PCA. Σε αυτό το πλαίσιο, έδειξε υψηλή υπολογιστική αποτελεσματικότητα και αξιοσημείωτο διαχωρισμό συστάδων, όπως υποδεικνύεται από τις ισχυρές βαθμολογίες Calinski-Harabasz και Silhouette, υποδηλώνοντας καλά καθορισμένες συστάδες με σημαντικό διαχωρισμό. Ωστόσο, όταν το MiniBatch KMeans εφαρμόστηκε σε δεδομένα μειωμένα μέσω t-SNE, η απόδοσή του μειώθηκε. Αυτή η αλλαγή υποδεικνύει την ευαισθησία του MiniBatch KMeans στη δομή των δεδομένων εισόδου, ιδιαίτερα σε μη γραμμικούς μετασχηματισμούς από t-SNE που εστιάζουν στις τοπικές σχέσεις δεδομένων, ενδεχομένως σε βάρος της καθολικής δομής.

Αντίθετα, η απόδοση του DBSCAN ανέδειξε τα δυνατά του σημεία και την καταλληλότητά του για ορισμένους τύπους δεδομένων. Σε συνδυασμό με δεδομένα μειωμένα από t-SNE, το DBSCAN εντόπισε αποτελεσματικά συμπλέγματα διαφόρων σχημάτων και μεγεθών. Αυτή η επιτυχία οφείλεται στην ικανότητα του DBSCAN να διαχειρίζεται ακραίες τιμές και στην προσέγγιση ομαδοποίησης που βασίζεται στην πυκνότητα, η οποία συμπληρώνει τη διατήρηση των τοπικών δομών του t-SNE. Αυτός ο συνδυασμός υπογραμμίζει την ικανότητα του DBSCAN για πλοήγηση σε πολύπλοκους, μη γραμμικούς χώρους δεδομένων, καθιστώντας το κατάλληλο για σύνολα δεδομένων με περίπλοκες χωρικές σχέσεις.

Η συγκεντρωτική ομαδοποίηση, που χρησιμοποιείται με δεδομένα μειωμένα από το SAE, έδειξε ισορροπημένη απόδοση σε όλες τις μετρήσεις. Αυτός ο συνδυασμός ανέδειξε την προσαρμοστικότητα του Agglomerative Clustering και την αποτελεσματικότητά του στη δημιουργία μιας διαφοροποιημένης ιεραρχίας συστάδων, ιδιαίτερα με αναπαραστάσεις δεδομένων που προέρχονται από μεθόδους μείωσης διαστάσεων που βασίζονται σε νευρωνικά δίκτυα, όπως το SAE. Η ισορροπημένη μετρική απόδοση δείχνει ότι η ομαδοποίηση, όταν ενσωματώνεται με μια τεχνική που καταγράφει μη γραμμικές σχέσεις όπως το SAE, μπορεί να προσφέρει μια ολοκληρωμένη άποψη της δομής δεδομένων.

Αυτά τα αποτελέσματα υπογραμμίζουν τη σημασία της επιλογής του σωστού συνδυασμού τεχνικής μείωσης διαστάσεων και αλγορίθμου ομαδοποίησης. Η αλληλεπίδραση μεταξύ αυτών των μεθόδων επηρεάζει σημαντικά την αποτελεσματικότητα και την ακρίβεια της ομαδοποίησης, κάτι που είναι κρίσιμο για πρακτικές εφαρμογές. Αυτή η μελέτη όχι μόνο αποκαλύπτει τα πλεονεκτήματα και τους περιορισμούς διαφόρων συνδυασμών, αλλά προκαλεί επίσης μια ευρύτερη συζήτηση για την εξισορρόπηση της υπολογιστικής απόδοσης με την ακρίβεια ομαδοποίησης, ζωτικής σημασίας για την ενημερωμένη εφαρμογή τεχνικών μηχανικής μάθησης σε πολύπλοκα σύνολα δεδομένων εικόνων.

Επιπλέον, δημιουργήθηκε ένα DataFrame (results_df), το οποίο περιλάμβανε λεπτομερείς πληροφορίες για κάθε συνδυασμό τεχνικής μείωσης διαστάσεων και αλγόριθμου ομαδοποίησης, συμπεριλαμβανομένων των υπολογισμένων μετρήσεων απόδοσης για το καθένα.

| Dimensionality_Reduction_Method | Clustering_Method | Training_Time | Execution_Time | Num_Clusters | Calinski_Harabasz | Davies_Bouldin | Silhouette |
|---|---|---|---|---|---|---|---|
| PCA | MiniBatchKMeans | 2.388.932.943.344.110 | 0.293759584426 8799 | 10 | 19.953.582.743.529.300 | 15.535.612.514.876.900 | 0.20609 6621030 45703 |
| PCA | DBSCAN | 2.388.932.943.344.110 | 0.29123 6877441 40625 | 10 | 3.658.584.569.189.050 | 21.639.059.529.551.400 | -0.18184 8382827 053 |
| PCA | GMM | 2.388.932.943.344.110 | 3.632.476.329.803.460 | 10 | 11.909.972.788.327.900 | 24.843.010.126.279.000 | 0.12396 8941523 66363 |
| SAE | MiniBatchKMeans | 2.964.507.269.859.310 | 0.13924 2887496 94824 | 10 | 32.522.858.606.884.600 | 1.345.917.609.597.960 | 0.21683 0149292 94586 |
| SAE | DBSCAN | 2.964.507.269.859.310 | 0.21299 6959686 2793 | 10 | 1.832.170.815.343.800 | 13.858.972.654.752.900 | -0.50302 0107746 1243 |
| SAE | GMM | 2.964.507.269.859.310 | 34.258.623.123.168.900 | 10 | 15.720.461.165.227.300 | 18.698.831.077.965.600 | 0.08136 0682845 11566 |
| t-SNE | MiniBatchKMeans | 541.009.881.734.848 | 0.16055 4885864 2578 | 10 | 11.894.619.584.384.100 | 0.79290 1473519 8553 | 0.42484 0122461 31897 |
| t-SNE | DBSCAN | 541.009.881.734.848 | 0.09484 7440719 60449 | 10 | 26.512.483.787.139.500 | 19.248.662.261.373.900 | -0.05547 2020059 82399 |
| t-SNE | GMM | 541.009.881.734.848 | 0.27274 7993469 2383 | 10 | 9.801.527.263.981.180 | 0.89303 1363512 9037 | 0.37328 2909393 31055 |
| Raw | MiniBatchKMeans | 0.0 | 0.14140 7728195 19043 | 10 | 1.161.676.789.234.320 | 20.914.422.959.362.600 | 0.13271 7047263 11673 |
| Raw | DBSCAN | 0.0 | 24.445.550.441.741.900 | 10 | 29.815.018.439.345.800 | 15.561.973.469.207.800 | -0.35358 8673509 5865 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Raw | GMM | 0.0 | 157.208.975.315.094 | 10 | 10.314.820.511.992.700 | 2.062.825.379.225.450 | 0.11869260889496788 |

**Πίνακας I:Τελικά Αποτελέσματα μετρικών,μεθόδων μείωσης διαστάσεων και τεχνικών ομαδοποίησης**

## 3.6 Περιορισμούς Καλύτερου Μοντέλου:

Αυτή η έρευνα, περιεκτική στη μεθοδολογία της, συνάντησε αρκετές προκλήσεις και περιορισμούς που αξίζουν αναγνώρισης. Μια πρωταρχική πρόκληση ήταν η διαχείριση της υπολογιστικής πολυπλοκότητας, ιδιαίτερα αξιοσημείωτη με τη χρήση του t-SNE για μείωση διαστάσεων. Ενώ το t-SNE είναι αποτελεσματικό στη διατήρηση τοπικών δομών σε δεδομένα υψηλών διαστάσεων, απαραίτητα για τον εντοπισμό προτύπων σε πολύπλοκα σύνολα δεδομένων όπως το Fashion MNIST, απαιτεί σημαντικούς υπολογιστικούς πόρους. Αυτό το σενάριο υπογραμμίζει μια θεμελιώδη ανταλλαγή στην επιστήμη δεδομένων: εξισορρόπηση λεπτομερούς, ακριβούς αναπαράστασης δεδομένων έναντι πρακτικών υπολογιστικών περιορισμών.

Ένας άλλος περιορισμός ήταν η ευαισθησία των αλγορίθμων ομαδοποίησης στις υπερπαραμέτρους τους, ιδιαίτερα εμφανής με το DBSCAN. Η απόδοση του DBSCAN βασίζεται σε μεγάλο βαθμό στον ακριβή συντονισμό των παραμέτρων πυκνότητάς του, όπου η εσφαλμένη επιλογή παραμέτρων μπορεί να οδηγήσει σε μη βέλτιστα αποτελέσματα ομαδοποίησης. Αυτό το ζήτημα υπογραμμίζει την κρίσιμη φύση του συντονισμού υπερπαραμέτρων στους αλγόριθμους ομαδοποίησης και την ανάγκη για εκτεταμένους πειραματισμούς για τον καθορισμό των βέλτιστων ρυθμίσεων.

Η υψηλή διάσταση του ίδιου του συνόλου δεδομένων Fashion MNIST αποτελούσε μια σημαντική πρόκληση. Αυτό απαιτούσε τη χρήση ισχυρών τεχνικών μείωσης διαστάσεων για την αποτελεσματική απλοποίηση των δεδομένων διατηρώντας τα πολύπλοκα χαρακτηριστικά τους. Η διαχείριση αυτής της πτυχής ήταν ζωτικής σημασίας για την επιτυχία της φάσης ομαδοποίησης.

Επιπλέον, η εστίαση στο σύνολο δεδομένων Fashion MNIST εγείρει ερωτήματα σχετικά με τη γενίκευση των ευρημάτων. Αν και το σύνολο δεδομένων είναι ποικίλο και πολύπλοκο, αντιπροσωπεύει μια συγκεκριμένη κατηγορία δεδομένων εικόνας. Η επέκταση της δυνατότητας εφαρμογής αυτών των αποτελεσμάτων σε άλλα σύνολα δεδομένων ή ευρύτερα περιβάλλοντα δεδομένων εικόνας απαιτεί περαιτέρω εξερεύνηση.

Αυτές οι προκλήσεις υπογραμμίζουν τη σημασία της προσεκτικής επιλογής και συντονισμού των τεχνικών μείωσης διαστάσεων και των αλγορίθμων ομαδοποίησης. Η μελέτη υποστηρίζει τις συνεχείς προσπάθειες για την ανάπτυξη πιο αποτελεσματικών και ευέλικτων μεθόδων για το χειρισμό δεδομένων υψηλών διαστάσεων. Η μελλοντική έρευνα θα μπορούσε να διερευνήσει ένα ευρύτερο φάσμα

συνόλων δεδομένων με ποικίλα χαρακτηριστικά και προκλήσεις και να διερευνήσει νέες, πιο αποδοτικές υπολογιστικά τεχνικές μείωσης διαστάσεων. Τέτοιες πρωτοβουλίες στοχεύουν να ξεπεράσουν τους περιορισμούς που προσδιορίζονται σε αυτή τη μελέτη και να ενισχύσουν την κατανόηση και την εφαρμογή των μεθόδων ομαδοποίησης στην ανάλυση δεδομένων εικόνας ευρύτερα.

# 4. Μελλοντικές Επεκτάσεις Έρευνας:

## 4.1 Μελλοντική Έρευνα για Τεχνικές Μείωσης Διαστάσεων:

Το πεδίο της μείωσης διαστάσεων προσφέρει σημαντικές δυνατότητες για περαιτέρω διερεύνηση, ιδιαίτερα όσον αφορά τα σύνολα δεδομένων εικόνας υψηλών διαστάσεων όπως το Fashion MNIST. Μελλοντικές μελέτες θα μπορούσαν να εξετάσουν αναδυόμενες τεχνικές όπως η Ενιαία Προσέγγιση και Προβολή πολλαπλών (UMAP). Το UMAP είναι αξιοσημείωτο για την ικανότητά του να διατηρεί τοπικές και παγκόσμιες δομές δεδομένων, παρέχοντας ταυτόχρονα υπολογιστική απόδοση. Επιπλέον, οι εξελίξεις στις αρχιτεκτονικές αυτόματων κωδικοποιητών, συμπεριλαμβανομένων των μεταβλητών αυτοκωδικοποιητών (VAEs) και των αυτοκωδικοποιητών που χρησιμοποιούν γενετικά αντίπαλα δίκτυα (GANs), παρουσιάζουν πολλά υποσχόμενες οδούς. Αυτά τα προηγμένα μοντέλα νευρωνικών δικτύων μπορεί να προσφέρουν νέες μεθόδους για την εξαγωγή ουσιαστικών, συμπιεσμένων αναπαραστάσεων δεδομένων εικόνας, αποκαλύπτοντας πιθανώς πιο περίπλοκα μοτίβα και συστάδες. Επιπλέον, η ενσωμάτωση αυτών των εξελιγμένων τεχνικών μείωσης διαστάσεων με συνελικτικά νευρωνικά δίκτυα (CNN) θα μπορούσε να οδηγήσει σε πιο αποτελεσματικές διαδικασίες εξαγωγής χαρακτηριστικών, ειδικά σχεδιασμένες για δεδομένα εικόνας. Μια τέτοια έρευνα θα μπορούσε να βελτιώσει την κατανόηση της μείωσης διαστάσεων στην ανάλυση εικόνας και να συμβάλει στην ανάπτυξη πιο αποτελεσματικών και ακριβών μοντέλων ομαδοποίησης, προάγοντας τη μάθηση χωρίς επίβλεψη στην ανάλυση δεδομένων εικόνας.

## 4.2 Βελτιώσεις σε αλγόριθμους ομαδοποίησης και βελτιστοποίηση υπερπαραμέτρων:

Όσον αφορά τους αλγόριθμους ομαδοποίησης και τη βελτιστοποίηση υπερπαραμέτρων, το εξελισσόμενο τοπίο των μεθοδολογιών ομαδοποίησης παρουσιάζει πολλές ευκαιρίες για έρευνα. Η διερεύνηση προηγμένων τεχνικών ομαδοποίησης, όπως η φασματική ομαδοποίηση ή το βελτιωμένο DBSCAN με βελτιστοποιημένες παραμέτρους, θα μπορούσε να βελτιώσει τον χειρισμό της πολυπλοκότητας δεδομένων εικόνας. Η ενσωμάτωση της ομαδοποίησης βασισμένης στη βαθιά μάθηση, η οποία συγχωνεύει τη μάθηση χαρακτηριστικών με την ομαδοποίηση, μπορεί να προσφέρει μια πιο ολιστική και αποτελεσματική προσέγγιση. Ταυτόχρονα, το πεδίο της βελτιστοποίησης υπερπαραμέτρων στη ομαδοποίηση απαιτεί περαιτέρω εξερεύνηση. Οι αυτοματοποιημένες μέθοδοι όπως η αναζήτηση πλέγματος, η τυχαία αναζήτηση ή προηγμένες τεχνικές όπως η Bayesian βελτιστοποίηση θα μπορούσαν να βελτιώσουν τις παραμέτρους του αλγορίθμου ομαδοποίησης, ενισχύοντας ενδεχομένως την απόδοσή τους. Η διερεύνηση προσαρμοστικών αλγορίθμων που προσαρμόζουν τις παραμέτρους τους σε απόκριση στα χαρακτηριστικά των δεδομένων θα μπορούσε να οδηγήσει σε πιο ευέλικτες και καθολικά εφαρμόσιμες μεθόδους ομαδοποίησης. Οι εξελίξεις σε αυτόν τον τομέα έχουν τη δυνατότητα να αυξήσουν σημαντικά την ακρίβεια και την

αποτελεσματικότητα των προσεγγίσεων ομαδοποίησης, αυξάνοντας την προσαρμοστικότητά τους σε διάφορους τύπους δεδομένων υψηλών διαστάσεων και επεκτείνοντας έτσι τη χρησιμότητά τους στο δυναμικό πεδίο της επιστήμης δεδομένων.

# 5. Σύνοψη:

Αυτή η έρευνα σηματοδοτεί μια σημαντική συμβολή στον τομέα της μηχανικής μάθησης, ιδιαίτερα στη ομαδοποίηση δεδομένων εικόνας υψηλών διαστάσεων. Εστιάζοντας στο σύνολο δεδομένων Fashion MNIST, η μελέτη διευκρίνισε την περίπλοκη αλληλεπίδραση μεταξύ των τεχνικών μείωσης διαστάσεων και των αλγορίθμων ομαδοποίησης. Η εκτεταμένη ανάλυσή μας δείχνει ότι ενώ απλούστερες τεχνικές όπως το PCA προσφέρουν υπολογιστική απόδοση, μπορεί να μην καταγράφουν πάντα αποτελεσματικά πολύπλοκες σχέσεις στα δεδομένα, σε αντίθεση με πιο προηγμένες μεθόδους. Αντίθετα, εξελιγμένες μέθοδοι όπως το t-SNE, παρά την υπολογιστική τους ένταση, έχουν επιδείξει αξιοσημείωτη ικανότητα στην ενεργοποίηση ουσιαστικών αποτελεσμάτων ομαδοποίησης, ειδικά όταν συνδυάζονται με αλγόριθμους που αξιοποιούν την ικανότητα του t-SNE να διατηρεί τις τοπικές δομές.

Η μελέτη υπογραμμίζει τη σημασία της προσεκτικής επιλογής και συντονισμού τεχνικών μείωσης διαστάσεων με αλγόριθμους ομαδοποίησης. Παρατηρήσαμε ότι η αλληλεπίδραση μεταξύ αυτών των στοιχείων επηρεάζει σημαντικά την αποτελεσματικότητα της ομαδοποίησης. Αυτό υποδηλώνει ότι μια καθολική προσέγγιση στην ανάλυση δεδομένων δεν είναι εφικτή. Η έρευνά μας πλοηγείται σε αυτές τις πολυπλοκότητες και παρέχει πληροφορίες για τη βελτιστοποίηση διαφορετικών συνδυασμών για πιο ακριβή και αποτελεσματικά αποτελέσματα ομαδοποίησης.

Ωστόσο, η έρευνα έχει περιορισμούς. Οι υπολογιστικές απαιτήσεις, ιδιαίτερα που σχετίζονται με μεθόδους όπως το t-SNE, αποτελούν σημαντική πρόκληση και χώρο για μελλοντική βελτίωση. Επιπλέον, η εστίαση στο σύνολο δεδομένων Fashion MNIST, ενώ είναι ενημερωτική, εγείρει ερωτήματα σχετικά με τη δυνατότητα εφαρμογής των ευρημάτων μας σε άλλα σύνολα δεδομένων εικόνων ή ευρύτερα περιβάλλοντα. Ωστόσο, οι γνώσεις και η πρόοδος που επιτεύχθηκε σε αυτήν τη μελέτη δημιουργούν μια σταθερή βάση για περαιτέρω εξερεύνηση, προάγοντας το πεδίο και καθοδηγώντας τη μελλοντική έρευνα για την αντιμετώπιση αυτών των προκλήσεων και τη διεύρυνση του πεδίου εφαρμογής των τεχνικών ομαδοποίησης στην ανάλυση δεδομένων εικόνας.

Οι εξελίξεις από αυτήν την έρευνα εμπλουτίζουν σημαντικά τον ευρύτερο κλάδο της μηχανικής μάθησης. Προσφέρουν ένα πλαίσιο για μελλοντικές μελέτες που στοχεύουν στις πολύπλοκες προκλήσεις της ανάλυσης δεδομένων εικόνας και ανοίγουν το δρόμο για την ανάπτυξη πιο εξελιγμένων και προσαρμόσιμων μεθόδων ομαδοποίησης, ευθυγραμμισμένες με τη δυναμική φύση του πεδίου της επιστήμης δεδομένων.

# Βιβλιογραφία:


1. Xiao, Han, Kashif Rasul, and Roland Vollgraf. "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms." arXiv preprint arXiv:1708.07747 (2017).

2. Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing Data using t-SNE." Journal of Machine Learning Research 9.Nov (2008): 2579-2605.

3. Abdi, H., and L.J. Williams. "Principal component analysis." Wiley Interdisciplinary Reviews: Computational Statistics 2.4 (2010): 433-459.

4. Vincent, Pascal, et al. "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion." Journal of Machine Learning Research 11.Dec (2010): 3371-3408.

5. Sculley, D. "Web-Scale K-Means Clustering." Proceedings of the 19th International Conference on World Wide Web (2010): 1177-1178.

6. Ester, Martin, et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." Kdd Vol. 96. No. 34. 1996.

7. Rousseeuw, Peter J. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.
8. Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." Communications in Statistics 3.1 (1974): 1-27.

9. Davies, David L., and Donald W. Bouldin. "A Cluster Separation Measure." IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1.2 (1979): 224-227.

10. Hubert, Lawrence, and Phipps Arabie. "Comparing partitions." Journal of Classification 2.1 (1985): 193-218.

11. Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research 12 (2011): 2825-2830.

12. Abadi, Martín, et al. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." (2015). Software available from tensorflow.org.

13. McInnes, Leland, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv preprint arXiv:1802.03426 (2018).

# Παραρτήματα:

## Κατάλογος Εικονών:



## Κατάλογος Πινάκων: