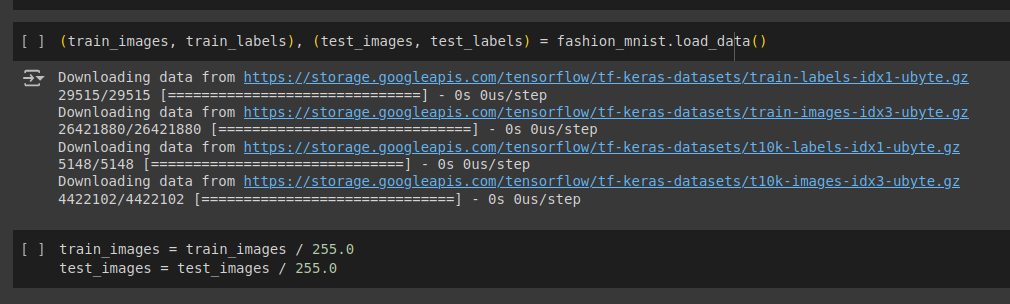
Η δεύτερη εργασία στο πλαίσιο του προγράμματος μεταπτυχιακών σπουδών στην "Εφαρμοσμένη Πληροφορική", με θέμα τις μεθόδους και εργαλεία τεχνητής νοημοσύνης, ασχολείται με την αξιολόγηση διαφορετικών συνδυαστικών μοντέλων μείωσης διάστασης χώρου και συσταδοποίησης επί εικόνων του fashion-mnist dataset. Η εργασία απαιτεί την υλοποίηση διάφορων τεχνικών μέσω γλώσσας Python, συμπεριλαμβανομένων των Principal Component Analysis (PCA) και Stacked Autoencoder (SAE) για τη μείωση διάστασης, καθώς και τη χρήση διαφορετικών τεχνικών συσταδοποίησης και αξιολόγησης.

Ξεκινώντας θα χωρίσουμε την εργασία σε 3 κεφάλαια, ότι αφορά τα Raw δεδομένα την τεχνική PCA και SAE. Ας δούμε όμως πριν από τις υλοποιήσεις πως έγινε η φόρτωση των δεδομένων.



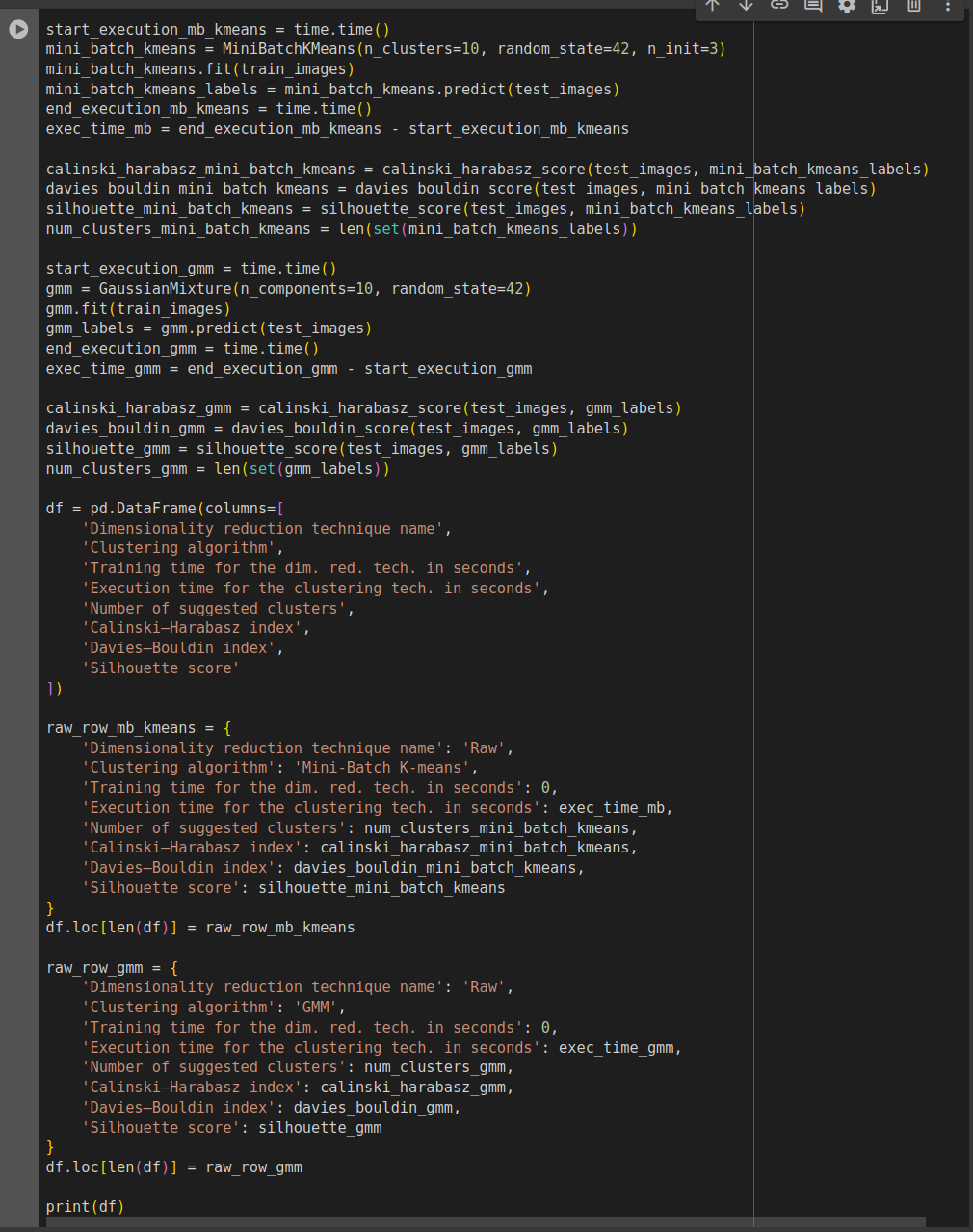
Αυτό το τμήμα κώδικα φορτώνει το σύνολο δεδομένων fashion-mnist και κανονικοποιεί τις τιμές των pixel των εικόνων, ώστε να είναι στο εύρος [0,1]. Η κανονικοποίηση αυτή είναι σημαντική για την επεξεργασία των δεδομένων πριν την εφαρμογή των αλγορίθμων μηχανικής μάθησης, καθώς βοηθά στη σύγκλιση του μοντέλου και στην αποφυγή της υπερ-εκπαίδευσης.

# **Raw Data**

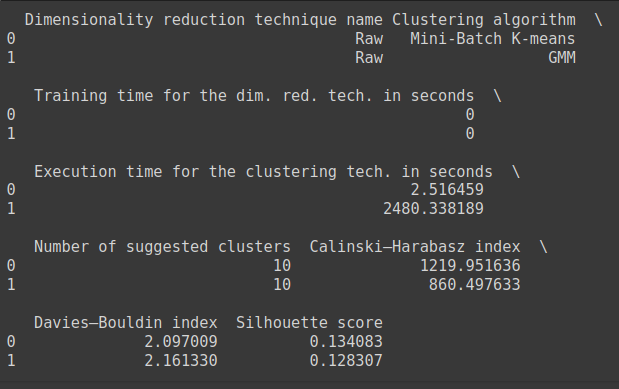
Στην παρακάτω κώδικα εκτελούνται διάφορες διαδικασίες για την αξιολόγηση των αλγορίθμων συσταδοποίησης στο dataset εικόνων. Αρχικά, χρησιμοποιείται ο αλγόριθμος Mini-Batch K-means για τη συσταδοποίηση των εικόνων εκπαίδευσης και την εφαρμογή του αλγορίθμου στα δεδομένα ελέγχου. Έπειτα, υπολογίζονται διάφορες μετρικές απόδοσης για τον Mini-Batch K-means, όπως ο δείκτης Calinski–Harabasz, ο δείκτης Davies–Bouldin και ο Silhouette score.

Στη συνέχεια, εφαρμόζεται ο αλγόριθμος Gaussian Mixture Model (GMM) στα δεδομένα εκπαίδευσης και ελέγχου, και υπολογίζονται αντίστοιχα οι μετρικές απόδοσης για τον GMM.

Τέλος, τα αποτελέσματα των μετρικών απόδοσης για κάθε αλγόριθμο καταχωρούνται σε ένα πίνακα δεδομένων (dataframe) του Pandas, προκειμένου να γίνει η σύγκρισή τους. Ο πίνακας περιλαμβάνει πληροφορίες όπως τον χρόνο εκτέλεσης των αλγορίθμων, τον αριθμό των προτεινόμενων clusters και τις μετρικές απόδοσης.



Η ανάλυση των αποτελεσμάτων από την εφαρμογή των τεχνικών συσταδοποίησης Mini-Batch K-means και Gaussian Mixture Model (GMM) πάνω στο raw σύνολο δεδομένων fashion-mnist αποκαλύπτει σημαντικές διαφορές στην απόδοση και στην αποδοτικότητα των δύο μεθόδων. Η τεχνική Mini-Batch K-means παρουσιάζει σημαντικά χαμηλότερο χρόνο εκτέλεσης (2.52 δευτερόλεπτα) σε σύγκριση με τον GMM (2480.34 δευτερόλεπτα), κάνοντας την πρώτη μέθοδο σαφώς πιο αποδοτική για μεγάλα σύνολα δεδομένων. Παράλληλα, οι μετρικές απόδοσης Calinski–Harabasz, Davies–Bouldin και Silhouette υποδεικνύουν ότι το Mini-Batch K-means επιτυγχάνει καλύτερη ποιότητα συσταδοποίησης με υψηλότερο Calinski–Harabasz index (1219.95 έναντι 860.50) και χαμηλότερο Davies–Bouldin index (2.10 έναντι 2.16), ενώ η διαφορά στο Silhouette score είναι ελάχιστη (0.134 έναντι 0.128). Συμπερασματικά, το Mini-Batch K-means όχι μόνο είναι πιο γρήγορο, αλλά και αποδίδει καλύτερα σε όρους ποιότητας συσταδοποίησης στο συγκεκριμένο dataset.



# **PCA**

Αρχικά, τα δεδομένα εικόνας του συνόλου Fashion MNIST χωρίζονται σε δύο σύνολα: ένα σύνολο εκπαίδευσης και ένα σύνολο επικύρωσης, με αναλογία 80%-20%. Αυτό γίνεται για να εξασφαλίσουμε ότι θα έχουμε επαρκή δεδομένα για την εκπαίδευση των μοντέλων μας και ταυτόχρονα να μπορούμε να αξιολογήσουμε την απόδοσή τους σε ένα ανεξάρτητο σύνολο δεδομένων.

Η μέθοδος PCA (Principal Component Analysis) χρησιμοποιείται για να μειώσει τις διαστάσεις των δεδομένων από 784 (28x28 pixels) σε 50. Αυτό βοηθά στη μείωση της πολυπλοκότητας των δεδομένων, ενώ παράλληλα διατηρεί τις πιο σημαντικές πληροφορίες. Με τη χρήση του PCA, καταγράφεται και ο χρόνος που απαιτείται για την εκπαίδευση του μοντέλου. Επιπλέον, αναπαρίσταται οπτικά η διαφορά μεταξύ των αρχικών εικόνων και των ανακατασκευασμένων εικόνων μετά την εφαρμογή του PCA, δείχνοντας πόσο καλά διατηρούνται οι κύριες πληροφορίες.

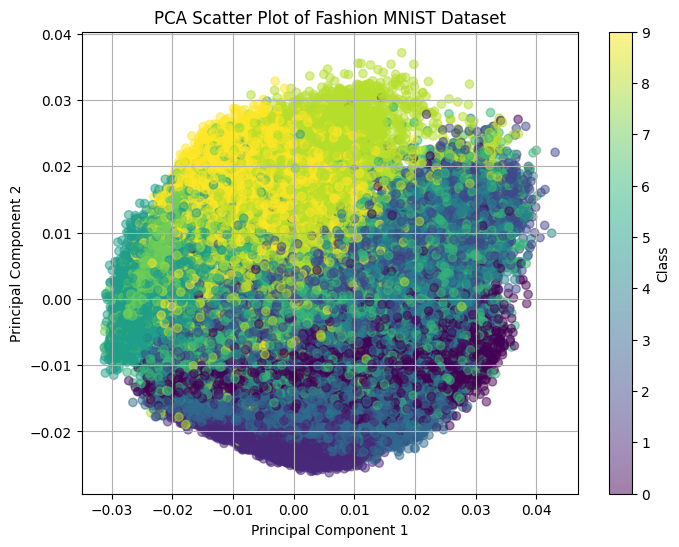
Μετά την εφαρμογή του PCA, οι δύο πρώτες κύριες συνιστώσες (principal components) χρησιμοποιούνται για να δημιουργηθεί ένα scatter plot. Αυτή η οπτικοποίηση βοηθά στο να δούμε πώς κατανέμονται τα δεδομένα στον νέο μειωμένο χώρο χαρακτηριστικών και αν υπάρχουν εμφανή clusters. Δύο διαφορετικές τεχνικές clustering, το Mini-Batch K-means και το Gaussian Mixture Model (GMM), εφαρμόζονται στα δεδομένα μετά το PCA. Οι αλγόριθμοι αυτοί προσπαθούν να αναγνωρίσουν φυσικά clusters στα δεδομένα. Ο χρόνος εκτέλεσης για κάθε αλγόριθμο καταγράφεται για να έχουμε μια σύγκριση της αποδοτικότητάς τους. Για την αξιολόγηση των αποτελεσμάτων των clusters χρησιμοποιούνται τρία metrics: το Calinski-Harabasz index, το Davies-Bouldin index και το Silhouette score. Αυτά τα metrics παρέχουν μια ποσοτική μέτρηση της ποιότητας των clusters που δημιουργήθηκαν από τους αλγόριθμους.

Τα τελικά αποτελέσματα, συμπεριλαμβανομένων των χρόνων εκτέλεσης, του αριθμού των προτεινόμενων clusters και των τιμών των αξιολογητικών metrics, καταγράφονται σε έναν πίνακα δεδομένων (DataFrame). Αυτό επιτρέπει μια οργανωμένη και συγκριτική ανάλυση των μεθόδων.

Tα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης (48,000 εικόνες) και επικύρωσης (12,000 εικόνες). Ο χρόνος εκπαίδευσης του PCA ήταν 5.55 δευτερόλεπτα και οι διαστάσεις των εικόνων μειώθηκαν από 784 σε 50 χαρακτηριστικά. Μετά την εφαρμογή του PCA, ορισμένες τυχαίες εικόνες ανακατασκευάστηκαν για να αξιολογηθεί η ικανότητα διατήρησης των κύριων πληροφοριών των εικόνων. Οι ανακατασκευασμένες εικόνες έδειξαν ότι το PCA μπορεί να διατηρήσει τις κύριες

Ο χρόνος εκτέλεσης για τον Mini-Batch K-means ήταν 0.49 δευτερόλεπτα, ενώ για τον GMM ήταν 36.35 δευτερόλεπτα. Οι ετικέτες των clusters που παράγονται από τους δύο αλγόριθμους παρουσίασαν διαφορετικές ταξινομήσεις των δεδομένων. Για την αξιολόγηση της ποιότητας των clusters χρησιμοποιήθηκαν ξανά οι τρεις δείκτες: ο Calinski-Harabasz index (1747.64 για Mini-Batch K-means και 1123.81 για GMM), ο Davies-Bouldin index (1.65 για Mini-Batch K-means και 2.44 για GMM), και ο Silhouette score (0.19 για Mini-Batch K-means και 0.16 για GMM).

Τα αποτελέσματα δείχνουν ότι ο Mini-Batch K-means παρήγαγε καλύτερα clusters σε σύγκριση με τον GMM, όπως φαίνεται από τους δείκτες αξιολόγησης. Ο Mini-Batch K-means είχε υψηλότερο Calinski-Harabasz index και χαμηλότερο Davies-Bouldin index, που υποδηλώνουν καλύτερη πυκνότητα και διαχωριστικότητα των clusters. Ο χρόνος εκτέλεσης του Mini-Batch K-means ήταν επίσης σημαντικά χαμηλότερος από τον χρόνο του GMM, καθιστώντας τον πιο αποδοτικό.



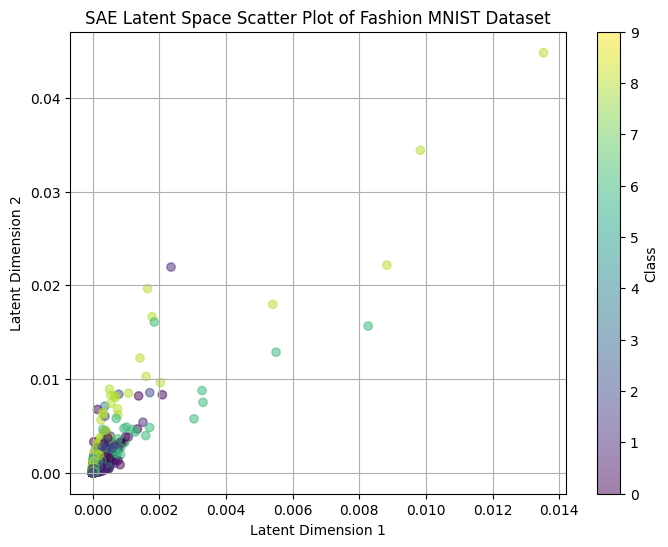
# **SAE**

Τα δεδομένα του συνόλου Fashion MNIST φορτώθηκαν και κανονικοποιήθηκαν ώστε ξανά οι τιμές των pixel να είναι μεταξύ 0 και 1. Τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης (48,000 εικόνες) και επικύρωσης (12,000 εικόνες). Έπειτα, σχεδιάστηκε και εκπαιδεύτηκε ένας αυτόματος κωδικοποιητής (SAE) για την ανάλυση των δεδομένων. Ο SAE περιλαμβάνει τρία επίπεδα κωδικοποίησης και τρία επίπεδα αποκωδικοποίησης, με τον συνολικό χρόνο εκπαίδευσης να εξαρτάται από τις εποχές και τα δεδομένα εκπαίδευσης. Η εκπαίδευση πραγματοποιήθηκε για 10 εποχές με batch size 128 και χρησιμοποιήθηκε η συνάρτηση απώλειας binary crossentropy.

Κατά τη διάρκεια της εκπαίδευσης, οι τιμές απώλειας και για τα σύνολα εκπαίδευσης και επικύρωσης μειώθηκαν, υποδηλώνοντας ότι ο SAE μάθαινε να αναπαράγει τις εικόνες με επιτυχία. Μετά την εκπαίδευση, οι εικόνες από το σύνολο δοκιμών κωδικοποιήθηκαν και αποκωδικοποιήθηκαν μέσω του SAE για να παρατηρηθούν οι ανακατασκευασμένες εικόνες. Οι ανακατασκευασμένες εικόνες παρουσίασαν αρκετά καλή ποιότητα, διατηρώντας τις κύριες πληροφορίες των αρχικών εικόνων.

Στη συνέχεια, τα δεδομένα εκπαίδευσης και δοκιμών κωδικοποιήθηκαν στον λανθάνοντα χώρο του SAE και πραγματοποιήθηκε clustering χρησιμοποιώντας τους αλγόριθμους Mini-Batch K-means και Gaussian Mixture Model (GMM).

Το μοντέλο περιλαμβάνει έναν αυτόματο κωδικοποιητή με τρία επίπεδα κωδικοποίησης και τρία επίπεδα αποκωδικοποίησης. Ο συνολικός αριθμός παραμέτρων είναι 222,384. Κατά την εκπαίδευση, η απώλεια μειώθηκε στο σύνολο εκπαίδευσης και το σύνολο επικύρωσης για 10 εποχές. Ο χρόνος εκπαίδευσης ανέρχεται σε περίπου 50 δευτερόλεπτα. Οι εκτιμήσεις απόδοσης του clustering δείχνουν ότι ο Mini-Batch K-means παρουσιάζει καλύτερα αποτελέσματα σε σχέση με τον GMM, με υψηλότερο Calinski-Harabasz index και χαμηλότερο Davies-Bouldin index. Επιπλέον, ο Mini-Batch K-means ολοκληρώθηκε σε περίπου 1.52 δευτερόλεπτα, ενώ ο GMM χρειάστηκε περίπου 1244.21 δευτερόλεπτα για την εκτέλεσή του.



# **Συμπεράσματα**

Με βάση την ανάλυση των αποτελεσμάτων της εφαρμογής των διαφόρων τεχνικών συσταδοποίησης στο σύνολο δεδομένων fashion-mnist, μπορούμε να συμπεράνουμε ότι ο αλγόριθμος Mini-Batch K-means εμφανίζει υψηλότερη απόδοση και αποδοτικότητα σε σχέση με τον αλγόριθμο Gaussian Mixture Model (GMM). Παρά το γεγονός ότι ο GMM παρουσιάζει υψηλότερο χρόνο εκτέλεσης, ο Mini-Batch K-means προσφέρει καλύτερη ποιότητα συσταδοποίησης, καθώς επιτυγχάνει υψηλότερες τιμές στους δείκτες Calinski-Harabasz και Davies-Bouldin, ενώ ο χρόνος εκτέλεσής του είναι σημαντικά χαμηλότερος. Τα αποτελέσματα αυτά υποδεικνύουν την αποτελεσματικότητα του Mini-Batch K-means στην επεξεργασία μεγάλων συνόλων δεδομένων, καθιστώντας τον προτιμότερη επιλογή για τη συσταδοποίηση εικόνων όπως αυτές του fashion-mnist dataset.

Στην περίπτωση αυτή, το PCA είναι καλύτερο από το SAE (Sparse Autoencoder) γιατί προσφέρει καλύτερή απόδοση στη διατήρηση της δομής των δεδομένων ενώ ταυτόχρονα μειώνει τη διαστατικότητα του χώρου χαρακτηριστικών. Το PCA είναι αποτελεσματικό στη μείωση των διαστάσεων των δεδομένων διατηρώντας σημαντικές πληροφορίες, ενώ το SAE μπορεί να έχει προβλήματα με την εκμάθηση αποδοτικών χαρακτηριστικών σε περιπτώσεις όπως αυτή του fashion-mnist dataset, όπου τα δεδομένα είναι λιγότερο δομημένα και δεν εκφράζονται εύκολα με σπάνιες αναπαραστάσεις. Επιπλέον, το PCA είναι γρηγορότερο στην εκτέλεσή του σε μεγάλα σύνολα δεδομένων σε σύγκριση με το SAE, κάνοντάς το πρακτικότερη επιλογή σε αυτήν την περίπτωση.