ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Αναφορά 2ης εργαστηριακής άσκησης 2025

ΙΩΑΝΝΗΣ ΜΠΟΥΖΑΣ
ΑΜ:5025

Εισαγωγή

Ο σκοπός αυτής της εργασίας είναι η διερεύνηση του τρόπου με τον οποίο η μείωση διαστάσεων επηρεάζει την λύση της ομαδοποίησης σε δεδομένα υψηλής διάστασης.

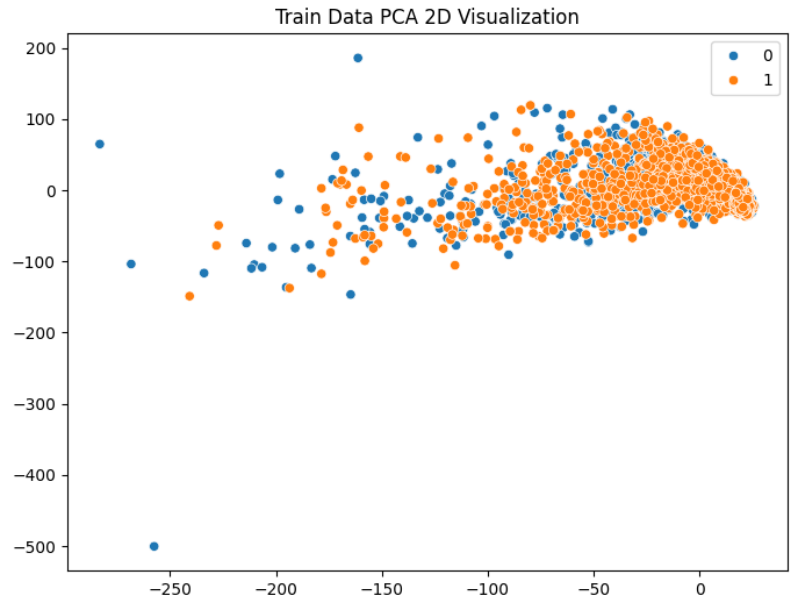
Μείωση Διαστάσεων

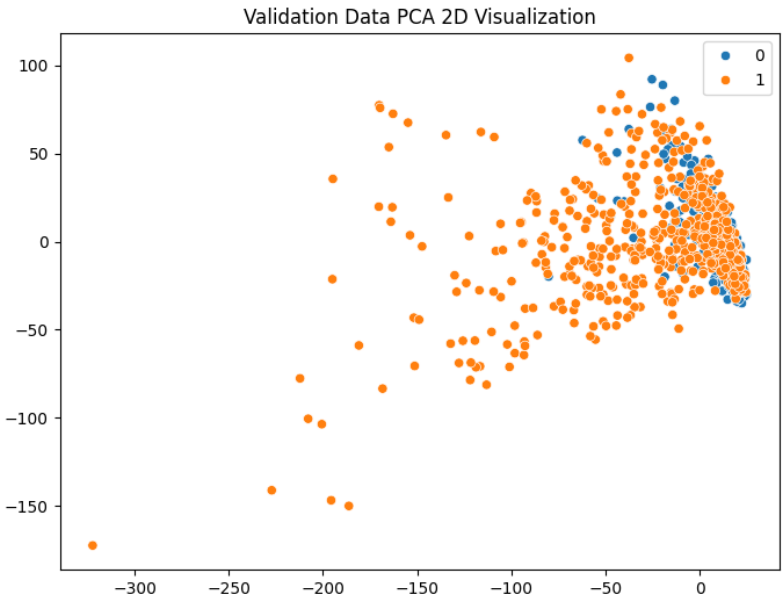
Εφαρμόσαμε τρεις μεθόδους μείωσης διάστασης:

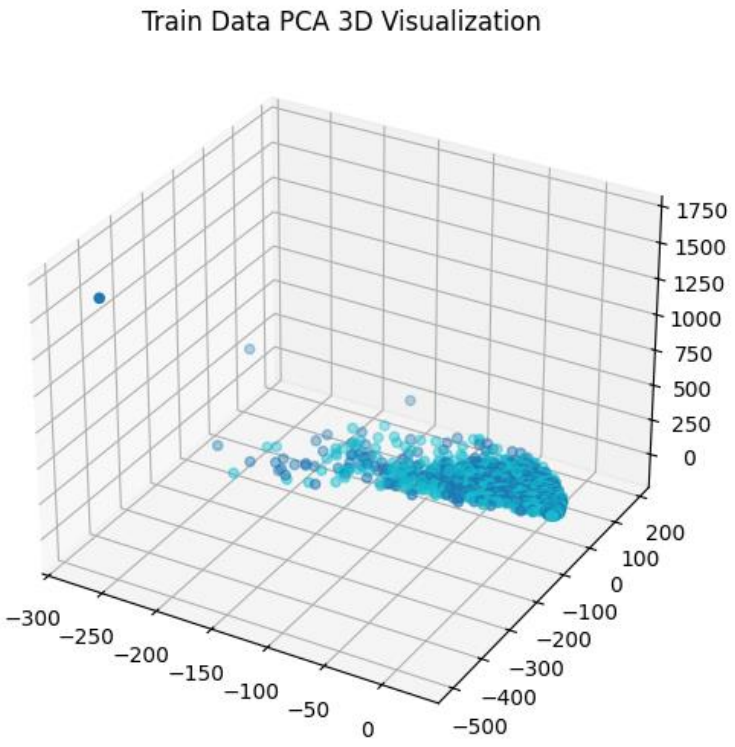
- PCA (Ανάλυση Πρωτευουσών Συνιστωσών)
- LLE (Τοπική Γραμμική Ενσωμάτωση)
- Autoencoders με Βαθιά Νευρωνικά Δίκτυα (με 2 κρυμμένα επίπεδα των 256 των 128 νευρώνων)

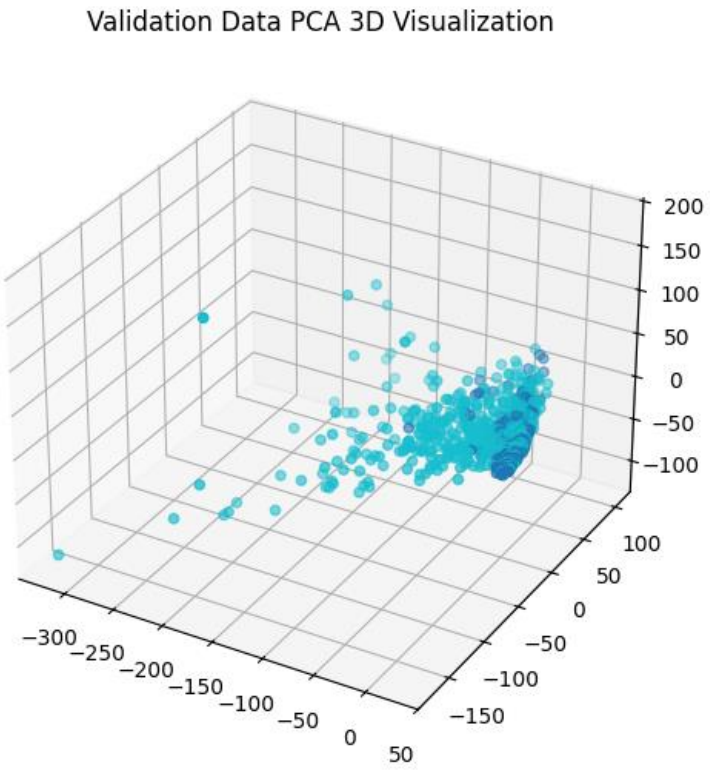
Με διαστάσεις της τάξεως M ∈ {2, 3, 5, 10, 20}. Τα δεδομένα μας προέρχονται από το προεκπαιδευμένο μοντέλο CNN ResNet50 με διαστάσεις input size 224x224. Τα δεδομένα μας έχουν χωριστεί σε 80% εκπαίδευσης και 20% επικύρωσης. Από το ResNet50 παίρνουμε δεδομένα διάστασης για την εκπαίδευση (5102, 100352) και για την επικύρωση (1274, 100352) (σ.σ. 100000+ χαρακτηρίστηκα). Επίσης κάνουμε scale τα δεδομένα μας προκειμένου να κανονικοποιηθούν και να έχουμε καλύτερα αποτελέσματα στην εκπαίδευση των μοντέλων.

PCA









Στις 2D απεικονίσεις, υπάρχουν δύο κλάσεις (0 και 1) που παρουσιάζουν μερικό διαχωρισμό αλλά με σημαντική επικάλυψη, ιδιαίτερα προς τη δεξιά πλευρά του διαγράμματος. Τα δεδομένα της εκπαίδευσης παρουσιάζουν μια συγκεντρωμένη συστάδα στη δεξιά πλευρά με πιο διάσπαρτα σημεία προς τα αριστερά. Τα δεδομένα επικύρωσης ακολουθούν παρόμοιο μοτίβο, υποδηλώνοντας συνεπή υποκείμενη δομή μεταξύ των συνόλων εκπαίδευσης και επικύρωσης.

Και τα δύο σύνολα δεδομένων περιέχουν ακραία σημεία, ιδιαίτερα ορατά στα δεδομένα εκπαίδευσης, όπου ορισμένα σημεία εμφανίζονται μακριά από την κύρια συστάδα (π.χ. το σημείο περίπου στο (-250, -500)).

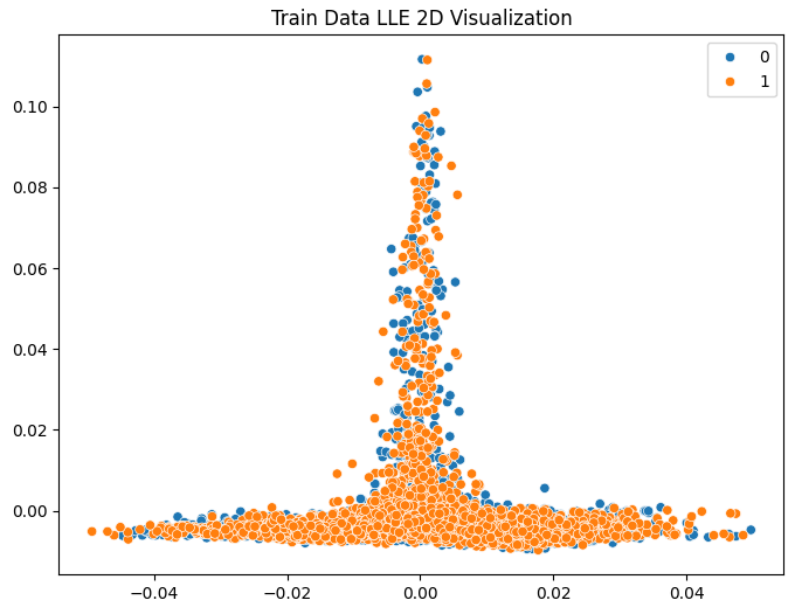
Οι τρισδιάστατες απεικονίσεις παρέχουν πρόσθετο διαχωρισμό κατά μήκος της τρίτης κύριας συνιστώσας, υποδηλώνοντας ότι η διατήρηση τριών συνιστωσών μπορεί να συλλάβει μεγαλύτερη διακύμανση από ό,τι μόνο δύο.

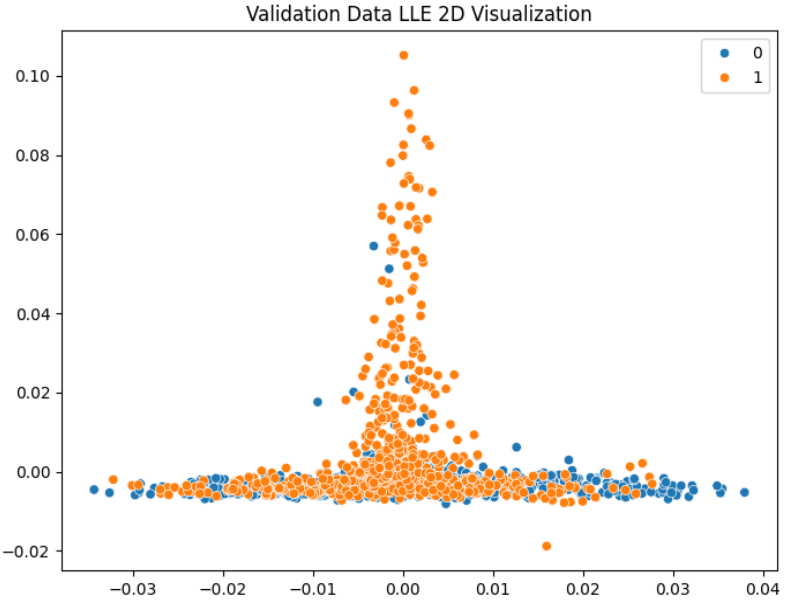
Το μεγαλύτερο μέρος της διακύμανσης των δεδομένων φαίνεται να συγκεντρώνεται κατά μήκος της πρώτης κύριας συνιστώσας (άξονας x), όπως υποδεικνύεται από τη μεγαλύτερη διασπορά των σημείων οριζόντια παρά κάθετα.

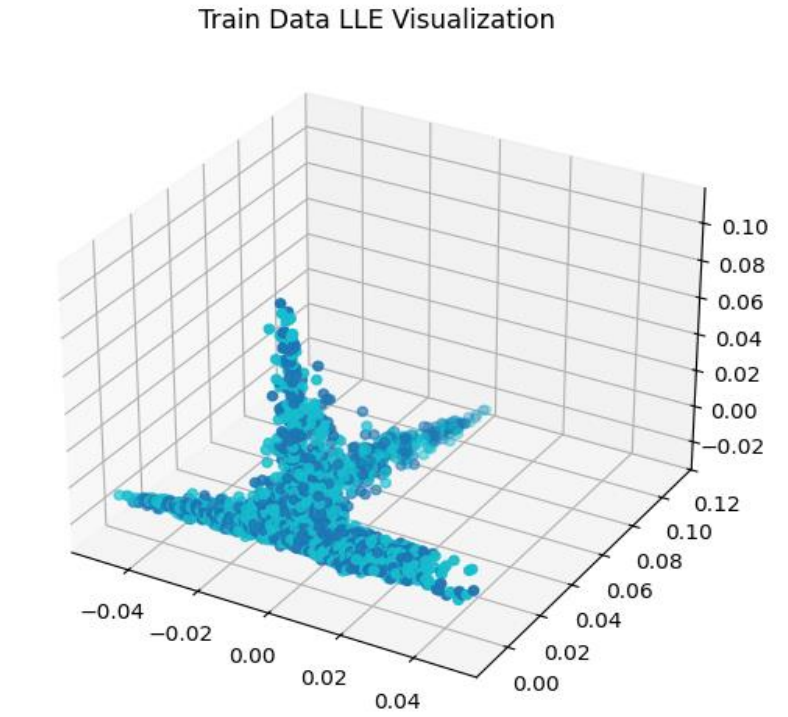
Η σημαντική αλληλοεπικάλυψη μεταξύ των κλάσεων και στα δύο σύνολα δεδομένων υποδηλώνει ότι ακόμη και μετά το μετασχηματισμό PCA, ένας απλός γραμμικός ταξινομητή μπορεί να δυσκολευτεί να διαχωρίσει τέλεια αυτές τις κλάσεις.

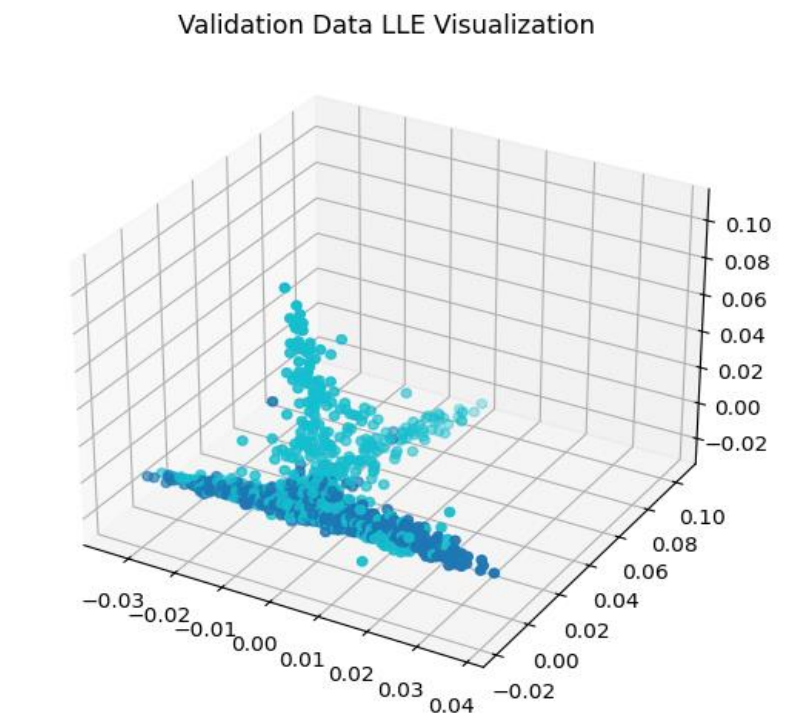
Η ανάλυση PCA αποκαλύπτει ότι, ενώ η μείωση της διάστασης βοηθά στην οπτικοποίηση της δομής των δεδομένων, παραμένει πολυπλοκότητα στα όρια των κλάσεων.

Locally Linear Embedding









Τόσο τα δεδομένα εκπαίδευσης όσο και τα δεδομένα επικύρωσης παρουσιάζουν ένα εμφανές μοτίβο σε σχήμα σταυρού ή "+" σε 2D, με έναν οριζόντιο άξονα και μια κάθετη προβολή σημείων. Αυτή η χαρακτηριστική δομή είναι αρκετά διαφορετική από τις απεικονίσεις PCA που μοιραστήκατε νωρίτερα.

Ο διαχωρισμός των κλάσεων είναι ελάχιστος στις προβολές LLE. Και οι δύο κλάσεις (0 και 1) εμφανίζονται κατανεμημένες σε όλη τη δομή του σταυρού χωρίς σαφή όρια διαχωρισμού.

Η LLE έχει αποτυπώσει μια μη γραμμική δομή που δεν ήταν εμφανής στις προβολές PCA. Αυτό υποδηλώνει ότι τα δεδομένα πιθανότατα βρίσκονται σε μια καμπυλωτή ή πτυχωτή πολλαπλότητα στον αρχικό χώρο υψηλής διάστασης.

Οι ενσωματώσεις LLE χρησιμοποιούν ένα πολύ μικρότερο εύρος κλίμακας (περίπου -0,04 έως 0,04 στον άξονα x, -0,02 έως 0,11 στον άξονα y) σε σύγκριση με τις προβολές PCA, το οποίο είναι τυπικό για την LLE.

Τα παρόμοια μοτίβα διασταύρωσης τόσο στα σύνολα εκπαίδευσης όσο και στα σύνολα επικύρωσης υποδεικνύουν ότι η LLE έχει βρει μια σταθερή ενσωμάτωση που γενικεύεται μεταξύ των συνόλων δεδομένων.

Οι τρισδιάστατες απεικονίσεις αποκαλύπτουν ότι η δομή του σταυρού έχει πράγματι κάποιο βάθος, καθώς εμφανίζεται ως τέμνοντα επίπεδα και όχι απλώς ως γραμμές. Αυτή η τρίτη διάσταση παρέχει πρόσθετη δομή που δεν είναι ορατή στις προβολές 2D.

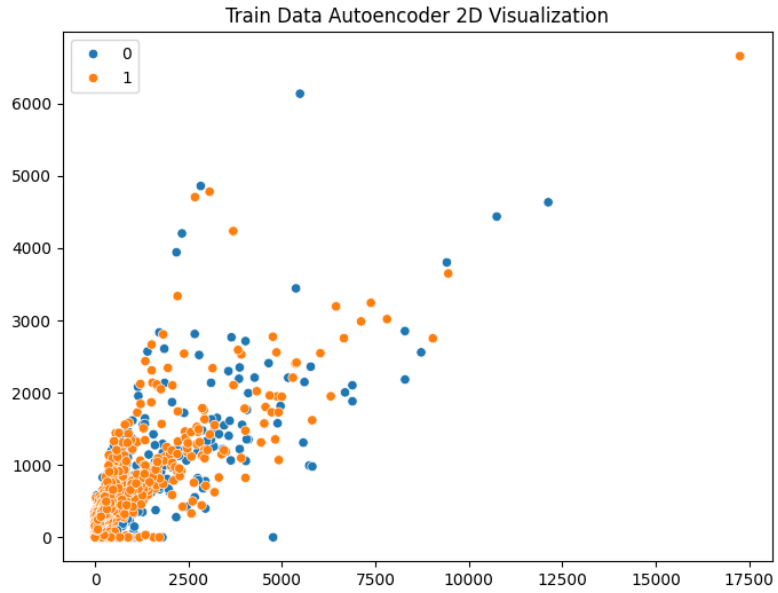
Η LLE έχει πιθανότατα διατηρήσει καλά τις σχέσεις των τοπικών γειτονιών, που είναι το κύριο πλεονέκτημά της, παρόλο που η συνολική δομή παίρνει αυτή τη χαρακτηριστική μορφή σταυρού. Υπάρχει υψηλότερη πυκνότητα σημείων στην τομή του σταυρού και κατά μήκος των αξόνων, με αραιότερες περιοχές στα άκρα.

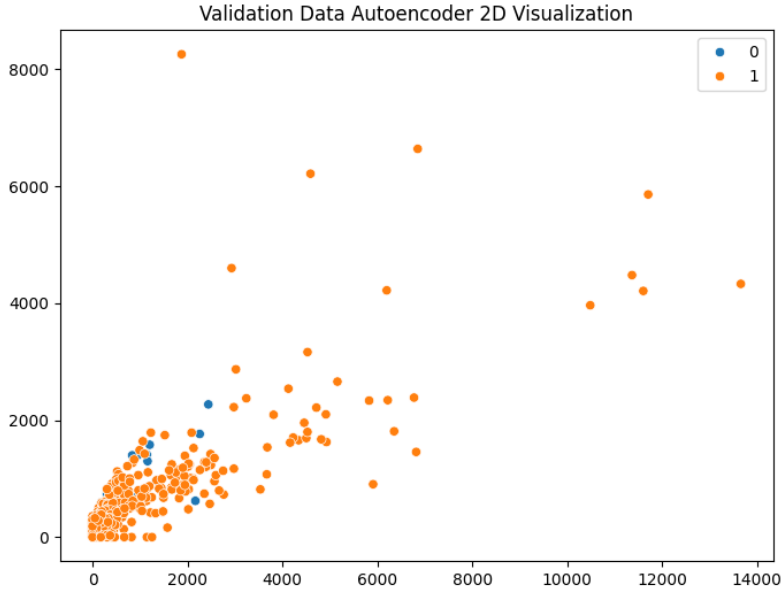
Σε σύγκριση με την PCA, η PCA έδειξε πιο διάσπαρτες, σταδιακά μεταβαλλόμενες συστάδες

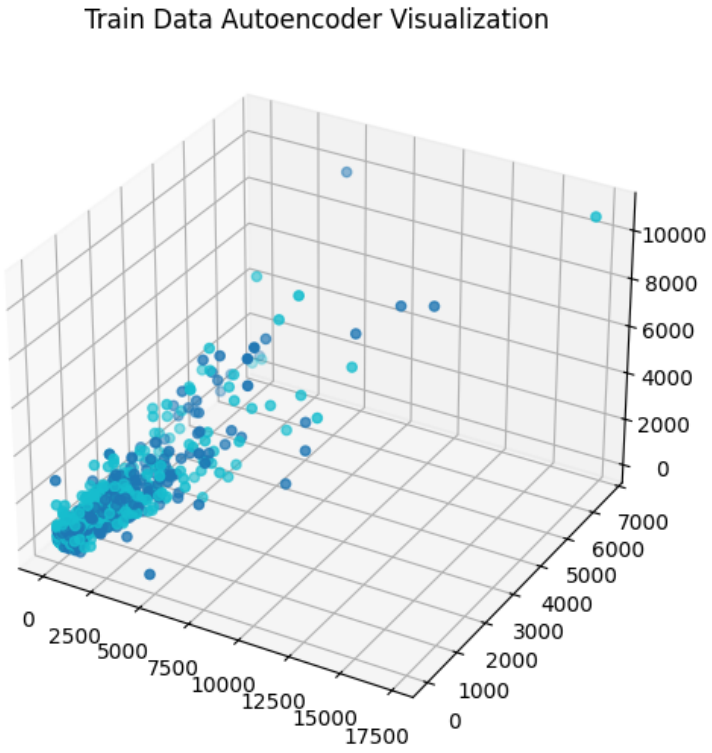
- Η LLE δείχνει μια πιο οξεία, πιο καθορισμένη γεωμετρική δομή
- Το μοτίβο του σταυρού υποδηλώνει ότι τα δεδομένα μπορεί να έχουν ορθογώνιους τρόπους μεταβολής που η LLE έχει διαχωρίσει
- Η LLE φαίνεται να έχει βρει χαμηλότερης διάστασης δομή που η PCA παρέλειψε

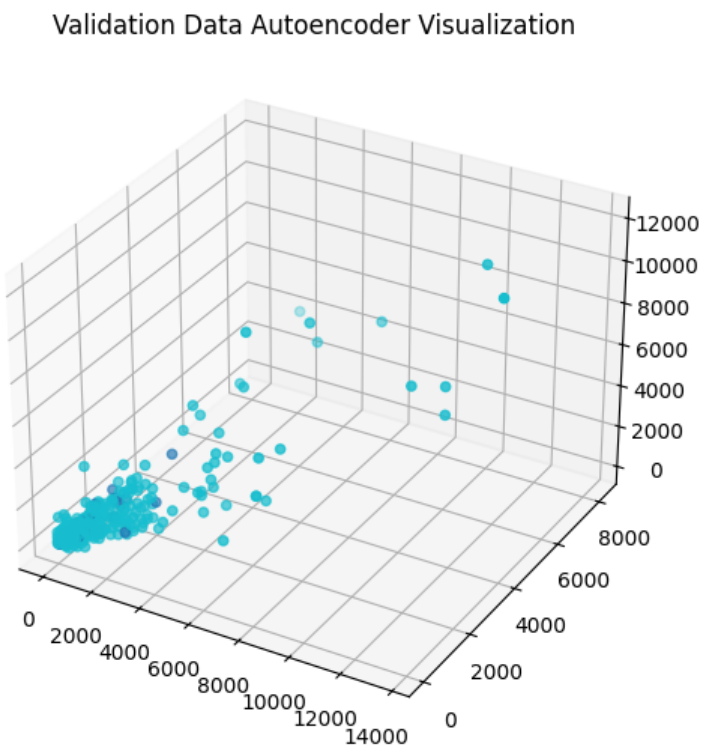
Αυτό υποδηλώνει ότι τα δεδομένα έχουν σημαντικά μη γραμμικά χαρακτηριστικά τα οποία οι γραμμικές μέθοδοι όπως η PCA δεν μπορούν να συλλάβουν πλήρως. Η απεικόνιση της LLE δείχνει ότι οι τοπικές σχέσεις στα δεδομένα μπορεί να είναι πιο σημαντικές από τις κατευθύνσεις της διακύμανσης

Autoencoder With Two Hidden Layers









Το πιο άμεσα εντυπωσιακό χαρακτηριστικό είναι η δραματικά μεγαλύτερη κλίμακα σε σύγκριση με την PCA και την LLE. Οι ενσωματώσεις του AutoEncoder εκτείνονται σε χιλιάδες μονάδες (0-17.000 στον άξονα x, 0-8.000 στον άξονα y), γεγονός που υποδηλώνει ότι το νευρωνικό δίκτυο έχει απεικονίσει τα δεδομένα σε ένα πολύ ευρύτερο αριθμητικό εύρος.

Τόσο τα σύνολα εκπαίδευσης όσο και τα σύνολα επικύρωσης παρουσιάζουν ένα περίπου "ορθογώνιο τρίγωνο" μοτίβο κατανομής, με τα σημεία να συγκεντρώνονται στο κάτω αριστερό μέρος και να εξαπλώνονται σταδιακά προς υψηλότερες τιμές. Αυτό υποδηλώνει ότι ο αυτόματος κωδικοποιητής έχει μάθει μια αναπαράσταση που δίνει έμφαση σε ορισμένες μη γραμμικές σχέσεις.

Υπάρχουν ευδιάκριτα σημεία ακραίων σημείων που είναι ορατά σε όλα τα διαγράμματα, ιδίως στις ανώτερες περιοχές και στο άκρο δεξιά των απεικονίσεων. Ο αυτόματος κωδικοποιητής φαίνεται να έχει ενισχύσει τον διαχωρισμό των ακραίων σημείων από τις κύριες συστάδες.

Ενώ δεν υπάρχει τέλειος διαχωρισμός μεταξύ των κλάσεων 0 και 1, φαίνεται να υπάρχει κάποια δομή στον τρόπο κατανομής τους. Τα σημεία της κλάσης 1 (πορτοκαλί) φαίνονται ελαφρώς πιο συγκεντρωμένα στις περιοχές χαμηλότερης πυκνότητας, με κάποιο διαχωρισμό ορατό σε συγκεκριμένες περιοχές του χώρου ενσωμάτωσης.

Τα δεδομένα επικύρωσης παρουσιάζουν παρόμοια συνολική δομή με τα δεδομένα εκπαίδευσης, υποδεικνύοντας ότι ο αυτόματος κωδικοποιητής έχει μάθει γενικεύσιμα χαρακτηριστικά και όχι υπερβολική προσαρμογή στις ιδιαιτερότητες των δεδομένων εκπαίδευσης.

Και τα δύο σύνολα παρουσιάζουν υψηλότερη πυκνότητα σημείων στις περιοχές κάτω αριστερά που μειώνεται σταδιακά προς τα έξω, γεγονός που υποδηλώνει ότι ο αυτόματος κωδικοποιητής έχει αντιστοιχίσει τα πιο κοινά μοτίβα σε αυτή την περιοχή.

Οι τρισδιάστατες απεικονίσεις αποκαλύπτουν πρόσθετη δομή που δεν είναι εμφανής στις προβολές 2D. Τα σημεία σχηματίζουν μια κάπως κωνική ή καμπυλωτή επιφάνεια στον τρισδιάστατο χώρο, υποδεικνύοντας ότι ο αυτόματος κωδικοποιητής έχει ανακαλύψει σημαντική διακύμανση στην τρίτη διάσταση.

Η προσέγγιση του αυτόματου κωδικοποιητή παρουσιάζει σημαντικές διαφορές από τις δύο προηγούμενες μεθόδους:

- Σε αντίθεση με την PCA: Ο αυτόματος κωδικοποιητής δεν εμφανίζει τα σαφή μοτίβα διακύμανσης κατεύθυνσης της PCA, γεγονός που υποδηλώνει ότι συλλαμβάνει διαφορετικές πτυχές της δομής των δεδομένων.
- Σε αντίθεση με την LLE: Ο αυτόματος κωδικοποιητής δεν παράγει το χαρακτηριστικό μοτίβο σταυρού που παρατηρείται στην LLE, αλλά δημιουργεί μια πιο συνεχή κατανομή με σταδιακές μεταβάσεις.
- Μοναδικές πτυχές: Ο αυτόματος κωδικοποιητής φαίνεται να έχει μάθει μια αναπαράσταση που δίνει έμφαση στις ακραίες τιμές και ενδεχομένως να συλλαμβάνει πιο σύνθετες μη γραμμικές σχέσεις στα δεδομένα.

Ομαδοποίηση στον Μειωμένο Χώρο

Για την ομαδοποίηση στον μειωμένο χώρο χρησιμοποιούμε 2 μεθόδους.

- K-Means με: Ευκλείδεια απόσταση κα Συνημίτονο απόστασης (Cosine Distance)
- Ιεραρχική Συνθετική Ομαδοποίηση (Agglomerative Clustering)

Πιο συγκεκριμένα για κάθε συνδυασμό μεθόδου μείωσης διάστασης, διάστασης M και μεθόδου ομαδοποίησης:

Εκτελέστε ομαδοποίηση με K = 2 έως $\boldsymbol{Kmax} = \mathbf{10}$ αριθμό ομάδων και υπολογίζουμε:

- Καθαρότητα (Purity)
- F-measure
- Silhouette score

Και εκτιμούμε τον βέλτιστο αριθμό συστάδων ($\boldsymbol{k}*$) με βάση τον δείκτη Silhouette

K-Means

Top 10 configurations by F1 Score:

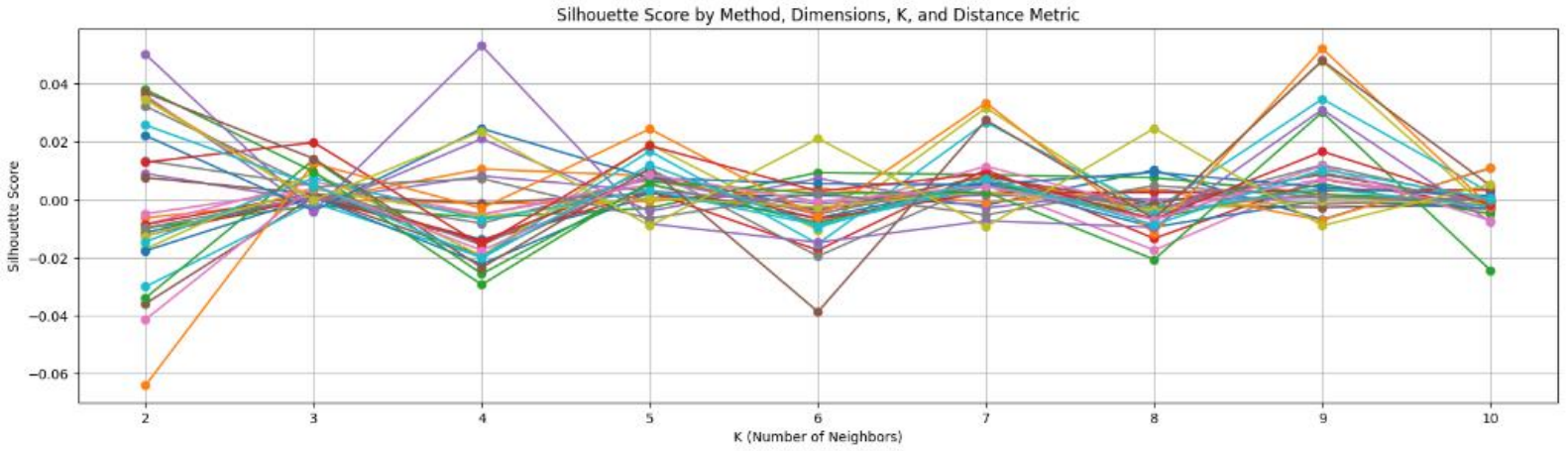
	Method	Dimensions	K	Distance	F1_Score	Purity	Silhouette
182	encoded	2	3	euclidean	0.537194	0.545526	-0.003862
114	lle	3	5	euclidean	0.534513	0.534537	0.005440
215	encoded	3	10	cosine	0.531508	0.531397	-0.001770
213	encoded	3	9	cosine	0.530041	0.574568	0.016643
146	lle	10	3	euclidean	0.527145	0.558085	0.003632
222	encoded	5	5	euclidean	0.526657	0.577708	-0.008307
116	lle	3	6	euclidean	0.526027	0.611460	-0.008074
265	encoded	20	8	cosine	0.524750	0.558870	-0.008601
224	encoded	5	6	euclidean	0.522915	0.585557	-0.014703
118	lle	3	7	euclidean	0.522757	0.551805	0.006154

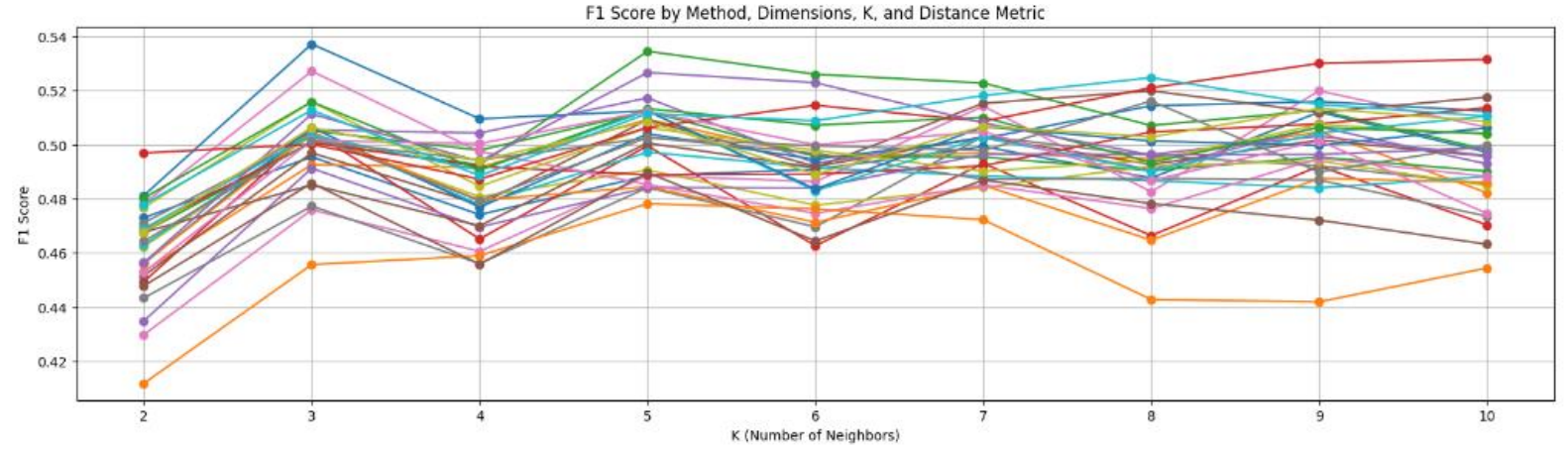
Top 10 configurations by Silhouette Score:

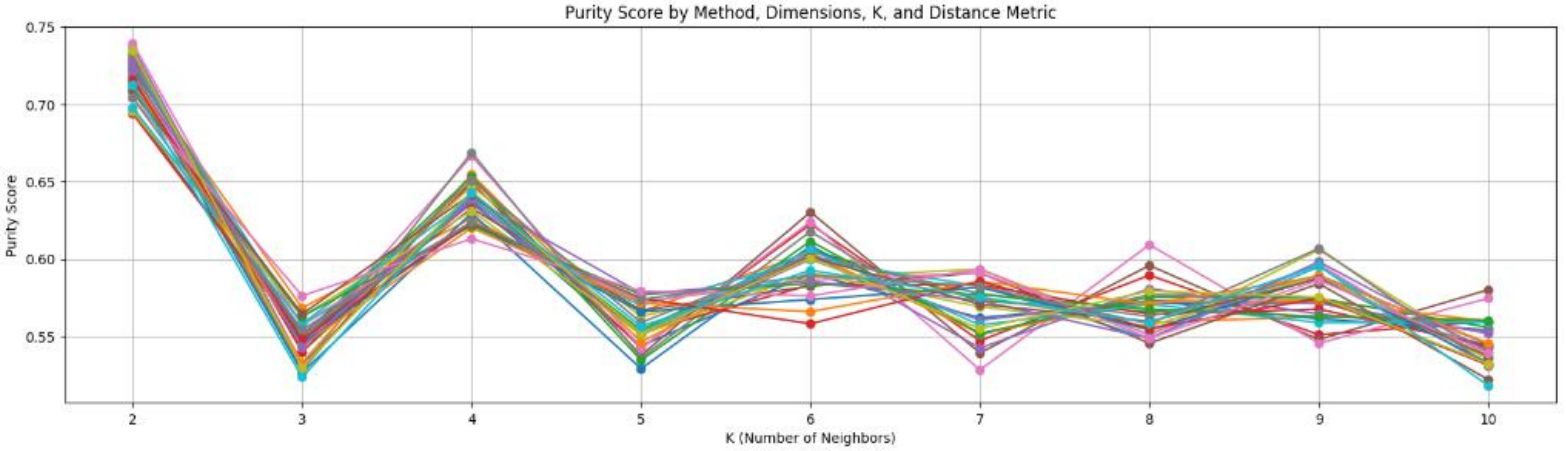
	Method	Dimensions	K	Distance	Silhouette	F1_Score	Purity
220	encoded	5	4	euclidean	0.053121	0.493848	0.639717
195	encoded	2	9	cosine	0.052141	0.441998	0.588697
216	encoded	5	2	euclidean	0.050123	0.456448	0.726845
231	encoded	5	9	cosine	0.048006	0.472180	0.587127
86	pca	20	9	euclidean	0.047877	0.494011	0.605965
198	encoded	3	2	euclidean	0.038002	0.480319	0.709576
217	encoded	5	2	cosine	0.036898	0.447903	0.708006
36	pca	5	2	euclidean	0.035894	0.467893	0.737049
91	lle	2	2	cosine	0.034921	0.456373	0.720565
87	pca	20	9	cosine	0.034730	0.484072	0.594976

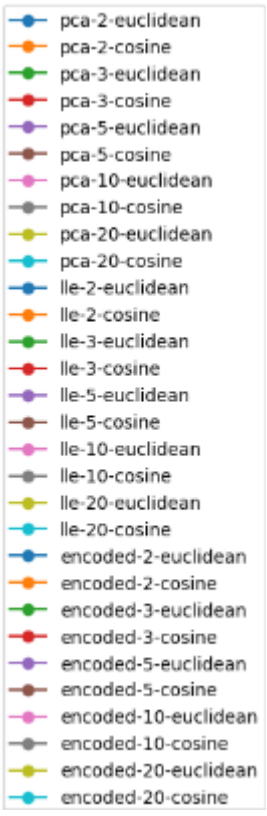
Top 10 configurations by Purity Score:

	Method	Dimensions	K	Distance	Purity	F1_Score	Silhouette
54	pca	10	2	euclidean	0.739403	0.429875	-0.008535
36	pca	5	2	euclidean	0.737049	0.467893	0.035894
252	encoded	20	2	euclidean	0.734694	0.467676	0.034490
18	pca	3	2	euclidean	0.733124	0.456481	-0.009814
145	lle	10	2	cosine	0.730769	0.471163	-0.009694
19	pca	3	2	cosine	0.728414	0.449434	-0.012351
73	pca	20	2	cosine	0.726845	0.463057	-0.029981
216	encoded	5	2	euclidean	0.726845	0.456448	0.050123
72	pca	20	2	euclidean	0.726060	0.462428	-0.016958
108	lle	3	2	euclidean	0.724490	0.467771	-0.033886






Purity Score by Method, Dimensions, K, and Distance Metric
Purity Score
0.75
0.70
0.65
0.60
0.55
2
3
4
5
6
7
8
9
10
K (Number of Neighbors)


pca-2-euclidean
pca-2-cosine
pca-3-euclidean
pca-3-cosine
pca-5-euclidean
pca-5-cosine
pca-10-euclidean
pca-10-cosine
pca-20-euclidean
pca-20-cosine
lle-2-euclidean
lle-2-cosine
lle-3-euclidean
lle-3-cosine
lle-5-euclidean
lle-5-cosine
lle-10-euclidean
lle-10-cosine
lle-20-euclidean
lle-20-cosine
encoded-2-euclidean
encoded-2-cosine
encoded-3-euclidean
encoded-3-cosine
encoded-5-euclidean
encoded-5-cosine
encoded-10-euclidean
encoded-10-cosine
encoded-20-euclidean
encoded-20-cosine

Στην παραπάνω φωτογραφίες παρουσιάζεται μια αξιολόγηση της απόδοσης της ομαδοποίησης σε διαφορετικές μεθόδους ενσωμάτωσης, μετρικές απόστασης και αριθμούς συστάδων K. Πιο συγκεκριμένα:

Silhouette Score

Σε όλες τις διαμορφώσεις είναι σχετικά χαμηλές, κυρίως εντός του εύρους -0,05 έως 0,05, υποδεικνύοντας ανεπαρκώς καθορισμένες ή επικαλυπτόμενες συστάδες. Αυτό υποδηλώνει ότι τα δεδομένα μπορεί να μην είναι εγγενώς διαχωρίσιμα με τις τρέχουσες μεθόδους ενσωμάτωσης ή ότι ο αλγόριθμος K-means δεν είναι ο βέλτιστος αλγόριθμος για αυτό το σύνολο δεδομένων.

Οι καλύτερες επιδόσεις παρατηρούνται με βάση την PCA.

F1 Score

Οι βαθμολογίες F1 παρουσιάζουν μέτρια διακύμανση μεταξύ των διαμορφώσεων, με τιμές γενικά μεταξύ 0,45 και 0,53. Αξίζει να σημειωθεί ότι οι βαθμολογίες είναι πιο σταθερές σε σχέση με το K, με αιχμές γύρω από το K = 3 και το K = 5.

Οι υψηλότερες βαθμολογίες F1 επιτυγχάνονται συνήθως από τις pc-2-euclidean και pc-3-euclidean, υποδεικνύοντας ότι οι μέθοδοι που βασίζονται στην PCA διατηρούν καλά την πληροφορία που σχετίζεται με τις τάξεις.

Η απόσταση συνημίτονου επιτυγχάνουν περιστασιακά υψηλές βαθμολογίες F1, αλλά είναι ασυνεπείς.

Purity Score

Εμφανίζουν υψηλές τιμές σε K=2 (έως ~0,73), στη συνέχεια σταθεροποιούνται μεταξύ 0,52 και 0,63 για υψηλότερες τιμές του K. Η δυαδική ομαδοποίηση (K=2) μπορεί να συλλάβει κάποια κυρίαρχη δομή, αλλά οι πιο λεπτομερείς διακρίσεις είναι πιο δύσκολο να εντοπιστούν. Παρόμοια με F1 και Silhouette, οι προβλέψεις με βάση την PCA και την ευκλείδεια απόσταση επιτυγχάνουν κορυφαίες επιδόσεις.

Βλέπουμε ότι τα με βάση Purities Score οι διαφορετικές διαμορφώσεις είναι σχετικά κοντά, υποδεικνύοντας μικρή διακύμανση των επιδόσεων.

AgglomerativeClustering

Top 10 configurations by Silhouette Score:

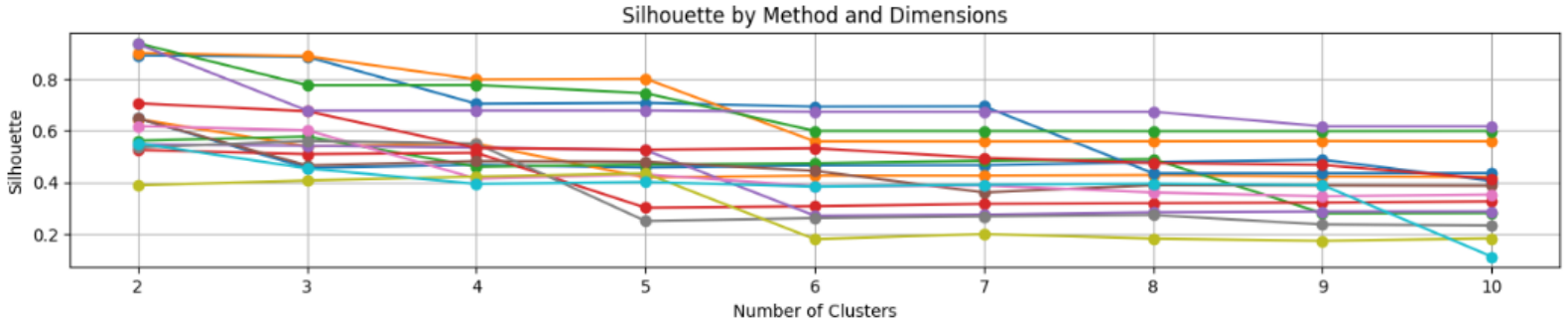
	Method	Dimensions	Clusters	Silhouette	F1_Score	Purity
108	encoded	5	2	0.938915	0.344650	0.525903
126	encoded	20	2	0.936139	0.344313	0.525118
99	encoded	3	2	0.902048	0.335073	0.503925
90	encoded	2	2	0.891901	0.337493	0.509419
100	encoded	3	3	0.888863	0.223382	0.503925
91	encoded	2	3	0.886786	0.224996	0.509419
102	encoded	3	5	0.801565	0.174091	0.520408
101	encoded	3	4	0.799048	0.217614	0.520408
110	encoded	5	4	0.777562	0.245525	0.558085
109	encoded	5	3	0.776633	0.327367	0.558085

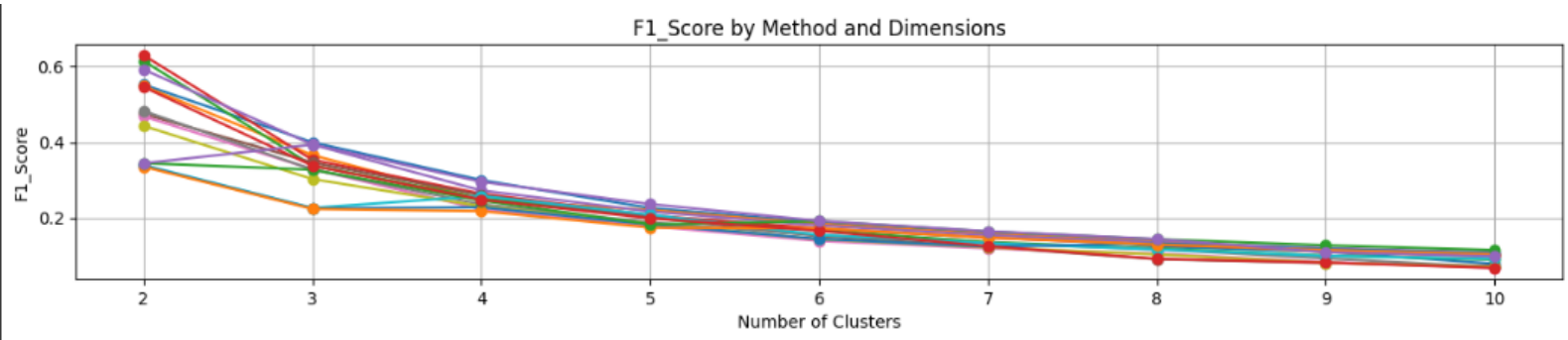
Top 10 configurations by F1 Score:

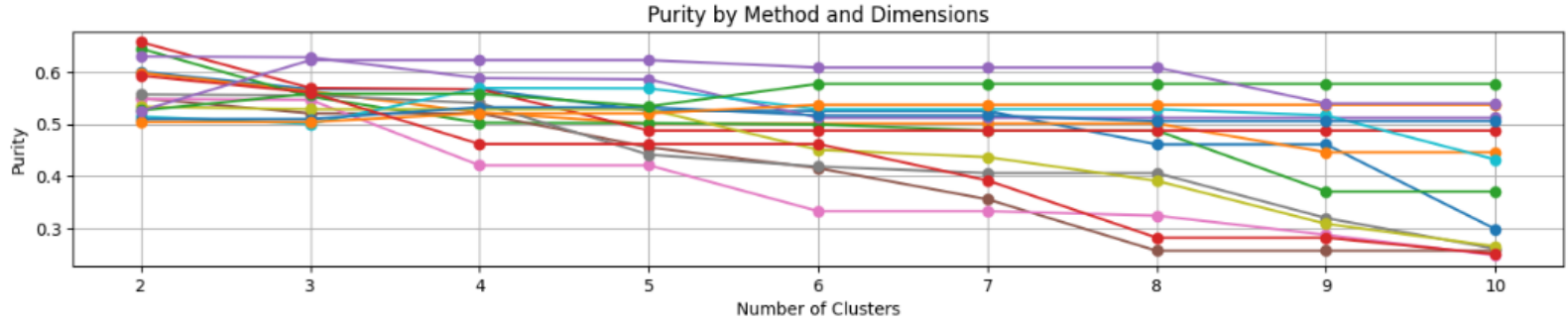
	Method	Dimensions	Clusters	F1_Score	Purity	Silhouette
27	pca	10	2	0.630179	0.656986	0.525647
18	pca	5	2	0.615220	0.644427	0.561636
36	pca	20	2	0.592491	0.629513	0.547337
0	pca	2	2	0.552531	0.600471	0.648690
9	pca	3	2	0.548587	0.598116	0.645423
117	encoded	10	2	0.546320	0.592622	0.706359
63	lle	5	2	0.482789	0.556515	0.535438
45	lle	2	2	0.474913	0.549451	0.647754
54	lle	3	2	0.468633	0.547881	0.617940
72	lle	10	2	0.443425	0.532967	0.388500

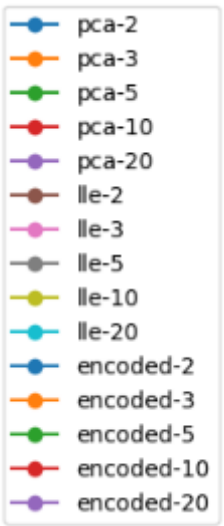
Top 10 configurations by Purity Score:

	Method	Dimensions	Clusters	Purity	F1_Score	Silhouette
27	pca	10	2	0.656986	0.630179	0.525647
18	pca	5	2	0.644427	0.615220	0.561636
36	pca	20	2	0.629513	0.592491	0.547337
37	pca	20	3	0.627943	0.393845	0.541930
127	encoded	20	3	0.622449	0.393759	0.678423
129	encoded	20	5	0.622449	0.236255	0.678766
128	encoded	20	4	0.622449	0.295319	0.678696
131	encoded	20	7	0.608320	0.164341	0.673739
132	encoded	20	8	0.608320	0.143798	0.673785
130	encoded	20	6	0.608320	0.191731	0.673925


Silhouette by Method and Dimensions
Silhouette
0.8
0.6
0.4
0.2
2
3
4
5
6
7
8
9
10
Number of Clusters


F1_Score by Method and Dimensions
F1_Score
0.6
0.4
0.2
2
3
4
5
6
7
8
9
10
Number of Clusters


Purity by Method and Dimensions
Purity
0.6
0.5
0.4
0.3
2
3
4
5
6
7
8
9
10
Number of Clusters


pca-2
pca-3
pca-5
pca-10
pca-20
lle-2
lle-3
lle-5
lle-10
lle-20
encoded-2
encoded-3
encoded-5
encoded-10
encoded-20

Στις παραπάνω φωτογραφίες παρουσιάζονται τα αποτελέσματα αξιολόγησης για τη Ιεραρχική Συνθετική Ομαδοποίηση (Agglomerative Clustering) που εφαρμόζεται σε διάφορες μεθόδους μείωσης διαστάσεων (PCA, LLE και AutoEncoders)

Πιο συγκεκριμένα:

Silhouette Score

Οι περισσότεροί συνδυασμοί ξεκινούν με υψηλό σκορ για K = 2 ειδικά εκείνοι που χρησιμοποιούν PCA και LLE, και στη συνέχεια μειώνονται καθώς αυξάνεται K.

Κορυφαίες επιδόσεις:

Για PCA-2, PCA-3 και LLE-2 διατηρούν υψηλό σκορ πάνω από 0,6 μέχρι K=5, υποδεικνύοντας καλά διαχωρισμένες συστάδες.

Η Ιεραρχική Συνθετική Ομαδοποίηση επωφελείται από τις γραμμικές ενσωματώσεις χαμηλών διαστάσεων, ιδιαίτερα από την PCA με 2-3 συνιστώσες.

F1 Score

Τα F1 scores κορυφώνονται στο K=2 και πέφτουν απότομα με την αύξηση του K

Κορυφαίες επιδόσεις:

Για PCA-2, LLE-2 και ENCODED-2 επιτυγχάνουν τα υψηλοτέρα σκορ F1 (πάνω από 0,6) σε K=2.Ωστόσο, όλες οι μέθοδοι συγκλίνουν σε παρόμοιες τιμές (~0,15-0,2) για K≥6. Η απότομη πτώση υποδηλώνει ότι η ευθυγράμμιση των labels με τις ground truth υποβαθμίζεται με περισσότερες συστάδες, πιθανώς λόγω υπερβολικής τμηματοποίησης.

Purity Score

Τα purity σκορ είναι υψηλότερα σε K=2 (έως 0,68) και μειώνονται με υψηλότερα K, ιδίως για τα LLE και τα AutoEncoders.

Κορυφαίες επιδόσεις:

Οι μέθοδοι που βασίζονται στην PCA (pca-2, pca-3, pca-10) διατηρούν σχετικά υψηλά και σταθερά σκορ στις περισσότερες μετρήσεις.

Encoded Embeddings παρουσιάζουν απότομη πτώση στο purity score μετά το K=4, γεγονός που υποδηλώνει overfitting ή κακό διαχωρισμό.

Η μέθοδος PCA παρέχει πιο συνεπή ευθυγράμμιση μεταξύ κλάσεων και συστάδων, ενώ τα Encoded Embeddings και τα LLE υποβαθμίζονται πιο γρήγορα.

Η Ιεραρχική Συνθετική Ομαδοποίηση αποδίδει καλύτερα με την μέθοδο PCA χαμηλής διάστασης (ιδίως 2-3 συνιστώσες), υποδεικνύοντας ότι οι απλές γραμμικές προβολές έχουν ουσιαστικότερη δομή.

Το K = 2 έως 4 φαίνεται να είναι το “sweet spot” για ουσιαστική ομαδοποίηση σε όλες τις μετρικές.

Οι μη γραμμικές μέθοδοι (LLE, Encoded) είναι υποσχόμενες σε χαμηλό K, αλλά δεν έχουν συνοχή καθώς αυξάνεται το K.

Σε σύγκριση με το K-Means , η Ιεραρχική Συνθετική Ομαδοποίηση παρέχει υψηλότερο Silhouette και Purity Score σε διάφορες διαμορφώσεις, ιδίως με την PCA.