Πανεπιστήμιο Μακεδονίας
Τμήμα Εφαρμοσμένης Πληροφορικής

Τίτλος Εργασίας:

«2η Εργασία – Clustering problems»

Συντάκτες:

Ευθυμίου Βασίλειος[1], **Παπαδόπουλος Νικόλαος**[2], **Τζελαλής Γεώργιος**[3] **& Τσώνη Σταυρούλα**[4]

Διδάσκων: Πρωτοπαπαδάκης Ευτύχιος
Μάθημα: Μηχανική Μάθηση
Εξάμηνο Μαθήματος: 7ο
Ημερομηνία: 22/01/2024

[1]**ics21035**@uom.edu.gr
[2]**ics21048**@uom.edu.gr
[3]**ics21032**@uom.edu.gr
[4]**ics21074**@uom.edu.gr

Πίνακας Περιεχομένων

1. Εισαγωγή

Στην παρούσα έρευνα χρησιμοποιήθηκαν οι τεχνικές Principal Component Analysis (PCA), Stacked Autoencoder (SAE) και Linear Discriminant Analysis (LDA) για τη μείωση των διαστάσεων των δεδομένων, ενώ εφαρμόστηκε clustering με τους αλγόριθμους Mini Batch K-means, DBSCAN και Agglomerative Clustering. Τα αποτελέσματα αξιολογήθηκαν με βάση τέσσερις μετρικές απόδοσης, τις οποίες θα δούμε στη συνέχεια.

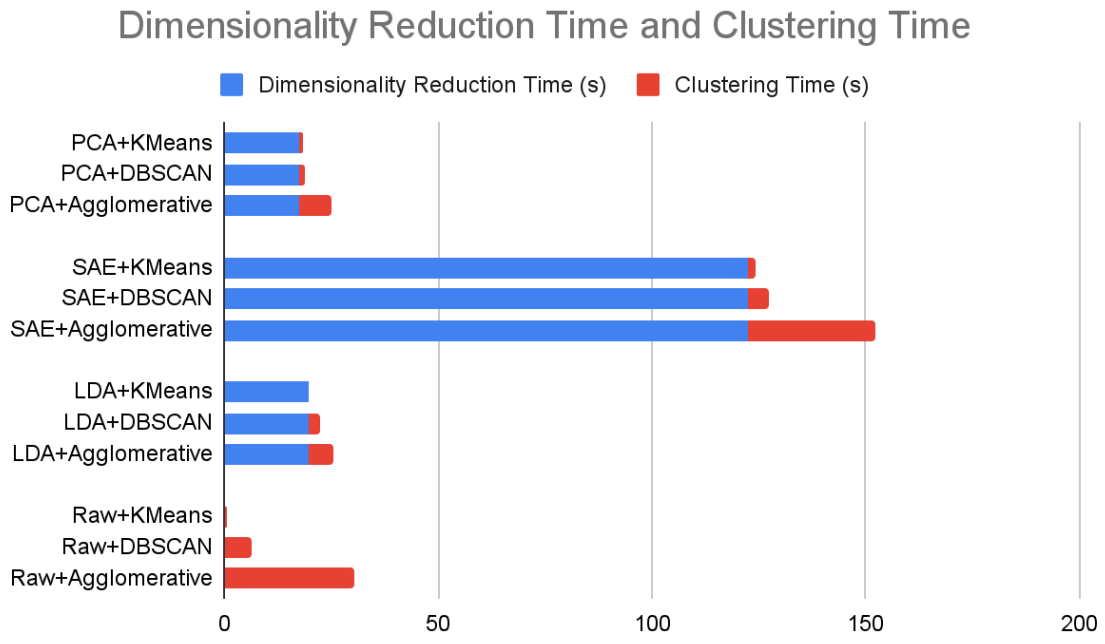
Έπεται μια σύντομη περιγραφή των κεντρικών σημείων της αναφοράς που θα αναπτυχθούν στις επόμενες ενότητες. Η αναφορά ακολουθεί μια ολοκληρωμένη δομή, προσφέροντας στον αναγνώστη μια συνολική εικόνα της εργασίας μας. Η επόμενη ενότητα, «*Πειραματικά Αποτελέσματα*», προσφέρει μία εκτενή επισκόπηση των συλλεγέντων δεδομένων, παρουσιάζοντας τα αποτελέσματα με τη μορφή πινάκων, γραφικών παραστάσεων και εικόνων. Στην τελευταία ενότητα «*Βέλτιστος Συνδυασμός Τεχνικών*» εξερευνάται η ύπαρξη ενός συνδυασμού τεχνικών, ο οποίος αποδίδει καλύτερα σε κάθε μετρική αξιολόγησης.

2. Πειραματικά Αποτελέσματα

Τα δεδομένα που προέκυψαν από την εκτέλεση του κώδικα και οι μετρικές που υπολογίστηκαν (και αξιοποιήθηκαν για την κατασκευή των διαγραμμάτων που ακολουθούν) είναι διαθέσιμα σε αυτό το Google Spreadsheet.

2.1. Διάρκεια Τεχνικών

Ένα αξιοσημείωτο μέτρο σύγκρισης πριν αναλύσουμε τις μετρικές απόδοσης είναι ο απαραίτητος χρόνος για την εφαρμογή των τεχνικών. Όπως βλέπουμε και στην *Εικόνα 2.1.1*, ανάλογα με τον συνδυασμό των τεχνικών, ο χρόνος μπορεί να διαφέρει σημαντικά. Παρατηρούμε πως ο μέγιστος χρόνος για το dimensionality reduction αντιστοιχεί στην τεχνική SAE, απέχοντας πολύ από τις υπόλοιπες περιπτώσεις, ενώ ο ελάχιστος χρόνος αντιστοιχεί στην τεχνική PCA, με μικρή διαφορά από την LDA. Προφανώς, στη συγκεκριμένη σύγκριση δεν λαμβάνονται υπ' όψιν τα ακατέργαστα δεδομένα (Raw) στα οποία δεν πραγματοποιείται μείωση διαστάσεων. Αντίστοιχα, οι χρόνοι για κάθε τεχνική ομαδοποίησης φαίνεται να ακολουθούν μία σειρά, με συντομότερο τον χρόνο της Mini Batch K-means και μεγαλύτερο αυτόν της Agglomerative Clustering. Επιπλέον, η Agglomerative παρουσιάζει τη μεγαλύτερη μεταβλητότητα στη διάρκειά.



Εικόνα 2.1.1: Χρονική Διάρκεια Τεχνικών

2.2. Πλήθος Ομάδων

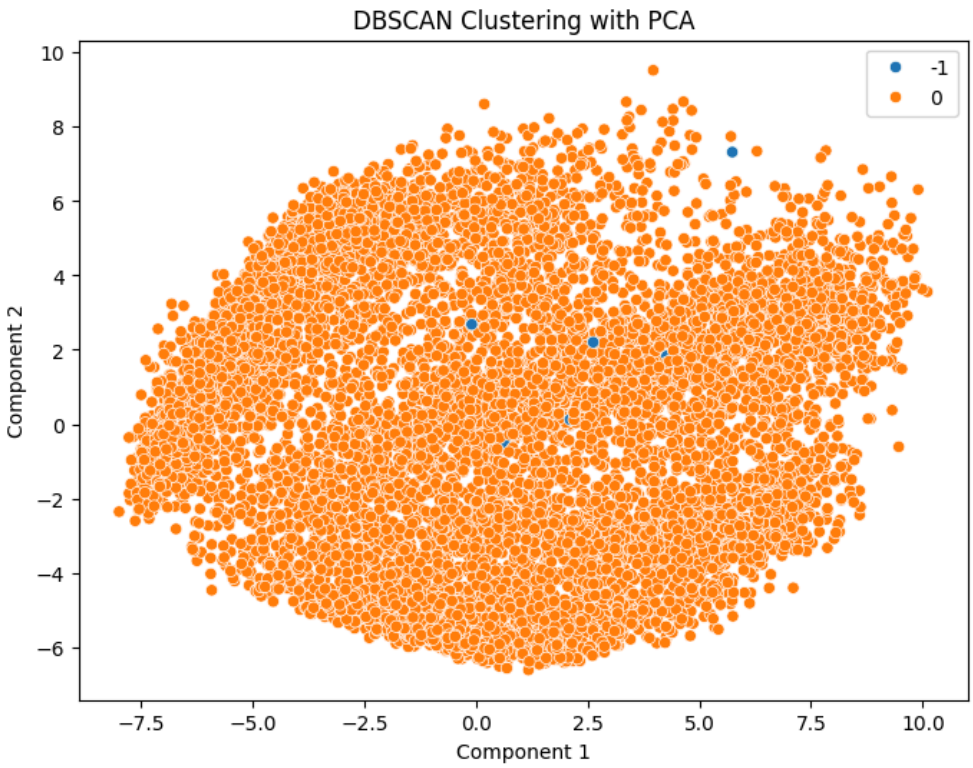
Σε κάθε συνδυασμό τεχνικών προτείνεται το πλήθος των ομάδων (clusters) στις οποίες είναι καλύτερο να διαχωριστούν τα δεδομένα. Στην *Εικόνα 2.2.1* βλέπουμε αυτόν τον προτεινόμενο αριθμό για κάθε συνδυασμό τεχνικών. Παρατηρούμε πως στις περισσότερες περιπτώσεις το προτεινόμενο πλήθος ομάδων είναι δέκα (10), το οποίο είναι και το σωστό πλήθος. Οι εξαιρέσεις προκύπτουν από την τεχνική ομαδοποίησης DBSCAN, η οποία ανεξαρτήτως της τεχνικής μείωσης διαστάσεων που χρησιμοποιήθηκε καταλήγει σε λανθασμένα συμπεράσματα.

Dimensionality Reduction Technique	Clustering Algorithm	Number of Suggested Clusters
PCA	KMeans	10
PCA	DBSCAN	1
PCA	Agglomerative	10
Stacked Autoencoder	KMeans	10
Stacked Autoencoder	DBSCAN	1
Stacked Autoencoder	Agglomerative	10
LDA	KMeans	10
LDA	DBSCAN	1
LDA	Agglomerative	10
Raw	KMeans	10
Raw	DBSCAN	4
Raw	Agglomerative	10

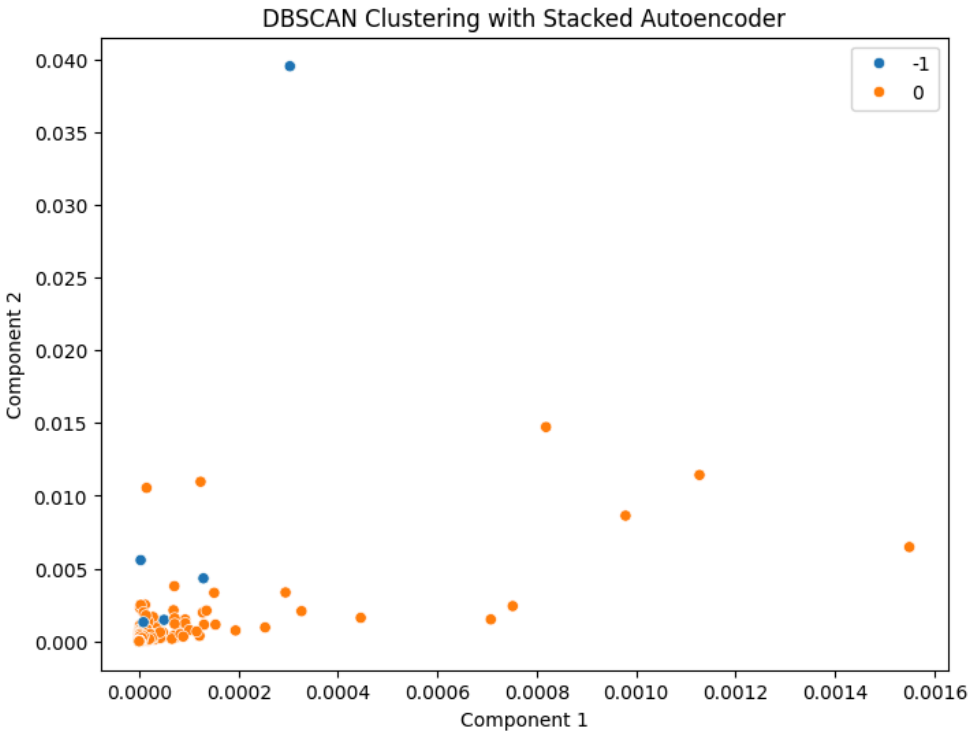
Εικόνα 2.2.1: Αριθμός προτεινόμενων ομάδων

2.2.a. DBSCAN

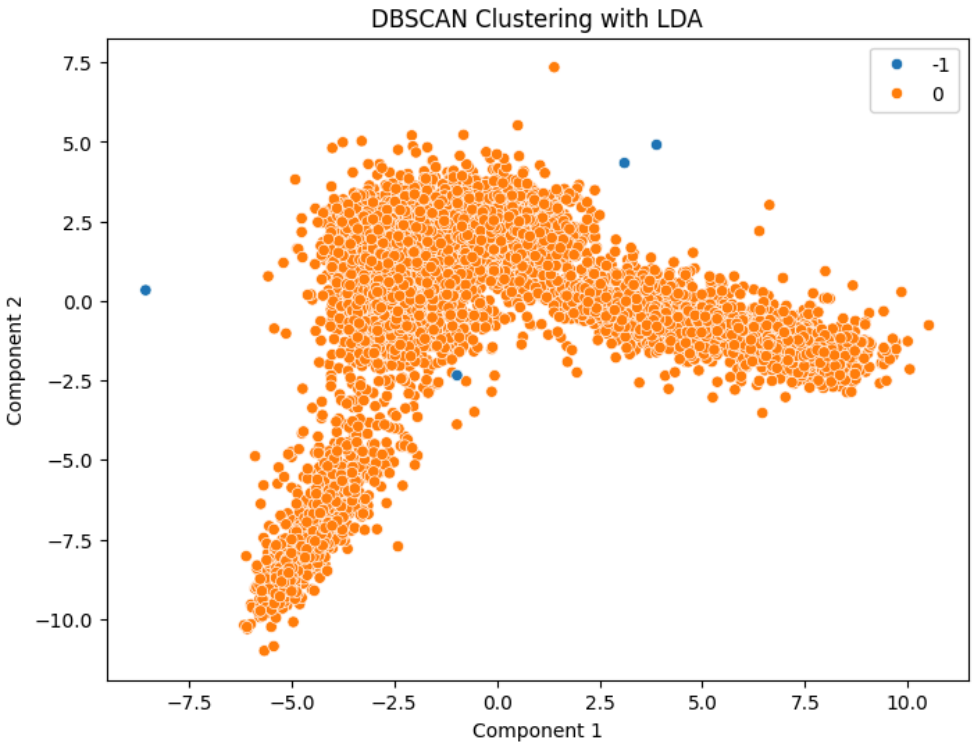
Ενδεικτικά scatterplots μεταξύ των πρώτων δύο (2) components που προκύπτουν κάθε φορά από την εκάστοτε τεχνική μείωσης διάστασης και εφαρμογή της τεχνικής DBSCAN (*Εικόνες 2.2.2, 2.2.3 και 2.2.4*) δείχνουν ότι η συγκεκριμένη τεχνική clustering αδυνατεί να διακρίνει τις διαφορετικές κλάσεις του dataset. Αυτό το φαινόμενο μπορεί να αποδοθεί στην πυκνή κατανομή των σημείων που φαίνεται να σχηματίζουν μια ενιαία μάζα, καθώς και στο υψηλό πλήθος διαστάσεων των δεδομένων.



Εικόνα 2.2.2: Εφαρμογή PCA στο test set και συσταδοποίηση με DBSCAN



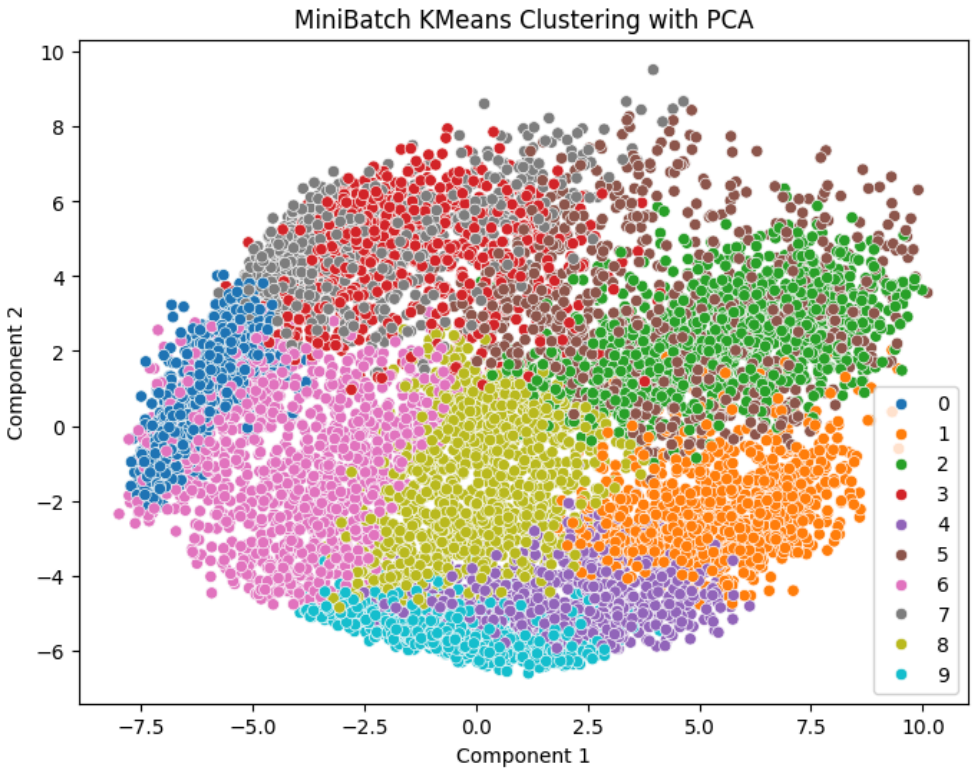
Εικόνα 2.2.3: Εφαρμογή Stacked Autoencoder στο test set και συσταδοποίηση με DBSCAN



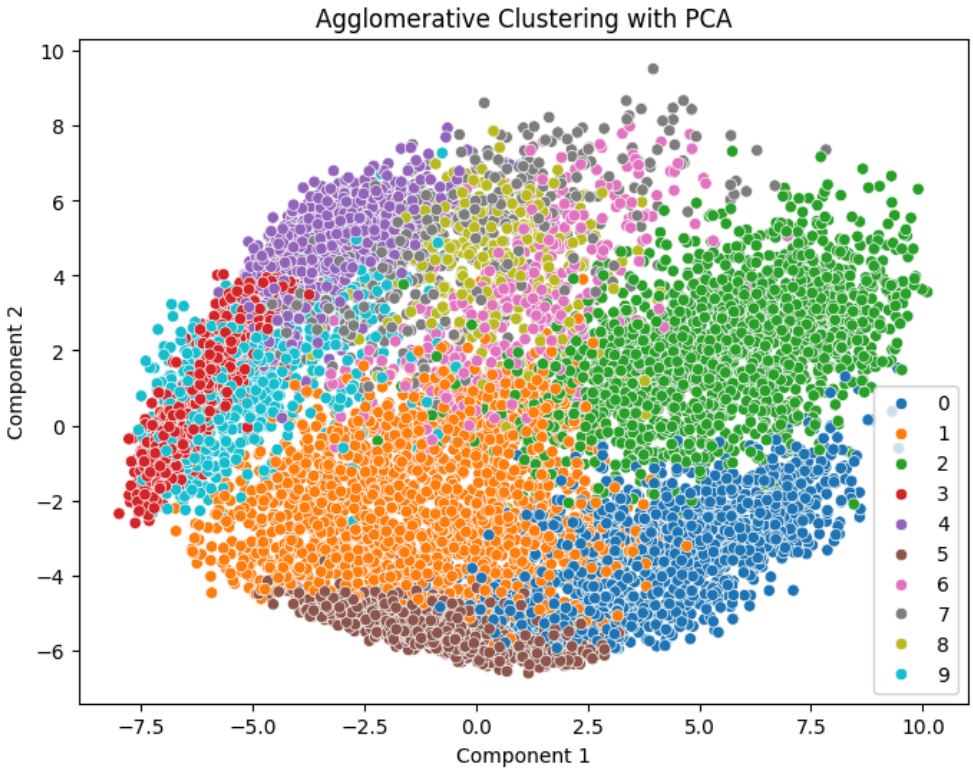
Εικόνα 2.2.4: Εφαρμογή LDA στο test set και συσταδοποίηση με DBSCAN

2.2.b. Mini Batch K-means & Agglomerative Clustering

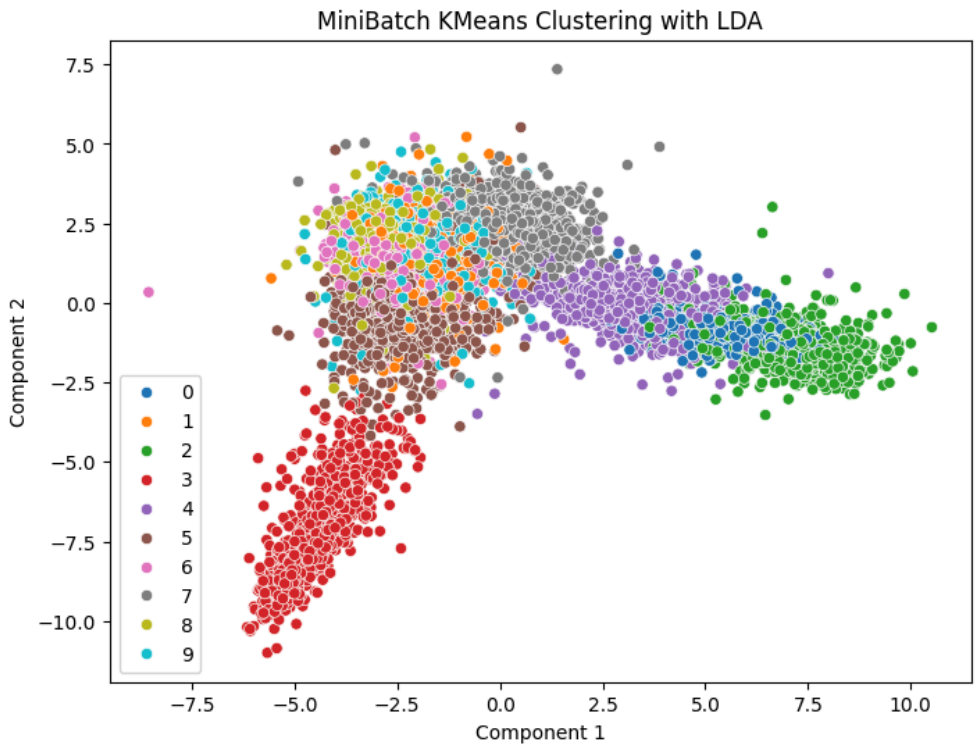
Από την άλλη, οι τεχνικές Mini Batch K-means και Agglomerative Clustering φαίνεται να αποδίδουν πολύ καλύτερα. Η αλληλοεπικάλυψη μεταξύ διαφορετικών κλάσεων εξακολουθεί να υφίσταται, με ορισμένες να είναι περισσότερο διαχωρίσιμες από άλλες, όπως διακρίνεται και στις *Εικόνες 2.2.5, 2.2.6, 2.2.7, 2.2.8*. Παρόλα αυτά, όπως επιβεβαιώνεται στη συνέχεια και από τις μετρικές, με τις τεχνικές αυτές επιτυγχάνεται καλύτερος διαχωρισμός των κλάσεων του dataset.



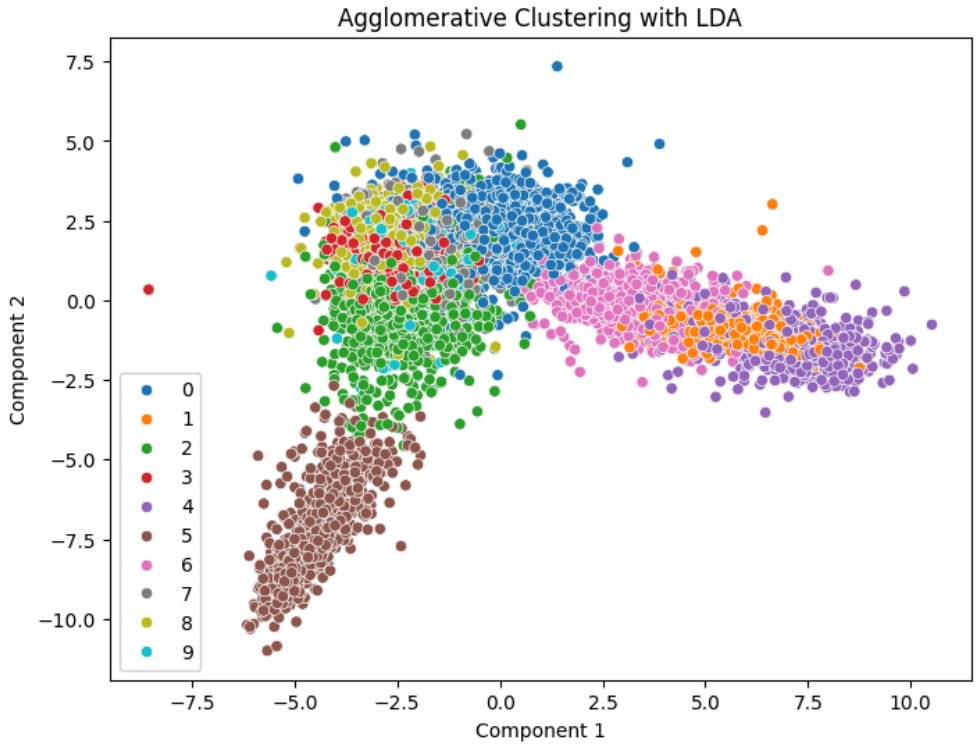
Εικόνα 2.2.5: Εφαρμογή PCA στο test set και συσταδοποίηση με Mini Batch K-Means



Εικόνα 2.2.6: Εφαρμογή PCA στο test set και συσταδοποίηση με Agglomerative Clustering



Εικόνα 2.2.7: Εφαρμογή LDA στο test set και συσταδοποίηση με Mini Batch K-Means

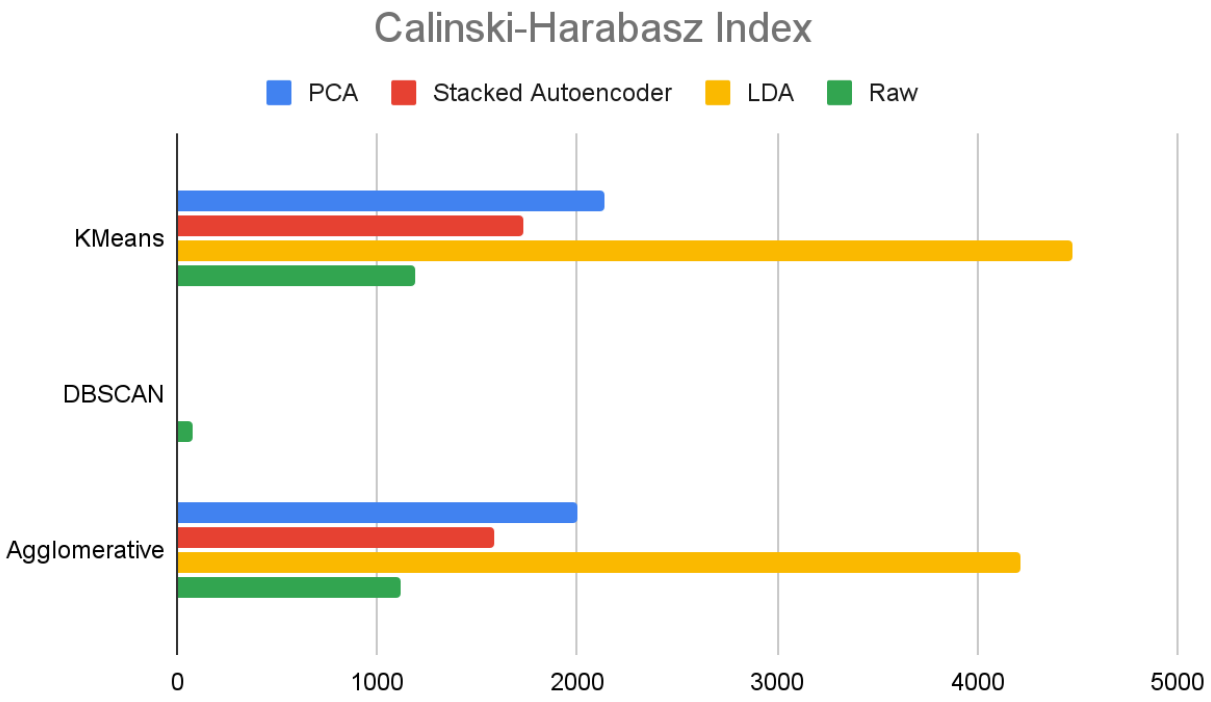


Εικόνα 2.2.8: Εφαρμογή LDA στο test set και συσταδοποίηση με Agglomerative Clustering

2.3. Μετρικές

Ακολουθεί αξιολόγηση της συσταδοποίησης σύμφωνα με τις μετρικές Calinski-Harabasz Index, Davies-Bouldin Index, Silhouette Score και Adjusted Rand Index. Συγκεκριμένα για κάθε μετρική παρουσιάζεται ένα διάγραμμα που απεικονίζει την απόδοση κάθε αλγόριθμου συσταδοποίησης (Mini Batch K-means, DBSCAN, Agglomerative Clustering) τόσο σε δεδομένα που έχουν υποστεί dimensionality reduction με κάποια εκ των τριών τεχνικών: Principal Component Analysis (PCA), Stacked Autoencoder και Linear Discriminant Analysis (LDA), όσο και σε ακατέργαστα δεδομένα (Raw).

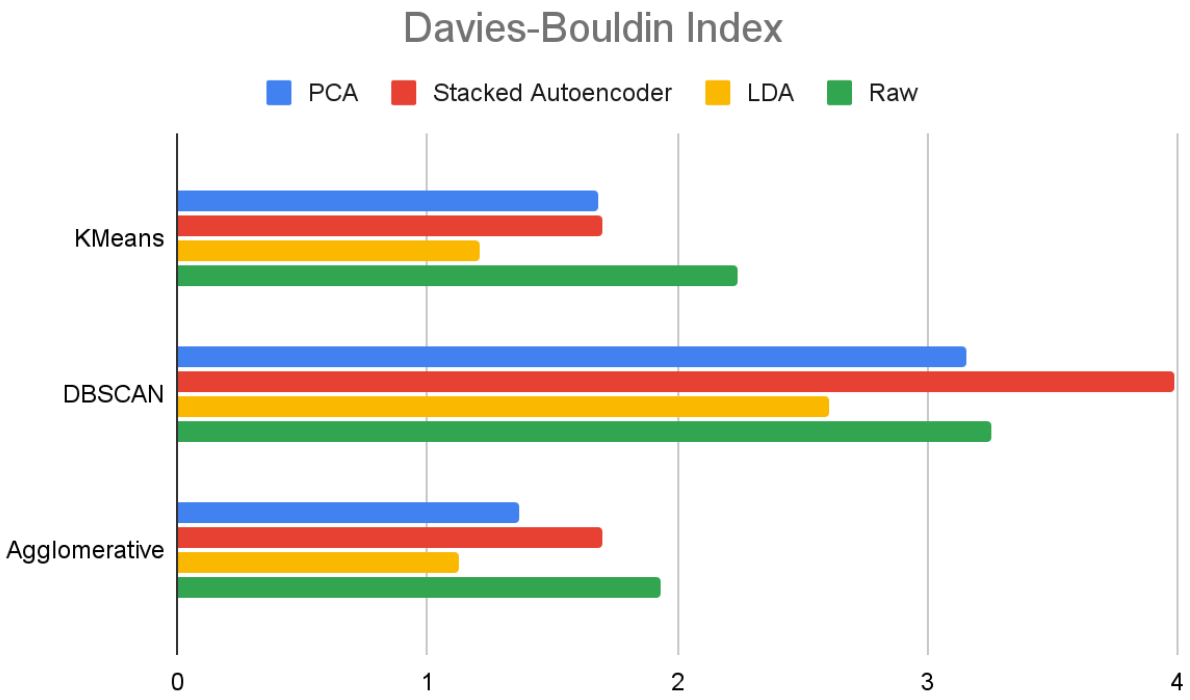
2.3.a. Calinski-Harabasz Index



Εικόνα 2.3.1: Τιμές Μετρικής Calinski-Harabasz Index

Το παραπάνω διάγραμμα απεικονίζει τις τιμές της μετρικής Calinski-Harabasz Index. Παρατηρούμε πως ο αλγόριθμος DBSCAN παρουσιάζει, σε καθε περιπτωση, πολύ χαμηλές τιμές, κοντά στο 0, το οποίο φανερώνει μη καλά διαχωρισμένες συστάδες. Αντίθετα, οι αλγόριθμοι K-means και Agglomerative για όλες τις τεχνικές dimensionality reduction παρουσιάζουν παρόμοια απόδοση, με τον K-means να κατέχει συνολικά ελαφρώς καλύτερες τιμές και την Linear Discriminant Analysis να κάνει το καλύτερο feature extraction. Οι υψηλές τιμές των δύο αυτών αλγορίθμων για την τεχνική Linear Discriminant Analysis δηλώνουν καλά διαχωρισμένες και συνεκτικές συστάδες.

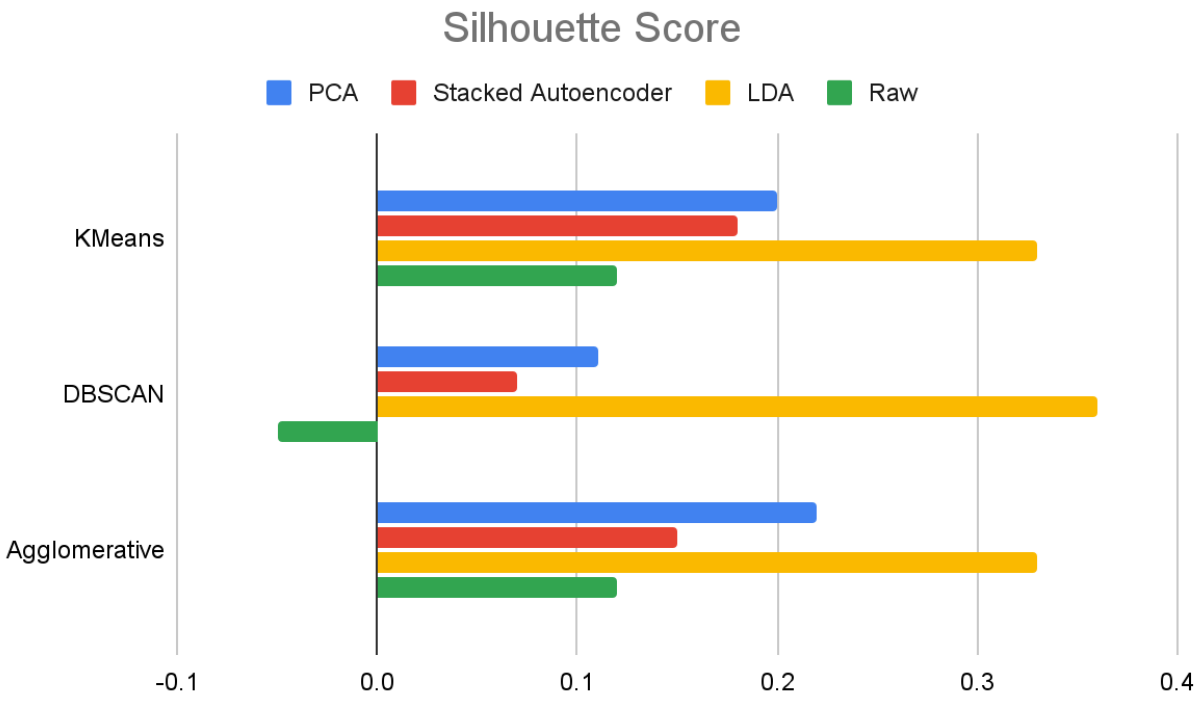
2.3.b. Davies-Bouldin Index



Εικόνα 2.3.2: Τιμές Μετρικής Davies-Bouldin Index

Στο συγκεκριμένο διάγραμμα διακρίνουμε τις τιμές της μετρικής Davies-Bouldin Index, για την οποία, σε αντίθεση με τις υπόλοιπες μετρικές, ισχύει πως όσο μικρότερη είναι η τιμή της τόσο πιο συμπαγής και καλά διαχωρισμένη θεωρείται μια κλάση. Παρατηρούμε πως οι αλγόριθμοι K-means και Agglomerative για όλες τις τεχνικές dimensionality reduction παρουσιάζουν παρόμοια απόδοση, με τον Agglomerative να είναι συνολικά ελαφρώς καλύτερος και την Linear Discriminant Analysis να κάνει το καλύτερο feature extraction. Ο αλγόριθμος DBSCAN σημειώνει τις υψηλότερες τιμές για όλες τις τεχνικές dimensionality reduction, με χειρότερη την τεχνική Stacked Autoencoder.

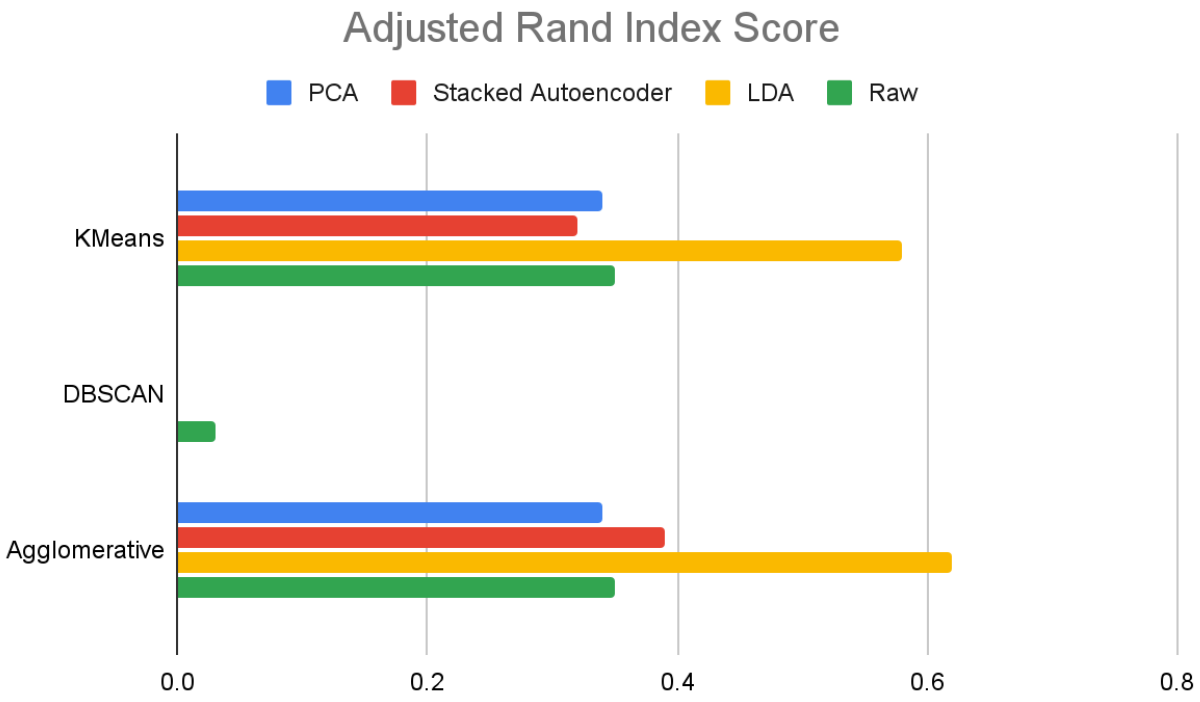
2.3.c. Silhouette Score



Εικόνα 2.3.3: Τιμές Μετρικής Silhouette Score

Στο παραπάνω διάγραμμα φαίνονται τα αποτελέσματα της μετρικής Silhouette Score. Αρχικά παρατηρούμε πως όταν ο αλγόριθμος DBSCAN έτρεξε σε ακατέργαστα δεδομένα, η μετρική Silhouette Score έλαβε αρνητική τιμή, γεγονός που υποδηλώνει μη καλό διαχωρισμό κλάσεων, δηλαδή αρκετά σημεία ενδέχεται να παρουσιάζουν μεγαλύτερη ομοιότητα με σημεία διαφορετικής κλάσης από αυτή στην οποία ανατέθηκαν. Οι αλγόριθμοι K-means και Agglomerative για όλες τις τεχνικές dimensionality reduction φαίνεται να παρουσιάζουν παρόμοια απόδοση, με την Linear Discriminant Analysis να κάνει το καλύτερο feature extraction, χωρίς παρ' όλα αυτά να πετυχαίνει υψηλή τιμή, υποδεικνύοντας όχι πλήρως διακριτές κλάσεις και μη πολύ καλή συσταδοποίηση. Οι αντίστοιχες τιμές για τον αλγόριθμο DBSCAN είναι χαμηλότερες, εκτός από την περίπτωση που εφαρμόζεται η Linear Discriminant Analysis ως τεχνική dimensionality reduction, που παρουσιάζει την υψηλότερη τιμή για τη μετρική Silhouette Score, το οποίο δείχνει πως τα σημεία στο dataset είναι πυκνά, ιδιαίτερα μετά την αφαίρεση γνωρισμάτων (δεδομένου των *Εικόνα 2.2.1* και *Εικόνα 2.2.4*).

2.3.d. Adjusted Rand Index Score



Εικόνα 2.3.4: Τιμές Μετρικής Adjusted Rand Index Score

Η μετρική Adjusted Rand Index παρουσιάζει ιδιαίτερο ενδιαφέρον, καθώς αξιοποιεί τα πραγματικά labels των δεδομένων προκειμένου να αξιολογήσει τη συσταδοποίηση. Συγκεκριμένα, αξιολογεί κατά πόσο τα labels που ανέθεσε ο clustering αλγόριθμος συμφωνούν με τα πραγματικά. Παρατηρούμε πως σε κανέναν αλγόριθμο και για καμία τεχνική dimensionality reduction η τιμή της Adjusted Rand Index δεν είναι μικρότερη του μηδενός, πράγμα που θα υποδείκνυε παραπλανητικά και αναξιόπιστα αποτελέσματα. Οι αλγόριθμοι K-means και Agglomerative για όλες τις τεχνικές dimensionality reduction παρουσιάζουν παρόμοια απόδοση, με την Linear Discriminant Analysis να κάνει την καλύτερη αφαίρεση γνωρισμάτων. Ο αλγόριθμος DBSCAN φαίνεται να μην δουλεύει καλά για τα συγκεκριμένα δεδομένα, καθώς για όλες τις τεχνικές dimensionality reduction η τιμή της μετρικής είναι 0, δηλαδή η συσταδοποίηση δεν είναι καλύτερη από τυχαία ανάθεση των σημείων σε συστάδες, ενώ για την αρχική μορφή των δεδομένων τα αποτελέσματα δεν παρουσιάζουν ιδιαίτερη βελτίωση, αφού η τιμή της μετρικής εξακολουθεί να βρίσκεται κοντά στο 0.

3. Βέλτιστος Συνδυασμός Τεχνικών

Η εύρεση ενός βέλτιστου συνδυασμού τεχνικών αντιστοιχεί στον συνδυασμό τεχνικών που επιδεικνύουν την καλύτερη απόδοση σε κάθε μετρική. Παρατηρούμε πως η καλύτερη τεχνική μείωσης διαστάσεων είναι ο LDA, ενώ οι αλγόριθμοι συσταδοποίησης αποδίδουν χειρότερα κατά τη χρήση ακατέργαστων (Raw) δεδομένων. Οι χαμηλές τιμές των μετρικών κατά τη χρήση των προαναφερθέντων τεχνικών dimensionality reduction σε συνδυασμό με τον αλγόριθμο DBSCAN οφείλονται στην ακαταλληλότητα του αλγορίθμου για το συγκεκριμένο dataset.

Δυστυχώς, δεν παρατηρείται περίπτωση όπου ο LDA σε συνδυασμό με κάποιον αλγόριθμο ομαδοποίησης επιτυγχάνει την καλύτερη απόδοση σε κάθε μετρική. Οι βέλτιστες επιδόσεις παρατηρούνται με τον LDA συνδυασμένο με Mini Batch K-means, όπου παρουσιάζουν τις καλύτερες αποδόσεις στις μετρικές Calinski-Harabasz Index και Silhouette Score, καθώς και με τον LDA συνδυασμένο με Agglomerative, όπου επιτυγχάνονται τα καλύτερα αποτελέσματα στις μετρικές Davies-Bouldin Index και Adjusted Rand Index Score.

Ένα επιπλέον κριτήριο που μπορούμε να λάβουμε υπόψη είναι ο χρόνος που χρειάζονται οι συνδυασμοί τεχνικών. Αναφερόμενοι μόνο στους δύο προαναφερόμενους συνδυασμούς τεχνικών, γρηγορότερος είναι ο LDA με τον K-means, αν και ο LDA συνδυασμένος με τον Agglomerative είναι πιο αργός κατά μόνο 3 δευτερόλεπτα. Γενικότερα, και οι δύο συνδυασμοί μπορούν να θεωρηθούν αρκετά γρήγοροι, όπως είδαμε και στο τμήμα «*2.1 Διάρκεια Τεχνικών*».

Συνοψίζοντας, σύμφωνα με τα παραπάνω, αν και δεν εντοπίστηκε βέλτιστος συνδυασμός τεχνικών, η καλύτερη τεχνική μείωσης διαστάσεων εμφανίζεται να είναι ο Linear Discriminant Analysis (LDA), ενώ η χειρότερη τεχνική ομαδοποίησης είναι ο DBSCAN.