Παρουσίαση Αναφοράς

Τεχνητή Νοημοσύνη, Μηχανική Μάθηση και Εφαρμογές

Παναγιώτης Τόγιας | Μαριαλένα Ματσούκα

Εαρινό Εξάμηνο 2019
ΔΠΜΣ ΠΕΖ

Προεπεξεργασία

- Αφαίρεση γονιδίων με χαμηλή ή μηδέν διακύμανση

Συγκεκριμένα, έγινε αφαίρεση μεταβλητών με εξής χαρακτηριστικά:

  - Με μία μοναδική τιμή για κάθε δείγμα (μηδέν διακύμανση)
  - Ανήκουν σε μία από τις δύο παρακάτω περιπτώσεις
    - έχουν πολύ λίγες μοναδικές τιμές σε σχέση με τον αριθμό των δειγμάτων (το λιγότερο 10)
    - ο λόγος της συχνότητας της πιο συχνά εμφανιζόμενης τιμής με τη συχνότητα της δεύτερης πιο συχνά εμφανιζόμενης τιμής είναι μεγάλος

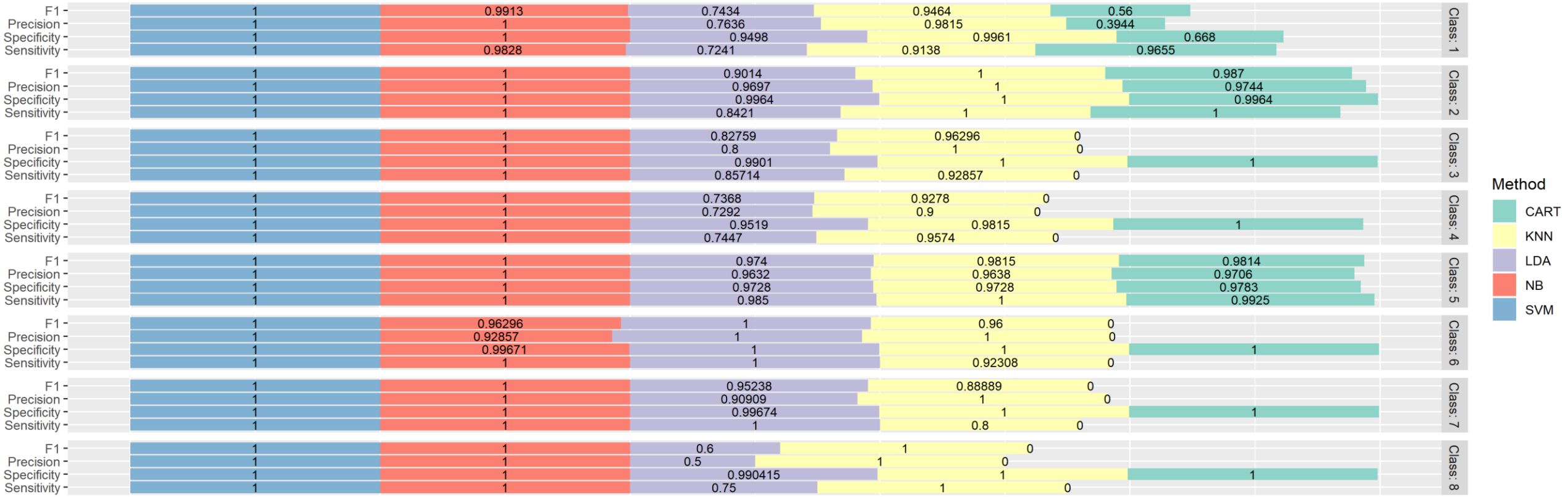
- Κανονικοποίηση z-score $z_i = \frac{x_i - \mu}{\sigma}$

από ~22000 σε ~15000 γονίδια με τις αντίστοιχες εγγραφές για κάθε κλάση:

1	2	3	4	5	6	7	8
58	38	14	47	133	13	10	4

Classification

10-fold Cross-Validation



Καλύτερη Επίδοση

SVM, Naïve Bayes

Οι παραπάνω αλγόριθμοι παρουσίασαν τιμές οι οποίες αγγίζουν το 100% σε κάθε κλάση

Μέση Επίδοση

LDA, KNN

Με καλύτερη απόδοση αυτή του KNN ειδικά στις κλάσεις με μικρό αριθμό εγγραφών

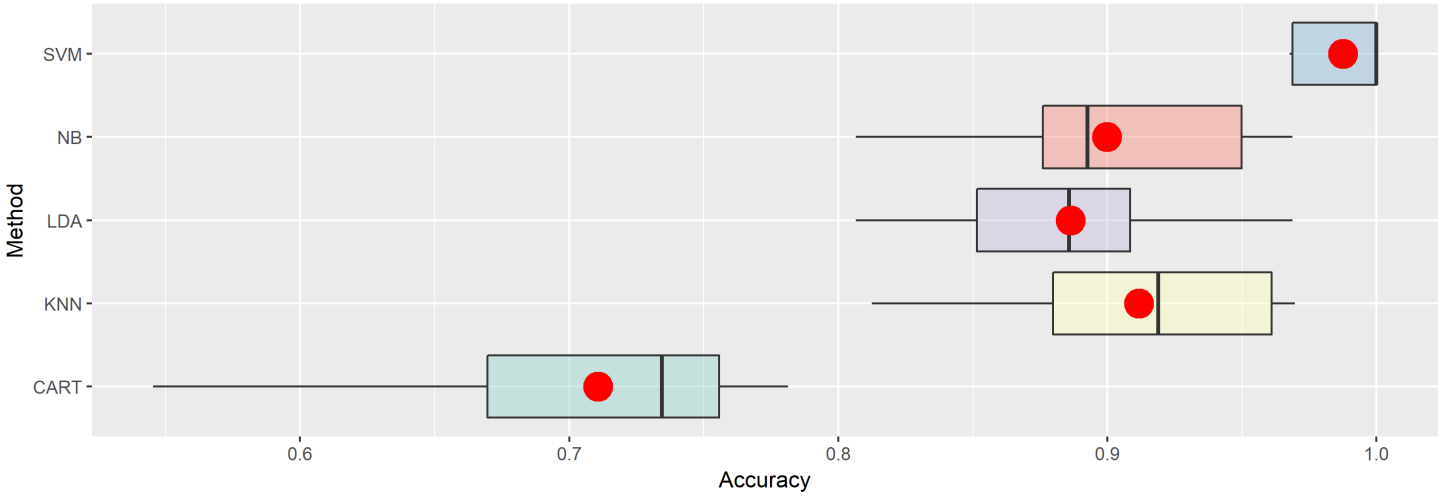
Χειρότερη Επίδοση

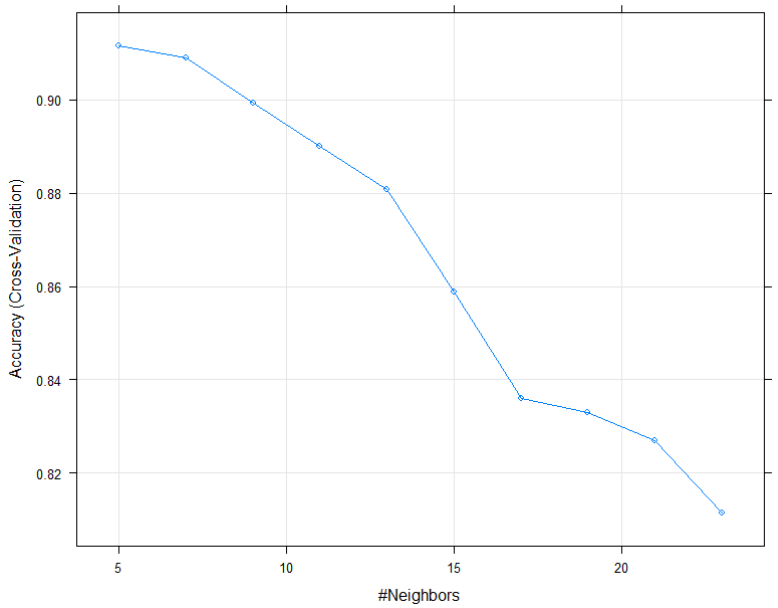
C. Trees

Τα δέντρα απόφασης, παρουσίασαν σημαντική αδυναμία στον χειρισμό των κλάσεων με μικρό αριθμό εγγραφών

Classification

10-fold Cross-Validation





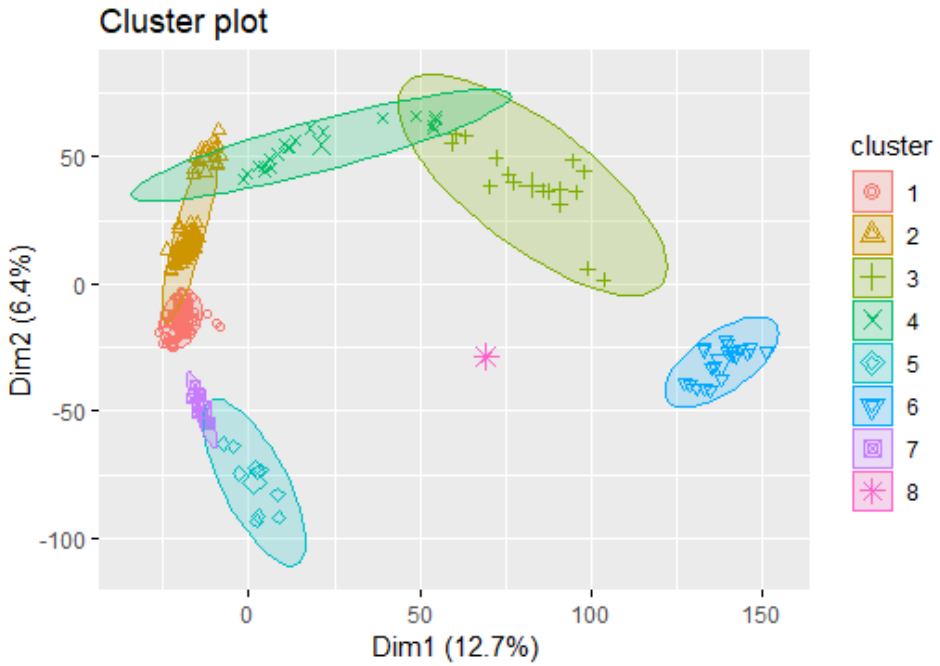
Το ίδιο ισχύει και στον υπολογισμό της μετρικής Accuracy για κάθε αλγόριθμο, χρησιμοποιώντας κάθε τιμή από την αντίστοιχη επανάληψη του Cross-Validation και την δημιουργία θηκογραμμάτων για την απεικόνισή της. Με σειρά καλύτερης επίδοσης, λοιπόν, υπάρχει η εξής ταξινόμηση:

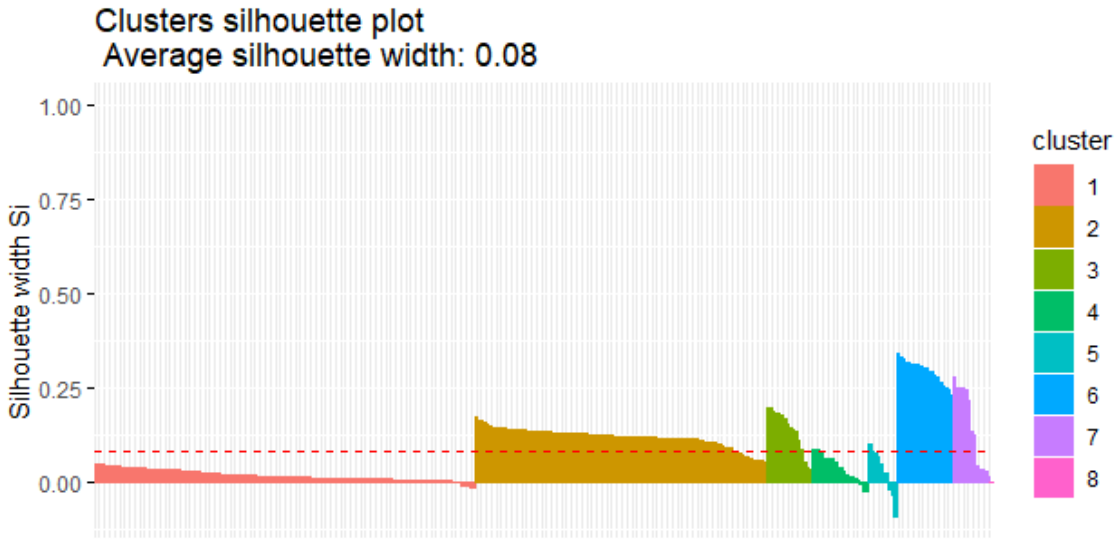
Support Vector Machines
Naïve Bayes
K-Nearest Neighbors
Linear Discriminant Analysis
Decision Trees

Συγκεκριμένα, για τον αλγόριθμο k-Nearest Neighbor, έγινε δοκιμή με διάφορες τιμές για k και επιλέχτηκε αυτή με το καλύτερο Accuracy (k=5 στη συγκεκριμένη περίπτωση)

Clustering

K-Means

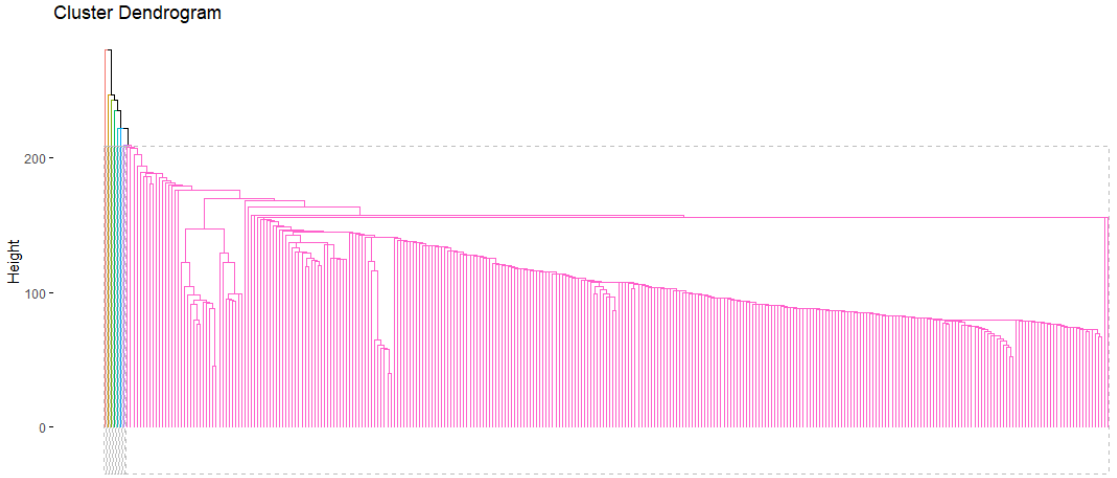


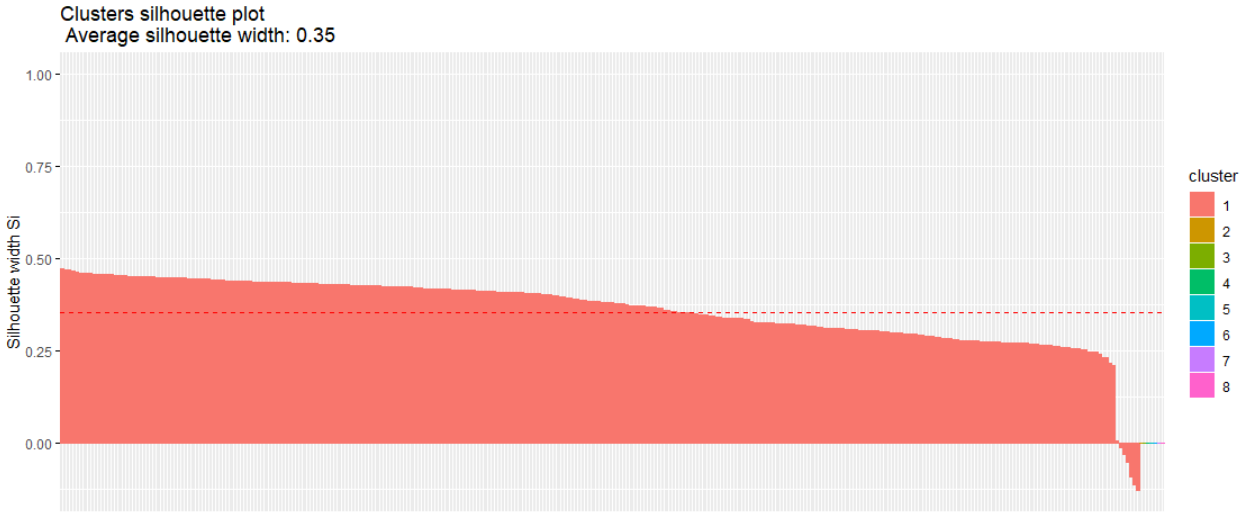


cluster	cluster size	Average silhouette width
1	134	0.02
2	103	0.12
3	16	0.14
4	20	0.04
5	10	0.03
6	20	0.29
7	13	0.14
8	0	0.00

Clustering

Hierarchical - Single Linkage

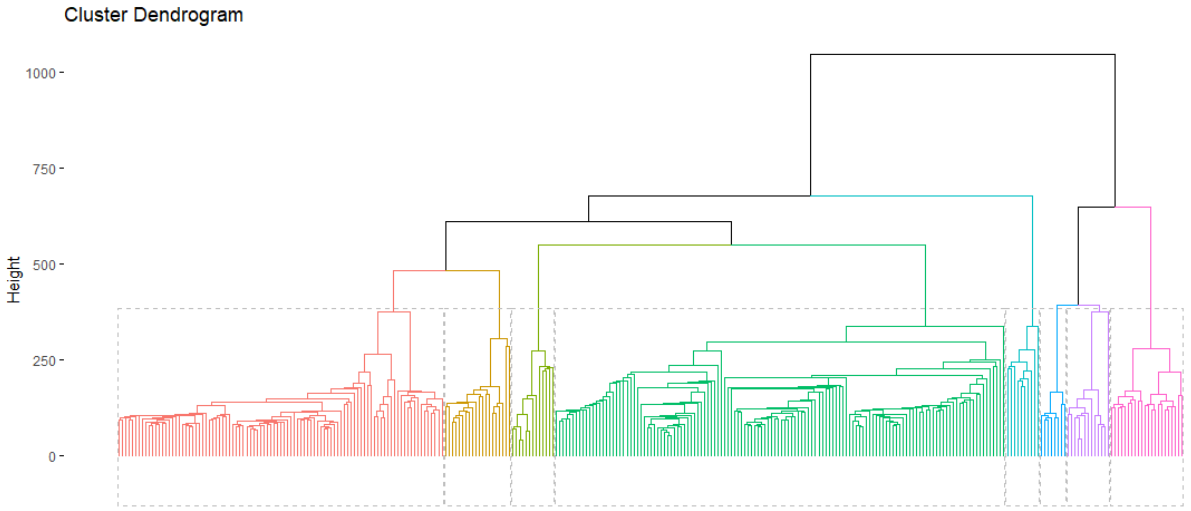


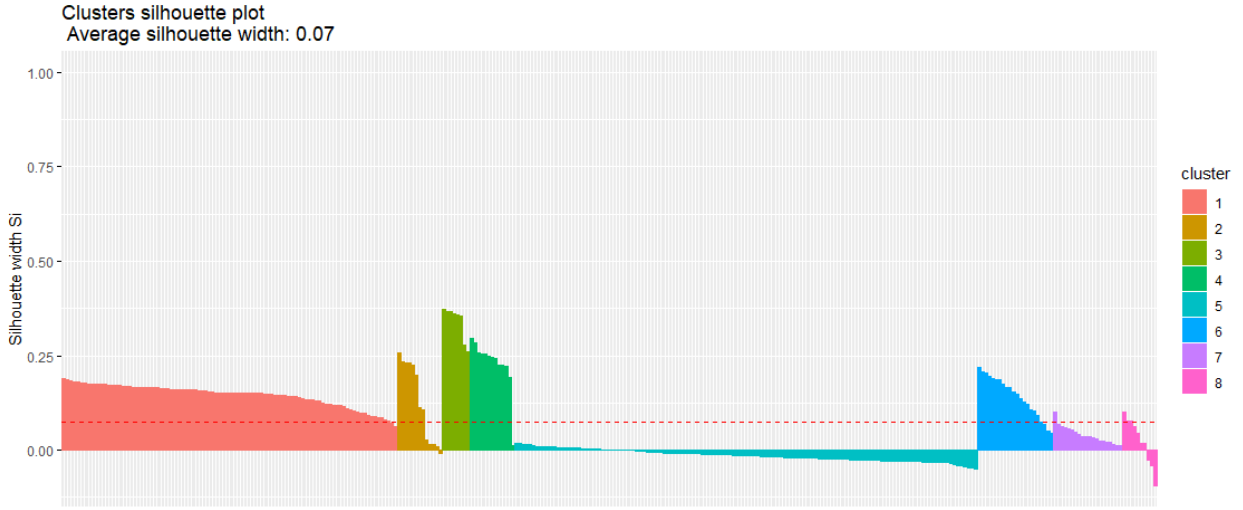


cluster	cluster size	Average silhouette width
1	310	0.36
2	1	0.00
3	1	0.00
4	1	0.00
5	1	0.00
6	1	0.00
7	1	0.00
8	1	0.00

Clustering

Hierarchical Agglomerative

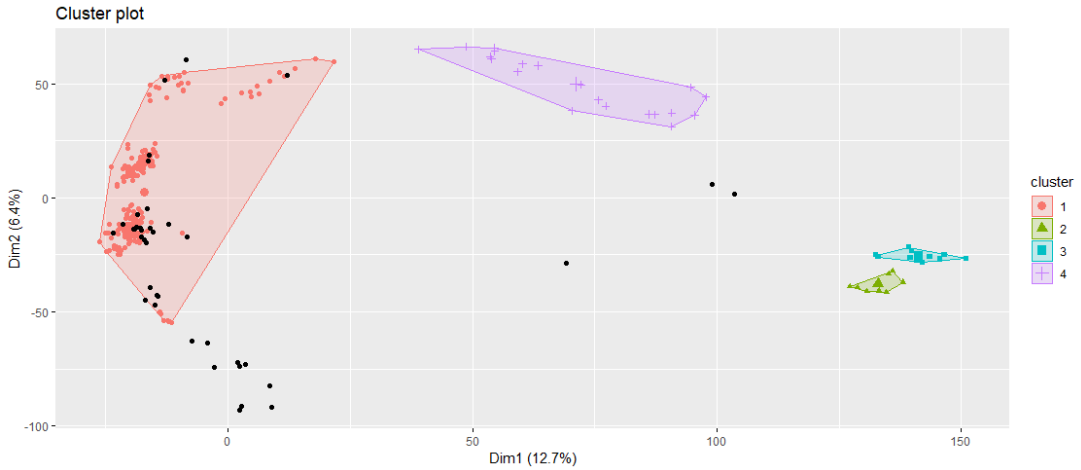


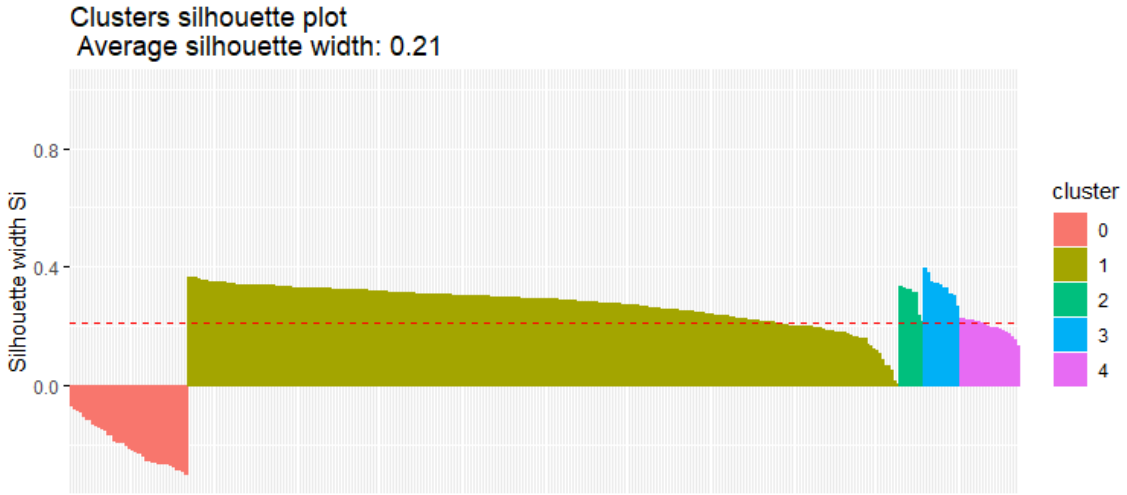


cluster	cluster size	Average silhouette width
1	97	0.15
2	13	0.13
3	8	0.34
4	13	0.23
5	134	-0.01
6	22	0.14
7	20	0.04
8	10	0.02

Clustering

DBSCAN

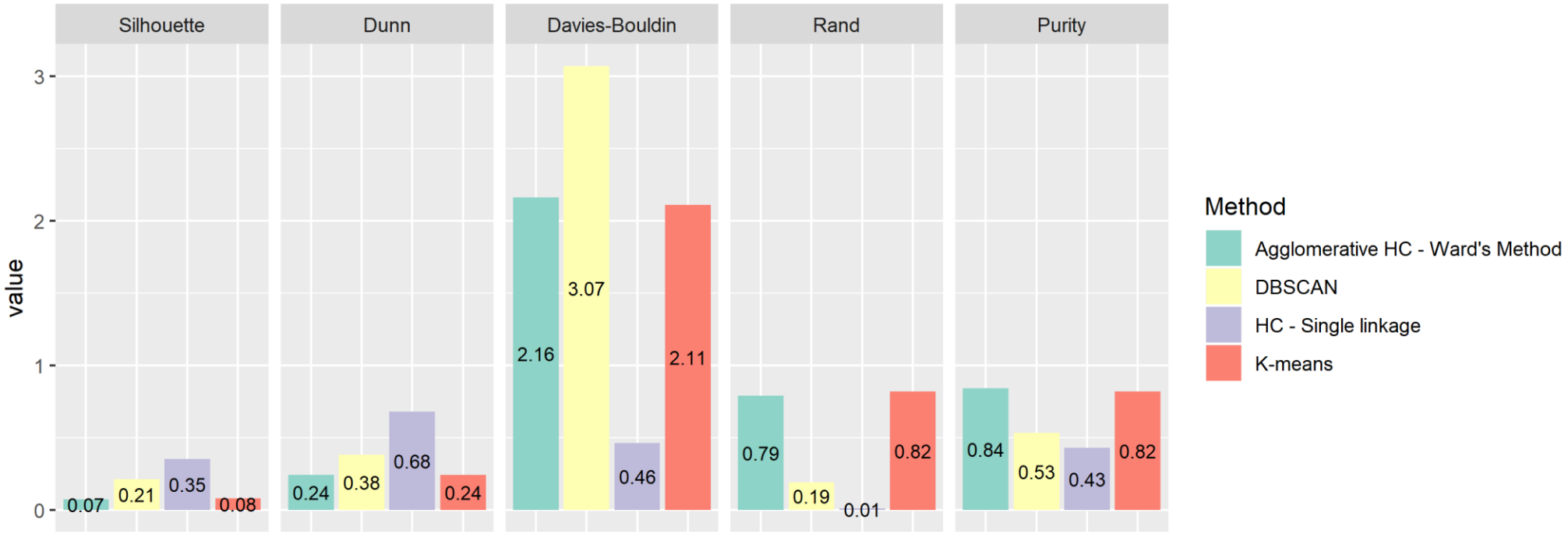




cluster	cluster size	Average silhouette width
0	39	-0.20
1	238	0.27
2	8	0.30
3	12	0.33
4	20	0,20

Clustering

Αποτελέσματα



Καθαρά με πληροφορία από τα εξωτερικά κριτήρια Rand και Purity, μπορούμε να επιβεβαιώσουμε και τις απεικονίσεις του κάθε αλγόριθμου παραπάνω και να καταλήξουμε στο συμπέρασμα ότι οι αλγόριθμοι k-means και AGNES απέδωσαν καλύτερα στην εύρεση των αρχικών 8 κλάσεων του συνόλου δεδομένων

Παρουσίαση Αναφοράς

Τεχνητή Νοημοσύνη, Μηχανική Μάθηση και Εφαρμογές

Παναγιώτης Τόγιας | Μαριαλένα Ματσούκα

Εαρινό Εξάμηνο 2019
ΔΠΜΣ ΠΕΖ