Studio e approfondimento del coefficiente Silhouette per la valutazione interna di risultati di clustering non-supervisionato

Federico Salvioni 845029

17 febbraio 2025

Clustering

Cluster

"A number of similar things that occur together"

Cluster analysis

"A statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics"

Quanti e quali sono i cluster?

- Gli algoritmi di clustering non supervisionato non sono in grado di determinare se il loro operato rispecchi la struttura di cluster, perché non vi è alcuna "ground truth";
- Utilizzare parametri errati induce un raggruppamento non rappresentativo del dataset, senza che l'algoritmo lo consideri un errore;
- Sono state proposte diverse metriche che sono in grado di stimare la qualità dell'operato di un algoritmo di clustering non supervisionato.

Introduzione di Silhouette

La metrica Silhouette si propone di rispondere alle seguenti domande:

- Il clustering è di buona qualità?
- Quali elementi sono stati ben classificati?
- Quali elementi sono inclassificabili?
- Il numero di cluster scelto è rappresentativo del dataset?

Questo è possibile a partire da una nozione di **distanza**.

Formula per la Silhouette

$a(i)$

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in \{A - \{i\}\}} d(i, j)$$

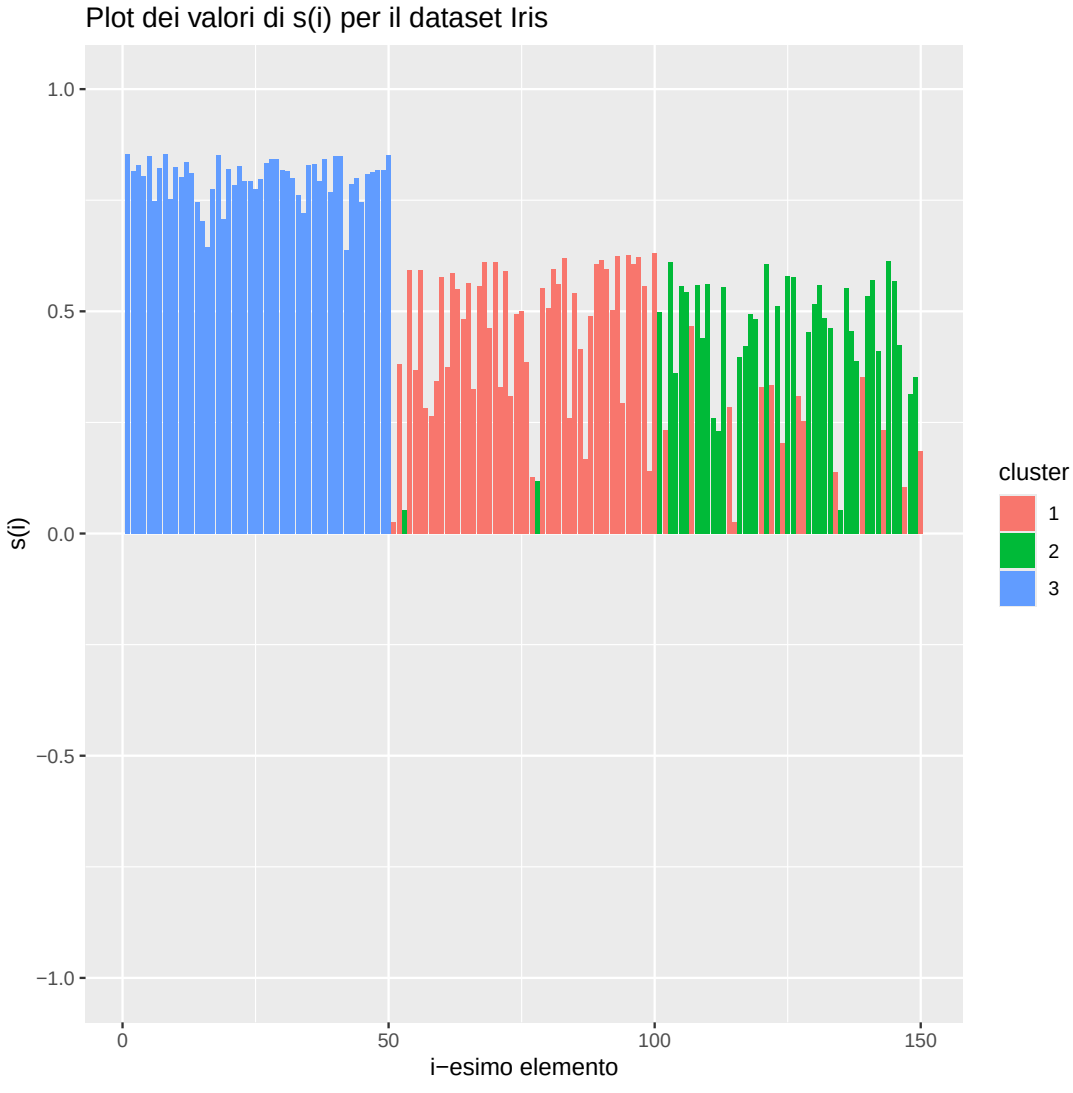
$b(i)$

$$b(i) = min_{C \neq A} \frac{1}{|C|} \sum_{j \in C} d(i, j)$$

$s(i)$

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Silhouette width



Algoritmi di clustering

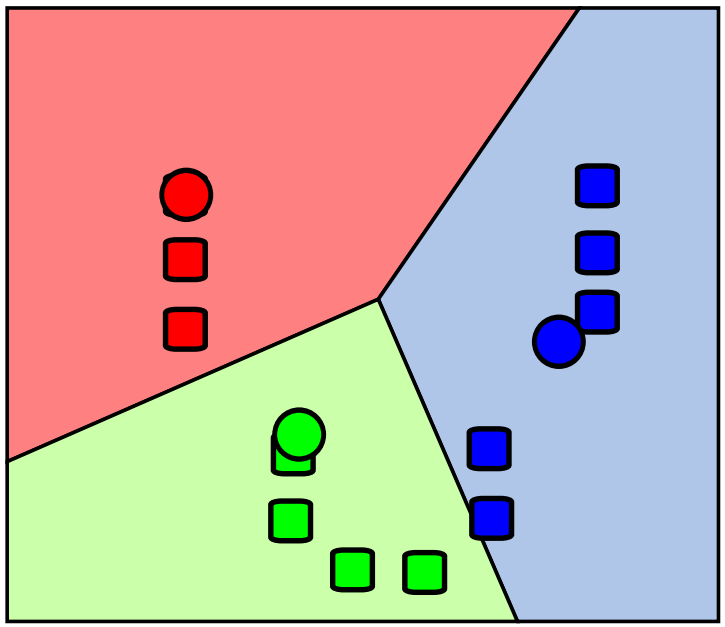


Figura: K-Means

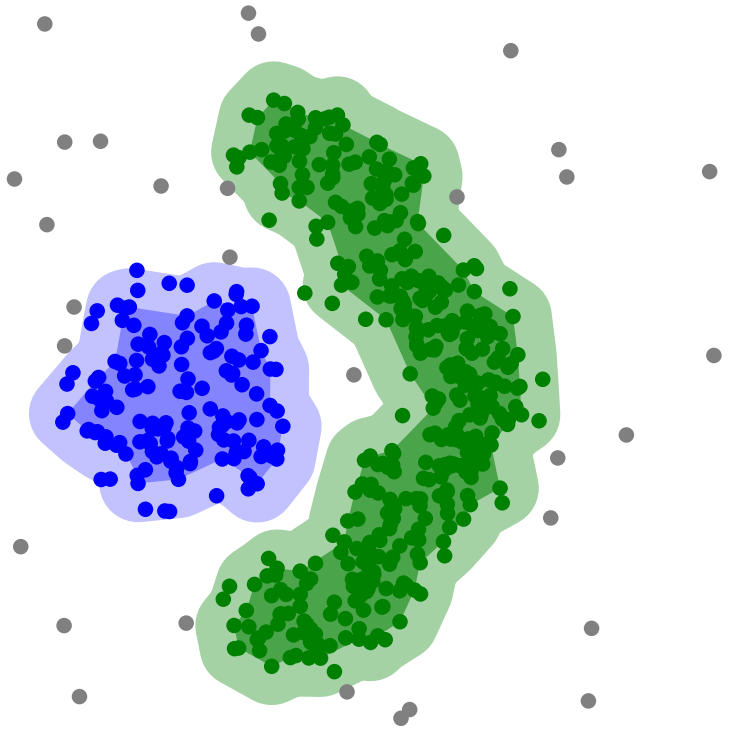


Figura: DBSCAN

Matrice binaria

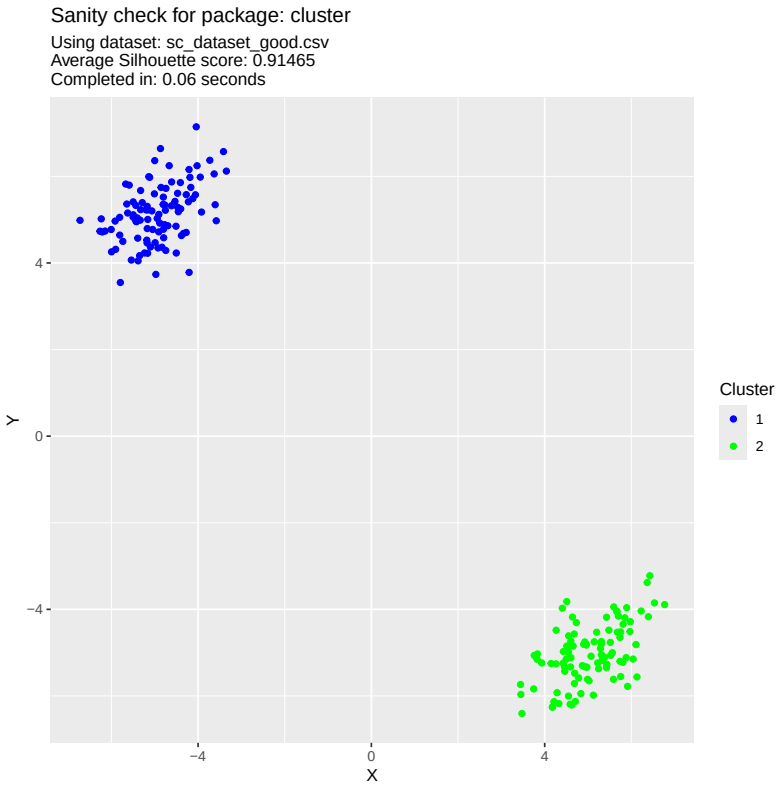


Figura: Sanity check (favorevole)

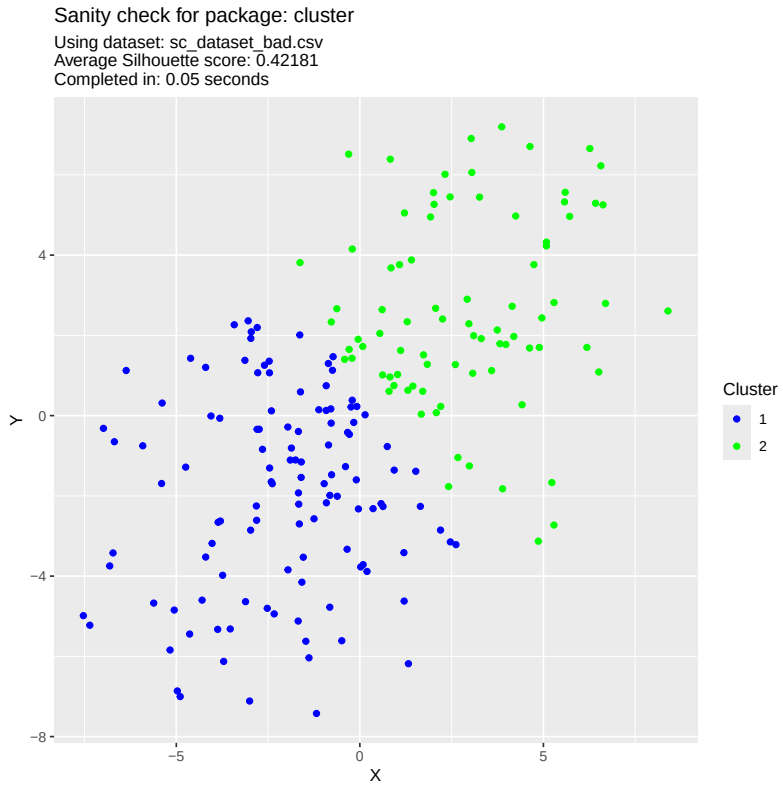


Figura: Sanity check (sfavorevole)

Sanity check

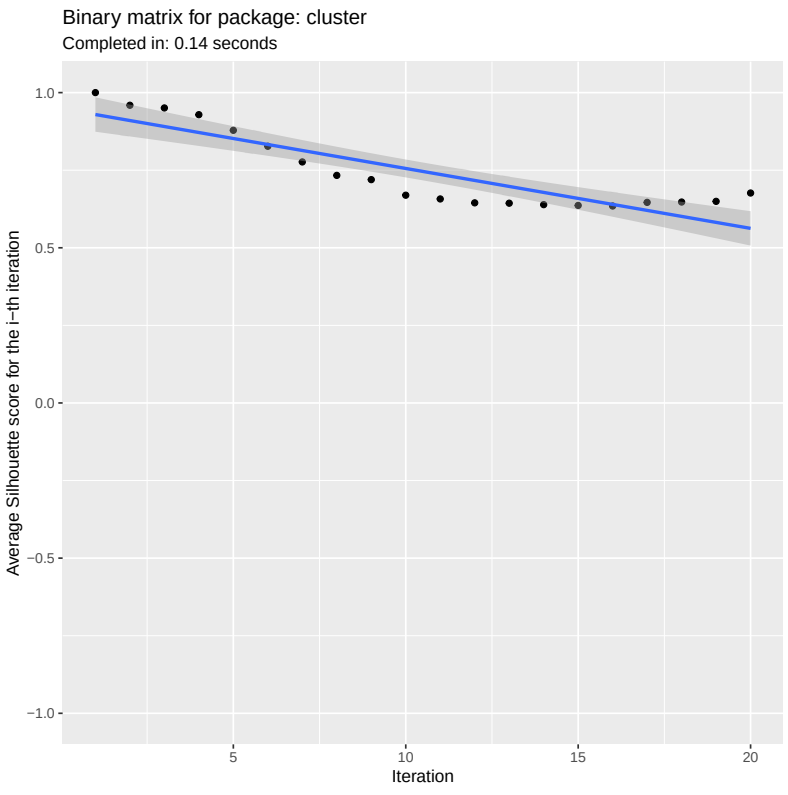


Figura: Matrice binaria

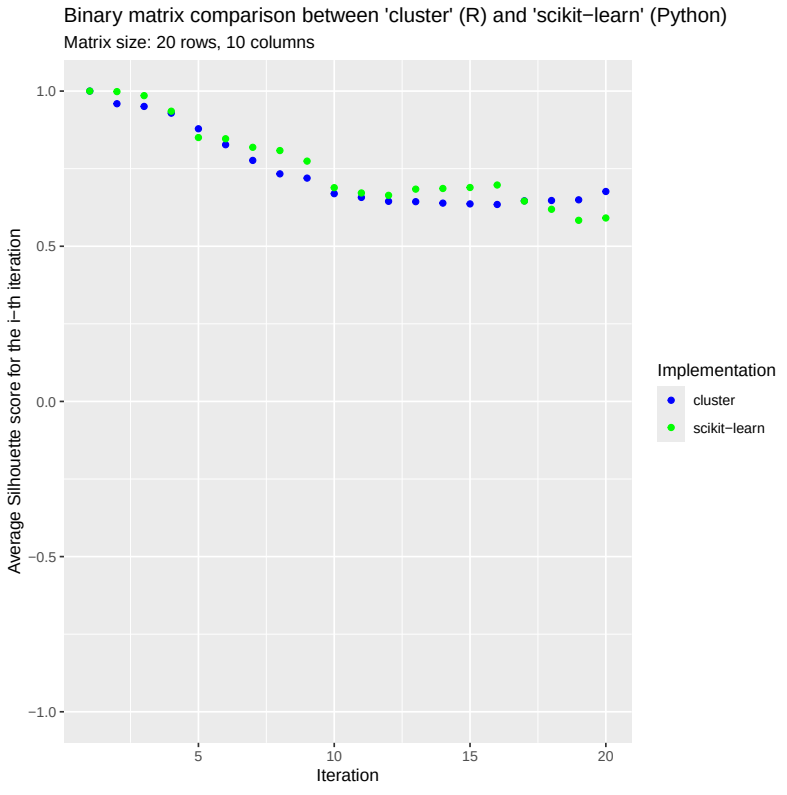


Figura: Comparazione R/Python

Clustering su dataset reali

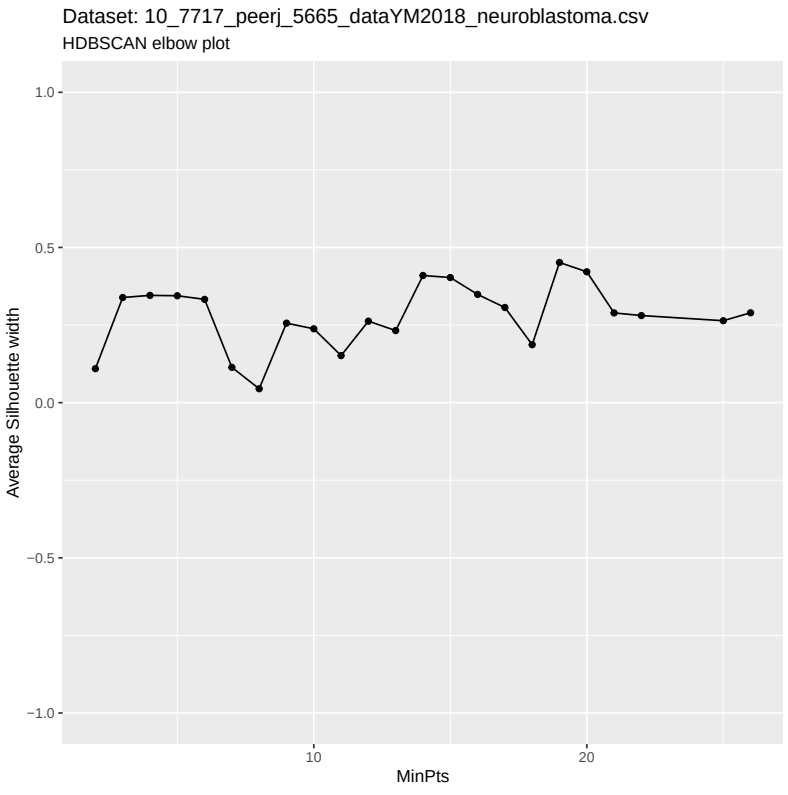


Figura: Elbow plot

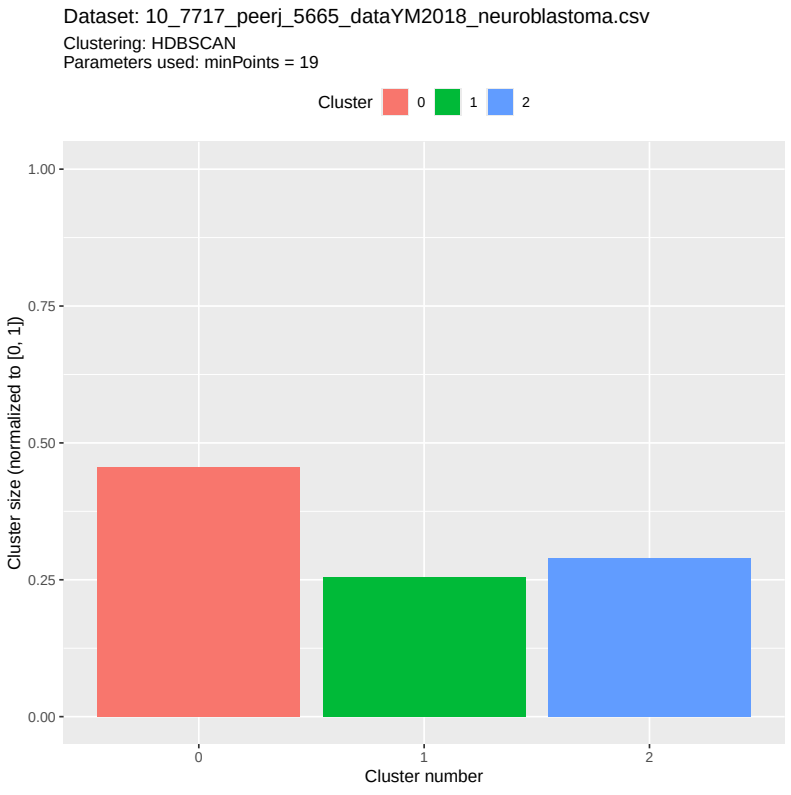


Figura: Clustering

Clustering su dataset reali

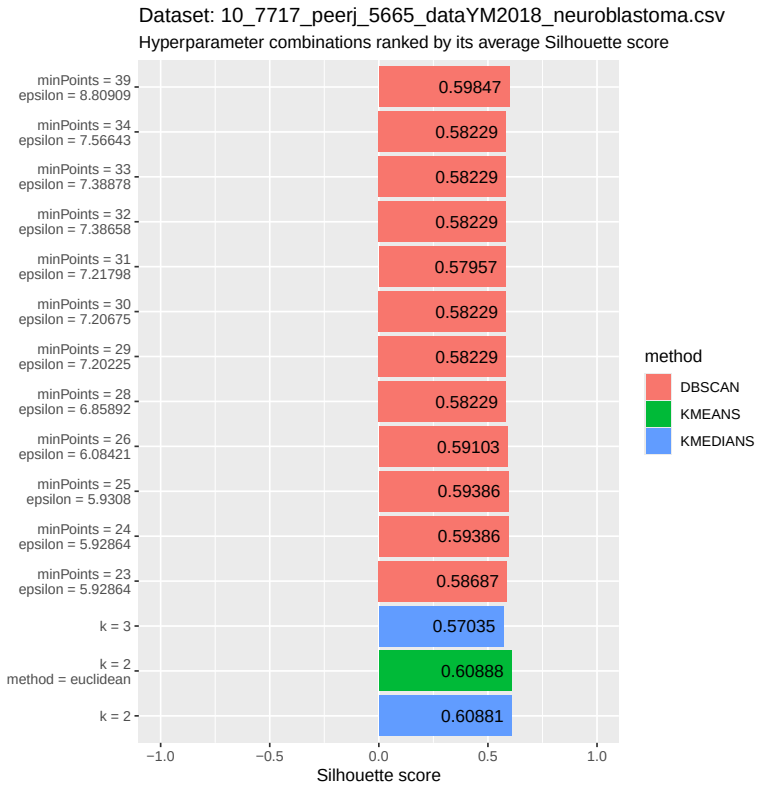


Figura: Ranking (1)

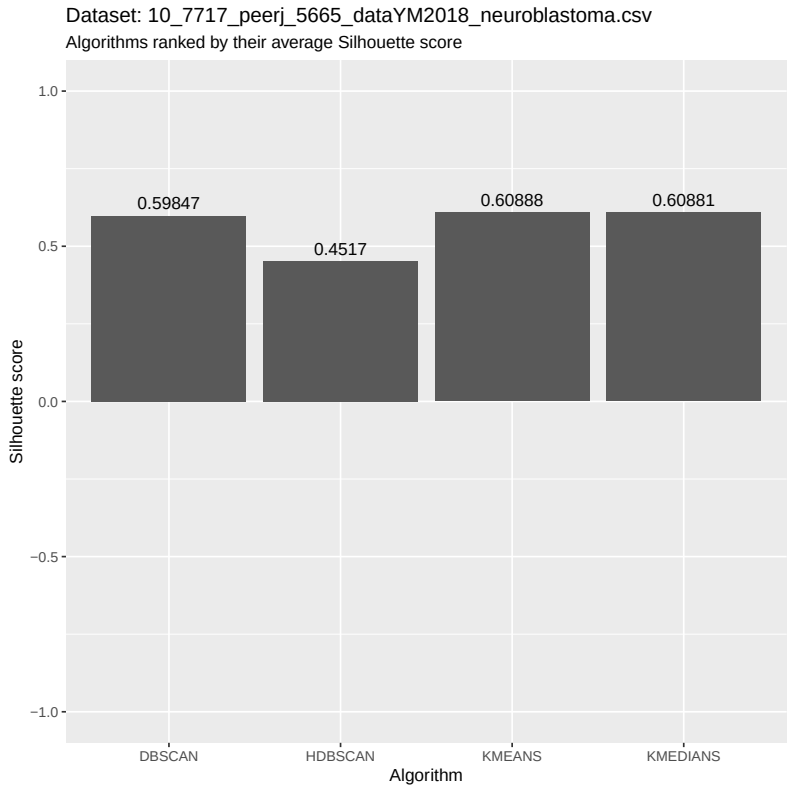


Figura: Ranking (2)

Sviluppi futuri

Possibili estensioni:

- Usare dataset che non siano EHR;
- Sostituire i valori ignoti con valori concreti, anziché scartare gli elementi manchevoli;
- Ridurre il numero di dimensioni (**principal component analysis**).

Fine

Domande?