**Università degli Studi di Milano - Bicocca**
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea in Informatica

# Analisi e confronto di metriche di validazione interna per il clustering

Tesi di laurea di: **Andrea Spagnolo**

Matricola: 879254

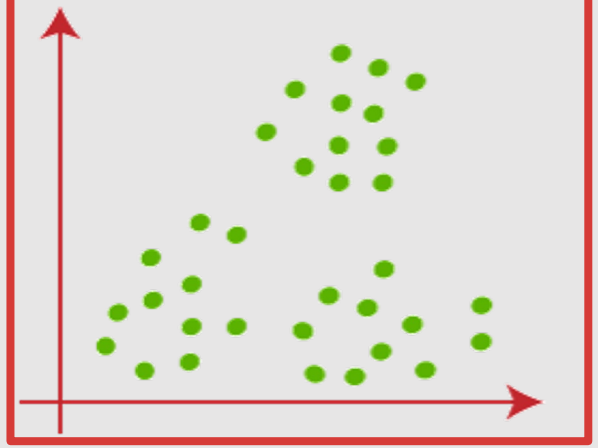
**Relatore**: Davide Ciucci
**Co-relatore**: Davide Chicco

**Anno accademico:** 2023-2024

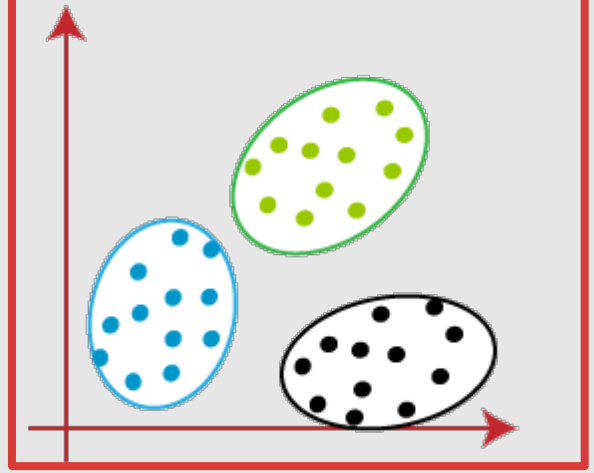
# Introduzione

- Il **clustering** è una tecnica di apprendimento non supervisionato utilizzata per raggruppare dati simili.
- Il **limite** principale del clustering è la difficoltà nel valutare la qualità dei risultati ottenuti.

Prima di k-means



Dopo k-means



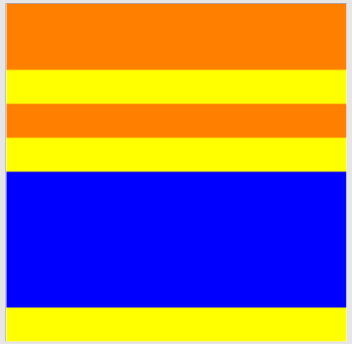
## Obiettivo

Confrontare le metriche Silhouette coefficient, Entropia, Davies-Bouldin index, Calinski-Harabasz index, Dunn index e Gap statistic per valutare quali siano le più efficaci.

# Metodi

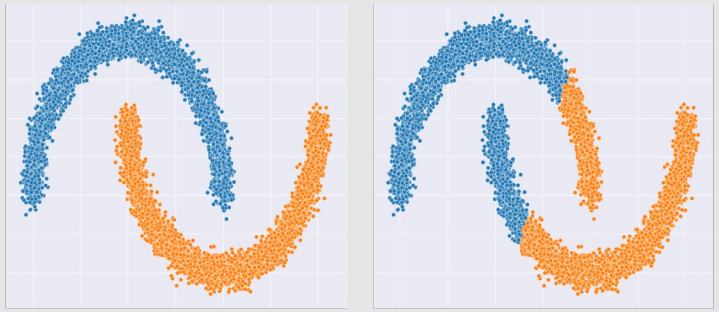
- Per determinare quali tra le 6 metriche performa meglio è stato svolto un test su diversi tipi di dataset.
- Questi dataset rappresentano casi dove il risultato del clustering va peggiorando, ed occorre verificare se e quali di queste metriche sono consistenti con questo peggioramento.

## Matrice di zeri e uni



Zeri Righe modificate Uni

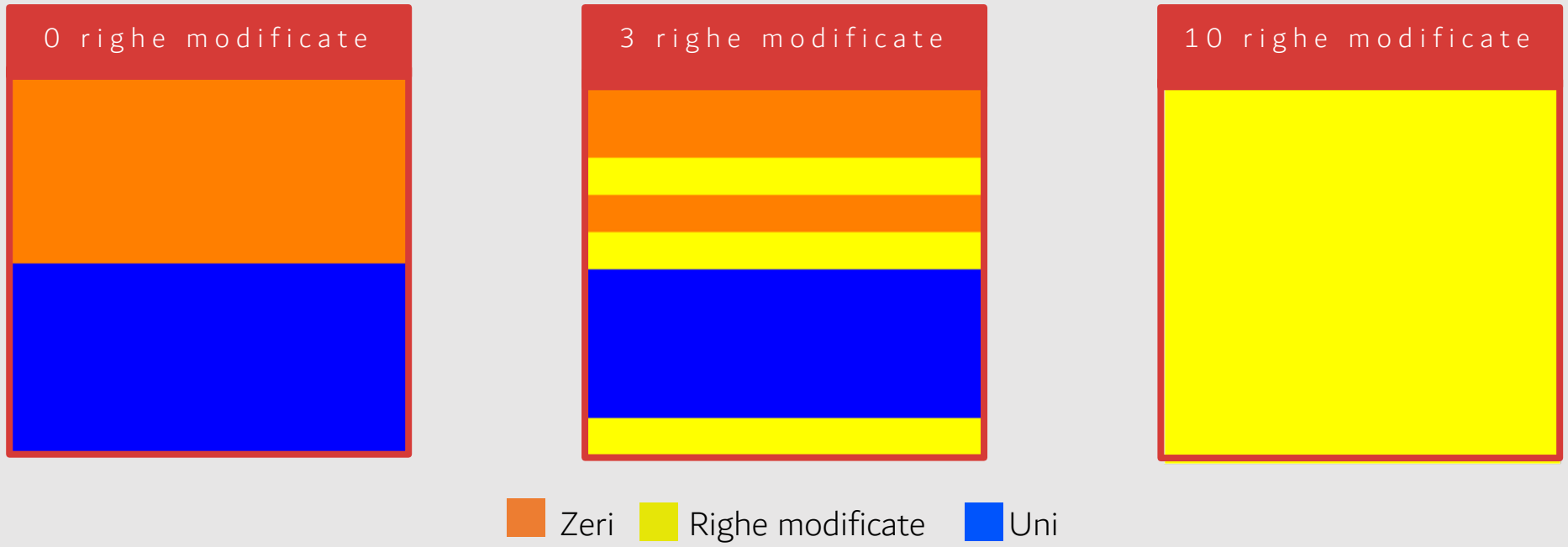
## Dataset artificiali



## Dataset reali EHRs

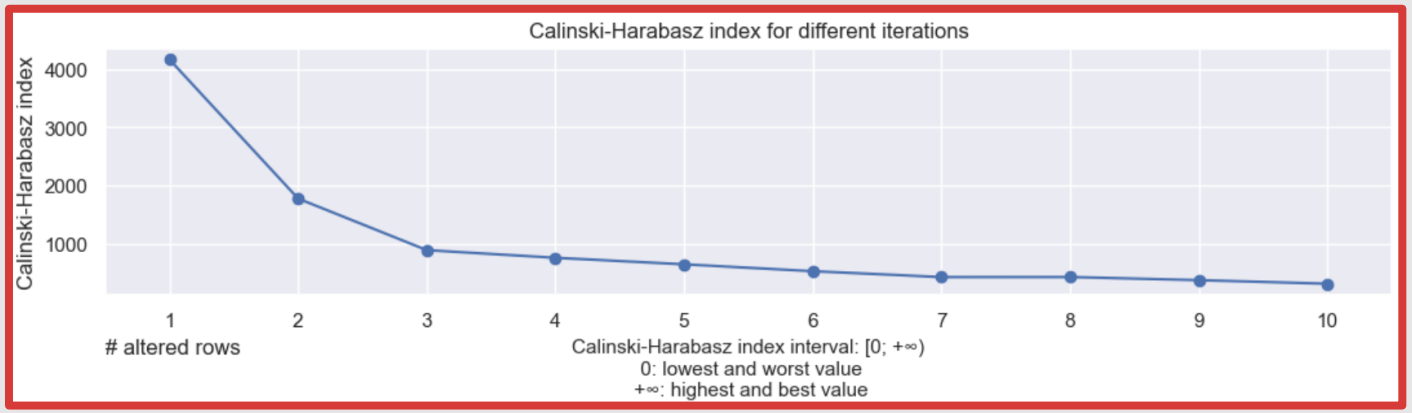
# Matrice di zeri e uni

- Ho applicato le sei metriche alla matrice di zeri e uni con una riga che viene manipolata ad ogni iterazione.
- L'**obiettivo** è capire quali metriche sono consistenti con il peggioramento della qualità del clustering.



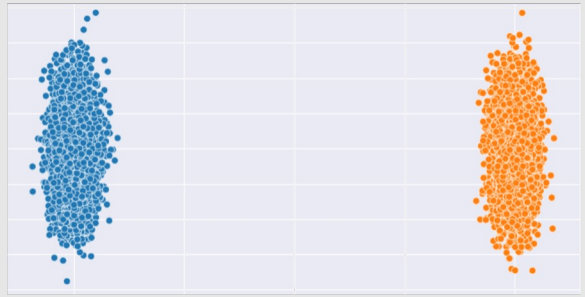
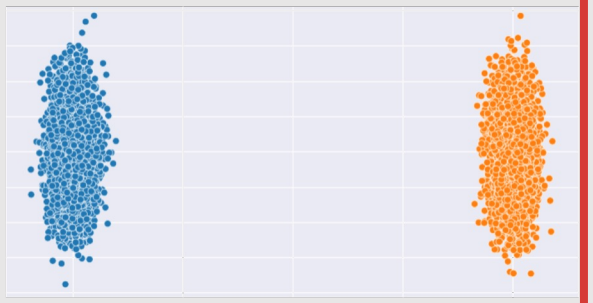
# Risultati dataset Matrice di zeri e uni

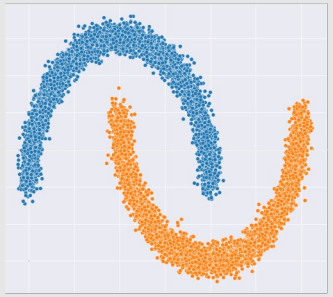
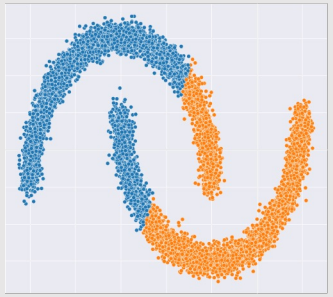
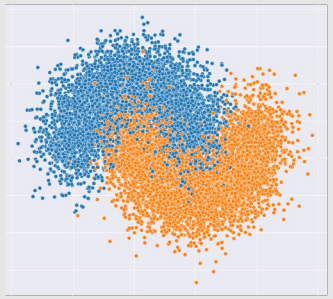
Dato questo dataset mi aspetto che le metriche peggiorino all'aumentare delle righe modificate.
Ciò è accaduto con tutte le metriche tranne che con l'Entropia che rimane sempre costante e con la Gap Statistic che ha un andamento irregolare.




| metric | result | time |
|---|---|---|
| Silhouette | correct | 0.901 ms |
| Complementary Entropy | wrong | 0.299 ms |
| Reciprocal Davies-Bouldin | correct | 0.435 ms |
| Dunn index | correct | 0.274 ms |
| Calinski-Harabasz | correct | 0.197 ms |
| Gap Statistic | wrong | 238.138 ms |

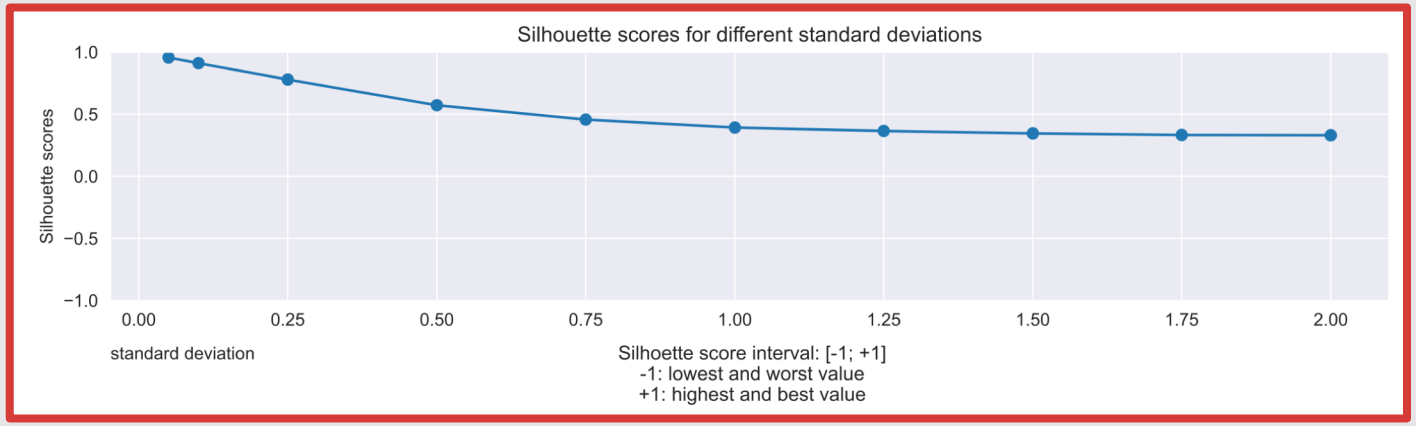
# Dataset artificiali

| Prima di k-means | Dopo di k-means |
| --- | --- |
|  |  |
|  |  |
|  |  |

| Prima di k-means | Dopo di k-means |
| --- | --- |
|  |  |
|  |  |
|  |  |

# Risultati dataset artificiali

È stato analizzato il comportamento delle metriche per ognuno dei 7 dataset. **Silhouette** e **Davies-Bouldin** sono le metriche che hanno registrato le migliori performance.




| metric | correct | wrong | % | average time |
|---|---|---|---|---|
| Silhouette | 6 | 1 | 85.71 | 0.284 sec |
| Reciprocal Davies-Bouldin | 6 | 1 | 85.71 | 0.0003 sec |
| Dunn index | 5 | 2 | 71.43 | 0.121 sec |
| Calinski-Harabasz | 4 | 3 | 57.14 | 0.0003 sec |
| Gap Statistic | 3 | 4 | 42.86 | 0.077 sec |
| Complementary Entropy | 1 | 6 | 14.29 | 0.0001 sec |

# Dataset reali

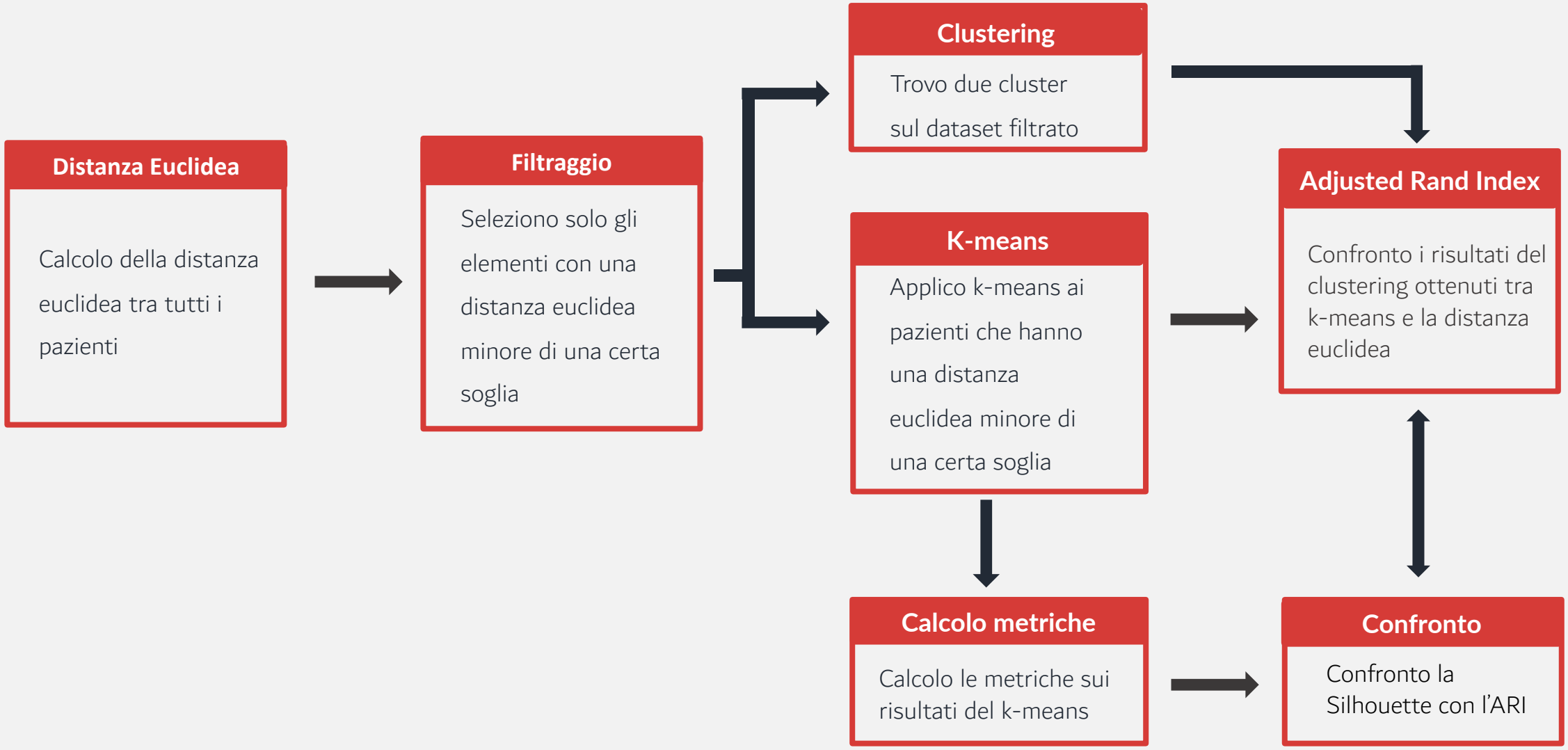
Ho preso in esame 5 diversi dataset medici reali riguardanti:

- la sepsi & SIRS
- la correlazione tra insufficienza cardiaca e depressione
- gli arresti cardiaci
- il neuroblastoma
- il diabete di tipo 1

Dataset artificiale cartella clinica

| age | sex_0male_1female | education | cancer_stage |
|---|---|---|---|
| 5 | 0 | 1 | 1 |
| 6 | 0 | 1 | 1 |
| 5 | 0 | 1 | 1 |
| 4 | 0 | 1 | 1 |
| 6 | 0 | 1 | 1 |
| 80 | 1 | 4 | 4 |
| 79 | 1 | 4 | 4 |
| 78 | 1 | 4 | 4 |
| 78 | 1 | 4 | 4 |
| 81 | 1 | 4 | 4 |

# Procedimento dataset reali



# Risultati dataset reali

È stato analizzato il comportamento delle metriche per ognuno dei 6 dataset.
**Silhouette**, **Davies-Bouldin** e **Dunn index** sono le metriche che hanno registrato le migliori performance.

| metric | correct | wrong | % | avarage time |
|---|---|---|---|---|
| Silhouette | 6 | 0 | 100.0 % | 3.346 ms |
| Reciprocal Davies-Bouldin | 6 | 0 | 100.0 % | 1.069 ms |
| Dunn index | 6 | 0 | 100.0 % | 1.095 ms |
| Calinski-Harabasz | 5 | 1 | 83.333 % | 0.806 ms |
| Gap Statistic | 4 | 2 | 66.667 % | 199.575 ms |
| Complementary Entropy | 3 | 3 | 50.0 % | 0.258 ms |

# Conclusione

- Analizzando i risultati, le metriche **Silhouette** e **Davies-Bouldin** sono risultate le **più performanti.**
- Questo è particolarmente interessante se si considera che queste metriche sono tra le più utilizzate e diffuse all'interno della comunità scientifica, suggerendo una conferma empirica dell'affidabilità.