



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA
Dipartimento di Informatica, Sistemistica e
Comunicazione
Corso di Laurea in Informatica

Analisi e validazione di metriche interne per il Clustering

Relatore: Prof. Davide Ciucci

Correlatore: Dott. Davide Chicco

Tesi di Laurea di:
Andrea Spagnolo
Matricola 879254

Anno Accademico 2023-2024

Abstract

Una delle problematiche principali degli algoritmi di clustering è valutare la qualità dei risultati ottenuti. Per affrontare questa problematica, si possono utilizzare diversi strumenti, tra cui le metriche di valutazione interna. Sebbene esistano diverse metriche di valutazione interna, sono pochi gli studi che ne confrontano le performance. Lo scopo del mio progetto di stage è quello di confrontare le metriche Silhouette coefficient, Entropia, Davies-Bouldin index, Calinski-Harabasz index, Dunn index e Gap statistic per valutare quali siano le più efficaci. Per determinare quali tra queste metriche sono le più significative ho creato dei dataset artificiali in cui i risultati del clustering vanno peggiorando, ed ho verificato se e quali di queste metriche sono consistenti con questo peggioramento. Successivamente ho testato le metriche anche su dati reali di cartelle cliniche elettroniche riguardanti la SIRS, la correlazione tra infarti e depressione, l'arresto cardiaco in Spagna, il neuroblastoma ed il diabete di tipo 1, sempre con lo scopo di confrontare le metriche e capire quali performano meglio. I risultati ottenuti hanno evidenziato come le metriche che sono risultate più efficaci nel mio studio, corrispondono anche a quelle attualmente più utilizzate e conosciute.

Contents

1	Introduzione	7
1.1	Algoritmi di clustering	8
1.1.1	K-means	8
1.2	Revisione letteratura	10
2	Metriche interne di validazione	11
2.1	Silhouette coefficient	11
2.2	Entropia	12
2.3	Davies-Bouldin index	13
2.4	Calinski-Harabasz index	14
2.5	Dunn index	14
2.6	Gap statistic	15
2.7	Riassunto metriche	16
3	Datasets	17
3.1	Dataset artificiale: matrice di zeri e uni	17
3.2	Dataset artificiali	17
3.2.1	Dataset nuvole	18
3.2.2	Dataset cerchi	19
3.2.3	Dataset semicerchi	20
3.2.4	Dataset parabole	21
3.2.5	Dataset sfere	22
3.2.6	Dataset pennellate	23
3.2.7	Dataset W	24
3.3	Dataset medici reali di cartelle cliniche elettroniche	25
3.3.1	Dataset medico artificiale	25
3.3.2	Dataset sepsi & SIRS	26
3.3.3	Dataset depressione ed insufficienza cardiaca	26
3.3.4	Dataset arresto cardiaco	27
3.3.5	Dataset neuroblastoma	27
3.3.6	Dataset diabete di tipo 1	27
4	Risultati	28
4.1	Risultati dataset artificiale: matrice di zeri e uni	28
4.2	Risultati dataset artificiali	30
4.2.1	risultati dataset nuvole	31
4.2.2	risultati dataset cerchi	33
4.2.3	risultati dataset semicerchi	35

4.2.4	risultati dataset parabole	37
4.2.5	risultati dataset sfere	39
4.2.6	risultati dataset pennellate	41
4.2.7	risultati dataset W	43
4.2.8	riassunto risultati dataset artificiali	45
4.3	Risultati dataset medici reali di cartelle cliniche elettroniche	46
4.3.1	dataset artificiale	46
4.3.2	risultati dataset sepsi & SIRS	49
4.3.3	risultati dataset depressione e insufficienza cardiaca	51
4.3.4	risultati dataset arresto cardiaco	52
4.3.5	risultati dataset neuroblastoma	54
4.3.6	risultati dataset diabete di tipo 1	56
4.3.7	riassunto risultati dataset reali di cartelle cliniche mediche	58
5	Conclusione e discussione	60
6	Dettagli tecnici	62
6.1	dettagli tecnici codice	62
6.2	dettagli tecnici hardware utilizzato	62
6.3	contatti	63

List of Figures

2.1	spiegazione Silhouette	12
3.1	matrice di zeri e uni con 0 righe modificate	17
3.2	matrice di zeri e uni con 3 righe modificate	17
3.3	matrice di zeri e uni con 10 righe modificate	17
3.4	cluster dataset nuvole	18
3.5	cluster dataset cerchi	19
3.6	cluster dataset semicerchi	20
3.7	cluster dataset parabole	21
3.8	cluster dataset sfere	22
3.9	cluster dataset pennellate	23
3.10	cluster dataset W	24
3.11	differenze cause tra Sepsi e SIRS	26
4.1	risultati dataset artificiale numerico	28
4.2	risultati dataset nuvole	31
4.3	risultati dataset cerchi	33
4.4	risultati dataset semicerchi	35
4.5	risultati dataset parabole	37
4.6	risultati dataset sfere	39
4.7	risultati dataset pennellate	41
4.8	risultati dataset W	43
5.1	istogramma sui numero di articoli delle metriche presenti su google scholar	61

List of Tables

2.1	riassunto metriche	16
3.1	dataset medico artificiale	25
4.1	risultati metriche dataset matrice di 0 e 1	30
4.2	risultati metriche dataset nuvole	32
4.3	risultati metriche dataset cerchi	34
4.4	risultati metriche dataset semicerchi	36
4.5	risultati metriche dataset parabole	38
4.6	risultati metriche dataset sfere	40
4.7	risultati metriche dataset pennellate	42
4.8	tabella risultati metriche dataset W	44
4.9	tabella riassuntiva dell'andamento delle metriche nei dataset artificiali	45
4.10	tabella dei cluster ottenuti con la distanza euclidea sul dataset artificiale con soglia del 20%	47
4.11	tabella dei cluster ottenuti k-means sul dataset artificiale con soglia del 20%	47
4.12	tabella dei cluster ottenuti con la distanza euclidea sul dataset artificiale con soglia del 33%	47
4.13	tabella dei cluster ottenuti k-means sul dataset artificiale con soglia del 33%	48
4.14	risultati metriche su dataset medico artificiale	48
4.15	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante la sepsi & SIRS con soglia del 20%	49
4.16	tabella dei cluster ottenuti k-means sul dataset riguardante la sepsi & SIRS con soglia del 20%	49
4.17	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante la sepsi & SIRS con soglia del 33%	50
4.18	tabella dei cluster ottenuti con k-means sul dataset riguardante la sepsi & SIRS con soglia del 33%	50
4.19	risultati metriche su dataset riguardante sepsi & SIRS	50
4.20	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante depressione & insufficienza cardiaca con soglia del 20%	51
4.21	tabella dei cluster ottenuti k-means sul dataset riguardante depressione & insufficienza cardiaca con soglia del 20%	51
4.22	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante depressione & insufficienza cardiaca con soglia del 33%	52

4.23	tabella dei cluster ottenuti con k-means sul dataset riguardante depressione & insufficienza cardiaca con soglia del 33%	52
4.24	risultati metriche sul dataset degli riguardante depressione & insufficienza cardiaca	52
4.25	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante l'arresto cardiaco con soglia del 20%	53
4.26	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante l'arresto cardiaco con soglia del 20%	53
4.27	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante l'arresto cardiaco con soglia del 33%	54
4.28	tabella dei cluster ottenuti con k-means sul dataset riguardante l'arresto cardiaco con soglia del 33%	54
4.29	tabella risultati metriche sul dataset riguardante l'arresto cardiaco	54
4.30	tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante il neuroblastoma con soglia del 20%	55
4.31	tabella dei cluster ottenuti con k-means sul dataset riguardante il neuroblastoma con soglia del 20%	55
4.32	tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il neuroblastoma con soglia del 33%	56
4.33	tabella dei cluster ottenuti con k-means sul dataset riguardante il neuroblastoma con soglia del 33%	56
4.34	tabella risultati metriche su dateset riguardante il neuroblastoma	56
4.35	tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il diabete di tipo 1 con soglia del 20%	57
4.36	tabella dei cluster ottenuti con k-means sul dataset riguardante il diabete di tipo 1 con soglia del 20%	57
4.37	tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il diabete di tipo 1 con soglia del 33%	58
4.38	tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il diabete di tipo 1 con soglia del 33%	58
4.39	risultati metriche su dataset riguardante il diabete di tipo 1	58
4.40	tabella risultati finali metriche dataset reali	59
4.41	tabella riassuntiva correlezione tra Silhouette e Adjusted Rand Index	59
5.1	tabelle con il numero di articoli pubblicati su google scholar per ogni metrica dall'anno 2000 ad oggi	60

Chapter 1

Introduzione

Il clustering è la tecnica di apprendimento automatico non supervisionato più utilizzata. Lo scopo degli algoritmi di clustering è quello di dividere i dati in gruppi (o cluster) in modo tale che gli oggetti presenti nello stesso cluster siano simili tra loro, mentre gli oggetti in cluster differenti siano diversi. Una delle sfide maggiori legate agli algoritmi di clustering è stabilire la qualità del cluster effettuato. Ci sono tre approcci possibili per calcolare la qualità di un clustering: la validazione esterna del clustering, la validazione interna e la validazione relativa. Nel mio progetto di stage mi sono occupato dello studio delle metriche di validazione interna sui cluster ottenuti dall'algoritmo k-means con k=2 (ovvero con 2 cluster). In particolare ho studiato le metriche di Silhouette coefficient [7], Entropia[1] , Davies-Bouldin index[6], Calinski-Harabasz index[5], Dunn index[4] e Gap statistic [9]. L'obiettivo del mio lavoro era di confrontare queste metriche, studiare il loro comportamento e valutare quali erano le più efficaci nel valutare la qualità dei cluster.

Il progetto, svolto con il linguaggio di programmazione Python, è articolato nei seguenti punti:

1. **Studio delle formule e del significato delle sei metriche:** Ho analizzato le formule matematiche che definiscono ciascuna metrica, approfondendo il loro significato e il modo in cui queste metriche valutano la qualità dei cluster.
2. **Ricerca nella letteratura scientifica:** Tramite Google Scholar [24], ho esaminato i pochi articoli scientifici che confrontano le varie metriche di validazione interna, ottenendo una panoramica delle opinioni e dei risultati precedentemente ottenuti nella comunità scientifica.
3. **Creazione di un dataset artificiale numerico su cui valutare le metriche:** Ho creato una matrice formata da una prima metà di soli zeri e da una seconda metà di soli uni. Questa matrice viene modificata una riga per volta sostituendo una riga di zeri o di uni selezionata casualmente, con una riga composta da numeri decimali compresi tra 0 e 1 e successivamente ho applicato alla matrice l'algoritmo k-means [2] con k=2. Per ognuna di queste iterazioni ho calcolato le metriche e per ognuna di esse, ho valutato se il suo andamento è consistente con il peggioramento della qualità del clustering.

4. **Creazione di dataset artificiali su cui valutare le metriche:** ho scritto il codice per generare un insieme di punti su un piano cartesiano a due dimensioni. Ho applicato k-means con $k=2$ a questi punti appena generati. Partendo da situazioni ideali, in cui k-means separa correttamente i punti nei due cluster, ho introdotto progressivamente una deviazione standard crescente per rendere più difficile il compito dell'algoritmo. Per ogni iterazione di k-means ho calcolato le metriche e per ognuna di esse, ho valutato se il suo andamento è stato consistente con il peggioramento della qualità del clustering.
5. **Applicazione delle metriche su dati reali:** ho preso in esame dataset medici reali e per ognuno di essi ho applicato un algoritmo che calcola la distanza euclidea che separa ciascun paziente da tutti gli altri pazienti presenti nel dataset. Dai risultati di questo algoritmo ho preso in considerazione solo i pazienti che hanno una distanza euclidea minore di una certa soglia. Su questi pazienti selezionati ho calcolato anche l'algoritmo k-means e ho confrontato i risultati ottenuti da k-means con quelli ottenuti dall'algoritmo della distanza euclidea attraverso Adjusted Rand Index [3]. Questo confronto permette di valutare quanto i pazienti selezionati siano classificati in maniera simile dai due diversi algoritmi. Ho ripetuto queste operazioni per una serie di soglie diverse. Inoltre per ogni iterazione di k-means ho calcolato le metriche e ho verificato se le metriche peggiorano al crescere della soglia.

I risultati ottenuti da questo studio hanno evidenziato come le metriche attualmente più diffuse e citate nella comunità scientifica sono anche le più efficaci nel valutare la qualità degli algoritmi di clustering.

1.1 Algoritmi di clustering

Nell'era dei big data trovare schemi e pattern all'interno di quantità enormi di dati è uno delle sfide maggiori nel campo dell'informatica. Il clustering, che fa parte della famiglia degli algoritmi di apprendimento automatico non supervisionato, sono un insieme di tecniche che cercano di risolvere questo problema. Il clustering è un processo fondamentale di esplorazione dei dati che mira a raggruppare insiemi di oggetti in modo tale che gli oggetti all'interno dello stesso gruppo (o cluster) siano più simili tra loro rispetto agli oggetti degli altri cluster. Il clustering trova applicazione in molteplici settori, tra cui la biologia per classificare le specie, il marketing per segmentare i clienti, la sicurezza informatica per individuare anomalie e molte altre aree in cui l'analisi di grandi volumi di dati è essenziale. La sua utilità risiede nella capacità di rivelare la struttura intrinseca dei dati, facilitandone la comprensione e l'interpretazione. Esistono diverse algoritmi di clustering tra cui i più noti sono il k-means, il DBSCAN [8] e Hierarchical Clustering [17]. Nel mio progetto di stage mi sono concentrato sullo studio dell'algoritmo k-means.

1.1.1 K-means

L'algoritmo K-means [2] è uno degli algoritmi più utilizzati per il clustering dei dati. L'obiettivo di questo algoritmo è determinare la partizione dei dati in k

cluster, con k specificato in input.

L'algoritmo k-means può essere riassunto in quattro passaggi:

1. **Inizializzazione dei centroidi:** Si selezionano casualmente k punti, detti seeds, che vengono usati come centroidi iniziali.
2. **Assegnazione dei punti ai cluster:** Per ciascun elemento del dataset viene calcolato la distanza con ogni centroide. L'elemento viene quindi assegnato al cluster più vicino
3. **Aggiornamento dei centroidi:** Una volta assegnati tutti i punti ai cluster, si calcola il nuovo centroide di ciascun cluster, prendendo la media delle coordinate dei punti assegnati al cluster.
4. **Ripetizione dei passaggi 2 e 3:** Se non è stato raggiunto nessun criterio di convergenza riparto dal punto 2

criteri di convergenza:

- non ci sono stati riassegnamenti di punti a nuovi cluster
- non c'è cambiamento, o è minimo, nelle posizioni dei centroidi
- la somma degli errori (SSE) quadratici è minima

$$SSE = \sum_{j=1}^k \sum_{x \in C_j} dist(x, m_j)^2$$

Con C_j j-esimo cluster, m_j centroide del cluster C_j e $dist(x, m_j)$ distanza tra il punto x e il centroide m_j

vantaggi k-means:

- Semplicità: l'algoritmo è facile da comprendere ed implementare
- Efficienza: l'algoritmo è molto efficiente. Ha una complessità di $O(tkn)$ con t numero di iterazioni, k numero di cluster e n numero di elementi del dataset

svantaggi k-means:

- Algoritmo sensibile ai seeds iniziali: La scelta dei centroidi iniziali può influenzare notevolmente i risultati finali.
- Richiede la specificazione del numero di cluster: È necessario specificare a priori il numero di cluster, il che può essere difficile quando non si conoscono le caratteristiche dei dati.
- Sensibile alla presenza di outliers: Gli outliers (valori notevolmente distanti dagli altri) possono influenzare negativamente i risultati di K-Means, spostando i centroidi e influenzando la forma e le dimensioni dei cluster.

1.2 Revisione letteratura

Il confronto tra metriche interne di validazione per il clustering è un problema di grande rilevanza nell'ambito dell'apprendimento automatico non supervisionato. Tuttavia, nonostante la sua importanza, questo argomento è stato relativamente poco esplorato nella letteratura scientifica.

Analizzando gli articoli della letteratura scientifica presenti su Google Scholar[24] con le seguenti query:

- "internal validation" AND "clustering" AND "comparison"
- best internal metric for clustering
- clustering validity indices
- comparison internal metrics for clustering
- cluster validity indices
- analysis internal validation metrics
- comparison of internal validation metrics

ho ottenuto pochi articoli inerenti le tematiche di mio interesse. Uno degli studi più significativi che ho trovato è l'articolo [13]. In questo articolo, gli autori confrontano trenta metriche interne di validazione testate su numerosi dataset con caratteristiche diverse. I risultati hanno evidenziato come non esistano metriche che si distinguono chiaramente come le migliori in assoluto. Tuttavia, è stato individuato un gruppo di 10 metriche che sembrano performare leggermente meglio delle altre. Tra queste Silhouette, Davies-Bouldin e Calinski-Harabasz sono le metriche che sono risultate essere le più efficaci. Più numerosi invece sono gli studi nella letteratura scientifica che confrontano le metriche interne di validazione con l'obiettivo di determinare il numero ottimale di cluster presenti nei dati come [12] o [10]. L'articolo [10] confronta sedici diverse metriche interne di validazione con l'obiettivo di trovare il numero ottimale di cluster al fine di migliorare l'accuratezza del clustering gerarchico. Lo studio ha indicato che l'indice Calinski-Harabasz ha ottenuto i risultati migliori tra i sedici indici testati. Tuttavia ci sono diversi altri studi che avevano lo stesso obiettivo che sono arrivati a risultati differenti.

Dalla revisione della letteratura scientifica emerge che non esistono metriche che performano nettamente meglio delle altre in tutte le situazioni. Gli studi condotti sono relativamente pochi e spesso giungono a conclusioni differenti, evidenziando la necessità di ulteriori ricerche per stabilire delle linee guida più significative.

Chapter 2

Metriche interne di validazione

La valutazione della qualità del clustering è un aspetto cruciale nei problemi di apprendimento automatico non supervisionato. Questa infatti permette di verificare l'affidabilità dei risultati garantendo che siano statisticamente significativi e non dovuti al caso. Esistono diversi approcci per valutare la qualità di un clustering.

- Metriche interne di validazione: La validazione interna si basa esclusivamente sui dati e sulla struttura del clustering ottenuto, senza fare riferimento a informazioni esterne o etichette predefinite. Le metriche interne valutano principalmente la coesione e la separazione dei cluster
- Metriche esterne di validazione: La validazione esterna confronta i risultati del clustering con una partizione di riferimento predefinita (ground truth), utilizzando etichette esterne conosciute. Questo approccio è utile quando si dispone di dati etichettati e si desidera valutare quanto i cluster ottenuti corrispondano alle etichette reali.
- Metriche relative di validazione: Viene eseguito il clustering più volte in modo da individuare la configurazione che massimizza la qualità del clustering

Nel mio progetto di stage mi sono soffermato sullo studio delle più note metriche interne di validazione

2.1 Silhouette coefficient

Silhouette coefficient[7] o Silhouette index è la metrica interna di validazione più nota e diffusa. Il valore di Silhouette coefficient è la misura di quanto un oggetto è simile agli altri elementi dello stesso cluster rispetto a quelli di un altro cluster. Il valore di Silhouette coefficient varia da -1 a 1 , dove 1 rappresenta il caso ottimale, mentre -1 il caso peggiore. La formula per calcolare il coefficiente di silhouette $s(i)$ di un singolo punto i è data da:

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

dove:

- $a(i)$ è la distanza media tra il punto i e tutti gli altri punti del suo stesso cluster.
- $b(i)$ è la distanza media tra il punto i e tutti i punti del cluster più vicino, ovvero il cluster con la distanza media minima.

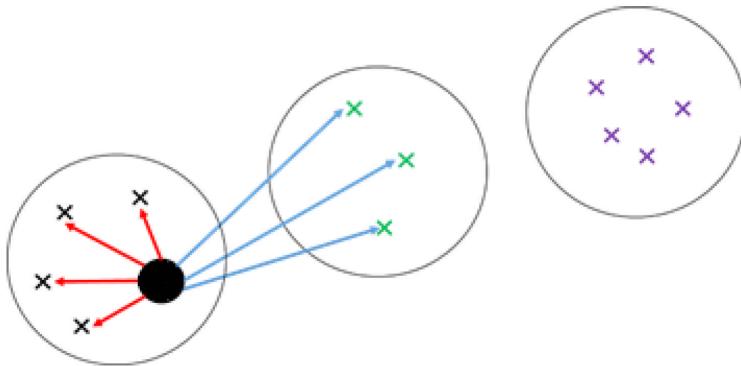


Figure 2.1: Spiegazione Silhouette

Nell'immagine le linee rosse rappresentano la distanza fra il punto i e tutti gli altri punti del suo stesso cluster. Le linee blu rappresentano la distanza tra i e tutti i punti del cluster più vicino al cluster di i

Fonte: [22]

2.2 Entropia

L'entropia di Shannon [1], derivata dalla teoria dell'informazione, è una misura che quantifica l'incertezza associata a una distribuzione di probabilità. Nella valutazione della qualità del clustering, l'entropia di Shannon viene utilizzata per misurare quanto bene i cluster separano i dati. Un valore basso di Entropia indica che i dati sono ben separati nei cluster, suggerendo che il clustering è di buona qualità. Un valore alto di Entropia invece indica che i dati sono distribuiti più uniformemente tra i cluster, suggerendo che il clustering potrebbe non essere efficace nel separare i dati in gruppi distinti. Per calcolare l'entropia di Shannon per un clustering, si procede come segue:

1. **Assegnazione dei Dati ai Cluster:** Supponiamo di avere k cluster e n dati. Ogni dato i è assegnato a un cluster C_j .
2. **Distribuzione di Probabilità:** Calcoliamo la proporzione di dati in ciascun cluster. La probabilità p_j che un dato appartenga al cluster C_j è

data da:

$$p_j = \frac{n_j}{n}$$

dove n_j è il numero di dati nel cluster C_j e n è il numero totale di dati.

3. **Formula dell'Entropia di Shannon:** L'entropia di Shannon H per il clustering è data da:

$$H = - \sum_{j=1}^k p_j \log(p_j)$$

2.3 Davies-Bouldin index

L'indice di Davies-Bouldin [6] rappresenta il rapporto tra la somma della dispersione all'interno del cluster e la separazione tra cluster. Rispetto al Silhouette coefficient il Davies-Bouldin index non ha un range limitato di valori, bensì può assumere qualsiasi valore maggiore o uguale di 0. Più il valore è vicino allo 0 più la qualità del clustering è elevata, di conseguenza più il valore è grande minore sarà la qualità del clustering. La formula del *Davies-Bouldin Index* (DBI) è la seguente:

$$DBI = \frac{1}{k} \sum_{i=1}^k R_i$$

dove:

$$R_i = \max_{j \neq i} R_{ij}$$

e

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}$$

con:

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - \mu_i\|$$

dove x è un punto nel cluster C_i , μ_i è il centroide del cluster C_i , e $|C_i|$ è il numero di punti nel cluster C_i . Quindi S_i rappresenta la dispersione del cluster C_i .

$$M_{ij} = \|\mu_i - \mu_j\|$$

dove μ_i e μ_j sono i centroidi dei cluster C_i e C_j . Quindi M_{ij} rappresenta la distanza tra i due centroidi

2.4 Calinski-Harabasz index

l'indice Calinski-Harabasz (CH) [5] è definito come il rapporto tra la separazione tra cluster (BCSS) e la dispersione all'interno del cluster (WCSS), normalizzato dal numero di gradi di libertà. L'indice di Calinski-Harabasz non ha un range di valori limitato, ma bensì può assumere qualsiasi valore maggiore o uguale di 0. Più il valore tende a $+\infty$ più la qualità del cluster sarà elevata, di conseguenza più il valore è piccolo minore sarà la qualità del clustering. La formula del *Calinski-Harabasz Index (CH)* è la seguente:

$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

dove:

- *BCSS(between-cluster dispersion matrix)* è la distanza tra il centroide di ogni cluster c_i e il centroide generale c .

$$BCSS = \sum_{i=1}^k n_i \cdot \|c_i - c\|^2$$

- *WCSS(within-cluster dispersion matrix)* è la distanza euclidea tra i punti x e il centroide del loro cluster c_i .

$$WCSS = \sum_{i=1}^k \sum_{x \in C_i} \|x - c_i\|^2$$

- k è il numero di cluster.
- n è il numero totale di punti.

2.5 Dunn index

Dunn index (DI) [4] è un indice che misura il grado di compattezza dei cluster e il grado di separazione tra cluster. Dunn index non ha un range di valori limitato, ma bensì può assumere qualsiasi valore maggiore o uguale di 0. Più il valore tende a $+\infty$ più la qualità del cluster sarà elevata, di conseguenza più il valore è piccolo minore sarà la qualità del clustering. La formula del *Dunn Index* è la seguente:

$$DI = \frac{\min_{1 \leq i < j \leq k} \delta(C_i, C_j)}{\max_{1 \leq i \leq k} \Delta(C_i)}$$

dove:

- $\delta(C_i, C_j)$ è la distanza tra i cluster C_i e C_j . Questa distanza può essere definita in diversi modi, ad esempio la distanza minima tra due punti appartenenti a cluster diversi.
- $\Delta(C_i)$ è la distanza intra-cluster per il cluster C_i , ovvero la massima distanza tra due punti all'interno del cluster C_i .
- k è il numero di cluster.

2.6 Gap statistic

La Gap Statistic [9] è una tecnica comunemente usata per determinare il numero ottimale di cluster in un dataset, ma può anche essere interpretata per valutare la qualità del clustering. La Gap Statistic aiuta a capire quanto la dispersione intra-cluster dei dati analizzati sia migliore rispetto a quella attesa da un clustering casuale. Un valore elevato della Gap Statistic suggerisce che il clustering è di buona qualità, mentre un valore basso suggerisce un clustering di pessima qualità. Per calcolare la Gap Statistic (Gap), si procede come segue:

1. Si calcola la dispersione intra-cluster W_k per il clustering ottenuto sui dati reali. La dispersione intra-cluster per k cluster è definita come:

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} \sum_{i,i' \in C_r} d_{ii'}$$

dove:

- C_r rappresenta il cluster r
- n_r è il numero di punti nel cluster r
- $d_{ii'}$ è la distanza tra gli oggetti i e i' .

2. Si genera un numero B di set di dati di riferimento casuali con la stessa distribuzione uniforme e dimensione del dataset originale. Per ciascuno di questi set, si applica lo stesso algoritmo di clustering e si calcola la dispersione intra-cluster $W_k^{(b)}$.
3. Si calcola la Gap Statistic come la differenza tra il logaritmo della dispersione intra-cluster media dei dataset di riferimento e il logaritmo della dispersione intra-cluster del dataset reale:

$$\text{Gap}_n(k) = \frac{1}{B} \sum_{b=1}^B \log(W_k^{(b)}) - \log(W_k)$$

dove $\frac{1}{B} \sum_{b=1}^B \log(W_k^{(b)})$ è la media dei logaritmi delle dispersioni intra-cluster dei dataset di riferimento.

2.7 Riassunto metriche

Table 2.1: riassunto metriche

Metrica	Intervallo	significato
Silhouette [7]	$[-1, +1]$	-1 valore peggiore; +1 valore migliore
Entropia [1]	$[0, \log_2 k]$	0 valore migliore; $\log_2 k$ valore peggiore
Davies-Bouldin [6]	$[0, +\infty]$	0 valore migliore; $+\infty$ valore peggiore
Dunn index [4]	$[0, +\infty]$	0 valore peggiore; $+\infty$ valore migliore
Calinski-Harabasz [5]	$[0, +\infty]$	0 valore peggiore; $+\infty$ valore migliore
Gap statistic [9]	$[-\infty, +\infty]$	$-\infty$ valore peggiore; $+\infty$ valore migliore

Per rendere più immediata ed intuitiva la comprensione dell'andamento delle metriche ho utilizzato il complementare dell'Entropia e il reciproco del Davies-Bouldin. In questo modo per tutte le metriche un valore alto corrispondeva ad un buon clustering, mentre un valore basso ad un pessimo clustering. Inoltre per quanto riguarda l'Entropia visto che ho lavorato solo con k-means con k=2 il limite superiore dei valori è di +1.

Chapter 3

Datasets

3.1 Dataset artificiale: matrice di zeri e uni

Come primo test per valutare le metriche ho creato una matrice formata da una prima metà di soli zeri e da una seconda metà di soli uni. Questa matrice viene modificata una riga per volta sostituendo una riga di zeri o di uni selezionata casualmente, con una riga composta da numeri decimali compresi tra 0 e 1 e successivamente ho applicato alla matrice l'algoritmo k-means con k=2. Per ognuna di queste iterazioni ho calcolato le metriche e per ognuna di esse, ho valutato se il suo andamento è consistente con il peggioramento della qualità del clustering.

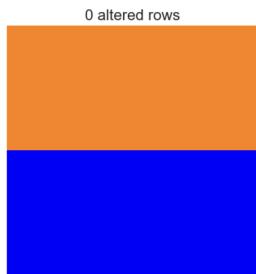


Figure 3.1: matrice di zeri e uni con 0 righe modificate



Figure 3.2: matrice di zeri e uni con 3 righe modificate

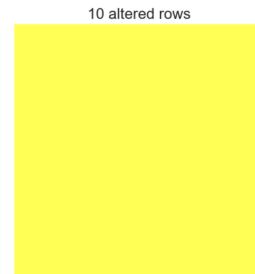
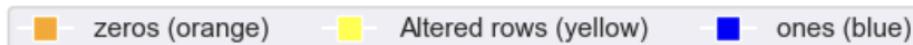


Figure 3.3: matrice di zeri e uni con 10 righe modificate



3.2 Dataset artificiali

Partendo dagli esempi presenti sulla sezione clustering del sito scikit-learn [28] ho generato 7 dataset artificiali contenenti un insieme di punti su un piano cartesiano a due dimensioni. Ho applicato k-means con k=2 a questi punti appena

generati. Partendo da situazioni ideali, in cui k-means separa correttamente i punti nei due cluster, ho introdotto progressivamente una deviazione standard crescente per rendere più difficile il compito dell'algoritmo. Per ogni iterazione di k-means ho calcolato le metriche e per ognuna di esse, ho valutato se il suo andamento è stato consistente con il peggioramento della qualità del clustering. Di seguito verranno mostrati degli estratti dei dataset scelti in modo tale da evidenziare l'aumento della deviazione standard e il conseguente peggioramento della qualità del clustering.

3.2.1 Dataset nuvole

Come primo dataset, ho scelto una rivisitazione a 2 cluster di uno degli esempi trovati su [28]. Inizialmente, i dati presentano due blob dalla forma allungata che il k-means riesce a distinguere facilmente. Successivamente, ho incrementato la deviazione standard fino a rendere i due blob indistinguibili, sia all'occhio umano che all'algoritmo k-means.

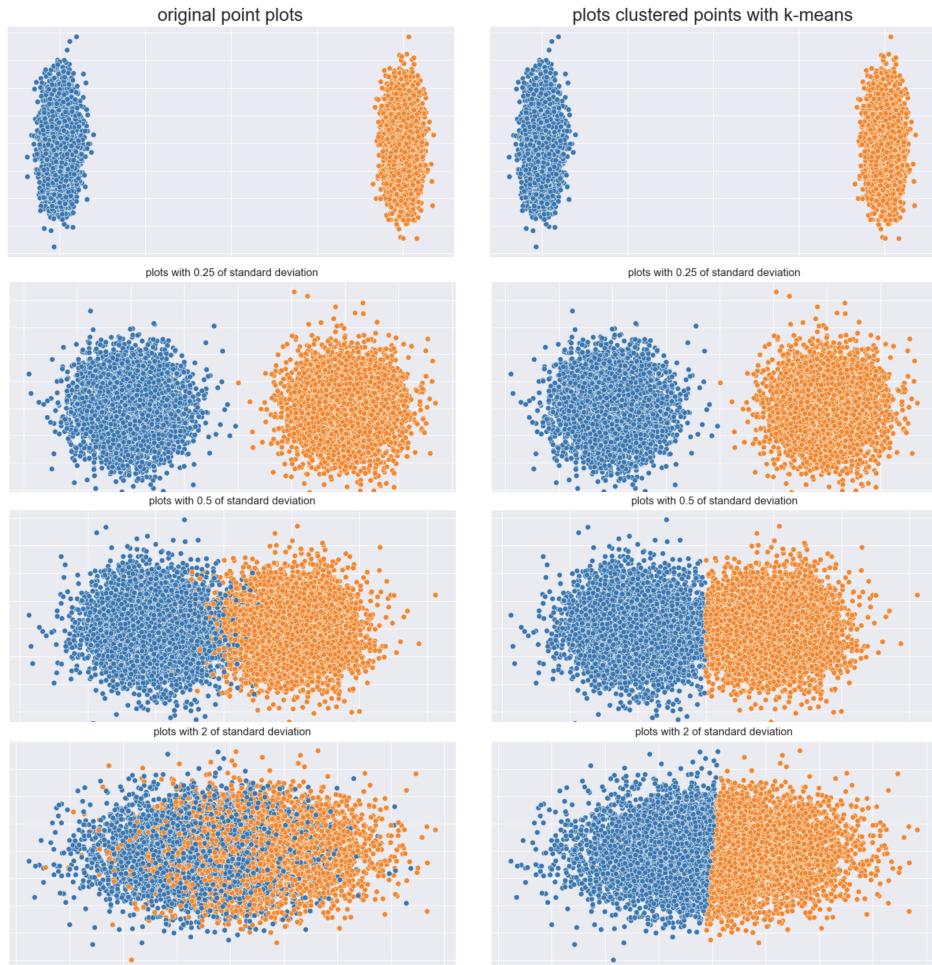


Figure 3.4: cluster dataset nuvole

3.2.2 Dataset cerchi

Come secondo dataset ho scelto il caso dei due cerchi concentrici preso da [28]. Inizialmente i dati presentano due cerchi uno dentro l'altro. Successivamente, con l'aumentare della deviazione standard, il cerchio all'interno si sposta verso destra fino ad uscire dal cerchio più grande.

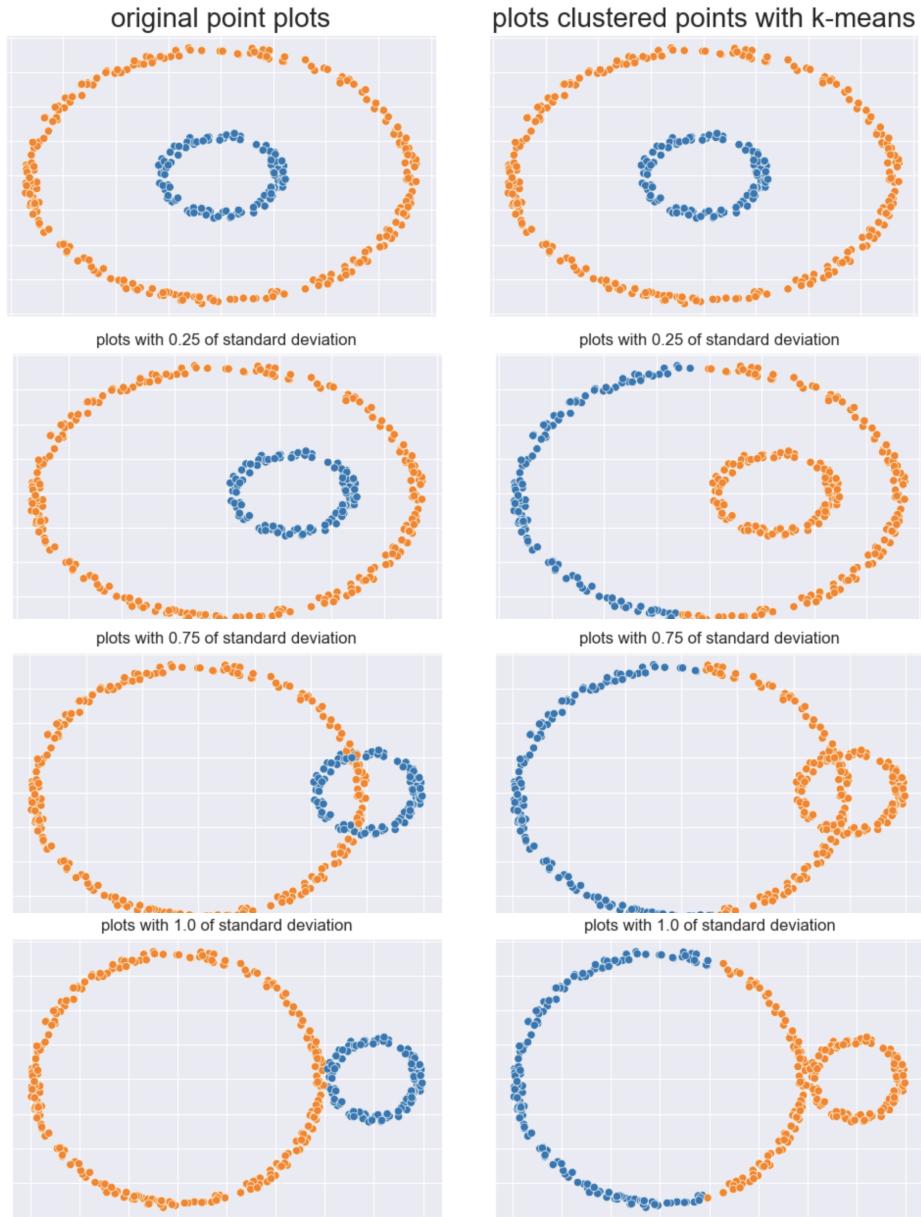


Figure 3.5: cluster dataset cerchi

3.2.3 Dataset semicerchi

Come terzo dataset ho scelto un caso in cui inizialmente ci sono due semicerchi che si intersecano. Successivamente all'aumentare della deviazione standard i due semicerchi si allontanano progressivamente in modo da essere sempre più facilmente distinguibili dall'algoritmo.

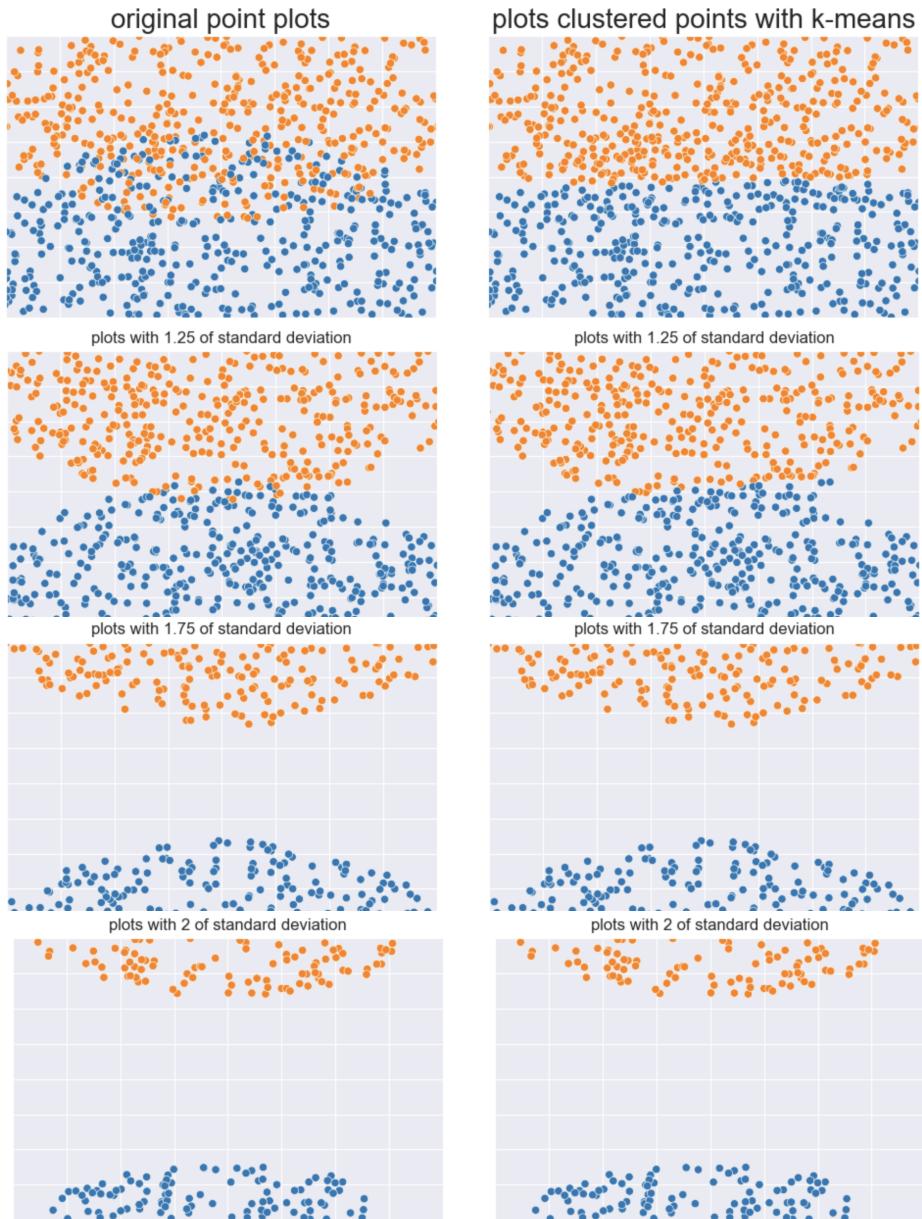


Figure 3.6: cluster dataset semicerchi

3.2.4 Dataset parabole

Come quarto dataset ho scelto il caso delle due parabole preso da [28]. Inizialmente le due parabole sono facilmente distinguibili all'occhio umano, anche se k-means fatica già a separarle. Successivamente all'aumentare della deviazione standard le due parabole progressivamente si uniscono fino a fondersi in un unico blob.

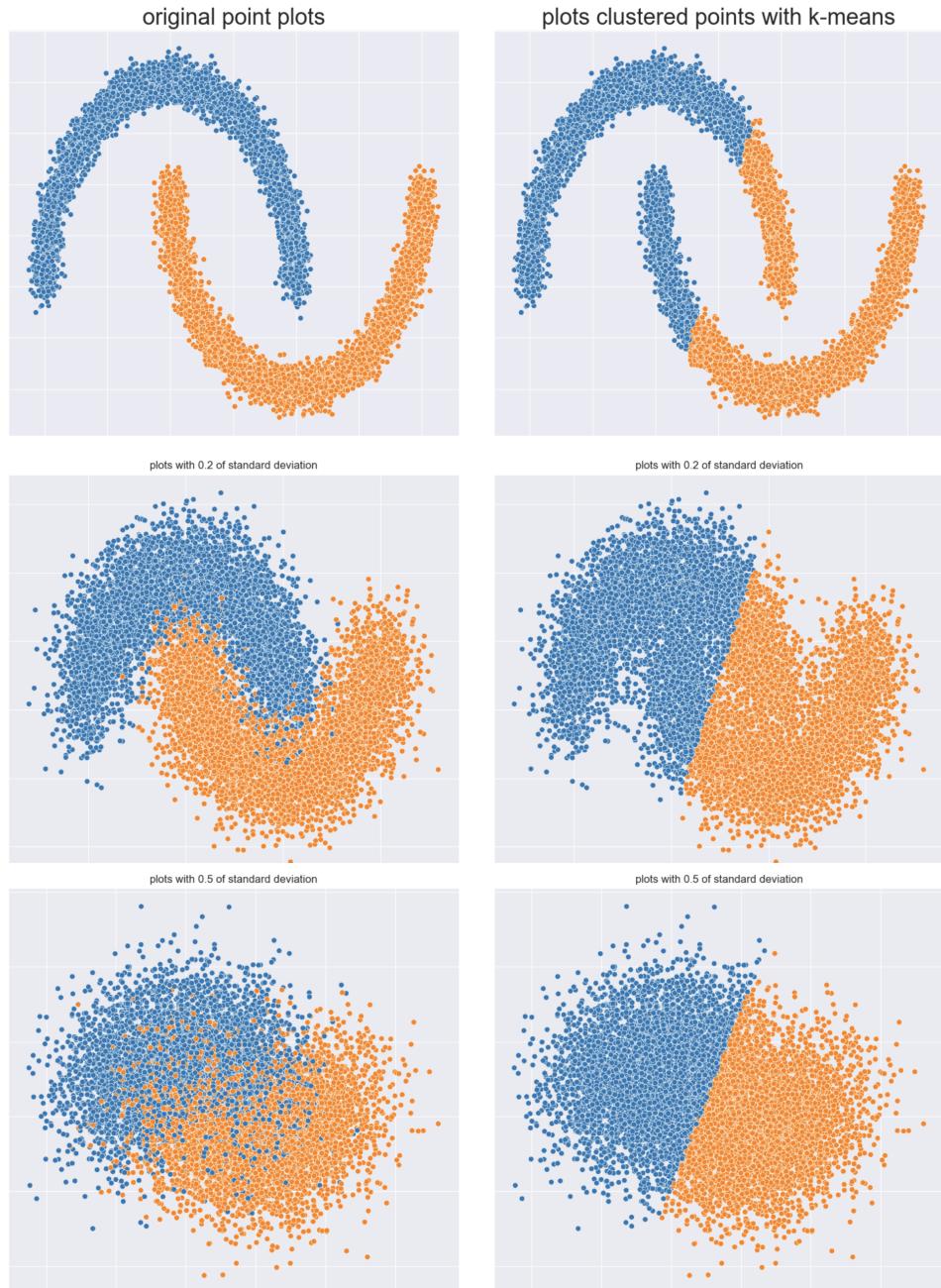


Figure 3.7: cluster dataset parabole

3.2.5 Dataset sfere

Come quinto dataset ho creato due sfere inizialmente molto simili. Successivamente ho applicato la deviazione standard solo ad una sfera facendola espandere progressivamente fino a inglobare anche l'altra sfera.



Figure 3.8: cluster dataset sfere

3.2.6 Dataset pennellate

Come sesto dataset ho scelto una rivisitazione a 2 cluster di uno degli esempi trovati su [28]. Inizialmente i dati presentano due blob dalla forma allungata che il k-means è in grado di distinguere facilmente. Successivamente, ho incrementato la deviazione standard andando progressivamente ad avvicinare i due blob fino a farli unire del tutto.

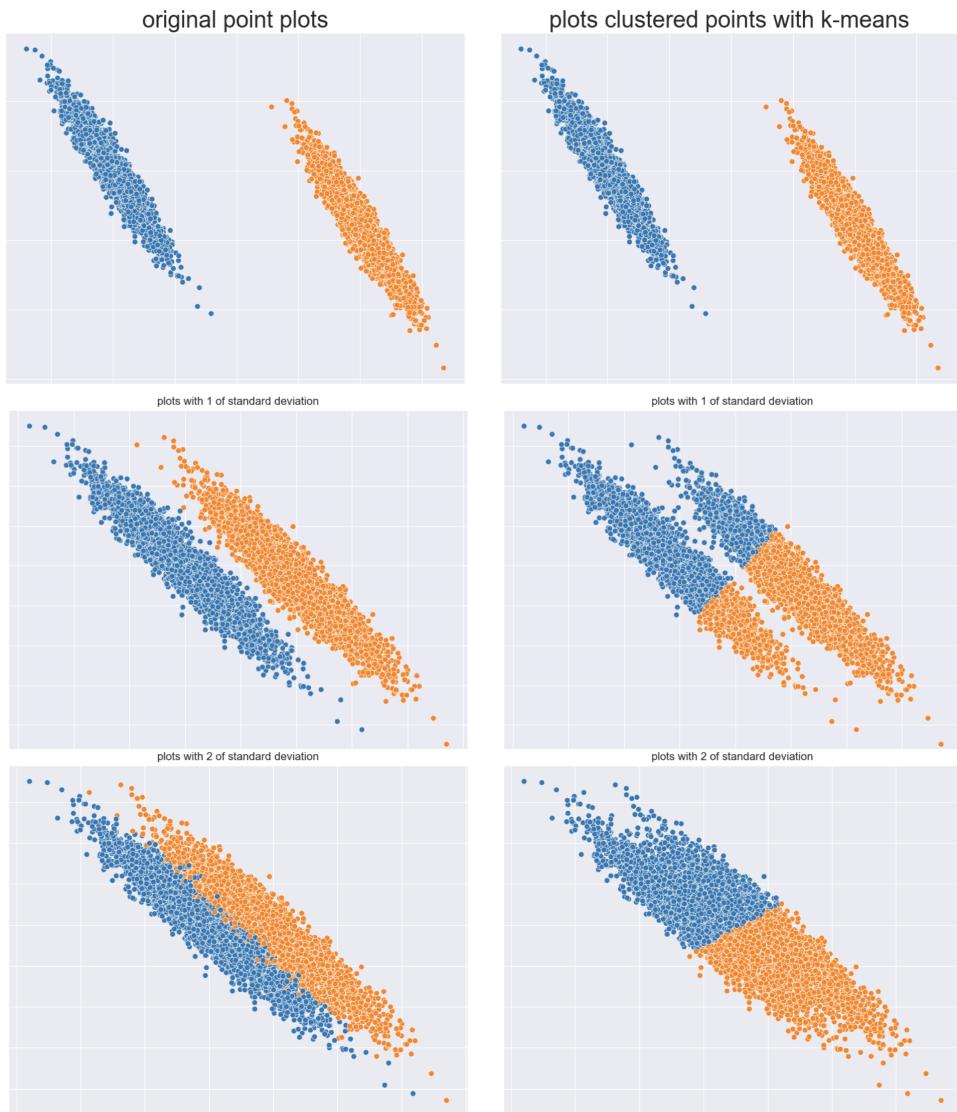


Figure 3.9: cluster dataset pennellate

3.2.7 Dataset W

Come settimo e ultimo dataset ho preso spunto da uno degli esempi del sito [23]. Inizialmente, i dati presentano due blob distinti: uno con una forma che ricorda una "W" e l'altro un blob sottile orizzontale sopra di esso. Successivamente, aumentando la deviazione standard, i due blob si disperdonano progressivamente fino a mescolarsi completamente, diventando indistinguibili.

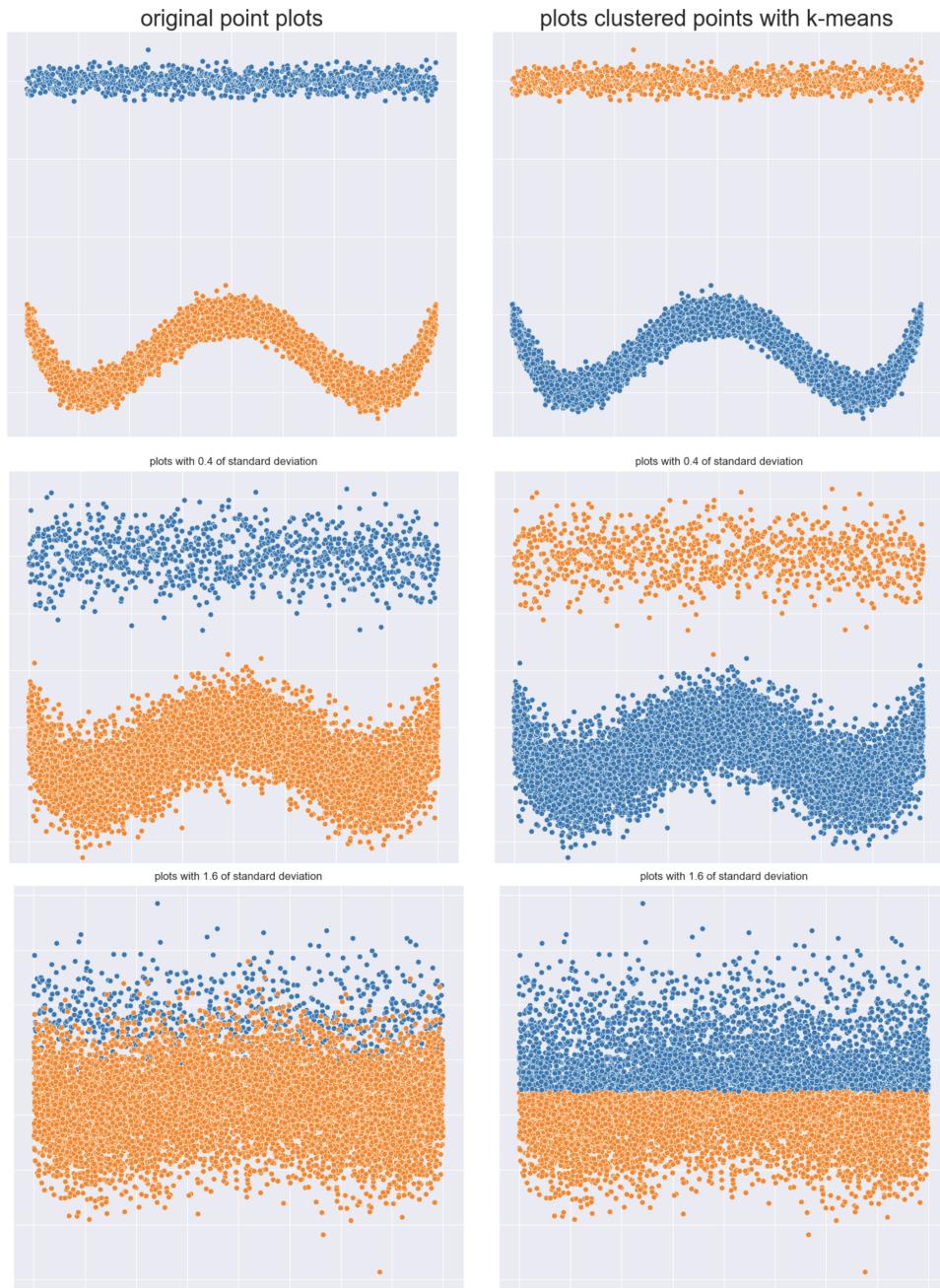


Figure 3.10: cluster dataset W

3.3 Dataset medici reali di cartelle cliniche elettroniche

Come ultimo test per valutare le metriche ho preso in esame 5 diversi dataset medici reali riguardanti la sepsi & SIRS [15], la correlazione tra insufficienza cardiaca e depressione [16], l'arresto cardiaco [18], il neuroblastoma[19] ed il diabete di tipo 1[20]. Per valutare il corretto funzionamento dei test ho prima sviluppato un piccolo dataset artificiale 3.1 che conteneva dei dati medici finti con valori molto estremi in modo che la classificazione in due cluster fosse molto facile. Su questi dati ho come prima cosa calcolato la distanza euclidea che separava ogni riga del dataset da tutte le altre, ovvero la "differenza" tra i dati di un paziente da quelli di tutti gli altri. Dai risultati di questo algoritmo ho preso in considerazione solo i pazienti che hanno una distanza euclidea maggiore di una certa soglia. Su questi pazienti selezionati ho calcolato l'algoritmo k-means e ho confrontato, tramite l'Adjusted Rand Index [3], la classificazione ottenuta con quella precedentemente ottenuta con la distanza euclidea. Ho ripetuto queste operazioni prima con una soglia del 20% e poi con una soglia del 33%. Per ogni iterazione di k-means ho anche calcolato le metriche per valutare se peggiorassero al crescere della soglia. Infatti è logico aspettarsi che con una soglia maggiore vengano classificati nello stesso cluster pazienti con una distanza euclidea maggiore tra di loro portando ad un peggioramento della qualità del clustering. Oltre a confrontare l'andamento delle metriche in base alla soglia ho anche confrontato il valore di Silhouette con il valore dell'Adjusted Rand Index per verificare se fossero coerenti. Ho svolto questo confronto solo con la Silhouette in quanto è l'unica metrica con un range limitato che ha ottenuto buone performance fino ad ora.

3.3.1 Dataset medico artificiale

Questo dataset artificiale rappresenta delle informazioni fintizie su un gruppo di 10 pazienti. Il dataset è composto da quattro colonne: età, sesso, livello di istruzione e stadio del cancro.

Table 3.1: dataset medico artificiale

età	sesso_0femmina_1maschio	educazione	stadio del cancro
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

3.3.2 Dataset sepsi & SIRS

Il dataset riguarda le cartelle cliniche di 1257 pazienti con SIRS (Sindrome da Risposta Infiammatoria Sistemica) e Sepsi, includendo variabili cliniche per analisi e confronto tra le due condizioni. La SIRS (Sindrome da Risposta Infiammatoria Sistemica) è una risposta infiammatoria generalizzata dell'organismo, che può essere causata da traumi, ustioni, pancreatite o altre condizioni. La Sepsi è una complicanza potenzialmente mortale di un'infezione, che provoca una risposta infiammatoria in tutto il corpo, simile alla SIRS, ma innescata da un'infezione. SIRS e Sepsi sono spesso messe a confronto perché entrambe le condizioni condividono una risposta infiammatoria sistemica con sintomi sovrapponibili, il che rende fondamentale distinguerele accuratamente per il trattamento appropriato.

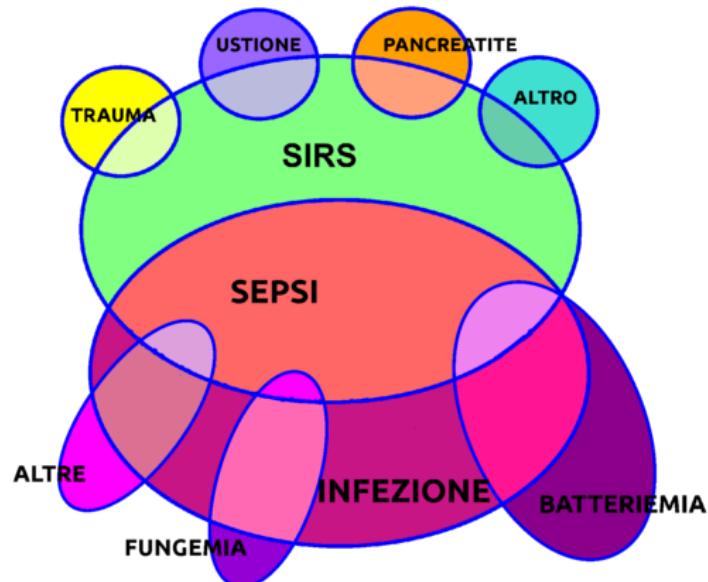


Figure 3.11: differenze cause tra Sepsi e SIRS

Fonte: [25]

3.3.3 Dataset depressione ed insufficienza cardiaca

Questo dataset riguarda le cartelle cliniche di 425 pazienti che con depressione e insufficienze cardiache. La depressione è un disturbo dell'umore caratterizzato da persistenti sentimenti di tristezza e perdita di interesse. È una condizione grave che può influenzare negativamente il funzionamento quotidiano di una persona. L'insufficienza cardiaca o scompenso cardiaco è una sindrome clinica complessa definita come l'incapacità del cuore di fornire il sangue in quantità adeguata rispetto all'effettiva richiesta dell'organismo o la capacità di soddisfare tale richiesta solamente a pressioni di riempimento ventricolari superiori alla norma.

3.3.4 Dataset arresto cardiaco

Questo dataset riguarda le cartelle cliniche di 422 pazienti soggetti ad arresto cardiaco extraospedaliero nella regione di Alicante in Spagna. L'arresto cardiaco è una condizione medica di emergenza in cui il cuore smette improvvisamente di battere in modo efficace, interrompendo il flusso di sangue al cervello e ad altri organi vitali. Senza intervento immediato, l'arresto cardiaco può portare rapidamente alla morte.

3.3.5 Dataset neuroblastoma

Il dataset riguarda le cartelle cliniche di 169 pazienti con neuroblastoma. Il neuroblastoma è un tumore maligno che si sviluppa dalle cellule nervose immature del sistema nervoso simpatico. È il tumore solido extracranico più comune nei bambini, rappresentando circa il 15% delle morti per cancro infantile.

3.3.6 Dataset diabete di tipo 1

Il diabete di tipo 1 è una malattia autoimmune cronica in cui il sistema immunitario attacca e distrugge le cellule beta del pancreas, che sono responsabili della produzione di insulina. L'insulina è un ormone essenziale per regolare i livelli di glucosio (zucchero) nel sangue. Senza insulina, il glucosio non può essere trasportato nelle cellule e rimane nel sangue, portando a livelli elevati di zucchero nel sangue.

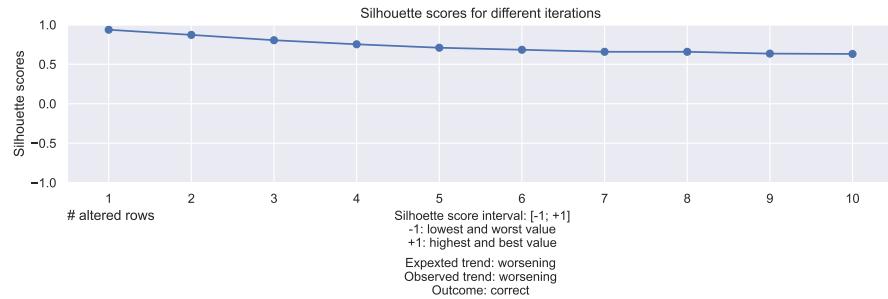
Chapter 4

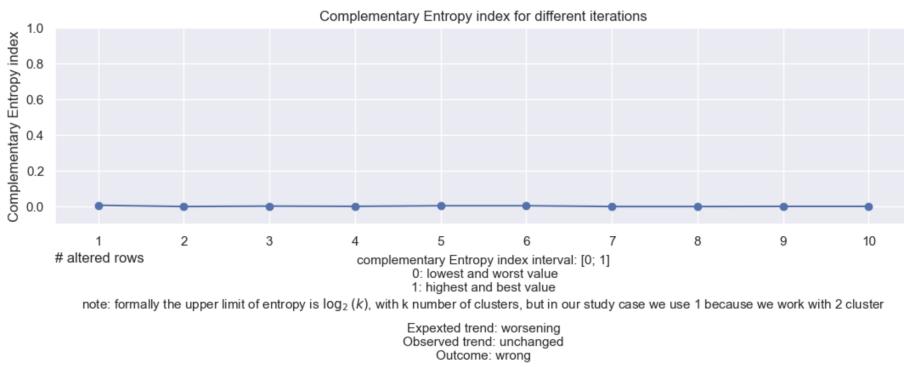
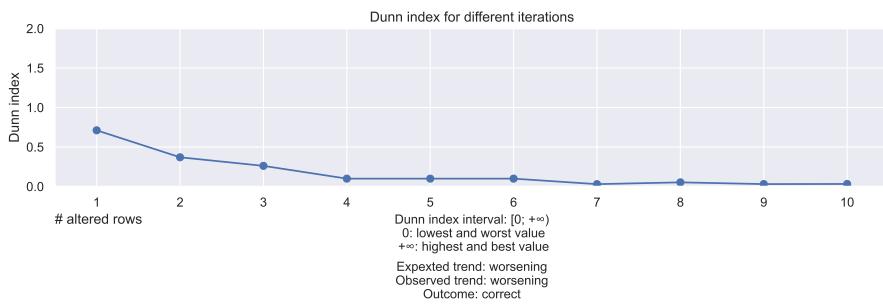
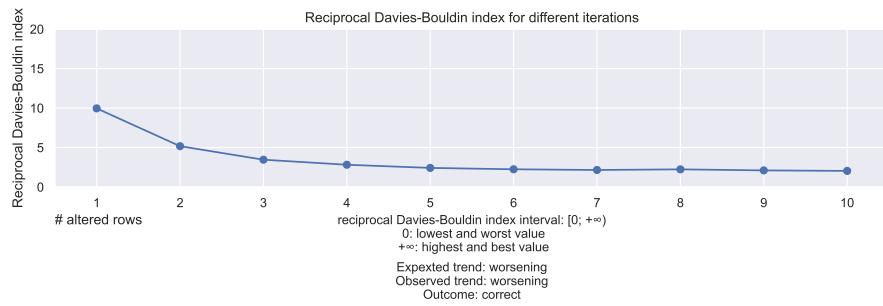
Risultati

4.1 Risultati dataset artificiale: 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 l'Entropia che rimane sempre costante e con la Gap statistic che ha un andamento irregolare. Di seguito sono presenti i grafici che mostrano l'andamento delle metriche all'aumentare delle righe modificate, e infine una tabella di recap che riassume il comportamento delle metriche.

Figure 4.1: risultati dataset artificiale numerico





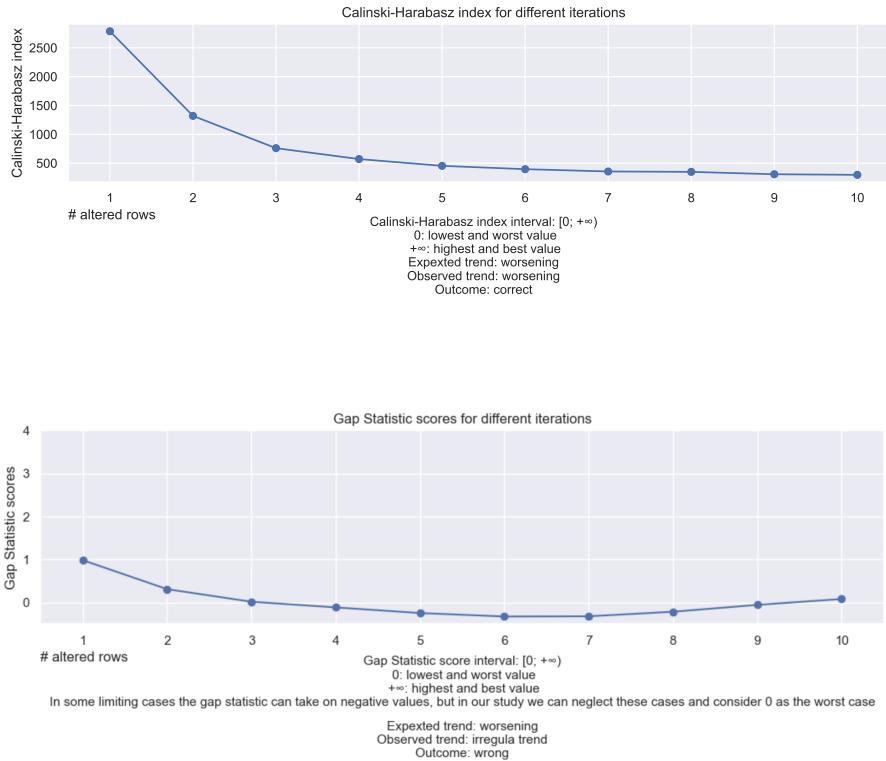


Table 4.1: risultati metriche dataset matrice di 0 e 1

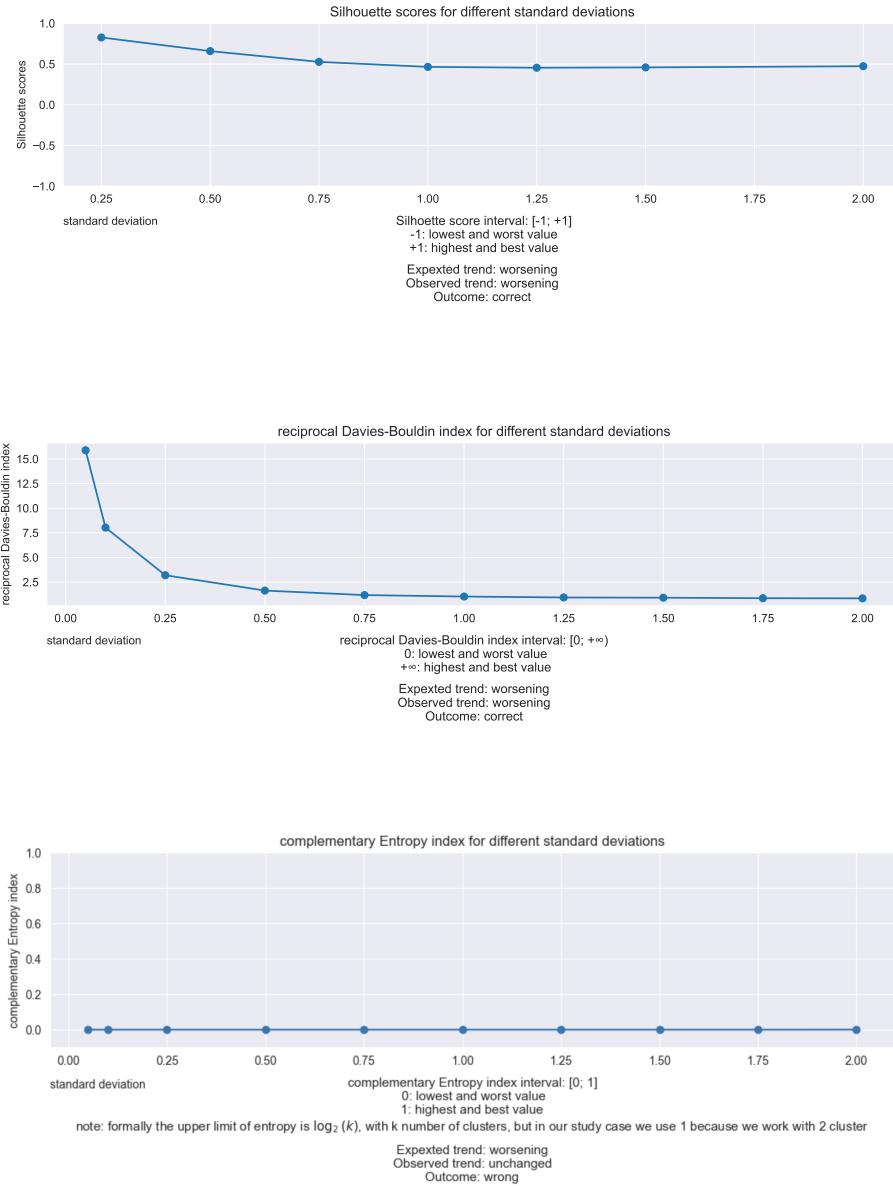
metriche	risultati	tempo
Silhouette	corretto	0.901 ms
Entropia complementare	errato	0.299 ms
Davies-Bouldin reciproco	corretto	0.435 ms
Dunn index	corretto	0.274 ms
Calinski-Harabasz	corretto	0.197 ms
Gap Statistic	errato	238.138 ms

4.2 Risultati dataset artificiali

Per ognuno dei 7 dataset verranno mostrati 6 grafici, ciascuno dei quali illustra l'andamento di una metrica all'aumentare della deviazione standard. Inoltre, sarà presentata una tabella finale riassuntiva che mostra quali metriche hanno funzionato correttamente e per ogni metrica il suo tempo medio di calcolo.

4.2.1 risultati dataset nuvole

Figure 4.2: risultati dataset nuvole



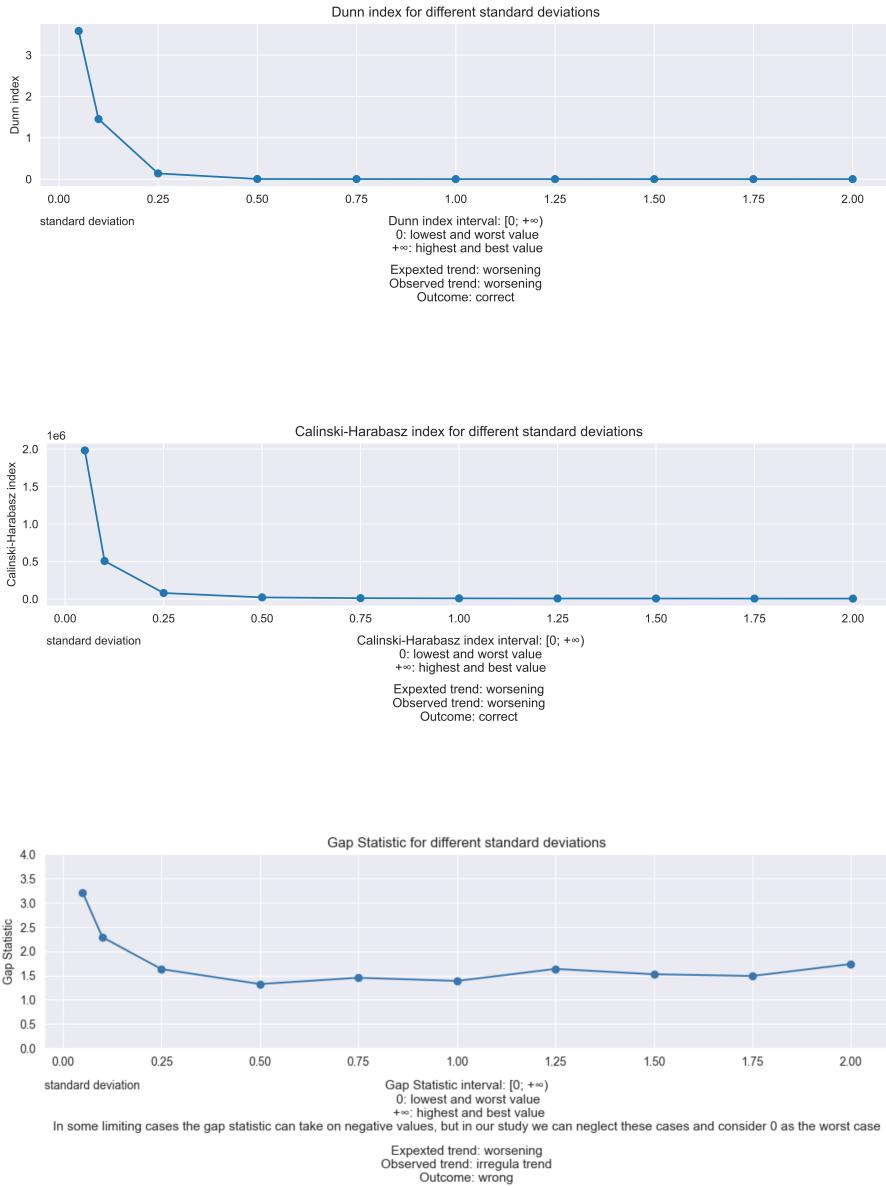


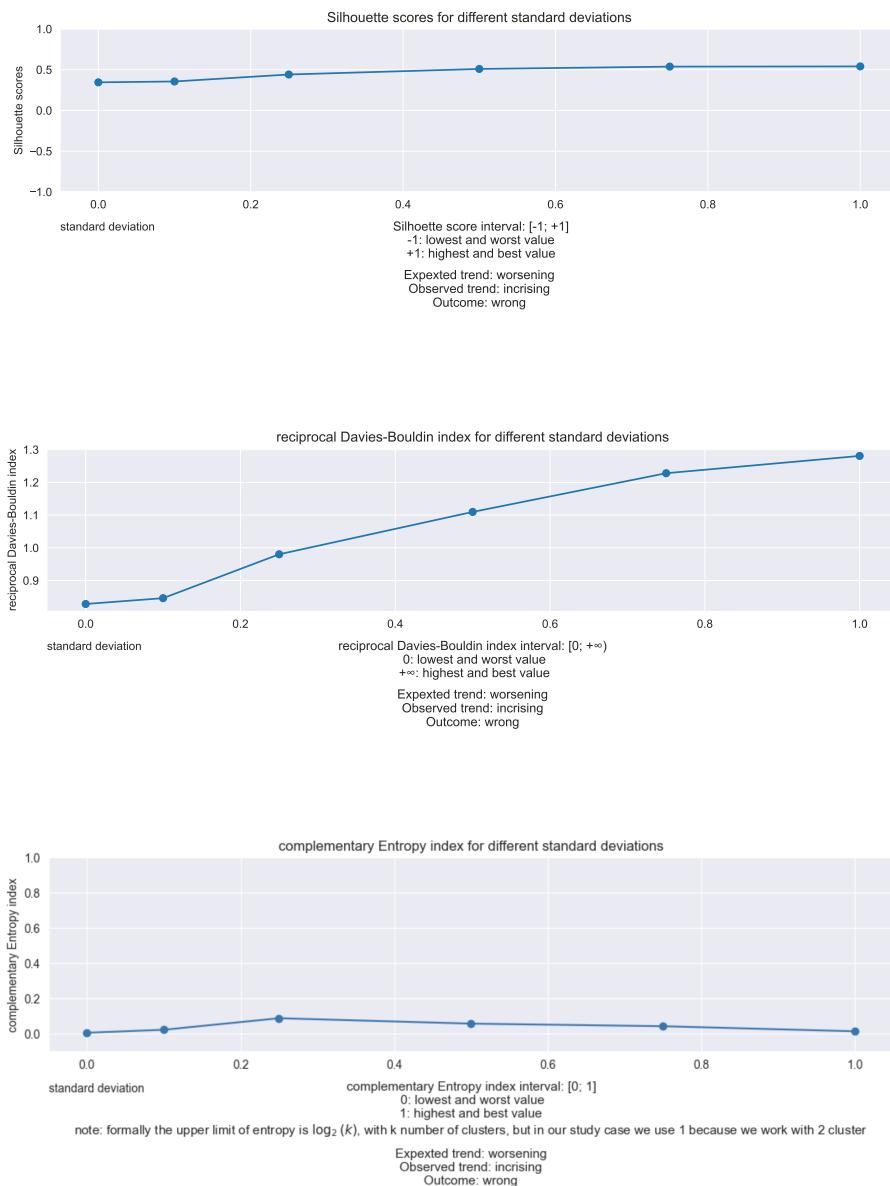
Table 4.2: risultati metriche dataset nuvole

metriche	risultati	tempo medio
Silhouette	corretto	1.02 sec
Entropia complementare	errato	0.0005 sec
Davies-Bouldin reciproco	corretto	0.001 sec
Dunn index	corretto	0.407 sec
Calinski-Harabasz	corretto	0.001 sec
Gap Statistic	corretto	0.353 sec

4.2.2 risultati dataset cerchi

Dato questo dataset mi aspetto un peggioramento delle metriche all'aumentare della deviazione standard. Tuttavia, i risultati mostrano un trend opposto: tutte le metriche migliorano con l'aumento della deviazione standard. Questo comportamento è imputabile all'algoritmo k-means, che non offre performance ottimali su questo tipo di dataset.

Figure 4.3: risultati dataset cerchi



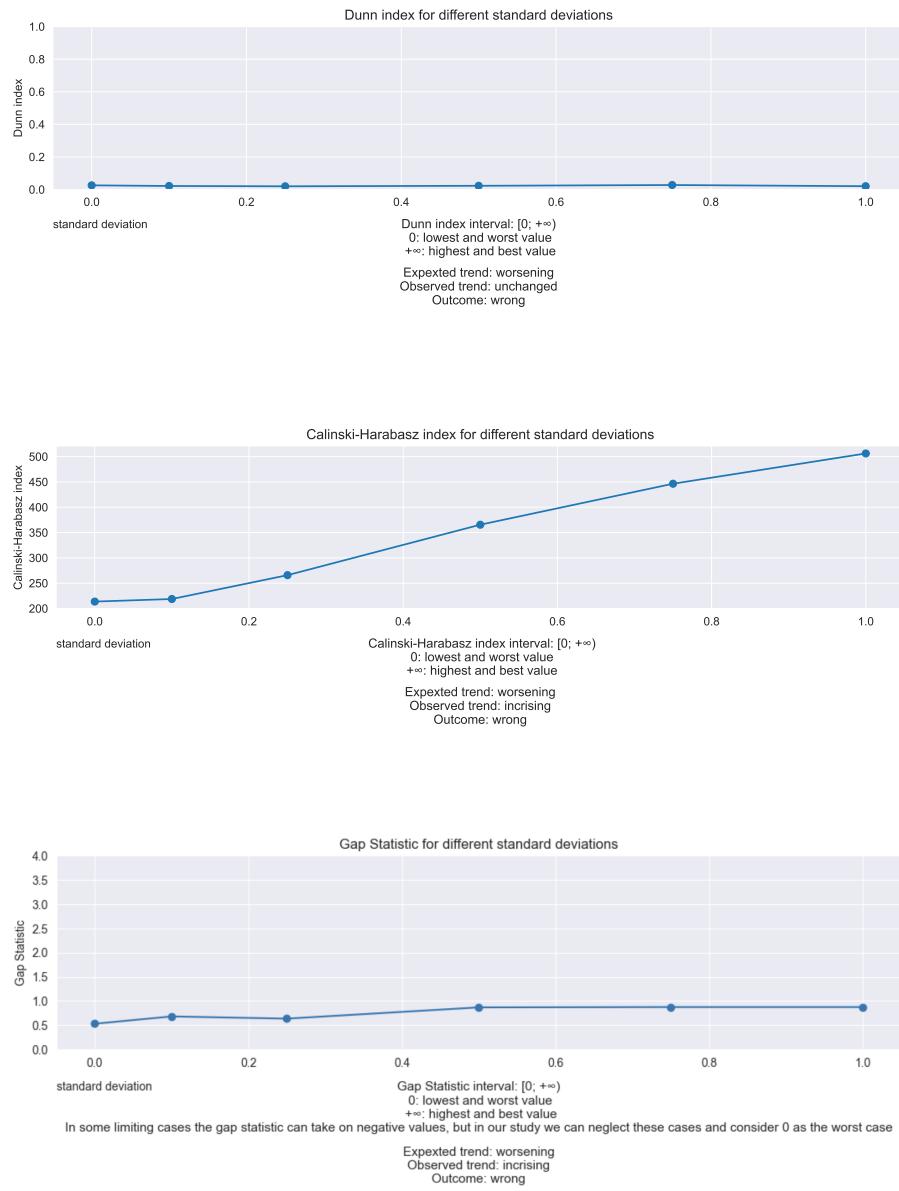


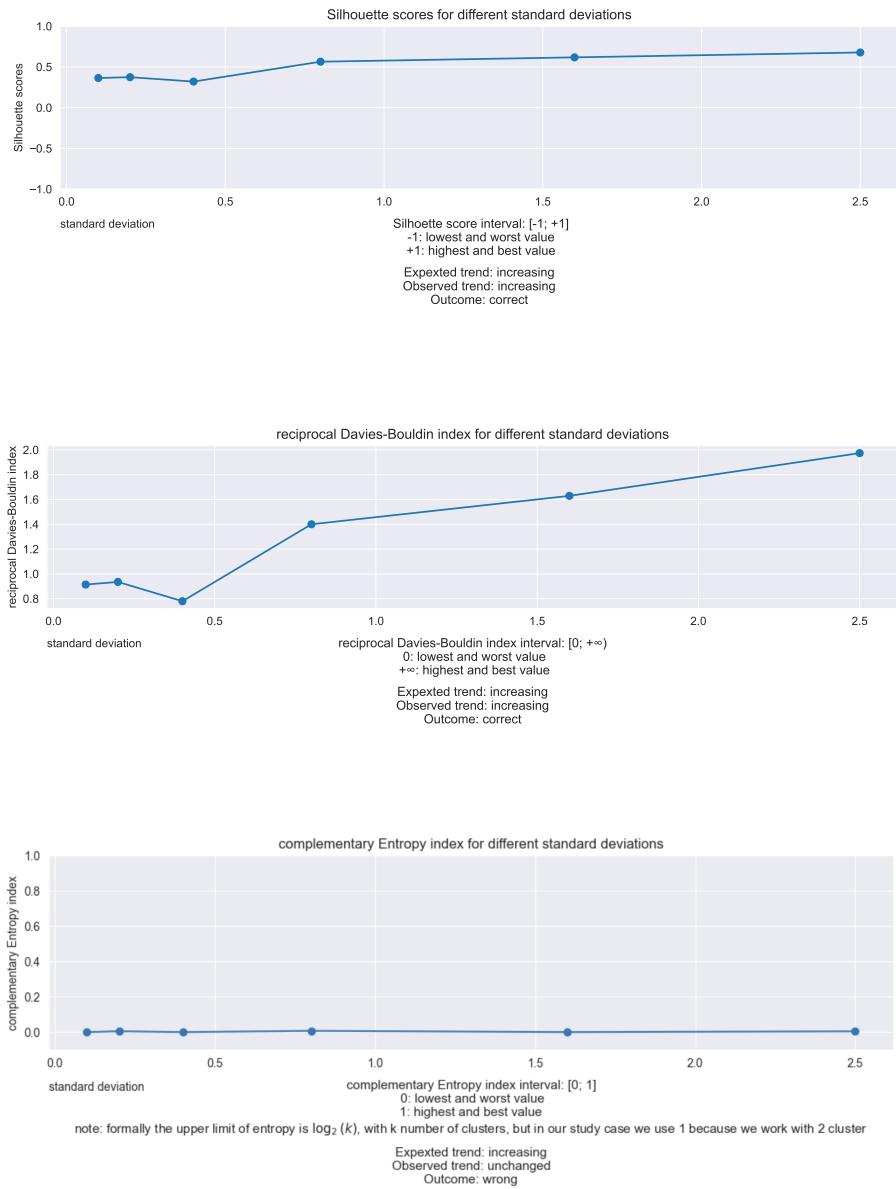
Table 4.3: risultati metriche dataset cerchi

metriche	risultati	tempo medio
Silhouette	errato	0.002 sec
Entropia complementare	errato	0.0002 sec
Davies-Bouldin reciproco	errato	0.0004 sec
Dunn index	errato	0.0007 sec
Calinski-Harabasz	errato	0.0002 sec
Gap Statistic	errato	0.039 sec

4.2.3 risultati dataset semicerchi

Dato questo dataset mi aspetto che le metriche migliorino all'aumentare della deviazione standard. Questo accade per tutte le metriche tranne che per l'Entropia, che rimane costante, e per Calinski-Harabasz che ha un andamento irregolare.

Figure 4.4: risultati dataset semicerchi



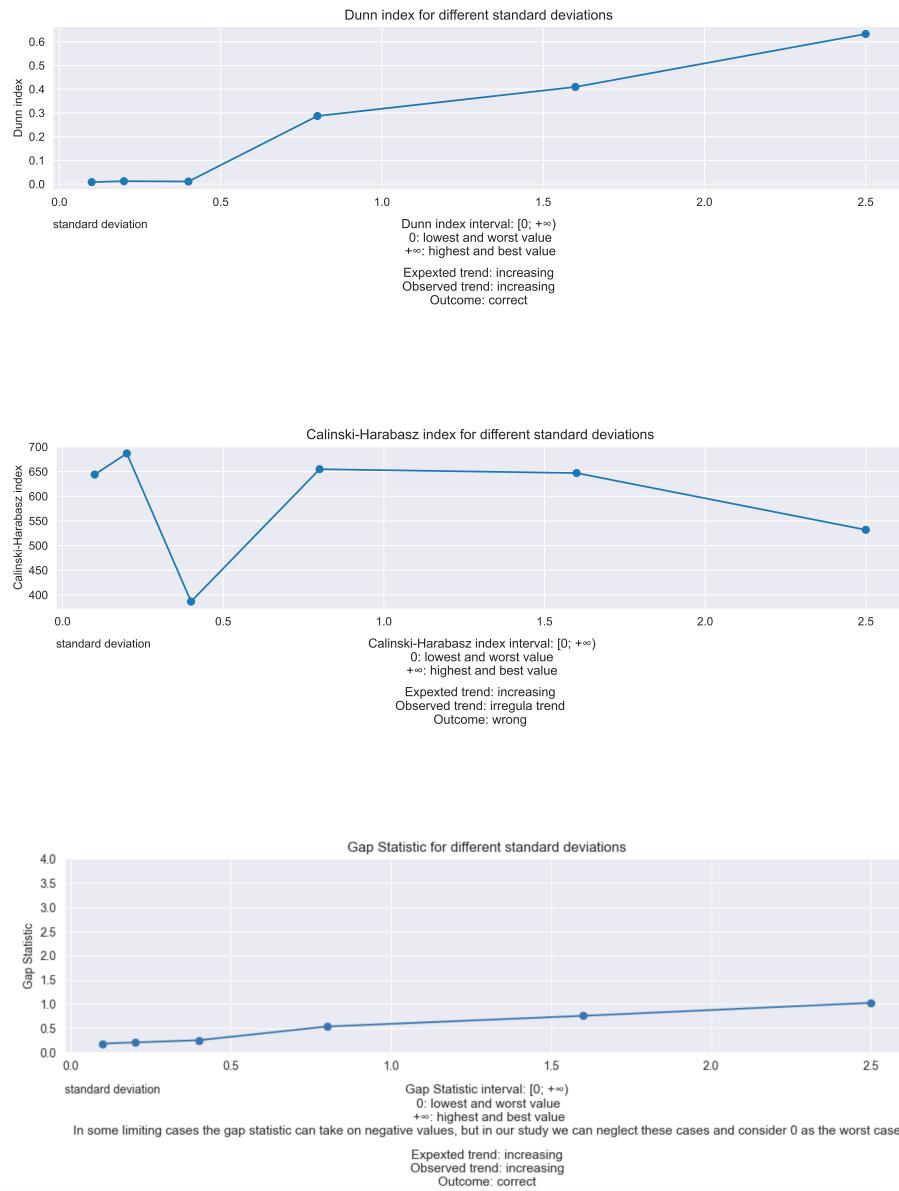


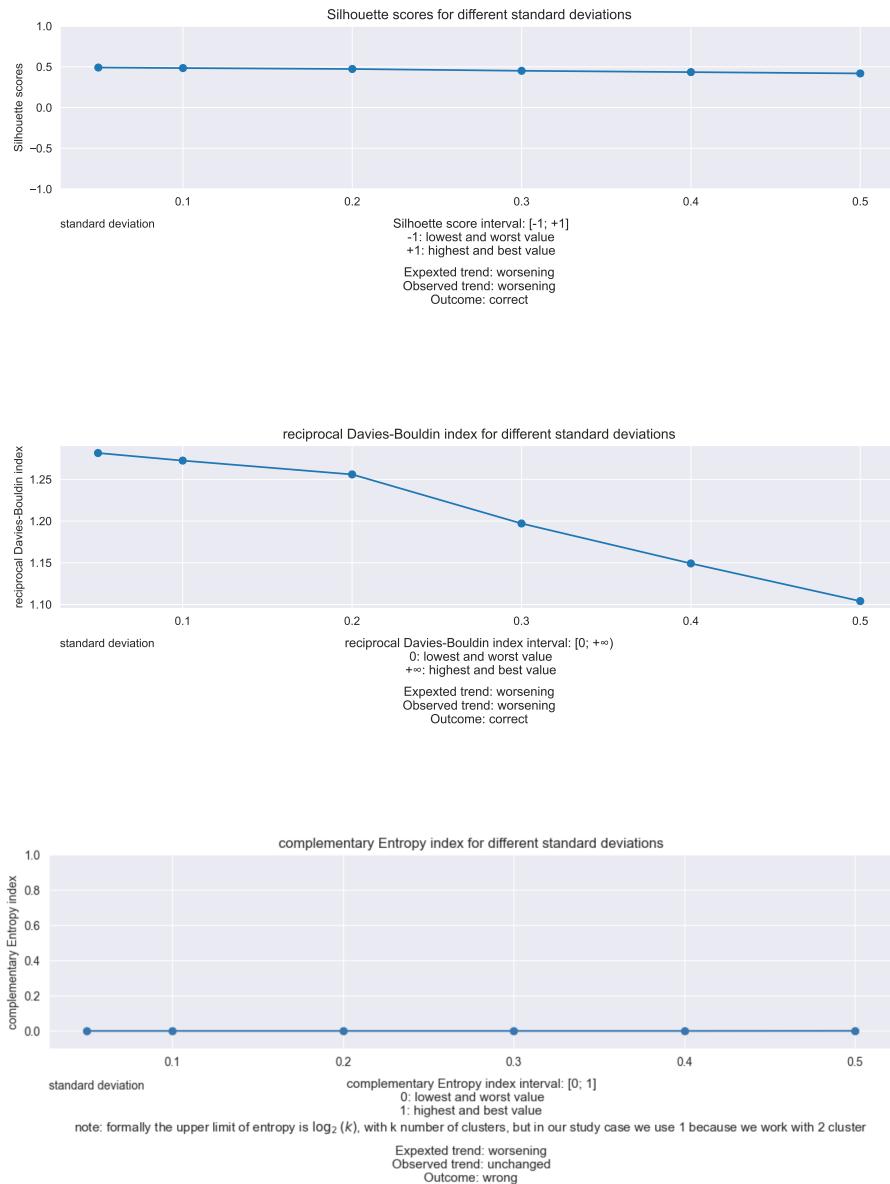
Table 4.4: risultati metriche dataset semicerchi

metriche	risultati	tempo
Silhouette	corretto	0.009 sec
Entropia complementare	errato	0.0003 sec
Davies-Bouldin reciproco	corretto	0.0006 sec
Dunn index	corretto	0.005 sec
Calinski-Harabasz	errato	0.0003 sec
Gap Statistic	corretto	0.073 sec

4.2.4 risultati dataset parbole

Dato questo dataset mi aspetto che le metriche peggiorino all'aumentare della deviazione standard. Questo accade per tutte le metriche tranne che per l'Entropia e il Dunn index, che rimangono costanti, e per la Gap Statistic che ha un andamento irregolare.

Figure 4.5: risultati dataset parbole



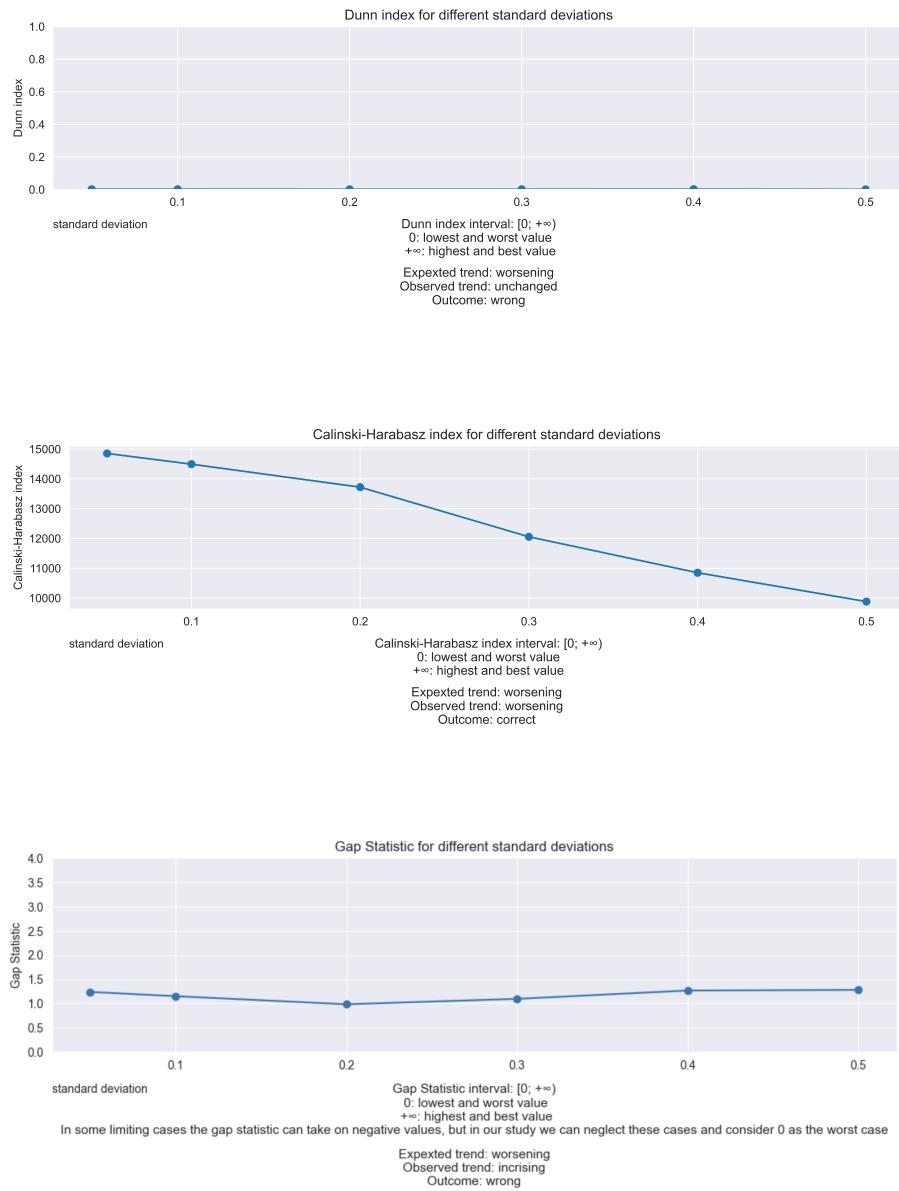


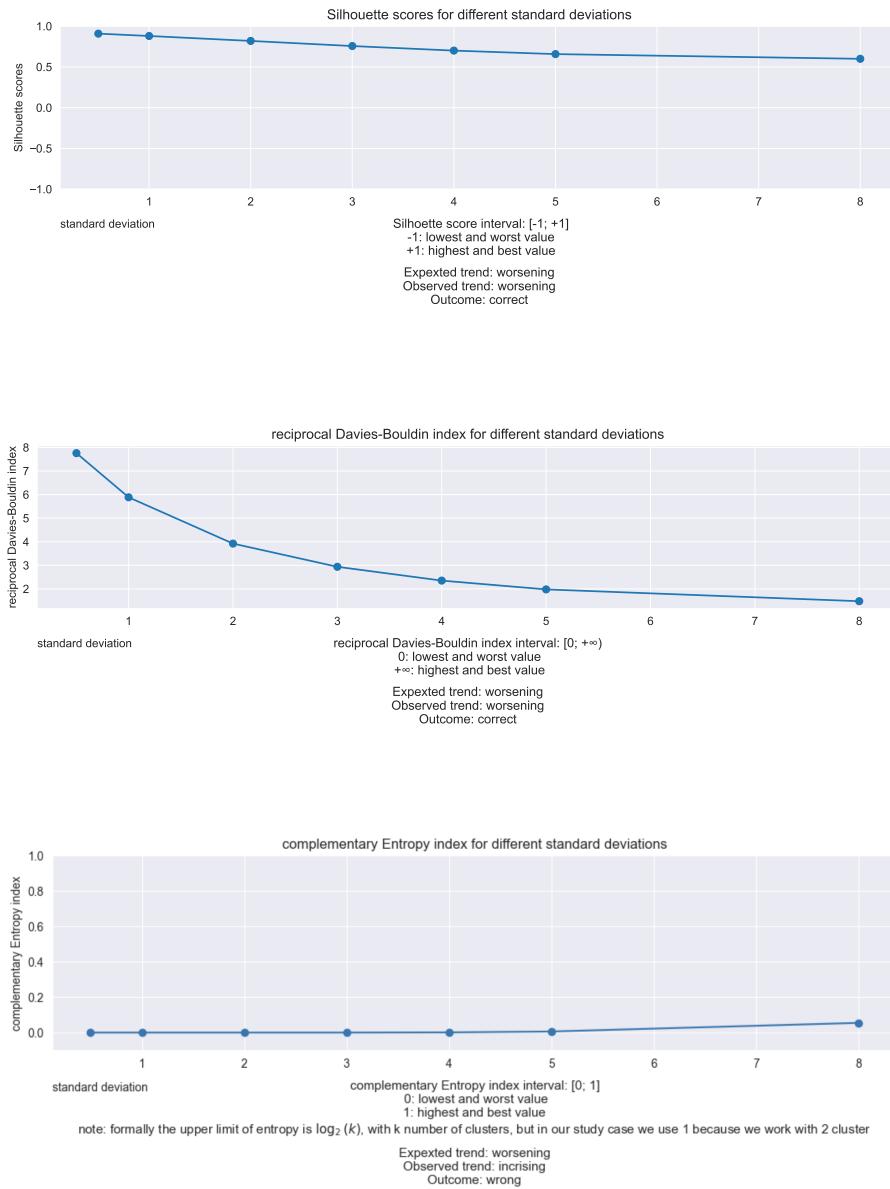
Table 4.5: risultati metriche dataset parbole

metriche	risultati	tempo
Silhouette	corretto	0.993 sec
Entropia complementare	errato	0.001 sec
Davies-Bouldin reciproco	corretto	0.002 sec
Dunn index	errato	0.42 sec
Calinski-Harabasz	corretto	0.42 sec
Gap Statistic	errato	0.255 sec

4.2.5 risultati dataset sfere

Dato questo dataset mi aspetto che le metriche peggiorino all'aumentare della deviazione standard. Questo accade per tutte le metriche tranne che per l'Entropia, che rimane costante, e per Gap Statistic che ha un andamento irregolare.

Figure 4.6: risultati dataset sfere



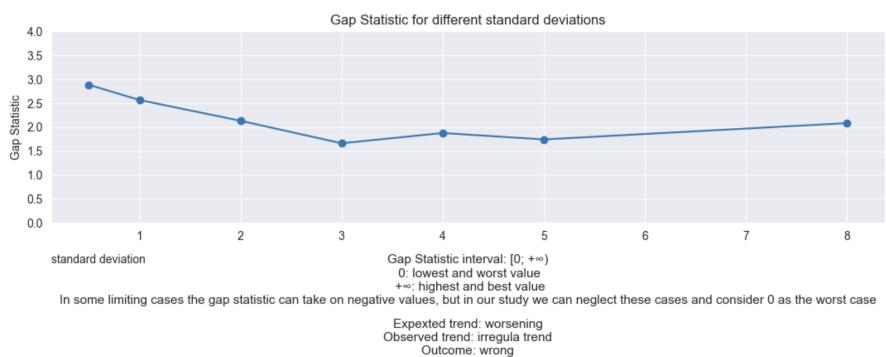
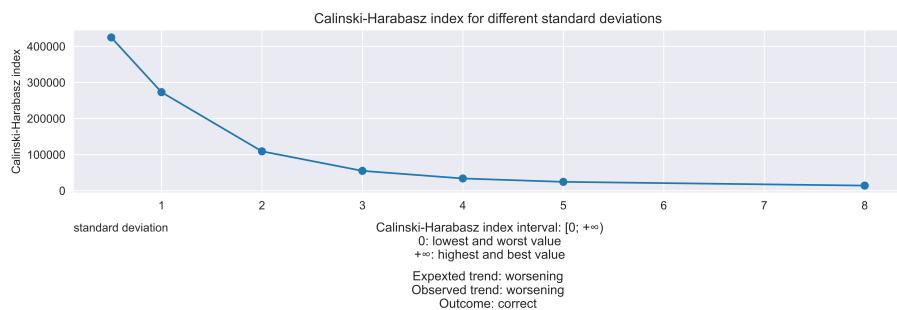
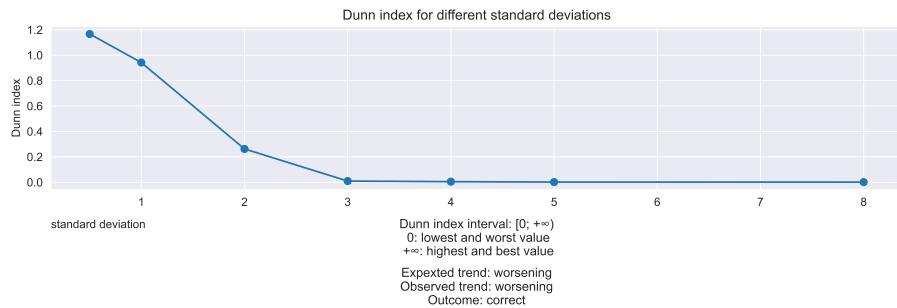


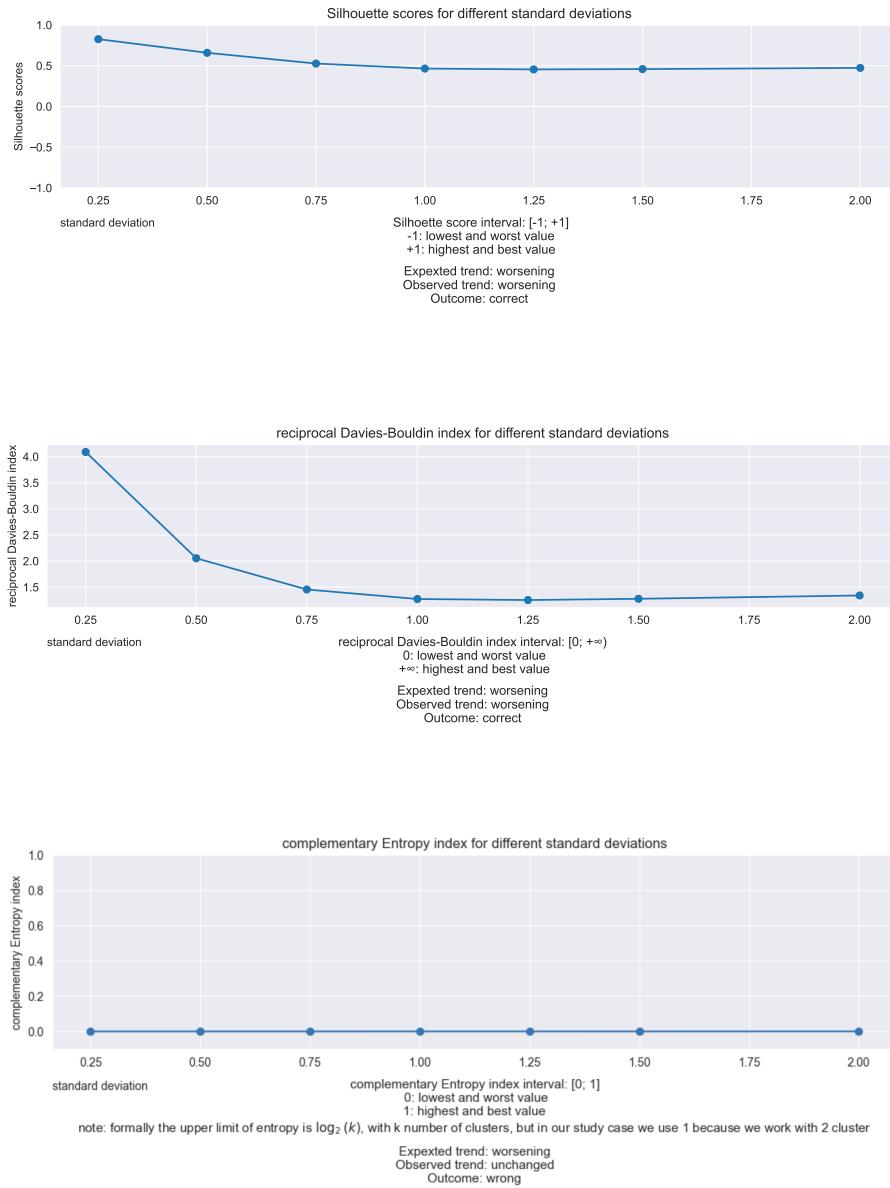
Table 4.6: risultati metriche dataset sfere

metriche	risultati	tempo
Silhouette	corretto	0.935 sec
Entropia complementare	errato	0.0005 sec
Davies-Bouldin reciproco	corretto	0.001 sec
Dunn index	corretto	0.412 sec
Calinski-Harabasz	corretto	0.001 sec
Gap Statistic	errato	0.267 sec

4.2.6 risultati dataset pennellate

Dato questo dataset mi aspetto che le metriche peggiorino all'aumentare della deviazione standard. Questo accade per tutte le metriche tranne che per l'Entropia, che rimane costante.

Figure 4.7: risultati dataset pennellate



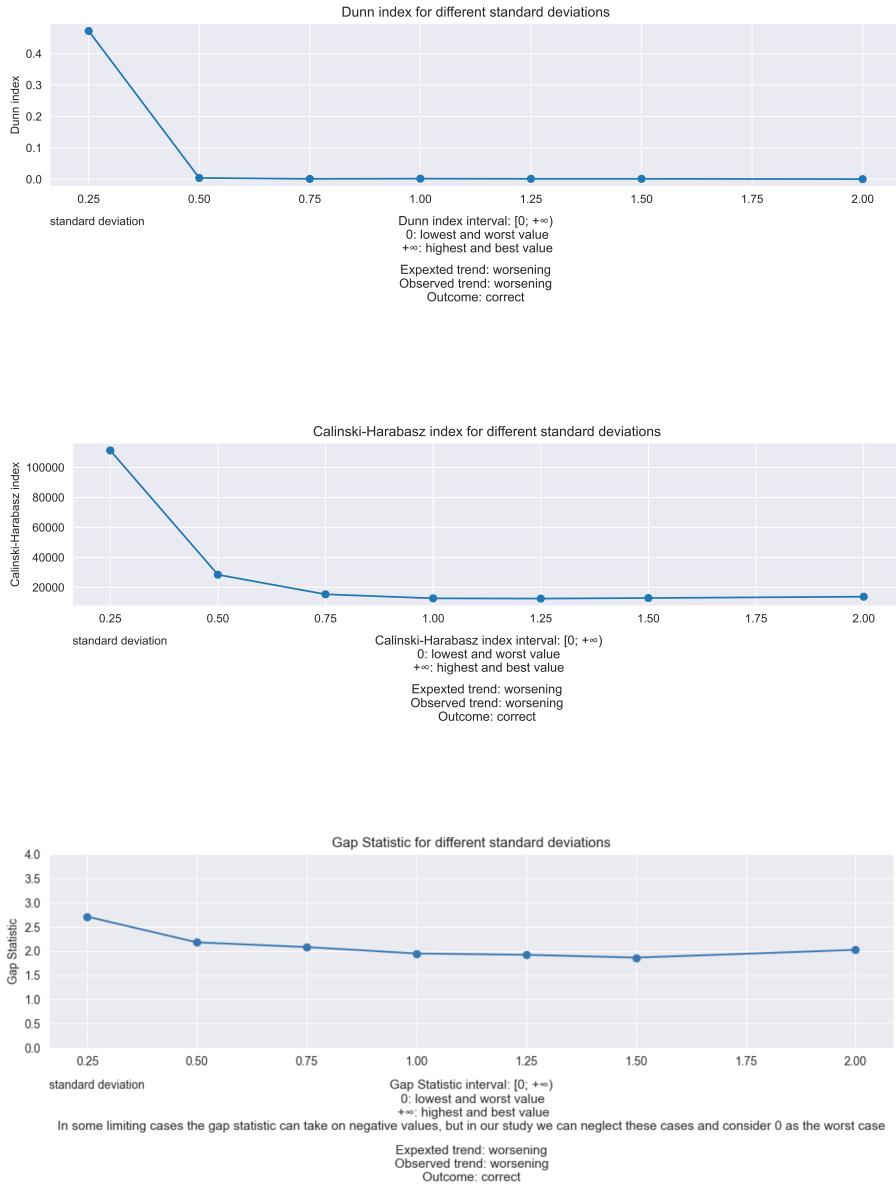


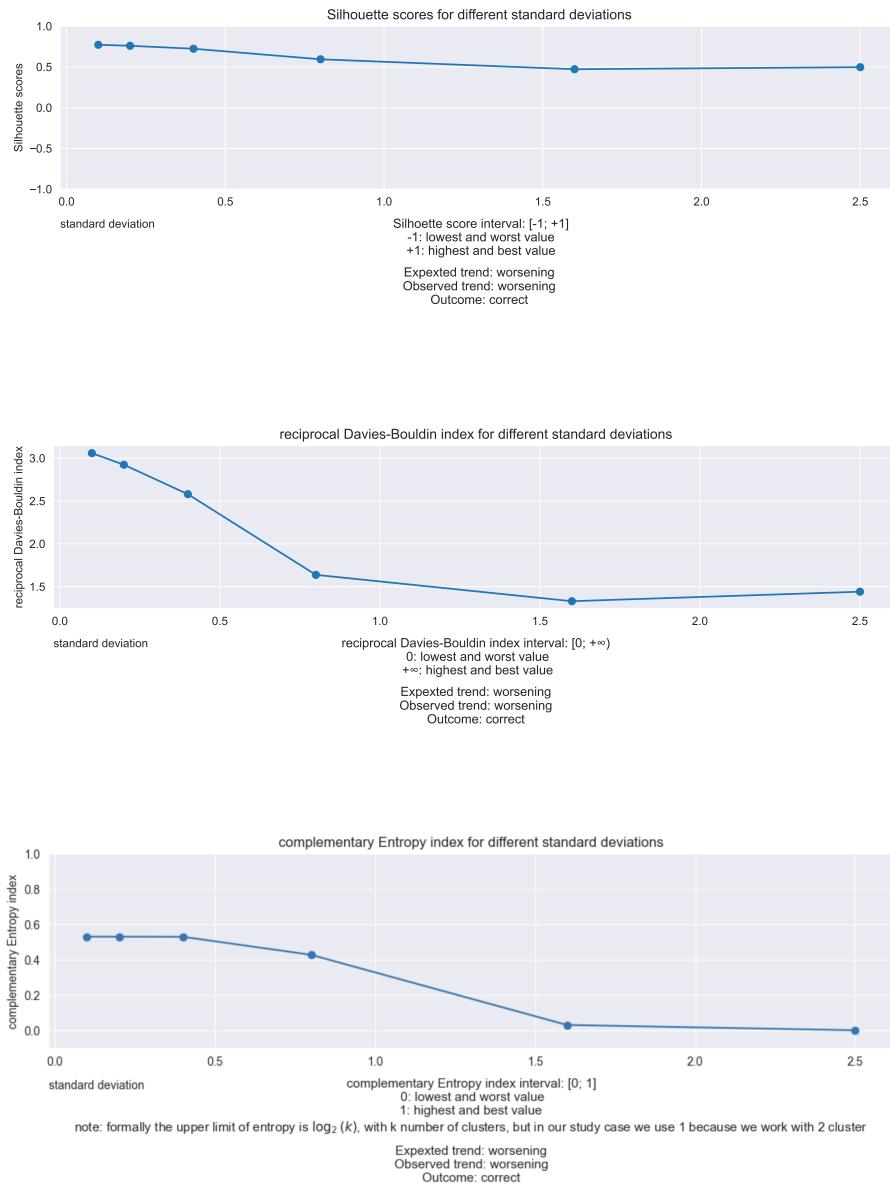
Table 4.7: risultati metriche dataset pennellate

metriche	risultati	tempo
Silhouette	corretto	0.959 sec
Entropia complementare	errato	0.0005 sec
Davies-Bouldin reciproco	corretto	0.001 sec
Dunn index	corretto	0.406 sec
Calinski-Harabasz	corretto	0.001 sec
Gap Statistic	corretto	0.259 sec

4.2.7 risultati dataset W

Dato questo dataset mi aspetto che le metriche peggiorino all'aumentare della deviazione standard. Questo accade per tutte le metriche tranne che per Gap Statistic che ha un andamento irregolare.

Figure 4.8: risultati dataset W



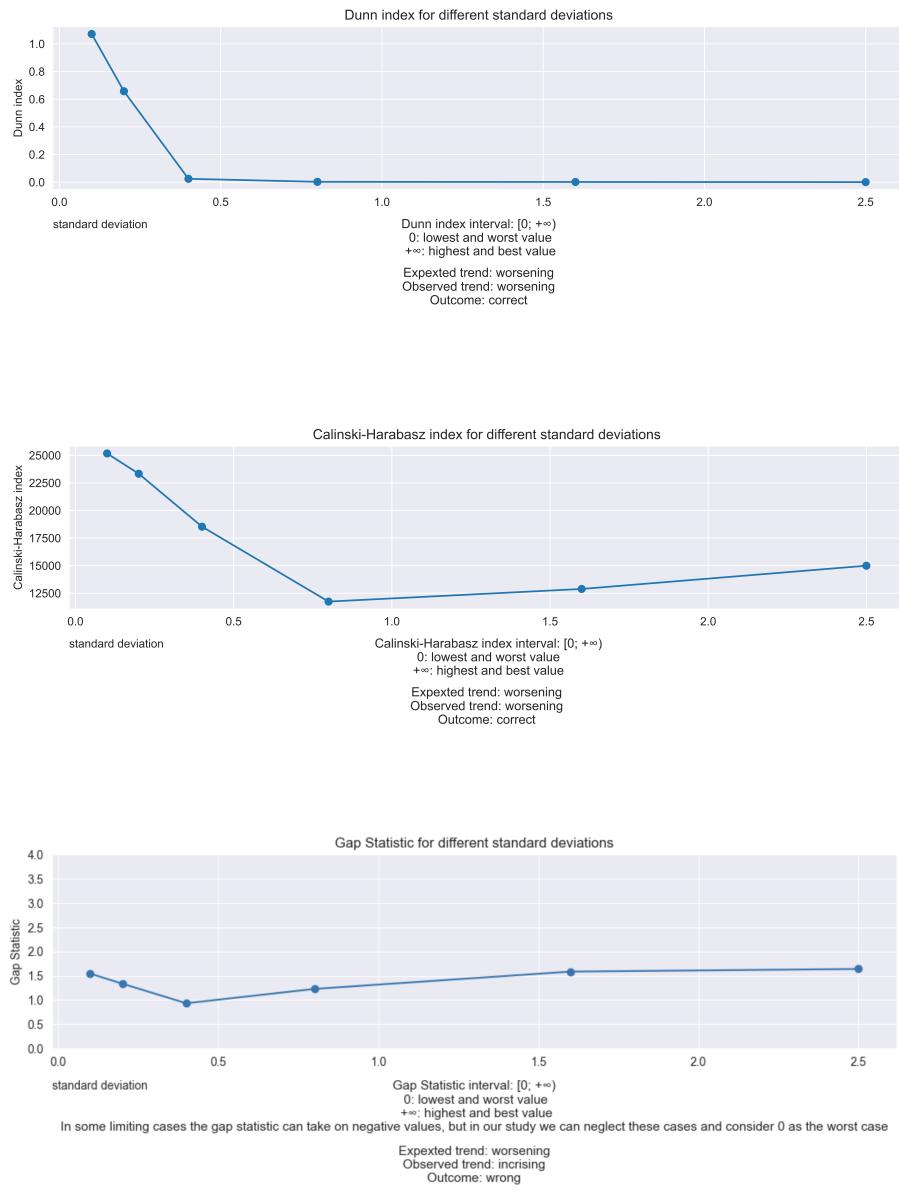


Table 4.8: tabella risultati metriche dataset W

metriche	risultati	tempo
Silhouette	corretto	1.029 sec
Entropia complementare	corretto	0.0005 sec
Davies-Bouldin reciproco	corretto	0.001 sec
Dunn index	corretto	0.441 sec
Calinski-Harabasz	errato	0.001 sec
Gap Statistic	errato	0.277 sec

4.2.8 riassunto risultati dataset artificiali

Dopo aver analizzato i risultati ottenuti sui 7 dataset artificiali, in questa sezione presenteremo una tabella riassuntiva che confronta le performance delle diverse metriche di validazione utilizzate. La tabella mostra, per ogni metrica, quante volte ha funzionato correttamente nel rilevare la qualità del clustering, indicando la sua affidabilità complessiva. Inoltre, la tabella riporta il tempo medio di calcolo per ciascuna metrica, fornendo un'indicazione delle risorse computazionali richieste.

Table 4.9: tabella riassuntiva dell'andamento delle metriche nei dataset artificiali

metriche	corretto	errato	%	tempo medio
Silhouette	6	1	85.71	0.284 sec
Davies-Bouldin reciproco	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
Entropia complementare	1	6	14.29	0.0001 sec

- La metrica Silhouette ha dimostrato un'elevata affidabilità, con un tasso di successo dell'85.71%. Tuttavia, ha un tempo medio di calcolo relativamente elevato rispetto alle altre metriche, richiedendo mediamente 0.284 secondi. Questa metrica è particolarmente utile quando si desidera una misura chiara della coesione e della separazione dei cluster, nonostante il costo computazionale più alto.
- La metrica Davies-Bouldin ha mostrato performance equivalenti alla Silhouette in termini di accuratezza, con un tasso di successo dell'85.71%. Tuttavia, il suo tempo medio di calcolo è significativamente inferiore, rendendola una scelta preferibile in scenari in cui le risorse computazionali sono limitate.
- Il Dunn Index ha ottenuto un discreto tasso di successo del 71.43%, ma ha un tempo medio di calcolo superiore a quello della metrica Davies-Bouldin reciproco. Quindi considerando che ci sono metriche che hanno una precisione maggiore e nel contempo un costo computazionale minore, dal mio studio si evince come il Dunn non sia la prima scelta come metrica interna per il calcolo della qualità del clustering.
- La metrica Calinski-Harabasz ha un tasso di successo del 57.14%, indicando una performance non particolarmente buone. Tuttavia, il suo tempo medio di calcolo è molto basso, rendendola una metrica efficiente dal punto di vista computazionale. Questa metrica può essere utile come misura supplementare quando si valutano altre metriche più accurate.
- La Gap Statistic ha mostrato una performance mediocre, con un tasso di successo del 42.86%. Il tempo medio di calcolo è moderato. Questa metrica potrebbe non essere la scelta ideale per la valutazione del clustering nei dataset utilizzati in questo studio, ma risulta sicuramente più utile nel

calcolo del numero di cluster ottimali che è lo scopo principali per cui è stata pensata.

- La Entropia complementare ha avuto la performance peggiore, con un tasso di successo del 14.29%. Nonostante il tempo medio di calcolo molto basso, questa metrica non sembra essere efficace per la valutazione del clustering nei dataset esaminati.

4.3 Risultati dataset medici reali di cartelle cliniche elettroniche

Per ognuno dei 6 dataset analizzati verranno mostrati i risultati per ognuna delle soglie con cui sono stati calcolati. In particolare per ogni soglia verranno mostrati:

- i risultati delle metriche
- la tabella contenente i cluster ottenuti con la distanza euclidea
- la tabella contenente i cluster ottenuti con k-means
- i risultati del calcolo dell'Adjusted Rand Index
- La consistenza tra la metrica Silhouette e l'Adjusted Rand Index. Di seguito i 3 casi possibili, dato

$$r = \frac{\text{metrica Silhouette}}{\text{l'Adjusted Rand Index}}$$

- $0 < r < 0.33$ allora si dirà che sono consistenti
- $0.33 < r < 1$ allora si dirà che sono sufficientemente consistenti
- $1 < r < 2$ allora si dirà che non sono consistenti

4.3.1 dataset artificiale

risultati con soglia del 20%

valori delle metriche:

- Silhouette Score: 0.991
- Entropia complementare Score: 0.0
- Calinski-Harabasz Score: 48464.583
- Davies-Bouldin reciproco Score: 90.787
- Dunn index Score: 75.781
- Gap Statistic Score: 7.278

Table 4.10: tabella dei cluster ottenuti con la distanza euclidea sul dataset artificiale con soglia del 20%

cluster	cluster elements	%
1° cluster	5 pazienti su 10	50.0
2° cluster	5 pazienti su 10	50.0
no cluster	0 pazienti su 10	0.0

Table 4.11: tabella dei cluster ottenuti k-means sul dataset artificiale con soglia del 20%

cluster	cluster elements	%
1° cluster	5 pazienti su 10	50.0
2° cluster	5 pazienti su 10	50.0
no cluster	0 pazienti su 10	0.0

il 100% degli elementi è stato classificato correttamente

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.

valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:

La Silhouette e l'Adjusted Rand index sono consistenti

risultati con soglia del 33%

valori delle metriche:

- Silhouette Score: 0.991
- Entropia complementare Score: 0.0
- Calinski-Harabasz Score: 48464.583
- Davies-Bouldin reciproco Score: 90.787
- Dunn index Score: 75.781
- Gap Statistic Score: 7.278

Table 4.12: tabella dei cluster ottenuti con la distanza euclidea sul dataset artificiale con soglia del 33%

cluster	cluster elements	%
1° cluster	5 pazienti su 10	50.0
2° cluster	5 pazienti su 10	50.0
no cluster	0 pazienti su 10	0.0

il 100% degli elementi è stato classificato correttamente

Table 4.13: tabella dei cluster ottenuti k-means sul dataset artificiale con soglia del 33%

cluster	cluster elements	%
1° cluster	5 pazienti su 10	50.0
2° cluster	5 pazienti su 10	50.0
no cluster	0 pazienti su 10	0.0

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.

valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:

La Silhouette e l'Adjusted Rand index sono consistenti

Table 4.14: risultati metriche su dataset medico artificiale

metriche	risultati	tempo
Silhouette	corretto	1.662 ms
Entropia complementare	corretto	0.313 ms
Davies-Bouldin reciproco	corretto	0.439 ms
Dunn index	corretto	0.708 ms
Calinski-Harabasz	corretto	0.553 ms
Gap Statistic	corretto	629.955 ms

4.3.2 risultati dataset sepsi & SIRS

risultati con soglia del 20%

valori delle metriche:

- Silhouette Score: 0.491
- Entropia complementare Score: 0.006
- Calinski-Harabasz Score: 499.571
- Davies-Bouldin reciproco Score: 1.155
- Dunn index Score: 0.729
- Gap Statistic Score: 0.987

Table 4.15: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante la sepsi & SIRS con soglia del 20%

cluster	cluster elements	%
1° cluster	226 pazienti su 1257	32.776
2° cluster	187 pazienti su 1257	1.193
no cluster	844 pazienti su 1257	66.030

Il 32.856% degli elementi è stato classificato correttamente nei due cluster. Il restante 67.144% degli elementi non è stato classificato

Table 4.16: tabella dei cluster ottenuti k-means sul dataset riguardante la sepsi & SIRS con soglia del 20%

cluster	cluster elements	%
1° cluster	226 pazienti su 1257	7.956
2° cluster	187 pazienti su 1257	26.014
no cluster	844 pazienti su 1257	66.030

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.
valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:
La Silhouette e l'Adjusted Rand index sono consistenti

risultati con soglia del 33%

valori delle metriche:

- Silhouette Score: 0.364

- Entropia complementare Score: 0.050
- Calinski-Harabasz Score: 560.611
- Davies-Bouldin reciproco Score: 0.816
- Dunn index Score: 0.044
- Gap Statistic Score: 0.731

Table 4.17: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante la sepsi & SIRS con soglia del 33%

cluster	cluster elements	%
1° cluster	883 pazienti su 1257	70.246
2° cluster	169 pazienti su 1257	13.444
no cluster	205 pazienti su 1257	16.309

83.691% degli elementi è stato classificato correttamente nei due cluster. Il restante 16.309% degli elementi non è stato classificato

Table 4.18: tabella dei cluster ottenuti con k-means sul dataset riguardante la sepsi & SIRS con soglia del 33%

cluster	cluster elements	%
1° cluster	388 pazienti su 1257	30.867
2° cluster	664 pazienti su 1257	52.824
no cluster	205 pazienti su 1257	16.309

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.
valore di Adjusted Rand index ottenuto: 0.318

Consistenza tra Silhouette e Adjusted Rand index:
La Silhouette e l'Adjusted Rand index sono consistenti

Table 4.19: risultati metriche su dataset riguardante sepsi & SIRS

metriche	risultati	tempo
Silhouette	corretto	14.579 ms
Entropia complementare	errato	0.406 ms
Davies-Bouldin reciproco	corretto	1.863 ms
Dunn index	corretto	2.346 ms
Calinski-Harabasz	errato	2.494 ms
Gap Statistic	corretto	286.183 ms

4.3.3 risultati dataset depressione e insufficienza cardiaca

risultati con soglia del 20%

valori delle metriche:

- Silhouette Score: 0.616
- Entropia complementare Score: 0.0543
- Calinski-Harabasz Score: 28.434
- Davies-Bouldin reciproco Score: 1.920
- Dunn index Score: 1.460
- Gap Statistic Score: 0.278

Table 4.20: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante depressione & insufficienza cardiaca con soglia del 20%

cluster	cluster elements	%
1° cluster	7 pazienti su 425	1.647
2° cluster	4 pazienti su 425	0.941
no cluster	414 pazienti su 425	97.412

2.588% degli elementi è stato classificato correttamente nei due cluster. Il restante 97.412% degli elementi non è stato classificato

Table 4.21: tabella dei cluster ottenuti k-means sul dataset riguardante depressione & insufficienza cardiaca con soglia del 20%

cluster	cluster elements	%
1° cluster	4 pazienti su 425	0.941
2° cluster	7 pazienti su 425	1.647
no cluster	414 pazienti su 425	97.412

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.

valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:

La Silhouette e l'Adjusted Rand index sono sufficientemente consistenti

risultati con soglia del 33%

valori delle metriche:

- Silhouette Score: 0.263
- Entropia complementare Score: 0.0350
- Calinski-Harabasz Score: 30.785

- Davies-Bouldin reciproco Score: 0.651
- Dunn index Score: 0.193
- Gap Statistic Score: 0.030

Table 4.22: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante depressione & insufficienza cardiaca con soglia del 33%

cluster	cluster elements	%
1° cluster	55 pazienti su 425	12.941
2° cluster	27 pazienti su 425	6.353
no cluster	343 pazienti su 425	80.706

19.294% degli elementi è stato classificato correttamente nei due cluster. Il restante 80.706% degli elementi non è stato classificato

Table 4.23: tabella dei cluster ottenuti con k-means sul dataset riguardante depressione & insufficienza cardiaca con soglia del 33%

cluster	cluster elements	%
1° cluster	50 pazienti su 425	11.765
2° cluster	32 pazienti su 425	7.529
no cluster	343 pazienti su 425	80.706

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.
valore di Adjusted Rand index ottenuto: 0.179

Consistenza tra Silhouette e Adjusted Rand index:
La Silhouette e l'Adjusted Rand index sono consistenti

Table 4.24: risultati metriche sul dataset degli riguardante depressione & insufficienza cardiaca

metriche	risultati	tempo
Silhouette	corretto	1.022 ms
Entropia complementare	corretto	0.218 ms
Davies-Bouldin reciproco	corretto	0.63 ms
Dunn index	corretto	0.864 ms
Calinski-Harabasz	errato	0.822 ms
Gap Statistic	corretto	15.972 ms

4.3.4 risultati dataset arresto cardiaco

risultati con soglia del 20%

valori metriche:

- Silhouette Score: 0.765
- Entropia complementare Score: 0.110
- Calinski-Harabasz Score: 229.912
- Davies-Bouldin reciproco Score: 3.131
- Dunn index Score: 1.776
- Gap Statistic Score: 0.878

Table 4.25: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante l'arresto cardiaco con soglia del 20%

cluster	cluster elements	%
1° cluster	27 pazienti su 422	6.398104
2° cluster	12 pazienti su 422	2.843602
no cluster	383 pazienti su 422	90.758294

9.242% degli elementi è stato classificato correttamente nei due cluster. Il restante 90.758% degli elementi non è stato classificato

Table 4.26: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante l'arresto cardiaco con soglia del 20%

cluster	cluster elements	%
1° cluster	27 pazienti su 422	6.398104
2° cluster	12 pazienti su 422	2.843602
no cluster	383 pazienti su 422	90.758294

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.
valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:
La Silhouette e l'Adjusted Rand index sono consistenti

risultati con soglia del 33%

valori metriche:

- Silhouette Score: 0.362
- Entropia complementare Score: 0.0197
- Calinski-Harabasz Score: 52.903
- Davies-Bouldin reciproco Score: 0.762
- Dunn index Score: 0.0351
- Gap Statistic Score: 0.099

Table 4.27: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante l'arresto cardiaco con soglia del 33%

cluster	cluster elements	%
1° cluster	60 pezienti su 422	14.218009
2° cluster	43 pezienti su 422	10.189573
no cluster	319 pezienti su 422	75.592417

24.408% degli elementi è stato classificato correttamente nei due cluster. Il restante 24.408% degli elementi non è stato classificato

Table 4.28: tabella dei cluster ottenuti con k-means sul dataset riguardante l'arresto cardiaco con soglia del 33%

cluster	cluster elements	%
1° cluster	60 pezienti su 422	14.218009
2° cluster	43 pezienti su 422	10.189573
no cluster	319 pezienti su 422	75.592417

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.
valore di Adjusted Rand index ottenuto: 0.470

Consistenza tra Silhouette e Adjusted Rand index:
La Silhouette e l'Adjusted Rand index sono consistenti

Table 4.29: tabella risultati metriche sul dataset riguardante l'arresto cardiaco

metriche	risultati	tempo
Silhouette	corretto	0.783 ms
Entropia complementare	corretto	0.198 ms
Davies-Bouldin reciproco	corretto	0.538 ms
Dunn index	corretto	0.805 ms
Calinski-Harabasz	corretto	1.149 ms
Gap Statistic	corretto	16.656 ms

4.3.5 risultati dataset neuroblastoma

risultati con soglia del 20%

valori metriche

- Silhouette Score: 0.528
- Entropia complementare Score: 0.082
- Calinski-Harabasz Score: 17.586
- Davies-Bouldin reciproco Score: 1.409

- Dunn index Score: 1.357
- Gap Statistic Score: -0.838

Table 4.30: tabella dei cluster ottenuti con la distanza euclidea sul dataset riguardante il neuroblastoma con soglia del 20%

cluster	cluster elements	%
1° cluster	4 pezienti su 169	2.366864
2° cluster	8 pezienti su 169	4.733728
no cluster	157 pezienti su 169	92.899408

7.101% degli elementi è stato classificato correttamente nei due cluster. Il restante 92.899% degli elementi non è stato classificato

Table 4.31: tabella dei cluster ottenuti con k-means sul dataset riguardante il neuroblastoma con soglia del 20%

cluster	cluster elements	%
1° cluster	8 pezienti su 169	4.733728
2° cluster	4 pezienti su 169	2.366864
no cluster	157 pezienti su 169	92.899408

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.

valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:

La Silhouette e l'Adjusted Rand index sono sufficientemente consistenti

risultati con soglia del 33%

valori delle metriche:

- Silhouette Score: 0.255
- Entropia complementare Score: 0.0290
- Calinski-Harabasz Score: 13.057
- Davies-Bouldin reciproco Score: 0.638
- Dunn index Score: 0.2513
- Gap Statistic Score: -0.269

Table 4.32: tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il neuroblastoma con soglia del 33%

cluster	cluster elements	%
1° cluster	25 pezienti su 169	14.792899
2° cluster	15 pezienti su 169	8.875740
no cluster	129 pezienti su 169	76.331361

23.669% degli elementi è stato classificato correttamente nei due cluster. Il restante 76.331% degli elementi non è stato classificato

Table 4.33: tabella dei cluster ottenuti con k-means sul dataset riguardante il neuroblastoma con soglia del 33%

cluster	cluster elements	%
1° cluster	16 pezienti su 169	9.467456
2° cluster	24 pezienti su 169	14.201183
no cluster	129 pezienti su 169	76.331361

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.

valore di Adjusted Rand index ottenuto: 0.182

Consistenza tra Silhouette e Adjusted Rand index:

La Silhouette e l'Adjusted Rand index sono consistenti

Table 4.34: tabella risultati metriche su dateset riguardante il neuroblastoma

metriche	risultati	tempo
Silhouette	corretto	1.168 ms
Entropia complementare	corretto	0.214 ms
Davies-Bouldin reciproco	corretto	0.796 ms
Dunn index	corretto	0.86 ms
Calinski-Harabasz	corretto	0.765 ms
Gap Statistic	corretto	16.107 ms

4.3.6 risultati dataset diabete di tipo 1

risultati con soglia del 20%

- valori delle metriche:
- Silhouette Score: 0.4100
- Entropia complementare Score: 0.145
- Calinski-Harabasz Score: 19.313
- Davies-Bouldin reciproco Score: 1.077

- Dunn index Score: 1.126
- Gap Statistic Score: 0.398

Table 4.35: tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il diabete di tipo 1 con soglia del 20%

cluster	cluster elements	%
1° cluster	7 pezienti su 67	10.447761
2° cluster	18 pezienti su 67	26.865672
no cluster	42 pezienti su 67	62.686567

37.313% degli elementi è stato classificato correttamente nei due cluster. Il restante 62.687% degli elementi non è stato classificato

Table 4.36: tabella dei cluster ottenuti con k-means sul dataset riguardante il diabete di tipo 1 con soglia del 20%

cluster	cluster elements	%
1° cluster	7 pezienti su 67	10.447761
2° cluster	18 pezienti su 67	26.865672
no cluster	42 pezienti su 67	62.686567

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.

valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:

La Silhouette e l'Adjusted Rand index sono sufficientemente consistenti

risultati con soglia del 33%

valori delle metriche:

- Silhouette Score: 0.3789028460610209
- Entropia complementare Score: 0.10787871942283167
- Calinski-Harabasz Score: 36.24251472884974
- Davies-Bouldin reciproco Score: 0.90894898134704
- Dunn index Score: 0.6536529591951651
- Gap Statistic Score: 0.6348699587056217

Table 4.37: tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il diabete di tipo 1 con soglia del 33%

cluster	cluster elements	%
1° cluster	38 pezienti su 67	56.716418
2° cluster	17 pezienti su 67	25.373134
no cluster	12 pezienti su 67	17.910448

82.09% degli elementi è stato classificato correttamente nei due cluster. Il restante 17.91% degli elementi non è stato classificato

Table 4.38: tabella dei cluster ottenuti con distanza euclidea sul dataset riguardante il diabete di tipo 1 con soglia del 33%

cluster	cluster elements	%
1° cluster	17 pezienti su 67	25.373134
2° cluster	38 pezienti su 67	56.716418
no cluster	12 pezienti su 67	17.910448

Per confrontare i risultati ottenuti con k-means da quelli ottenuti con la distanza euclidea utilizzo l'Adjusted Rand index.
valore di Adjusted Rand index ottenuto: 1.0

Consistenza tra Silhouette e Adjusted Rand index:
La Silhouette e l'Adjusted Rand index sono sufficientemente consistenti

Table 4.39: risultati metriche su dataset riguardante il diabete di tipo 1

metriche	risultati	tempo
Silhouette	corretto	0.864 ms
Entropia complementare	corretto	0.202 ms
Davies-Bouldin reciproco	corretto	0.567 ms
Dunn index	corretto	0.827 ms
Calinski-Harabasz	errato	0.789 ms
Gap Statistic	errato	232.578 ms

4.3.7 riassunto risultati dataset reali di cartelle cliniche mediche

Dopo aver analizzato i risultati ottenuti sui 6 dataset reali di cartelle cliniche mediche, in questa sezione presenteremo una tabella riassuntiva che confronta le performance delle diverse metriche di validazione utilizzate. La tabella mostra, per ogni metrica, quante volte ha funzionato correttamente nel rilevare la qualità del clustering, indicando la sua affidabilità complessiva. In questi casi diciamo che una metrica funziona correttamente se peggiora al crescere della soglia con cui vengono filtrati i risultati del cluster effettuato con la distanza euclidea. Inoltre, la tabella riporta il tempo medio di calcolo per ciascuna metrica, fornendo un'indicazione delle risorse computazionali richieste.

Table 4.40: tabella risultati finali metriche dataset reali

metriche	corretto	errato	%	avarage time
Silhouette	6	0	100.0 %	3.346 ms
Davies-Bouldin reciproco	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
Entropia complementare	3	3	50.0 %	0.258 ms

Silhouette, Davies-Bouldin reciproco, e Dunn Index emergono come le metriche più affidabili per la validazione del clustering, con una percentuale di successo del 100%. Davies-Bouldin si distingue per il suo ottimo equilibrio tra accuratezza e tempo di calcolo, rendendola la scelta preferita per scenari pratici. Calinski-Harabasz offre un buon compromesso tra accuratezza e velocità di calcolo, risultando utile in contesti dove la velocità è una priorità. Gap Statistic presenta una discreta accuratezza, ma comunque peggiore delle altre metriche precedentemente citate, ma inoltre è anche penalizzata dalla sua alta complessità computazionale. L'Entropia, sebbene estremamente veloce, non garantisce un'accuratezza elevata e non risulta affidabile per decretare la qualità di un clustering.

La seguente tabella riassume i risultati della correlazione tra la metrica Silhouette e l'Adjusted Rand Index (ARI) calcolati su vari dataset reali di cartelle cliniche mediche. La tabella presenta i risultati per diverse soglie di correlazione (20% e 30%) e classifica la consistenza tra le due metriche come "consistenti", "sufficientemente consistenti", o "non consistenti".

Table 4.41: tabella riassuntiva correlezione tra Silhouette e Adjusted Rand Index

database	threshold	result
dataset artificiale	20%	consistenti
dataset artificiale	30%	consistenti
diabete di tipo 1	20%	sufficientemente consistenti
diabete di tipo 1	30%	sufficientemente consistenti
arresto cardiaco	20%	consistenti
arresto cardiaco	30%	sufficientemente consistenti
depressione & insufficienza cardiaca	20%	sufficientemente consistenti
depressione & insufficienza cardiaca	30%	consistenti
sepsi & SIRS	20%	consistenti
sepsi & SIRS	30%	consistenti
neuroblastoma	20%	consistenti
neuroblastoma	30%	sufficientemente consistenti

I risultati della tabella indicano che, nella maggior parte dei casi, esiste una buona correlazione tra la metrica Silhouette e l'Adjusted Rand Index nei dataset di cartelle cliniche elettroniche.

Chapter 5

Conclusione e discussione

Questo studio contribuisce alla letteratura esistente offrendo una valutazione comparativa di alcune delle metriche interne di validazione del clustering più conosciute. Fino ad ora, tale argomento risultava essere poco trattato, e questo lavoro sottolinea la necessità di ulteriori ricerche e test per approfondire la comprensione e l'efficacia di queste metriche. Analizzando i risultati 4.9 4.40 le metriche Silhouette e Davies-Bouldin si sono dimostrate le più performanti, sia per quanto riguarda i dataset artificiali, sia per quanto riguarda i dataset reali. Questo è particolarmente interessante se si considera che queste metriche sono tra le più utilizzate e diffuse all'interno della comunità scientifica (come mostrato nella tabella e nel grafico seguente), suggerendo una conferma empirica della loro validità e affidabilità.

Table 5.1: tabelle con il numero di articoli pubblicati su google scholar per ogni metrica dall'anno 2000 ad oggi

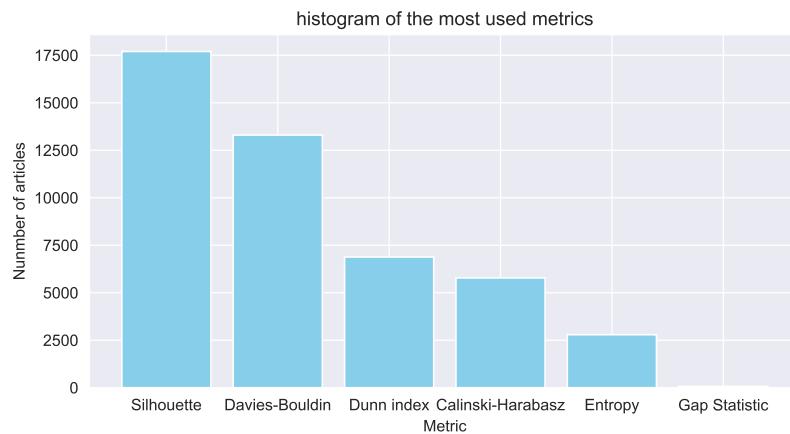
metrica	numero di articoli
Silhouette	17,700
Davies-Bouldin	13,300
Dunn index	6,880
Calinski-Harabasz	5,780
Entropia	2,790
Gap Statistic	69

il numero di articoli trovati e il risultato di una ricerca su google scholar con le seguenti queries:

- "Silhouette coefficient" OR "Silhouette index" OR "Silhouette score"
- "Davies-Bouldin coefficient" OR "Davies-Bouldin index" OR "Davies-Bouldin score"
- "Dunn coefficient" OR "Dunn index" OR "Dunn score"
- "Calinski-Harabasz coefficient" OR "Calinski-Harabasz index" OR "Calinski-Harabasz score"
- "Entropy clustering"

- "Gap Statistics score" OR "Gap Statistics index" OR "Gap Statistics coefficient" OR "Gap Statistic score" OR "Gap Statistic index" OR "Gap Statistic coefficient"

Figure 5.1: istrogramma sui numero di articoli delle metriche presenti su google scholar



Tuttavia, ci sono alcuni limiti nel mio lavoro che vanno evidenziati. Innanzitutto, il numero di dataset utilizzati per i test è relativamente limitato. Ampliando lo studio con nuovi test su una varietà più ampia di dataset, si potrebbe ottenere una visione più completa e accurata delle prestazioni delle metriche. Inoltre, il mio confronto si è concentrato su sei metriche tra le più note. Esistono metriche nuove, come quella proposta nell'articolo [21], che potrebbero rivelarsi più efficienti rispetto a quelle testate, suggerendo una direzione futura per ulteriori ricerche.

Chapter 6

Dettagli tecnici

Linee guida scrittura tesi

Per la strutturazione della mia tesi, ho seguito le indicazioni riportate nel seguenti articolo [11]. Questo articolo spiega come strutturare una tesi nell'ambito della bioinformatica. Nonostante la mia tesi riguardi un ambito differente dell'informatica le linee guida risultavano comunque valide. Inoltre in parallelo alla realizzazione della tesi ho scritto anche una documentazione accurata di tutti i passaggi che svolgevo seguendo le linee guida dell'articolo [14]

6.1 dettagli tecnici codice

Il progetto è stato svolto utilizzando Python 3.10.14 come linguaggio di programmazione, sfruttando l'ambiente Jupyter Notebook per lo sviluppo e l'esecuzione del codice.

Le principali librerie utilizzate nel progetto sono state:

- `seaborn` (versione: 0.13.2) [30] e `matplotlib` (versione: 3.8.4) [26] per la generazione dei grafici
- `scikit-learn` (versione: 1.4.2) [28] e `scipy` (versione: 1.13.0) [29] per il calcolo delle metriche di validazione del clustering.

Il codice è stato salvato e caricato progressivamente su GitHub, garantendo così una versione controllata e un backup costante del lavoro svolto. Di seguito il link alla repository di github in cui si può trovare il codice e tutti i risultati del mio stage [27]

6.2 dettagli tecnici hardware utilizzato

Per quanto riguarda i tempi di calcolo delle metriche, questi sono stati misurati su una computer con le seguenti specifiche tecniche:

- Modello: MacBook Air 2020
- Processore: Apple M1 (prima generazione)

- Memoria RAM: 8 GB
- Memoria di Archiviazione: 256 GB
- Sistema Operativo: macOS Sonoma, versione 14.2.1

6.3 contatti

Per eventuali dubbi e domande riguardo la mia tesi potete contattarmi al seguente indirizzo email: andreaspongolo99@gmail.com

Bibliography

- [1] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [2] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967.
- [3] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.
- [4] J. C. Dunn, “A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters,” *Journal of Cybernetics*, 1973.
- [5] C. Tadeusz and H. Jerzy, “A dendrite method for cluster analysis,” *Communications in Statistics*, 1974.
- [6] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.
- [7] P. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Computational and Applied Mathematics*, 1987.
- [8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *kdd*, vol. 96, 1996, pp. 226–231.
- [9] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *Journal of the Royal Statistical Society*, 2001.
- [10] Y. Shim, J. Chung, and I.-C. Choi, “A comparison study of cluster validity indices using a nonhierarchical clustering algorithm,” in *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06)*, IEEE, vol. 1, 2005, pp. 199–204.
- [11] W. S. Noble, “A quick guide to organizing computational biology projects,” *PLoS computational biology*, vol. 5, no. 7, e1000424, 2009.
- [12] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of internal clustering validation measures,” in *2010 IEEE international conference on data mining*, Ieee, 2010, pp. 911–916.

- [13] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, “An extensive comparative study of cluster validity indices,” *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.
- [14] S. Schnell, *Ten simple rules for a computational biologist’s laboratory notebook*, 2015.
- [15] B. Guçyetmez and H. K. Atalan, “C-reactive protein and hemogram parameters for the non-sepsis systemic inflammatory response syndrome and sepsis: What do they mean?” *PLOS One*, vol. 11, no. 2, e0148699, 2016.
- [16] B. D. Jani, F. S. Mair, V. L. Roger, S. A. Weston, R. Jiang, and A. M. Chamberlain, “Comorbid depression and heart failure: A community cohort study,” *PLOS One*, vol. 11, no. 6, e0158570, 2016.
- [17] F. Nielsen and F. Nielsen, “Hierarchical clustering,” *Introduction to HPC with MPI for Data Science*, pp. 195–211, 2016.
- [18] R. Requena-Morales, A. Palazón-Bru, M. M. Rizo-Baeza, J. M. Adsuar-Quesada, V. F. Gil-Guillén, and E. Cortés-Castell, “Mortality after out-of-hospital cardiac arrest in a spanish region,” *PLOS One*, vol. 12, no. 4, e0175818, 2017.
- [19] Y. Ma, J. Zheng, J. Feng, L. Chen, K. Dong, and X. Xiao, “Neuroblastomas in eastern china: A retrospective series study of 275 cases in a regional center,” *PeerJ*, vol. 6, e5665, 2018.
- [20] Y. Takashi, M. Ishizu, H. Mori, *et al.*, “Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes,” *PLOS One*, vol. 14, no. 5, e0216416, 2019.
- [21] X. Wang and Y. Xu, “An improved index for clustering validation based on silhouette index and calinski-harabasz index,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 569, 2019, p. 052024.
- [22] M. Shutaywi and N. N. Kachouie, “Silhouette analysis for performance evaluation in machine learning with applications to clustering,” *Entropy*, vol. 23, no. 6, p. 759, 2021.
- [23] *Blob di esempio*, https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg, ultima visita: 10-07-2024.
- [24] *Google scholar*, <https://scholar.google.com/>, ultima visita: 14-07-2024.
- [25] *Immagine sepsi e sirs*, https://it.wikipedia.org/wiki/Sindrome_da_risposta_infiammatoria_sistemica, ultima visita: 06-07-2024.
- [26] *Matplotlib*, <https://matplotlib.org/>, ultima visita: 11-07-2024.
- [27] *Repository github*, <https://github.com/andreaspagnolo/stage.git>, ultima visita: 14-07-2024.
- [28] *Scikit learn clustering*, <https://scikit-learn.org/stable/modules/clustering.html>, ultima visita: 13-07-2024.
- [29] *Scipy*, <https://scipy.org/>, ultima visita: 11-07-2024.
- [30] *Seaborn*, <https://seaborn.pydata.org/>, ultima visita: 11-07-2024.