Presentazione del Dunn Index

2) Il problema della valutazione di una o più tecniche di cluster analysis si occupa di determinare se e quanto un set di dati ha una struttura non randomica. Tale valutazione viene fatta tramite metriche di valutazione ed indici che sono tradizionalmente classificati in 2 categorie: esterni o supervisionati e interno o non supervisionati. Gli indici esterni valutano la similarità con una struttura esterna nota, quelli interni danno delle misure di coesione o separazione proprie delle strutture di ogni cluster.
Il dunn index appartiene alla categoria degli indici interni ed è definito come:

$$V_D = \min_{1 \le i \le k} \left\{ \min_{i+1 \le j \le k} \left(\frac{D(C_i, C_j)}{\max_{1 \le l \le k} diam(C_l)} \right) \right\},$$

k = il numero totale di gruppi formati dall'algoritmo di cluster

D = distanza tra due cluster → $D(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y),$

diam = diametro → $diam(C_l) = \max_{x,y \in C_l} d(x, y).$

d = $d(u, v) \triangleq \|u - v\| \triangleq \langle u - v | u - v \rangle^{1/2}$

Dunque, il Dunn Index è calcolato come il rapporto tra la minima distanza tra due diversi cluster e il più grande diametro tra i vari cluster. Valori alti dell'indice, ossia distanza minima tra cluster di gran lunga superiore al diametro massimo, significano che l'algoritmo produce CWS (compact and well-separated) clusters. Per capire il numero ottimale di cluster da utilizzare, è sufficiente vedere a quale k corrisponde il Dunn Index più alto. [1] *"Quantifica la relatività tra il grado di compattezza e il grado di separazione degli ammassi."* [3].

Gli aspetti negativi legati a questa metrica sono dovuti alla sua scarsissima scalabilità (più o meno si parla di 20 ore per dataset contenenti 2 milioni di osservazioni) e forte sensibilità agli outliers. Per questo motivo, spesso la distanza tra cluster viene approssimata alla distanza tra centroidi, definita come: $\delta ij = d\ (\ vi\ ,\ vj\ )$, *(2) con* vi *e* vj *sono rispettivamente i centroidi dei cluster* Ci *e* Cj *."*[3]

3) Articoli:

[1] https://www.crcpress.com/Data-Clustering-Algorithms-and-Applications/Aggarwal-Reddy/p/book/9781466558212

L'indice di Dunn viene descritto come un forte strumento di validazione, utilizzato per evidenziare la compattezza e la separazione dei cluster. I risultati mostrano che funziona bene con varie tecniche di clustering, ma può avere difficoltà con dati complessi e di grandi dimensioni.

[2] Clustering validation of CLARA and K-means using Silhuette & Dunn Measures on Iris Dataset

In questo articolo il Dunn Index viene applicato per convalidare il clustering del set di dati Iris. I risultati dimostrano che K-means si comporta meglio di CLARA in termini di separazione dei cluster, come dimostrato dai valori di Dunn più elevati.

[3] Parallel and scalable Dunn Index for the validation of big data clusters

La ricerca presenta una versione scalabile dell'indice di Dunn, che consente di migliorare l'efficienza nell'elaborazione di grandi insiemi di dati. È stato dimostrato che è in grado di convalidare efficacemente i cluster nei big data, con risultati paragonabili alle metriche di clustering tradizionali.

[4] BMS: An improved Dunn index for Document Clustering validation

L'indice BMS Dunn modificato offre una migliore differenziazione dei cluster di documenti. I risultati dimostrano che supera l'indice di Dunn tradizionale nell'individuazione di cluster ben separati e compatti nei dataset di testo.

Articoli ambito medico:

[5] Dunn's index for cluster tendency assessment of pharmacological data sets

L'indice di Dunn viene applicato per valutare la tendenza dei cluster nei dati farmacologici, distinguendo efficacemente i cluster significativi legati ai farmaci. I risultati rivelano che aiuta a scoprire modelli nascosti all'interno dei set di dati farmaceutici.

[6] Fuzzy and hard clustering analysis for thyroid disease

Lo studio utilizza l'indice di Dunn per confrontare il clustering fuzzy e quello hard su dati relativi a malattie della tiroide. I risultati indicano che il clustering fuzzy, con l'aiuto dell'indice di Dunn, cattura meglio i cluster sovrapposti, evidenziando i suoi vantaggi rispetto ai metodi di clustering hard.

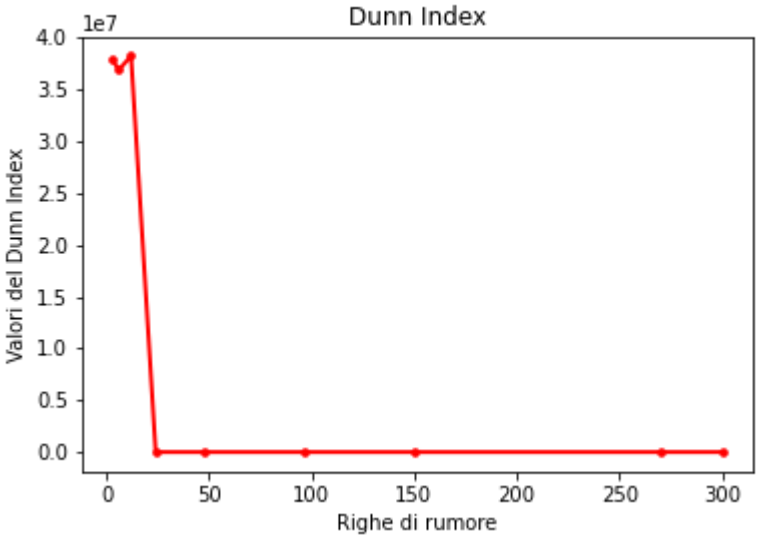
4) Pacchetti Python:

- Validclust

5) Per vedere il comportamento della metrica al variare della complessità di un dataset, la si è testata su un dataset artificiale costituito solo da 0 ed 1 che ad ogni test veniva disturbato con tuple di “disturbo” di numero sempre crescente. Le tuple sono costituite da numeri compresi tra 0 ed 1 e sostituite in maniera randomica all’interno del dataset originario.
Il dataset è creato unendo 150 tuple contenenti 150 0 ed altrettante tuple della stessa dimensione contenenti solo 1, dunque con una shape finale di 300x150.
La valutazione del primo test, con dataset senza disturbo e kmeans con k =2, ha come dunn index il valore **infinito**. Questo è il più alto valore che l’indice possa assumere. Siccome le tuple di rumore sono sostituite in maniera casuale all’interno del dataset, nonostante si sia impostato il random_state del k-means in modo tale che l’inizializzazione iniziale sia sempre la stessa, la metrica cambia leggermente perché il dataset è ogni volta leggermente diverso.
Impostando per ogni test un ciclo in cui il numero di cluster vari tra 2 e 15 si verifica quello che massimizza il dunn index ottenendo i seguenti risultati:

Rumore	N. Cluster	Dunn Index
0	2	inf
3	5	37900611.07446212
6	8	36916992.64750718,
12	14	38260272.08064264
24	3	1.1646336930145884
48	3	1.113308850758832
96	3	1.081215928860283
150	3	1.091815937563801
270	3	1.0533984377354868

300	3	0.693785309803108

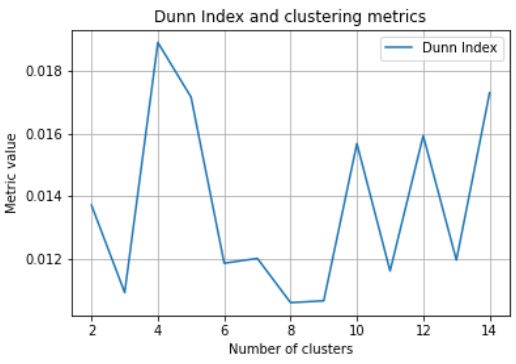


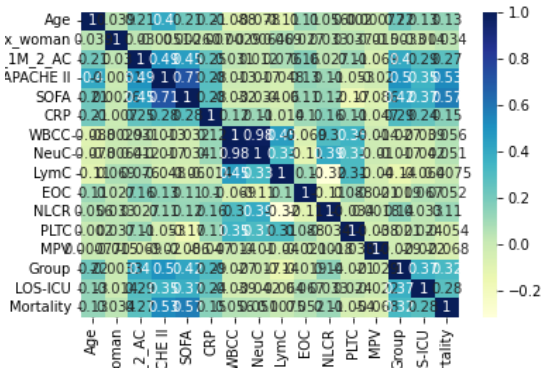
Si nota che la complessità del dataset cambia radicalmente il valore ottimo della metrica. In particolare, si ha un cambio molto evidente nel passaggio da 12 tuple di disturbo a 24.

6) Si è poi passati a test su dataset di complessità reale. I 5 dataset presi in esame sono relativi a dati pubblici di cartelle cliniche e differiscono in dimensioni e complessità.
Anche in questo caso, per ogni dataset, è stato selezionato il numero di cluster tra 2 e 15 che massimizzasse il dunn index.

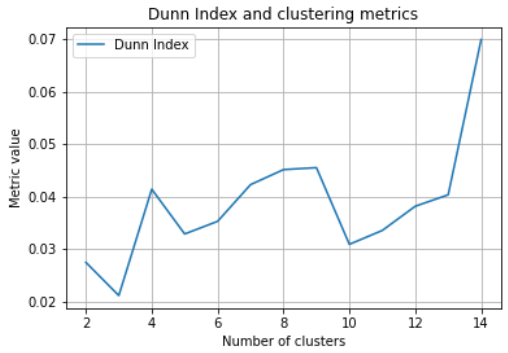
Il primo dataset di dimensioni 1257 × **16,** con le seguenti colonne: Age, Sex_woman, diagnosis_0EC_1M_2_AC, APACHE II, SOFA, CRP, WBCC, NeuC, LymC, EOC, NLCR, PLTC, MPV, Group, LOS-ICU, Mortality. La maggior parte di queste variabili sono binarie ().

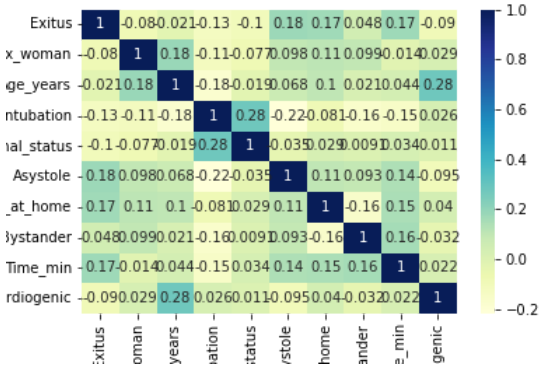
La heatmap mostra forte correlazioni solo tra 2 di queste 16 variabili. In questo caso, il numero di cluster che massimizzava il dunn index è 4 che corrisponde ad un dunn index pari a: 0.01889, ossia molto basso. Come si può vedere nella figura seguente l'andamento del valore dell'indice non è coerente al variare del numero di cluster.



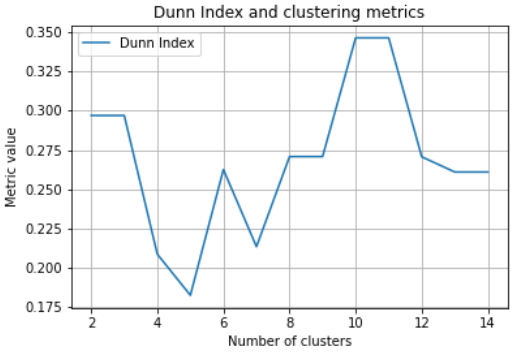


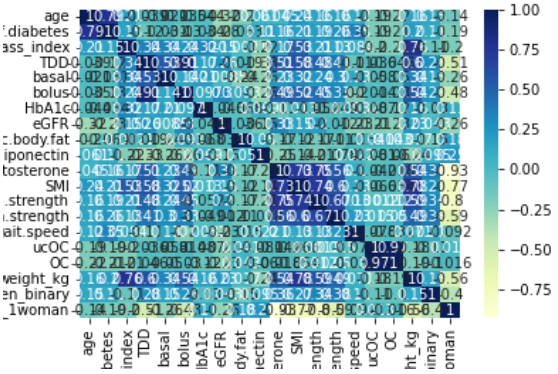
Il dataset successivo ha dimensioni pari a 422 × **10** contiene le seguenti colonne: Exitus, sex_woman, Age_years, Endotracheal_intubation, Asystole, Cardiac_arrest_at_home, Bystander, Time_min, Cardio genic.
Le dimensioni sono inferiori al precedente e correlazioni mai superiori a 0.28. In questo caso, con 14 cluster, si aveva un dunn index pari a 0.0698, leggermente più alto del primo caso, ma ancora molto molto vicino allo 0. In questo caso, il dunn index, globalmente, aumenta notevolmente all'aumentare del numero di cluster.





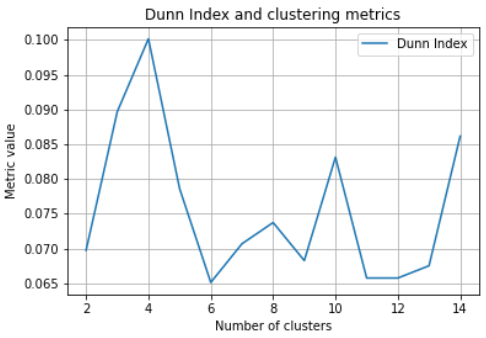
Il terzo dataset 67 × **20** riguarda dati clinici sul **diabete**: age, duration.of.diabetes, body_mass_index, TDD, basal, bolus, HbA1c, eGFR, perc.body.fat, adiponectin, free.testosterone, SMI, grip.strength, knee.extension.strength, gait.sp, ucOC, OC, weight, insulin_regimen_binary, sex_0man_1woman.
Le correlazioni di media basso valore, ma molto sparse restituisce il valore di dunn index più alto per 10 cluster: 0.346. In questo caso, il grafico mostra, similmente al primo, assenza di trend o comportamenti particolari al variare del numero di cluster.

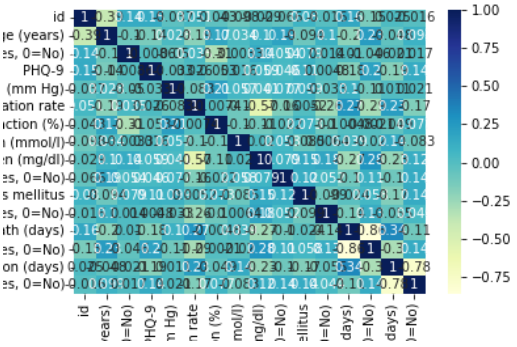




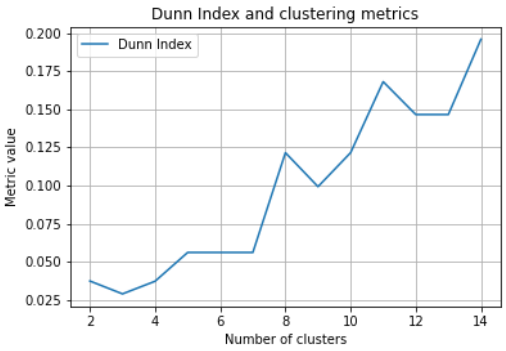
Il quarto dataset ha dimensioni simili al secondo 425 × **16**: id, Age, Male, PHQ-9, Systolic BP (mm Hg), Estimated glomerular filtration rate, Ejection fraction (%), Serum sodium (mmol/l), Blood urea nitrogen (mg/dl), Etiology HF(1=Yes, 0=No), Prior diabetes mellitus, Elevated level of BNP/NT-BNP (1=Yes, 0=No), Time from HF to Death (days), Death (1=Yes, 0=No), Time from HF to hospitalization (days), Hospitalized (1=Yes, 0=No),

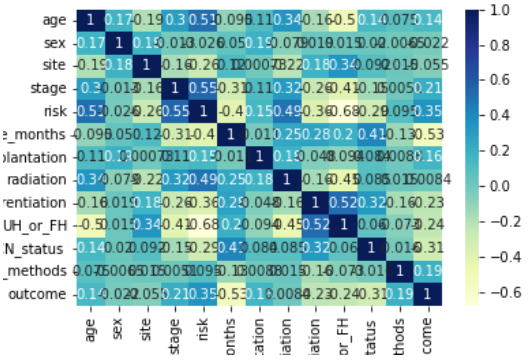
In questo caso, le correlazioni medio basse con tutte le variabili. In questo caso si ottengono 4 clusters con dunn index pari a 0.10. L'andamento in funzione del numero di cluster è fortemente instabile.





L'ultimo dataset ha dimensioni 173 x **13** : age, sex, site, stage, time_months, autologous_stem_cell_transplantation, radiation, degree_of_differentiation, UH_or_FH, MYCN_status, surgical_methods, outcome.
Le correlazioni sono basse. Il dunn index più alto si ha in corrispondenza di 14 cluster ed è pari a: 0.196. In questo caso, c'è un chiaro trend di miglioramento





La metrica del Dunn Index indica che il dataset sul quale il k-means funziona meglio è il terzo mentre quello su cui funziona peggio, il primo. Come suggerisce la letteratura, l'indice preferenzia dataset piccoli (il terzo dataset ha solo 67 osservazioni) mentre peggiora all'aumentare della loro grandezza (primo dataset con 1257 osservazioni).

7) per quanto riguarda possibile normalizzazione della matrica perché assuma valori tra 0 ed 1, si può certamente dire che il caso ideale, ossia il test 1 dei dati artificiali , fa riferimento al caso in cui i cluster siano iper compatti e quindi coincidano con un punto il cui diametro tende a 0. In questo caso, l'indice assume valore pari ad inf e non può rientrare nel range [0,1].