

Analisi di dati clinici relativi allo screening neonatale esteso e progettazione di metodologie di big data analysis e machine learning per la previsione delle malattie metaboliche nella popolazione pediatrica lombarda

Andrea Lucini Paioni - n° matricola 826578

Relatore: *Prof. Federico Cabitza*

Co-relatore: *Dott. Luca Marconi*

Anno accademico 2022/23



Il progetto Buzzi e lo screening neonatale esteso

Progetto nato all'interno del contesto **dell'ospedale Buzzi**, parte **dell'ASST** (Azienda Socio Sanitaria Territoriale) **Fatebenefratelli-Sacco**.

Obiettivo: analizzare uno dei più ampi e significativi database in Europa relativi allo **screening neonatale esteso**, comprendente i dati clinici e demografici di tutti i bambini nati in Regione Lombardia a partire da Giugno 2012 fino ad arrivare ad Aprile 2022.

Progetto per l'analisi dei dati per individuare correlazioni, patterns o signatures all'interno dei dati raccolti, e prevedere la comparsa di malattie metaboliche nella popolazione pediatrica lombarda.

Screening neonatale esteso: test per l'individuazione preventiva di patologie e condizioni cliniche nel neonato.


Screening attraverso il **prelievo di poche gocce di sangue** dal tallone del neonato, effettuato tra le **48 e le 72 ore** di vita; campioni analizzati nel Laboratorio specializzato per lo Screening, e risultati inviati all'Ospedale di nascita.



Fondazione Buzzi
PER L'OSPEDALE DEI BAMBINI



Le malattie metaboliche

- **Ipotiroidismo congenito (IC)**: condizione permanente, causata da fattori neonatali, con gravi **effetti** sul **sistema nervoso centrale**; incidenza di un caso ogni **3000** nati; necessità di **precoce terapia ormonale**.
 - **Fibrosi cistica (mucoviscidosi)**: **malattia ereditata** dai genitori, dovuta alla **mutazione** del **gene CTFR**; altera secrezioni di organi come polmoni e pancreas, causando danni gravi; necessità di intervenire con **farmaci dalla nascita** per limitarne i danni.
 - **Fenilchetonuria (PKU)**: raro **difetto metabolico**, porta a **gravi disturbi sul sistema nervoso**, fino a disabilità cognitive se non individuata e trattata; circa **50000** casi in tutto il mondo; per contrastarla, necessario rigoroso **regime alimentare** a **basso contenuto proteico**.
 - **Iperplasia surrenalica congenita**: gruppo di rare malattie genetiche, provocate da **insufficienza** di **cortisolo** o **aldosterone**; circa un caso ogni **16000** nati; trattamento **farmacologico** e **chirurgico** previsti.
 - **Atrofia Muscolare Spinale (SMA)**: rara **malattia neuromuscolare**, causata da perdita di motoneuroni; provoca **debolezza** e **atrofia muscolare** progressiva; circa un caso ogni **10000** nati.
- 

Il dataset

4 fasi di raccolta dei dati:

1. Raccolta ed “accettazione” (inserimento) dei dati nel sistema
2. Invio dei cartoncini al Buzzi
3. Analisi dei cartoncini al Buzzi
4. Analisi dei risultati e individuazione dei positivi



Dataset completo formato da **985792 records**, raccolti da circa **metà 2012** fino all'**Aprile 2022**, con un totale di **266 variabili**.

In questo progetto analizzato un **campione** del dataset completo (**30%** circa di osservazioni totali, ovvero **295738 records**), con **109** delle **266 features** originali così suddivise: 76 variabili quantitative (analiti, peso ed età gestazionale), 28 variabili qualitative, 3 variabili in formato data e 2 variabili necessarie per riconoscimento dei pazienti e dei campioni.

Metodologia e fasi del progetto

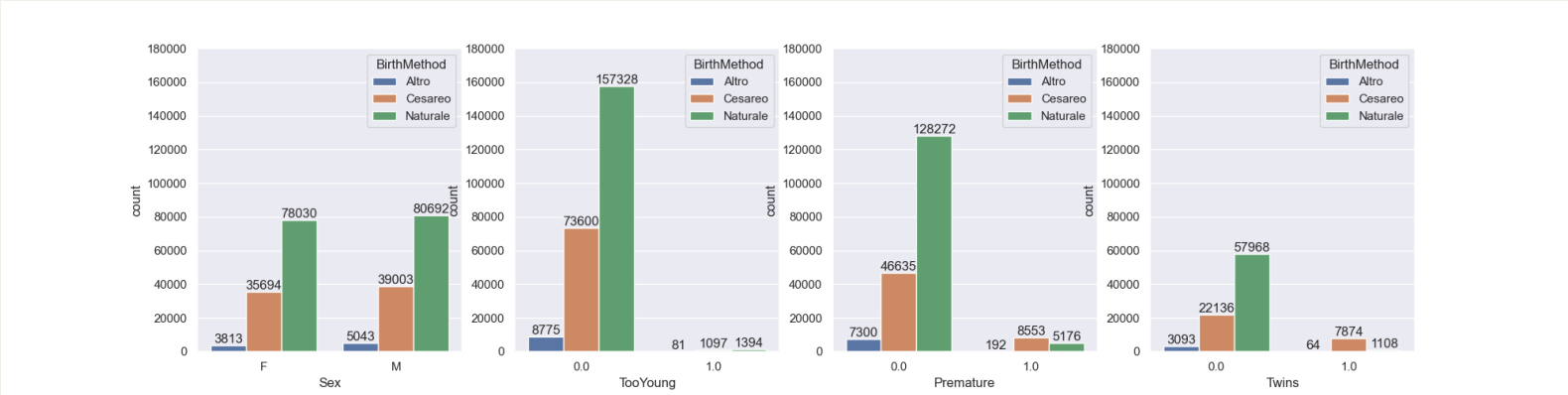
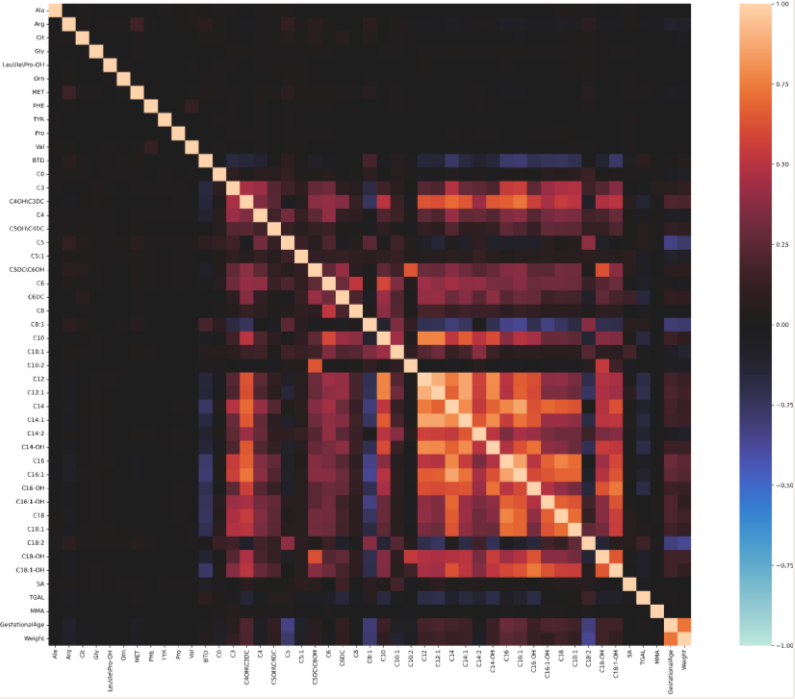
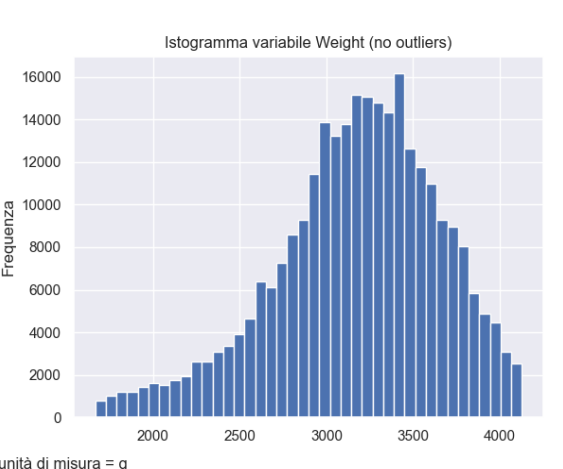
1. Fase di **import** e **pulizia dei dati**;
2. Fase di **analisi esplorativa**;
3. Fase di **analisi esplorativa stratificata** per la variabile “**Reparto**”;
4. Fase di applicazione delle tecniche di **riduzione della dimensionalità**;
5. Fase di applicazione delle tecniche di **cluster analysis**;
6. Fase di applicazione delle tecniche di **cluster analysis** ai dati ridotti con applicazione di tecniche di **riduzione della dimensionalità**;
7. Valutazione **risultati ottenuti**.



Pulizia dati ed analisi esplorativa

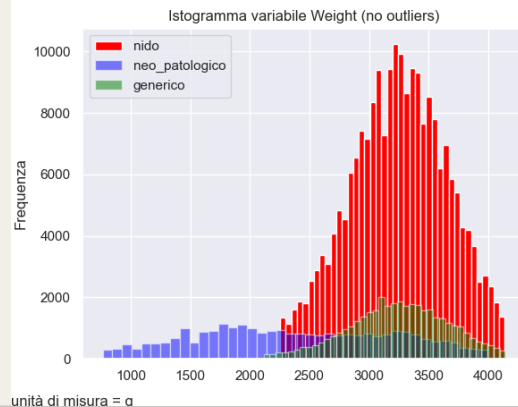
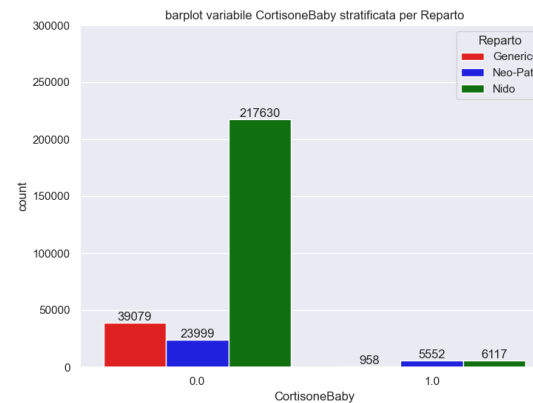
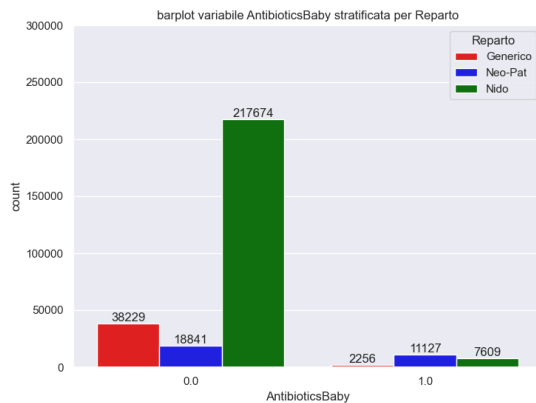
Tutti i passaggi sono svolti col linguaggio di programmazione **Python**, tramite utilizzo di metodi e librerie utili per l'analisi di dati, la creazione di infografiche e il calcolo di indici statistici.

	Unità misura	Count	% NaN	Min	2.5%	25%	50% or median	75%	97.5%	Max	Iqr	Skew	Kurt	Mean	Std	% outlier
ASATotal	μmol/L	20380	93%	0.03	0.1	0.18	0.23	0.32	0.72	2.77	0.14	4.06	28.51	0.27	0.17	5%
Ala	μmol/L	295057	0%	0.0	142.46	211.43	259.08	319.72	489.14	998529.9	108.29	89.44	10347.55	345.64	5143.8	5%
Allele 1		411	100%	2.0	9.0	31.0	31.0	55.0	194.0	224.0	24.0	3.11	10.91	46.0	37.21	5%
Arg	μmol/L	295057	0%	0.0	1.54	5.17	9.0	14.88	36.81	5352.47	9.71	96.99	11762.58	11.9	31.8	5%
Cit	μmol/L	295057	0%	0.0	6.04	10.93	14.03	17.92	29.67	6032.4	6.99	108.0	14840.54	15.42	33.2	5%
Glu	μmol/L	20380	3%	29.23	121.67	183.53	222.76	269.66	388.99	694.9	86.12	0.86	1.55	231.42	68.78	5%
Gly	μmol/L	295057	0%	0.0	170.03	304.83	388.03	483.54	756.47	926743.3	178.71	74.95	6167.81	520.44	7709.9	5%
Leu/Val/Pro-OH	μmol/L	295056	0%	0.0	87.89	122.58	146.71	177.66	264.72	348341.4	55.08	77.71	6521.68	193.58	2778.6	5%
Orn	μmol/L	295057	0%	0.0	50.34	80.19	102.62	133.2	236.93	401343.3	53.01	132.97	21080.83	129.72	1751.2	5%
MET	μmol/L	295056	0%	0.0	10.26	15.92	19.65	24.01	36.64	9288.54	8.09	104.64	13836.15	21.2	46.26	5%
PHE	μmol/L	295060	0%	0.12	33.35	45.58	53.1	61.86	86.53	197180.6	16.28	229.83	56721.54	59.11	640.63	5%
TYR	μmol/L	295060	0%	0.0	44.64	70.54	89.07	113.41	192.97	352349.8	42.87	159.84	29517.36	109.64	1537.1	5%
HCYS	μM	1002	100%	0.0	1.03	2.1	2.76	3.79	7.03	76.3	1.69	17.77	448.4	3.17	2.83	5%
Pro	μmol/L	295057	0%	0.0	108.64	150.31	176.92	209.34	300.74	527548.2	59.03	76.82	6960.62	238.57	3575.6	5%
Val	μmol/L	295057	0%	0.0	74.15	109.1	131.88	159.41	233.29	760742.1	50.31	118.55	19647.44	178.69	3217.1	5%
BTD	U/dl	96058	68%	11.01	130.51	213.75	262.77	303.03	355.97	486.7	89.28	-0.37	-0.38	256.31	60.75	5%
C0	μmol/L	295057	0%	0.0	8.18	13.94	18.62	24.62	42.44	7978.04	10.68	85.04	10328.36	20.87	39.29	5%
C3	μmol/L	295057	0%	0.0	0.43	1.11	1.69	2.34	4.27	77.52	1.23	8.47	395.37	1.84	1.12	5%
C4OH/C3D C	μmol/L	295057	0%	0.0	0.04	0.1	0.16	0.22	0.39	3.63	0.12	1.37	11.05	0.17	0.09	5%



Analisi esplorativa stratificata per reparto

Ripetizione di molti dei passaggi della prima fase di analisi esplorativa sul dataset **stratificato** per la variabile **Reparto** (indica il reparto in cui è ricoverato il neonato al momento dello screening, ovvero ‘**generico**’, ‘**neo-patologico**’ o ‘**nido**’).



		Grouped by Reparto				
		Missing	Overall	Generico	Neo-Pat	Nido
Twins	0.0	155920	129257 (92.4)	32438 (97.5)	6572 (62.8)	90247 (93.9)
	1.0		10560 (7.6)	837 (2.5)	3888 (37.2)	5835 (6.1)
TooYoung	0.0	0	292480 (98.9)	40005 (98.8)	28781 (96.0)	223694 (99.3)
	1.0		3257 (1.1)	480 (1.2)	1187 (4.0)	1590 (0.7)
TPNCARNFeed	0	0	295364 (99.9)	40485 (100.0)	29599 (98.8)	225280 (100.0)
	1		373 (0.1)	///	369 (1.2)	4 (0.0)
SampleQuality	OK	295733	4 (100.0)	1 (100.0)	///	3 (100.0)
BabyFed	0.0	172326	759 (0.6)	17 (0.1)	504 (4.5)	238 (0.3)
	1.0		122652 (99.4)	28395 (99.9)	10664 (95.5)	83593 (99.7)
AntibioticsMother	0.0	201061	78482 (82.9)	2553 (85.8)	8482 (76.8)	67447 (83.6)
	1.0		16194 (17.1)	423 (14.2)	2556 (23.2)	13215 (16.4)
TPNFeed	0	0	292390 (98.9)	40385 (99.8)	26812 (89.5)	225193 (100.0)
	1		3347 (1.1)	100 (0.2)	3156 (10.5)	91 (0.0)
MIXFeed	0	0	232808 (78.7)	30688 (75.8)	18363 (61.3)	183757 (81.6)
	1		62929 (21.3)	9797 (24.2)	11605 (38.7)	41527 (18.4)
CortisoneMother	0.0	200947	85832 (90.5)	2825 (94.6)	7342 (66.3)	75665 (93.7)
	1.0		8958 (9.5)	160 (5.4)	3729 (33.7)	5069 (6.3)

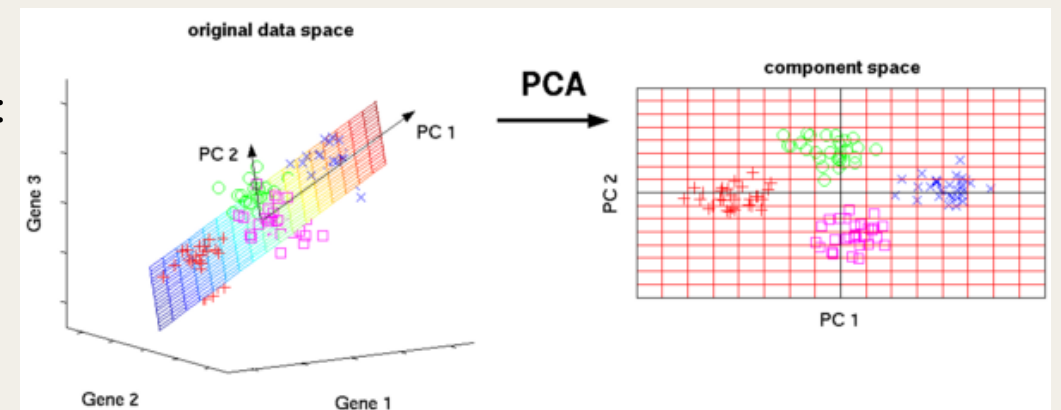
Riduzione di dimensionalità

Metodi di machine learning nati con l'obiettivo di **ridurre il numero di variabili** in un dataset di interesse, con **minima perdita di informazione** (variabilità).

Applicata in diversi ambiti (text mining, finanza, elaborazione di immagini, biostatistica), e spesso usata come **operazione preliminare** ad altre tecniche quali **cluster analysis**, **modelli di classificazione e modelli di regressione**.

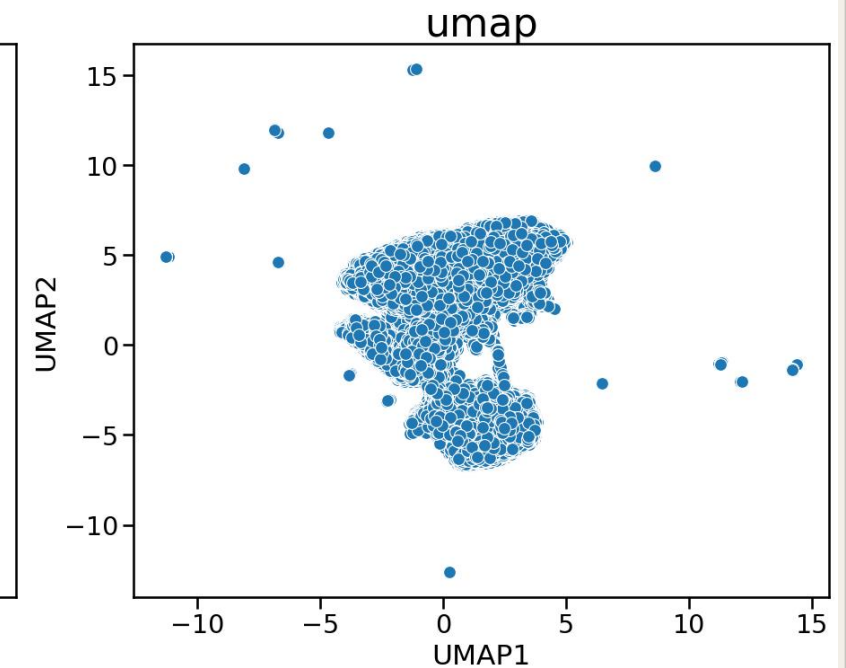
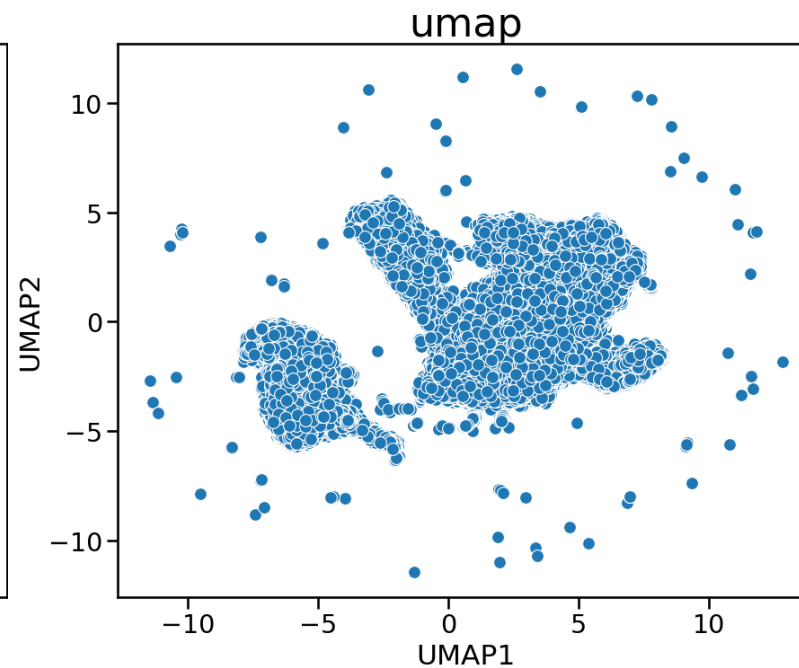
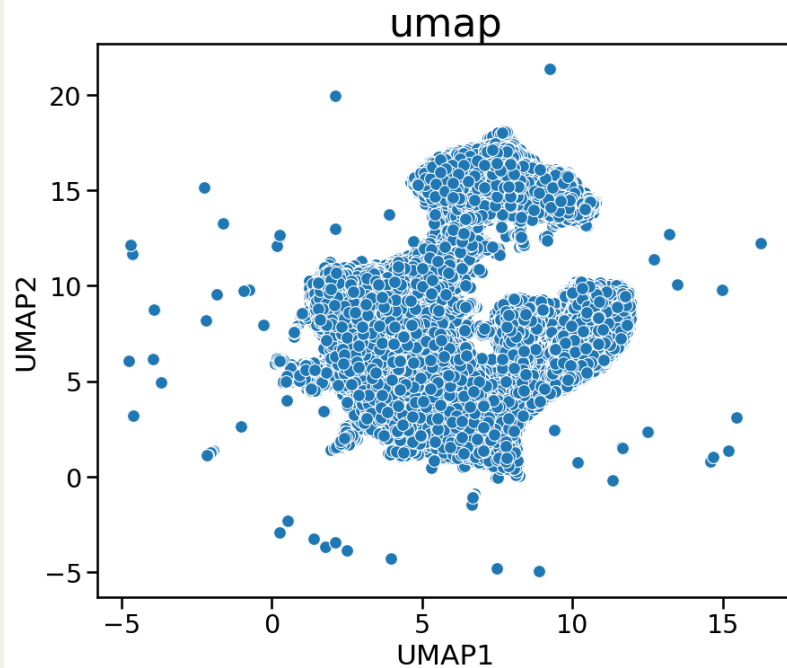
Fondamentali i valori dei **parametri** in ogni metodo applicato (scelta dei valori sempre delicata, si tratta di metodi di apprendimento non supervisionato, senza particolari metriche di valutazione della bontà dei risultati ottenuti).

Tra le principali tecniche applicate in questo progetto: **PCA**, **t-SNE** ed **UMAP**.



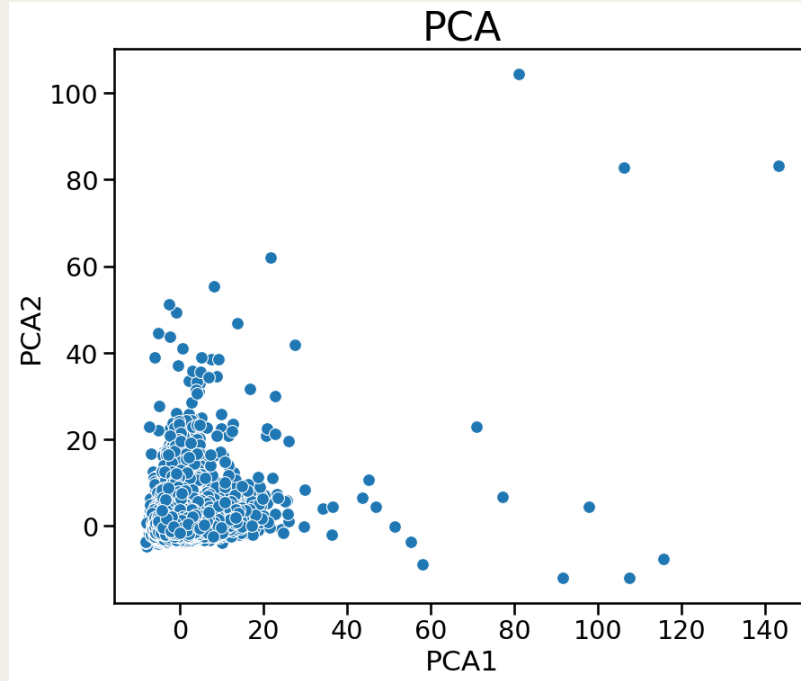
UMAP

Parametri fondamentali: **n_neighbors** (quanto UMAP mantiene della struttura locale), **min_dist** (quanto distano i punti in rappresentazioni a basse dimensioni), **n_components** (dimensionalità dello spazio risultante dall'algoritmo) e **metrics** (come viene calcolata la distanza tra punti).



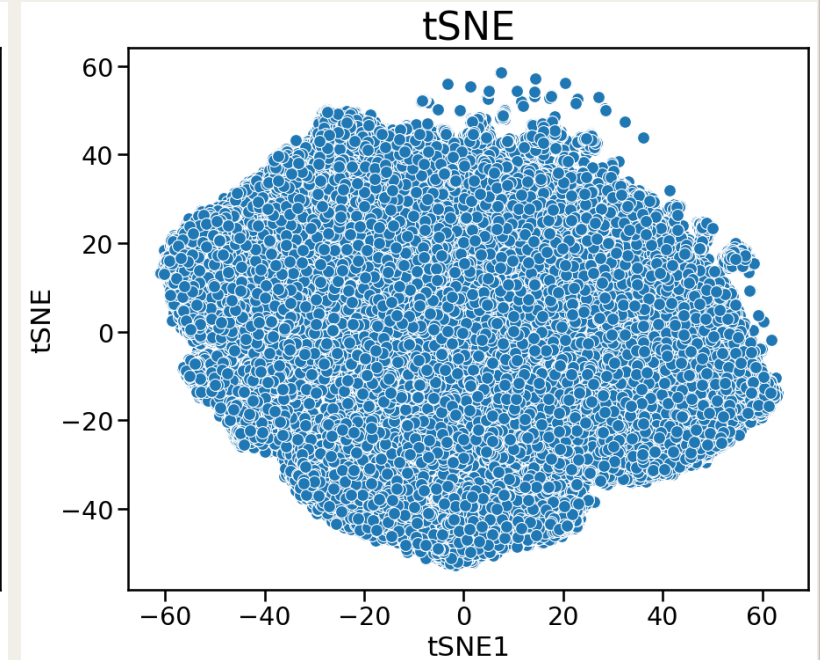
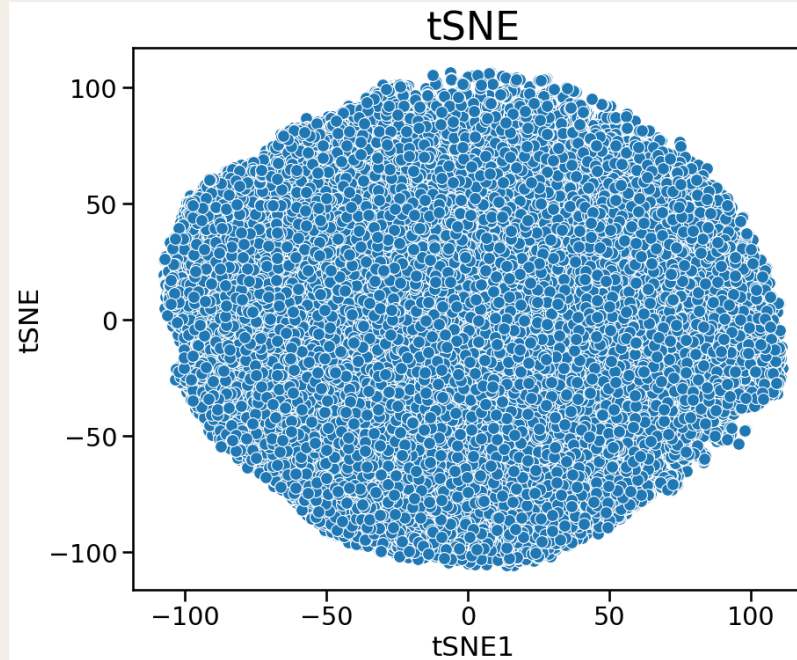
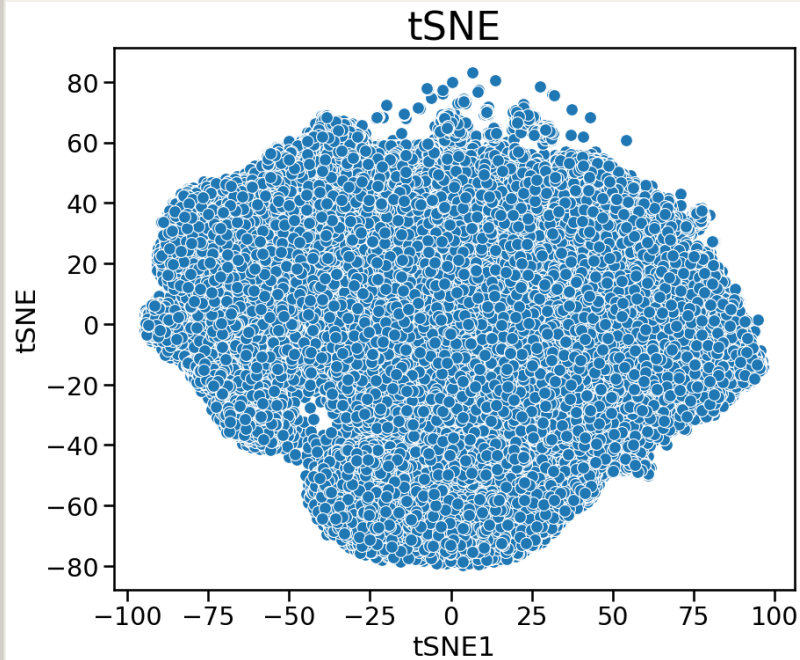
PCA

Diverse implementazioni con valori differenti per i parametri **whiten**, **svd_solver**, **tol**, **iterated_power**, **n_oversamples** e **power_iteration_normalizer**, senza particolari variazioni nei risultati.



t-SNE

Parametri fondamentali: **perplexity** (equilibrio tra conservazione di strutture globali e locali), **early_exageration** (quanto i campioni nei clusters vengono riprodotti vicini nello spazio a dimensionalità ridotta) ed **n_components** (dimensionalità dello spazio risultante dall'algoritmo).



Altri metodi non utilizzati

Altre tecniche di riduzione della dimensionalità escluse dai risultati finali:

- **Kernel PCA**: non implementata poiché prevedeva la creazione di uno spazio a dimensionalità superiore rispetto a quello in input (con creazione di una matrice di 600GB);
- **ICA (Independent Component Analysis)**, **SVD (Singular Value Decomposition)** e **NMF (Non-Negative Matrix Factorization)**: impossibili da eseguire richiedendo la decomposizione della matrice di dati in input (dimensioni vicine ai 300GB).

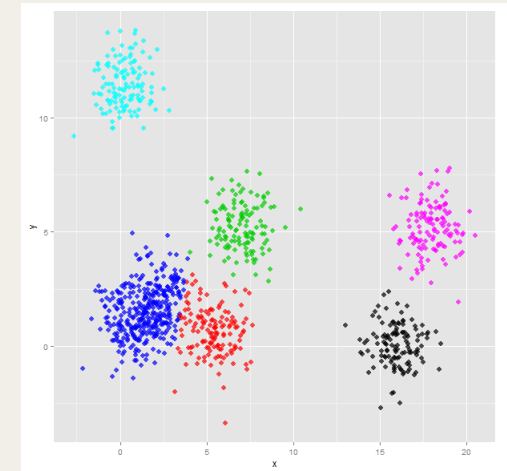
Cluster analysis

Metodi di machine learning nati con l'obiettivo di trovare **gruppi di osservazioni internamente molto compatti** e con **caratteristiche simili**, e **molto differenti** gli uni dagli altri.

Applicata in diversi ambiti (marketing, finanza, elaborazione di video ed immagini, text mining, biostatistica), e spesso legata ad altre metodologie, come tecniche di riduzione della dimensionalità.

Fondamentale porre particolare attenzione ai valori dei **parametri**, con molti metodi parecchio sensibili al variare degli stessi. L'efficacia dei risultati ottenuti può essere valutata sulla base di alcune metriche.


Tra le principali tecniche applicate in questo progetto: **K-means, clustering gerarchico agglomerativo, BIRCH, DBSCAN e spectral clustering**.





Metodi di valutazione cluster analysis

Sono stati usati una serie di **indici** per **valutare le performances** dei **metodi** di **cluster analysis**:

- **Il coefficiente di Silhouette** $(-1, +1)$: misura la qualità dell'assegnazione dei punti ai corretti clusters in un dataset; si ricercano valori vicini a $+1$;
 - **L'indice di Calinski-Harabasz** $(0, +\infty)$: rapporto tra la somma della dispersione tra i clusters e la dispersione all'interno dei clusters per tutti i gruppi creati; si ricercano valori elevati, tendenti all'infinito;
 - **L'indice di Davies-Bouldin** $(0, +\infty)$: rappresenta la "similitudine" media tra i clusters; si ricercano valori tendenti a 0 ;
 - **L'indice di Dunn** $(0, +\infty)$: rapporto tra minore distanza tra i due centroidi di una distribuzione e la massima distanza di due punti appartenenti a due clusters differenti; si ricercano valori elevati, tendenti all'infinito;
 - **L'entropy** $(0, \log_2(k))$: probabilità di associazione corretta di ogni elemento ai corretti clusters; utile per valutare la qualità e le caratteristiche del clustering ottenuto.
- 



Clustering gerarchico agglomerativo


Parametri fondamentali: **linkage** (criterio di collegamento tra punti), **n_clusters** (numero di clusters ricercati) e **metric** (metrica utilizzata per calcolare la distanza tra punti).

```
model_aggcl_2cl_ward_eucl = AgglomerativeClustering(n_clusters=2,  
linkage='ward', metric='euclidean')
```

Ottenuti **clusters con dimensioni significative** col **metodo di Ward** (parametro *linkage*), mentre un unico **cluster** comprendente la **maggior parte delle osservazioni** per **altri metodi di linkage**, a prescindere dai valori degli altri parametri.

Nei casi con **metriche diverse dall'euclidea**, risultati **simili ai metodi** del **legame singolo, medio e completo**, con piccoli clusters e un gruppo comprendente la maggioranza delle osservazioni.

Valori delle **metriche non molto elevati**, sinonimo di risultati del clustering non molto soddisfacenti. **Tempi** di esecuzione **non brevi**.






BIRCH

Parametri fondamentali: **threshold** (soglia sotto la quale avviene o meno l'unione di un punto ed un sottocampione), **n_clusters** (numero di clusters ricercati) e **branching_factor** (numero massimo di CFT sottoclusters in ogni nodo).

```
model_birch_2cl_05thr_50bra = Birch(threshold=0.5, branching_factor=50,  
n_clusters=2)
```

Ottenuti **clusters di dimensioni significative**, con grandi **differenze** al variare del **numero di clusters**, ma risultati abbastanza costanti modificando gli altri parametri.

Risultati delle **metriche con valori non molto elevati**, in linea con quelli ottenuti col clustering gerarchico agglomerativo, sinonimo di risultati del clustering non molto soddisfacenti. **Tempi** di esecuzione **lunghi**.






DBSCAN

Parametri fondamentali: **eps** (massima distanza tra due campioni considerati vicini), **min_samples** (numero di campioni nel vicinato di un punto perché questo possa essere “core point”) e **metric** (metrica utilizzata per calcolare la distanza tra punti).

```
model_dbscan_eps100_5mins = DBSCAN(eps=100, min_samples=5,  
metric='euclidean')
```

Risultati non soddisfacenti, il metodo individua un solo cluster formato dalla maggioranza delle osservazioni, ed associa quelle rimanenti a ‘rumore’.

Valori di alcune **metriche elevati** (come l’indice di Silhouette), altre **in linea** con quelle ottenute coi **metodi** di cluster analysis **precedenti**. **Tempi** di esecuzione **lunghi**.



K-means

Parametri fondamentali: **n_init** (numero di volte che viene rieseguito l'algoritmo), **n_clusters** (numero di clusters ricercati), **max_iter** (numero massimo di iterazioni per arrivare alla convergenza) e **tol** (tolleranza relativa alla norma di Frobenius della differenza tra due centroidi).

```
model_kmeans_2cl_100in_1000maxi_001tol = KMeans(n_clusters=2, n_init = 100, algorithm='lloyd', max_iter = 1000, tol = 0.01)
```

Risultati molto significativi, con **clusters di dimensioni significative**, in tutti i casi analizzati e con le diverse implementazioni dei parametri.

Valori delle **metriche non molto elevati**, ma tra i migliori ottenuti considerando tutti i metodi di cluster analysis. **Tempi** di esecuzione **molto brevi**, che permettono di esplorare il metodo a fondo e di eseguire più implementazioni in base ai valori dei parametri scelti.

Spectral clustering

Esecuzione dell'algoritmo **molto complessa**, sia per le scelte dei **numerosi parametri** a disposizione, sia perché **molto dispendioso computazionalmente** (ricostruisce uno spazio di dimensioni superiore allo spazio di input).

Anche con un campione ridotto di sole 5000 osservazioni, necessari **tempi di esecuzione superiori alle tre ore**, rispetto ad un massimo di 30 minuti per tutte le altre tecniche di cluster analysis.





Cluster analysis su dati ridotti

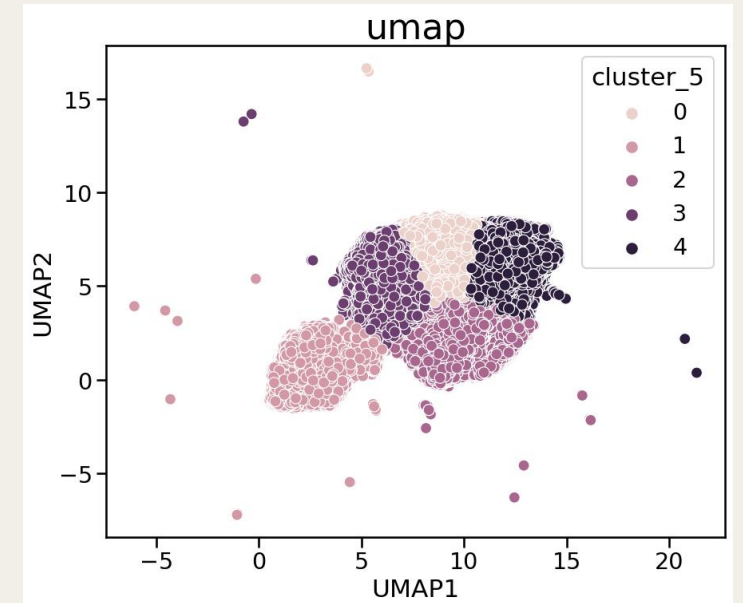
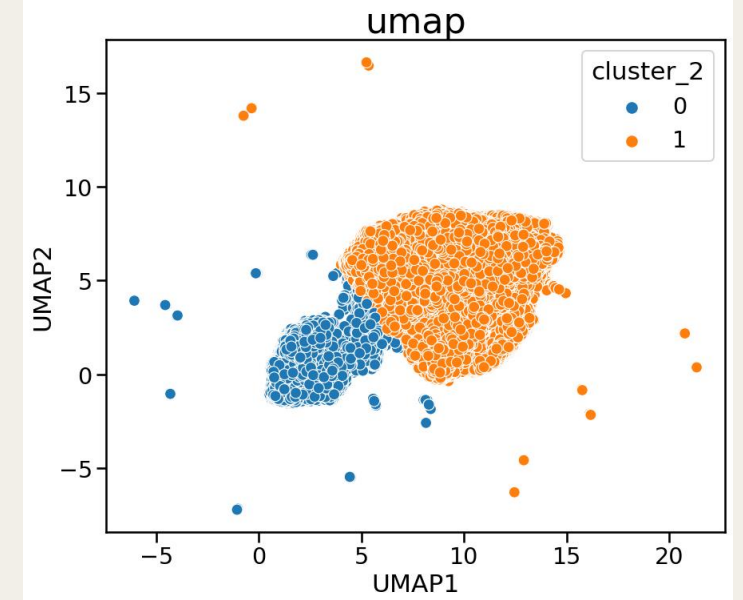
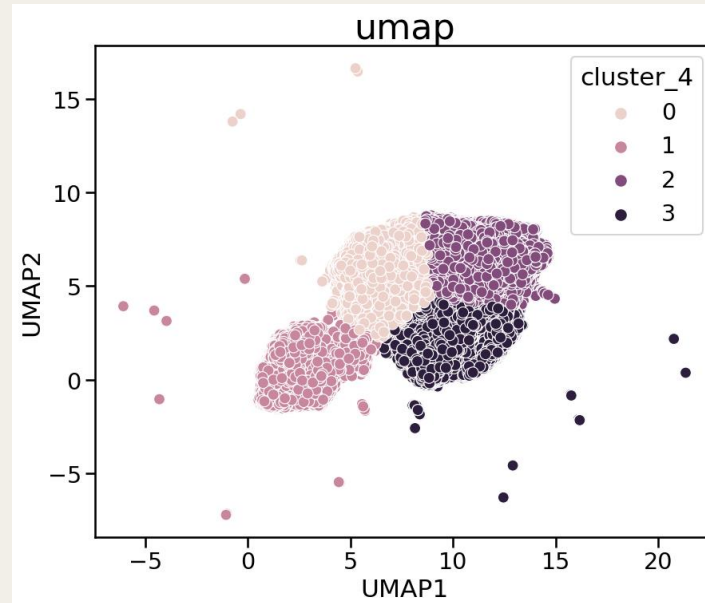
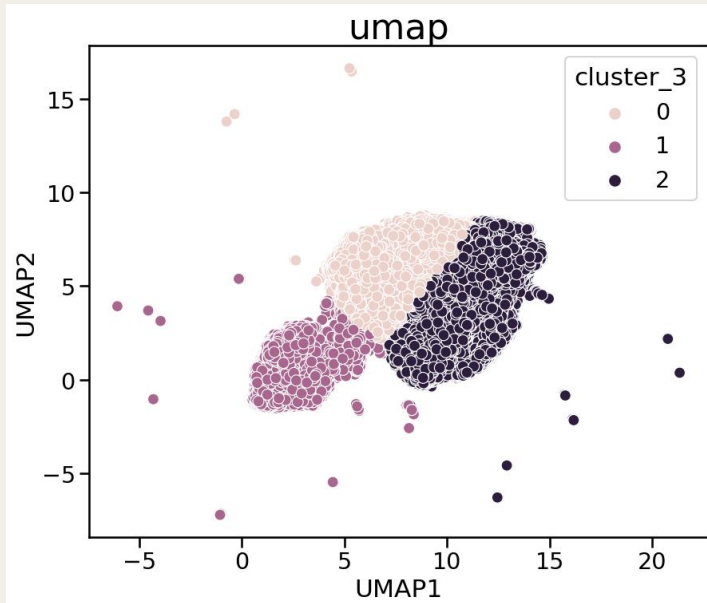
Obiettivo **riduzione della dimensionalità**: creare un nuovo dataset di input ridotto per i dati che mantenesse la capacità comunicativa (varianza) presente nei dati; obiettivo **cluster analysis**: creazione di clusters densi internamente e ben separati dagli altri esternamente.

Le due metodologie possono **lavorare in combinazione** per arrivare a risultati significativi, sfruttando a vicenda le caratteristiche e i punti di forza dell'altra: le informazioni ottenute dalle fasi precedenti del progetto sono utili per concentrarsi su metodi veloci, efficaci, altamente informativi e computazionalmente fattibili con i dispositivi a disposizione.

Effettuando riduzione di dimensionalità sui dati originali, la **cluster analysis** viene effettuata **sull'intero campione** a disposizione, non limitandosi più al campione di 50000 osservazioni.

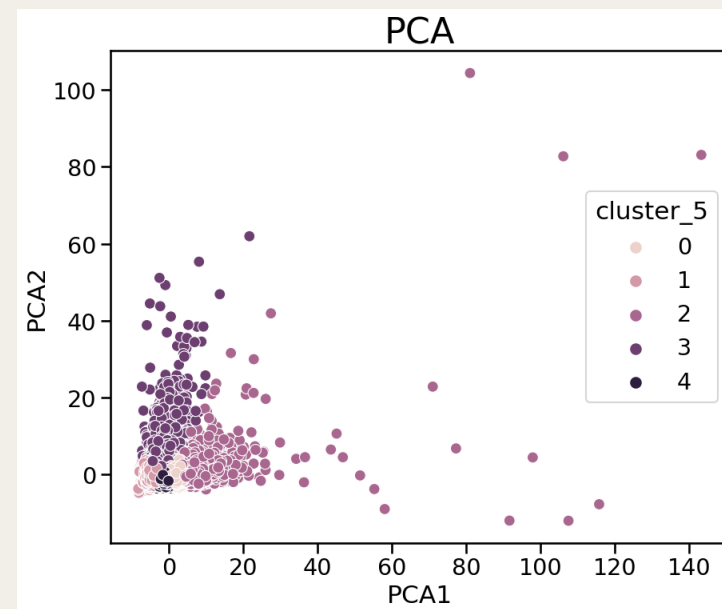
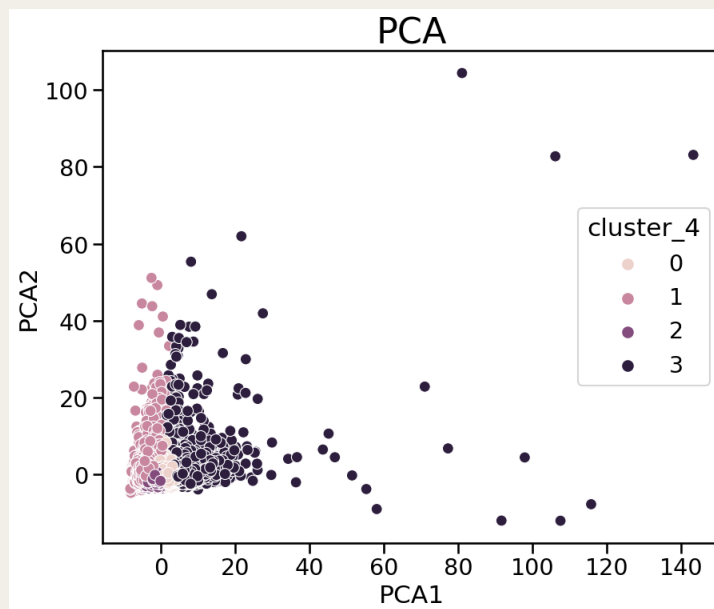
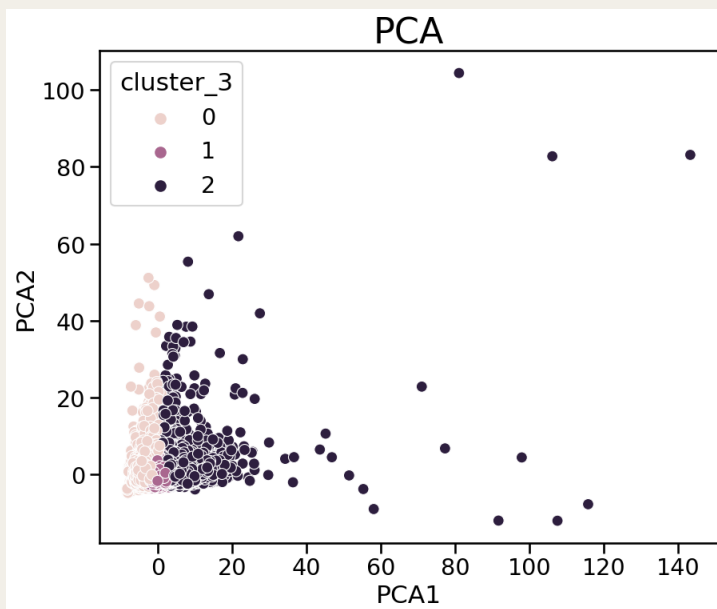
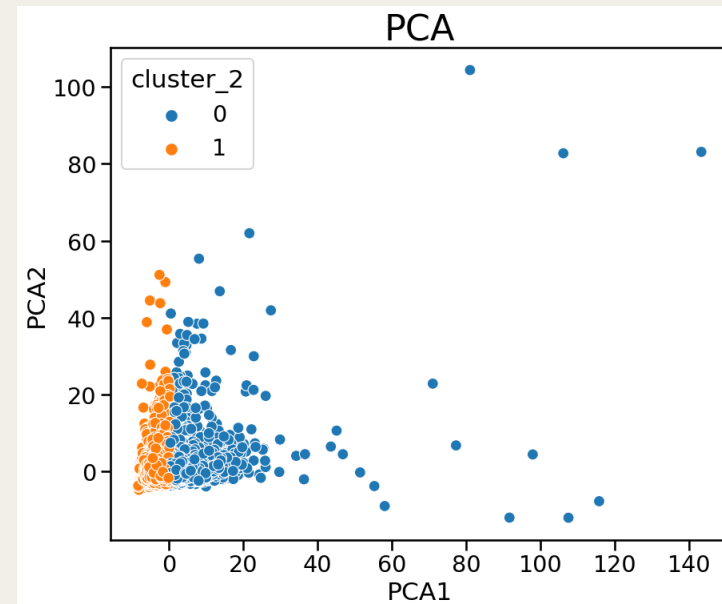
K-means dati ridotti UMAP

Cluster analysis col metodo **K-means** ($n_clusters = 2/3/4/5$) su dati ridotti con **UMAP** ($n_neighbors = 15$, $min_dist = 0.1$, $n_components = 2$, $metric = 'euclidean'$)



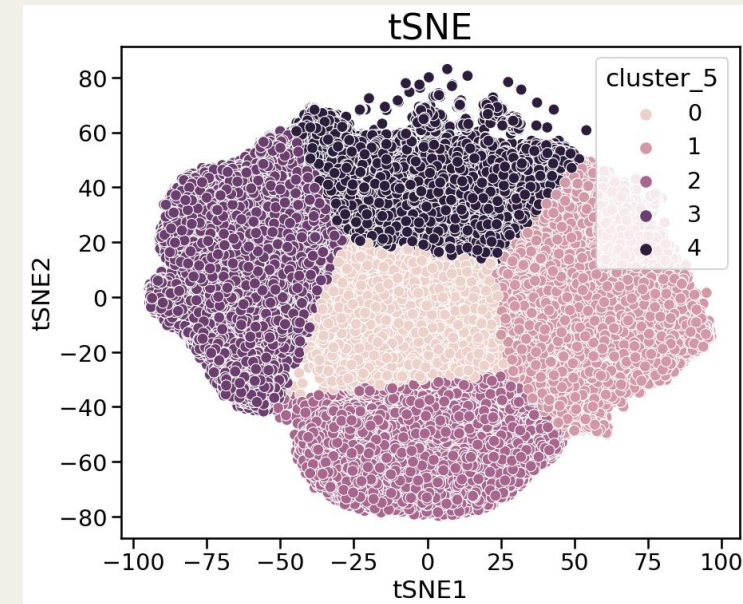
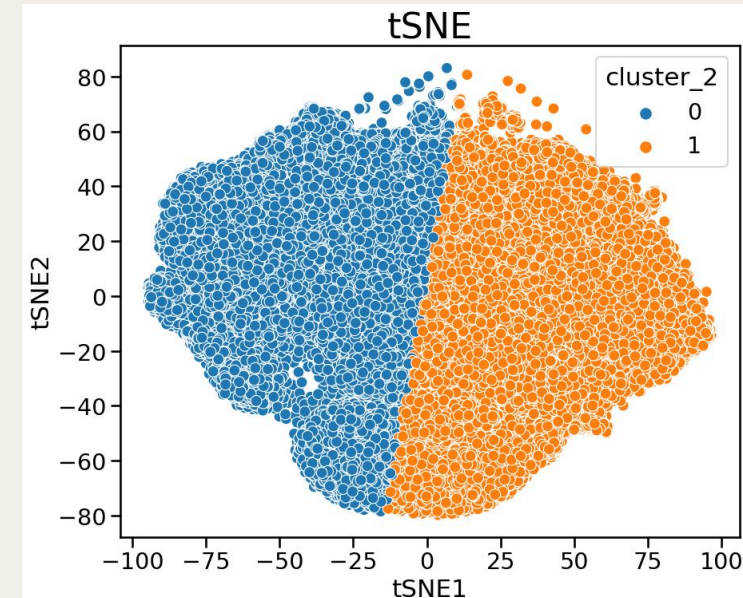
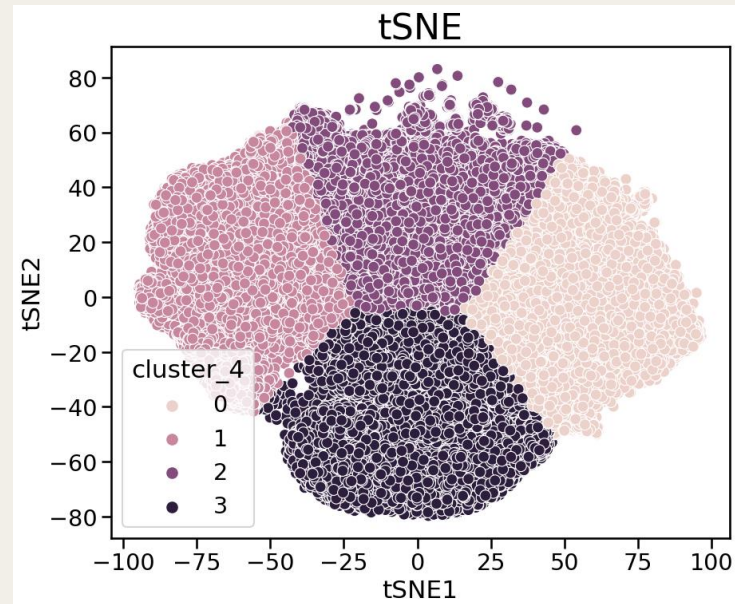
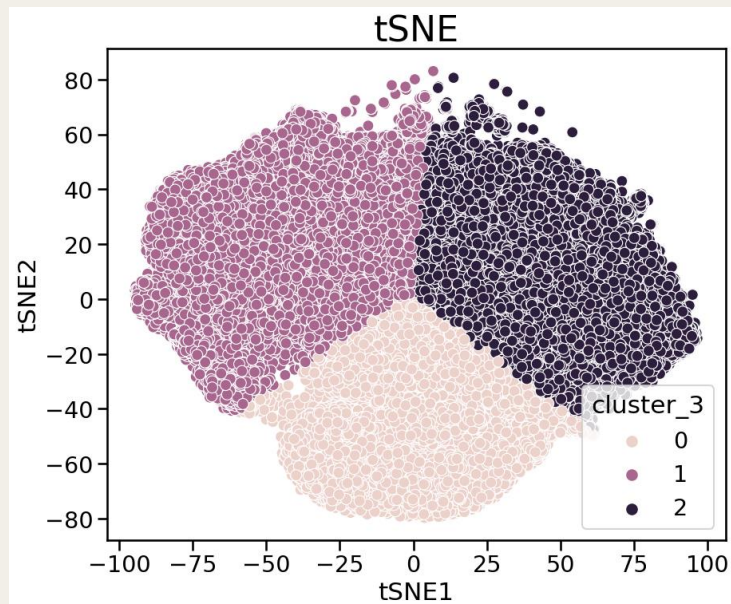
K-means dati ridotti PCA

Cluster analysis col metodo **K-means**(*n_clusters* = 2/3/4/5) su dati ridotti con **PCA()**



K-means dati ridotti t-SNE

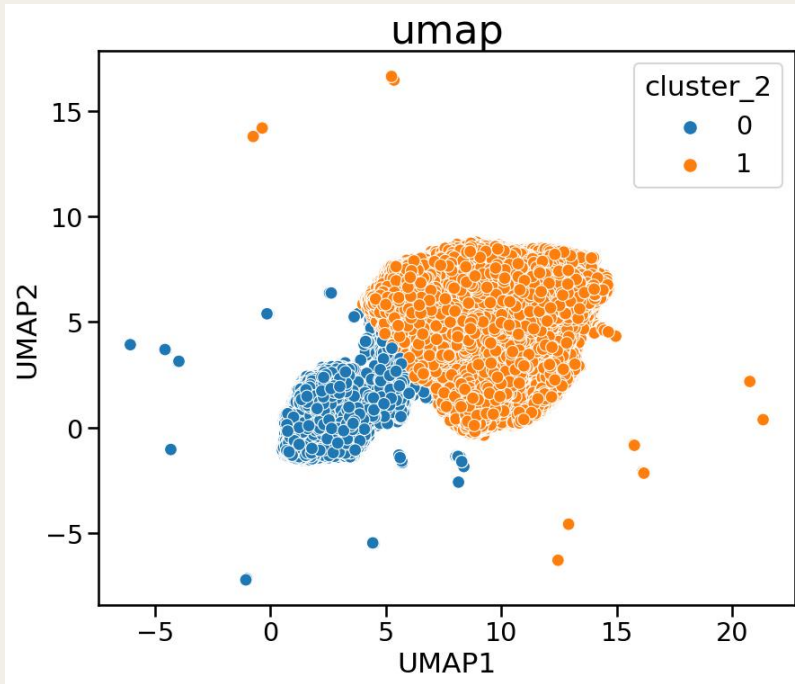
Cluster analysis col metodo **K-means** ($n_clusters = 2/3/4/5$) su dati ridotti con **t-SNE** ($perplexity = 50$, $early_exaggeration = 30$)



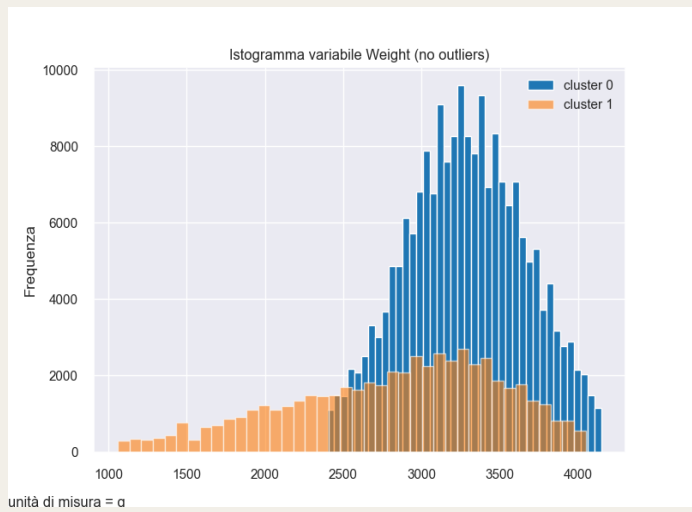
Miglior risultato cluster analysis su dati ridotti

Risultato migliore: **cluster analysis** con metodo **K-means**(*n_clusters*=2) su dati ridotti con **UMAP**(*n_neighbors*=15, *min_dist*=0.1, *n_components*=2, *metric*='euclidean').

Buoni risultati per tutti i **parametri** considerati della cluster analysis, e **UMAP** tra i metodi di **riduzione della dimensionalità maggiormente informativi** in ambito biomedico.



	Silhouette coefficient [-1,1]	Calinski-Harabasz index [0, +∞]	Davies-Bouldin index [0, +∞]	Dunn index [0, +∞]	Entropy [0, log2(K)]
K-means(2 clusters)	0.543	331062.561	0.571	0.241	12.273



		Grouped by cluster			
		Missing	Overall	0	1
AntibioticsBaby, n (%)	0.0	0	253444 (93.1)	54238 (93.2)	199206 (93.1)
	1.0		18670 (6.9)	3945 (6.8)	14725 (6.9)
ARTFeed, n (%)	0	0	246641 (90.6)	52753 (90.7)	193888 (90.6)
	1		25473 (9.4)	5430 (9.3)	20043 (9.4)
HUFeed, n (%)	0	0	92220 (33.9)	19717 (33.9)	72503 (33.9)
	1		179894 (66.1)	38466 (66.1)	141428 (66.1)
MIXFeed, n (%)	0	0	214641 (78.9)	45817 (78.7)	168824 (78.9)
	1		57473 (21.1)	12366 (21.3)	45107 (21.1)




Conclusioni e spunti di ricerca futuri

Riduzione di dimensionalità: migliori risultati con UMAP, strutture dei dati meno definite con t-SNE e PCA con poca possibilità di ottimizzare modificando i valori dei parametri.

Cluster analysis: ottimi risultati di K-means, altri metodi poco informativi e più dispendiosi computazionalmente (soprattutto DBSCAN e BIRCH).

Sviluppi futuri:

- **Verifiche** di altri metodi di **cluster analysis** (tecniche di fuzzy clustering, Mean Shift, Affinity Propagation...), ed altre tecniche di **riduzione della dimensionalità** (SOM, LDA, FA...).
 - **Ulteriori esplorazioni** relative a **variabili qualitative** di particolare interesse (Etnia, complicazioni nel parto...)
 - **Aggiunta** di dati relativi all'**insorgenza** di **malattie metaboliche**, da confrontare con i risultati delle tecniche di cluster analysis.
- 



Grazie per l'attenzione

