Analisi dei Cluster: Scelta ottimale di k e Interpretazione

Team di Data Science

May 26, 2025

1 Introduzione

In questo report presentiamo i risultati dell'analisi di clustering su un insieme di dati di risposte, con l'obiettivo di determinare il numero ottimale di cluster k e di interpretare le strutture emerse.

2 Metodi di Selezione di k

2.1 Silhouette Score

Abbiamo calcolato lo silhouette score medio per valori di k compresi tra 2 e 6: Il valore

k	2	3	4	5	6
Silhouette	0.62	**0.63**	0.52	0.45	0.40

Table 1: Silhouette score medio per ciascun numero di cluster.

massimo di silhouette score si ottiene per $k = 3$, suggerendo tre gruppi ben separati e compatti.

2.2 Elbow Method

La curva dell'inerzia (somma delle distanze quadratiche intra-cluster) mostra una diminuzione netta iniziale e un appiattimento a partire da $k = 3$: La riduzione di inerzia da $k = 2$ a $k = 3$ è significativa, mentre per numeri di cluster maggiori il guadagno in compattezza è minore.

3 Visualizzazioni 2D

3.1 PCA 2D

La proiezione PCA bidimensionale evidenzia i tre cluster:

- Cluster 1 (arancione): gruppo molto compatto in alto a sinistra.

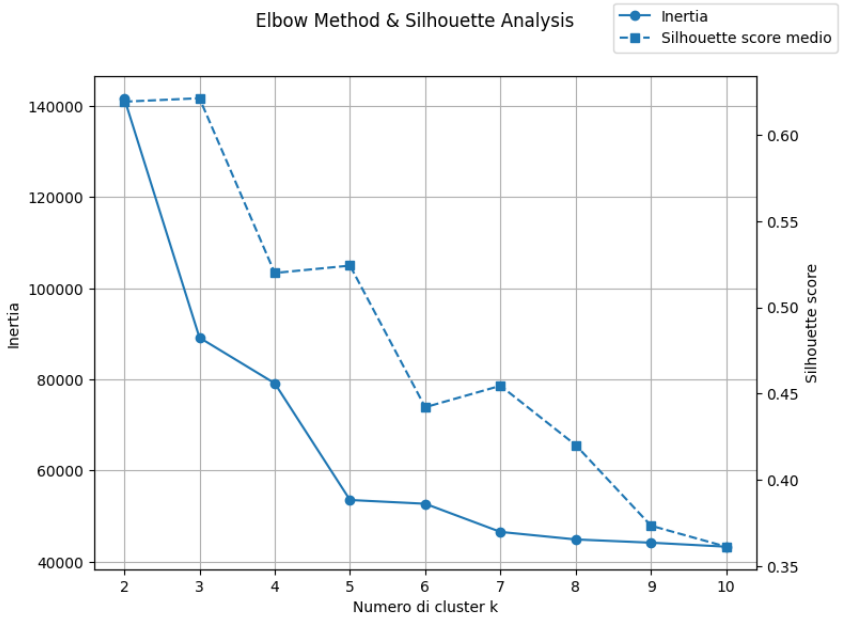

Figure 1: Elbow plot: inerzia in funzione di k.

Figura non trovata: pca2d.png

Figure 2: Distribuzione dei cluster (PCA 2D).

- Cluster 2 (verde): compatto nella parte bassa con forma a "U".
- Cluster 0 (blu): il più numeroso (60% dei punti), sparso e a forma di mezzaluna a destra.

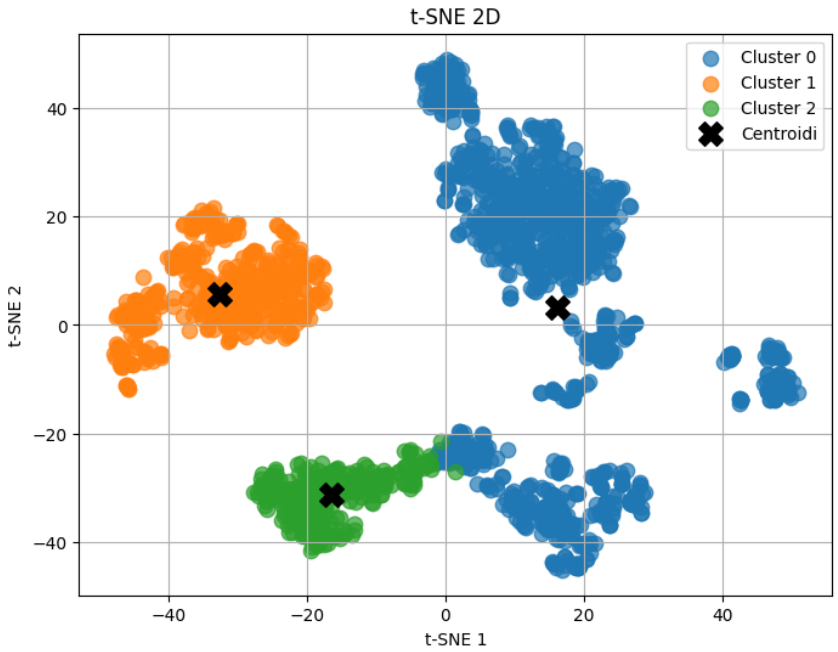

Figure 3: Distribuzione dei cluster (t-SNE 2D).

4 Distribuzione dei Cluster

La tabella seguente riassume la dimensione e la percentuale di ciascun cluster sul totale dei campioni: Il Cluster 0 è dominante e più eterogeneo; i Cluster 1 e 2 sono più piccoli

Cluster	N. elementi	% sul totale
0	1060	59.79%
1	453	25.55%
2	260	14.66%

Table 2: Dimensione e percentuale di ogni cluster.

e densi.

5 Conclusioni e Proposte

La scelta di $k = 3$ è giustificata dal massimo silhouette score e dall'appiattimento dell'Elbow plot. Tuttavia, il grande Cluster 0 potrebbe contenere sottogruppi di interesse:

- Applicare un secondo livello di clustering su Cluster 0 per distinguere sfumature interne.
- Valutare altri indici (Davies–Bouldin, Calinski–Harabasz) per una conferma alternativa.
- Se l'obiettivo è una netta distinzione tra risposte "lecite" e "jailbreak", considerare $k = 2$, con $k = 3$ riservato all'analisi di eccezioni.
- Validare le assegnazioni su un sottoinsieme etichettato manualmente per misurare precision/recall.