**CLUSTERING- algoritmo KMEANS**

Il clustering è un algoritmo di [apprendimento automatico](https://www.ibm.com/it-it/think/topics/machine-learning) non supervisionato che organizza e classifica oggetti, punti dati oppure osservazioni differenti in gruppi o cluster in base a somiglianze o modelli. A differenza dell'apprendimento supervisionato, in cui il modello apprende da dati etichettati, il clustering non utilizza classi predefinite: l'algoritmo analizza i dati e li raggruppa autonomamente in base a caratteristiche condivise.

L’**algoritmo K-Means** è una tecnica di **clustering non supervisionato** che suddivide un insieme di dati in **K gruppi** sulla base della somiglianza tra i punti. L'obiettivo è minimizzare la distanza tra i punti e il loro **centroide**, ossia il punto medio del cluster.

Il processo inizia con la scelta di **K centroidi iniziali**, assegnati casualmente, seguita dall’assegnazione di ogni punto al centroide più vicino (usando la distanza euclidea). Successivamente, i centroidi vengono ricalcolati come la media dei punti assegnati e il processo si ripete finché i centroidi non cambiano più in modo significativo o si raggiunge il numero massimo di iterazioni.

Per scegliere il numero ottimale di cluster **K**, si utilizzano metodi come il **Metodo del Gomito (SSE)**, che identifica il punto di "flesso" della curva degli errori, e il **Silhouette Score**, che misura la coesione dei cluster. **K-Means è veloce ed efficiente**, ma presenta alcune limitazioni: richiede la scelta preliminare di **K**, è sensibile agli **outlier** e funziona meglio con cluster di forma sferica.

**Obiettivo dell’Analisi:**

L'obiettivo di questa analisi tramite clustering è raggruppare i paesi del mondo in cluster sulla base di dati demografici e geografici, utilizzando l'algoritmo K-Means. In particolare, abbiamo preso in considerazione:

* Popolazione storica dal 1970 al 2022.
* Superficie del paese (Area km²).
* Densità di popolazione (abitanti per km²).
* Tasso di crescita della popolazione (%).

La normalizzazione dei dati è stata applicata per rendere confrontabili le variabili con scale diverse (ad esempio, la popolazione in milioni e il tasso di crescita in percentuale).

**Pre-processing dei dati:**

* Importazione del dataset.
* Selezione delle feature più significative per il clustering.
* **Standardizzazione** dei dati tramite **StandardScaler** per evitare che le variabili con valori molto grandi (es. popolazione) dominassero le altre.

**Determinazione del numero ottimale di cluster (K)**

* È stata utilizzata la **metrica SSE (Sum of Squared Errors)** con il **Metodo del Gomito**.
* È stato calcolato il **Silhouette Score**, che misura la coesione dei cluster.

**Esecuzione del Clustering K-Means**

* Il valore ottimale di K è stato scelto **basandosi sui grafici SSE e Silhouette Score**.
* Il modello K-Means è stato applicato sui dati normalizzati.
* È stato calcolato il **Silhouette Score finale**, che fornisce un’indicazione sulla qualità della separazione tra cluster.

**Visualizzazione dei Risultati**

* **Grafico del Metodo del Gomito**: mostra il punto in cui la riduzione dell'errore SSE rallenta, suggerendo il miglior numero di cluster.

Immagine che contiene testo, schermata, diagramma, Diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

* **Grafico del Silhouette Score**: indica quale valore di K garantisce i cluster più coerenti.

Immagine che contiene testo, schermata, linea, Diagramma

Il contenuto generato dall'IA potrebbe non essere corretto.

* **Grafico della distribuzione dei cluster**: rappresenta i cluster in base a **Popolazione 2022 e Tasso di Crescita (%)**.

Immagine che contiene testo, schermata

Il contenuto generato dall'IA potrebbe non essere corretto.

**ANALISI DEI RISULTATI:**

Dall'analisi, il numero ottimale di cluster è stato identificato come K = 6, come mostrato nei grafici:

* Metodo del Gomito: il punto di flesso della curva SSE si verifica attorno a K = 6, suggerendo che oltre questo valore l'aggiunta di cluster riduce poco l'errore.
* Silhouette Score: mostra un valore massimo attorno a K = 6, indicando che i cluster sono ben separati e coerenti.
* Distribuzione dei Cluster: il grafico mostra come i paesi sono stati raggruppati in base alla loro popolazione e al tasso di crescita.

L'algoritmo ha identificato 6 gruppi distinti che raggruppano paesi con caratteristiche simili. Ad esempio:

* Alcuni cluster possono rappresentare paesi molto popolosi con crescita elevata.
* Altri cluster potrebbero includere paesi con popolazione ridotta e crescita negativa.
* Altri ancora potrebbero contenere paesi ad alta densità, ma con una popolazione totale inferiore.

**Accuratezza del Modello**

L'accuratezza nel clustering viene misurata dal **Silhouette Score**, che varia tra -1 e 1:

* **Valore vicino a 1** → Cluster ben definiti e separati.
* **Valore vicino a 0** → Cluster sovrapposti e poco chiari.
* **Valore negativo** → Assegnazioni errate di punti al cluster sbagliato.

**Silhouette Score finale ottenuto:**

Silhouette Score: 0.424

Questo valore indica che il clustering è di **qualità moderata**, suggerendo che i cluster sono ben separati, ma con alcune sovrapposizioni.