Laurea Triennale in Informatica - Università di Salerno
Corso di Fondamenti di Intelligenza Artificiale - Prof. Fabio Palomba

PROGETTO FIA 2024-2025

MelodyMind

Vincenzo Medica
(matr. 0512116808)

Problema da risolvere

- Raggruppare un insieme di canzoni che hanno caratteristiche simili all'interno di playlist.
- Le canzoni devono essere selezionate da un dataset ottenuto da Spotify.
- Soluzione possa essere facilmente integrata in futuro con Spotify.

Come?

Creando un Agente Intelligente:

- Descrivendo l’Ambiente;
- Selezionando e analizzando il Dataset;
- Valutando vari Algoritmi di Clustering.

Utilizzando il linguaggio **JavaScript** con il runtime environment **Node.js**.

Descrizione dell’Ambiente - PEAS

PERFORMANCE

- Massimizzare la similarità intra-cluster;
- Minimizzare il numero di outlier.

ENVIRONMENT

- Dataset: Un set di dati contenente canzoni con caratteristiche estratte tramite le API di Spotify;
- Palylists: vuote che andranno ad accogliere le canzoni raggruppate per caratteristiche comuni.

ACTUATORS

- Algoritmi di clustering (DBScan, K-means);
- Funzione per generare playlist sulla base dei cluster.

SENSORS

- L’agente raccoglie le caratteristiche delle canzoni da un dataset preesistente, ottenuto tramite le API di Spotify.

Dataset - Selezione

- Trovato su Kaggle;
- Estratto da Spotify (API Spotify);
- Circa 2000 canzoni;
- Rilasciate tra il 1956 e il 2019;
- Canzoni variegate nel panorama musicale.

	B	C	D	E
1	Title	Artist	Top Genre	Year
2	Sunrise	Norah Jones	adult standards	2004
3	Black Night	Deep Purple	album rock	2000
4	Clint Eastwood	Gorillaz	alternative hip hop	2001
5	The Pretender	Foo Fighters	alternative metal	2007
6	Waitin' On A Sunny Day	Bruce Springsteen	classic rock	2002
7	The Road Ahead (Miles Of The Unknown)	City To City	alternative pop rock	2004
8	She Will Be Loved	Maroon 5	pop	2002
9	Knights of Cydonia	Muse	modern rock	2006
10	Mr. Brightside	The Killers	modern rock	2004
11	Without Me	Eminem	detroit hip hop	2002
12	Love Me Tender	Elvis Presley	adult standards	2002
13	Seven Nation Army	The White Stripes	alternative rock	2003
14	Als Het Golft	De Dijk	dutch indie	2000
15	I'm going home	Ten Years After	album rock	2005
16	Fluorescent Adolescent	Arctic Monkeys	garage rock	2007
17	Zonder Jou	Paul de Leeuw	dutch cabaret	2006
18	Speed of Sound	Coldplay	permanent wave	2005
19	Uninvited	Alanis Morissette	alternative rock	2005
20	Music	John Miles	classic uk pop	2004
21	Cry Me a River	Justin Timberlake	dance pop	2002
22	Fix You	Coldplay	permanent wave	2005
23	The Cave	Mumford & Sons	modern folk rock	2009
24	Als De Morgen Is Gekomen	Jan Smit	dutch pop	2006
25	Somebody Told Me	The Killers	modern rock	2004
26	Dichterbij Dan Ooit	BLØF	dutch pop	2002
27	Miracle	Ilse DeLange	dutch americana	2008
28	Smokers Outside the Hospital Doors	Editors	alternative dance	2007
29	Cleanin' Out My Closet	Eminem	detroit hip hop	2002
30	Der Weg	Herbert Grönemeyer	german pop	2008

Dataset - Features

- **Titolo, Artista, Anno;**
- **TopGenre**: genere predominante dell'artista;
- **Beats Per Minute** (BPM): battiti per minuto;
- **Energy** : [0, 100], misura l'intensità della canzone e l'energia trasmessa;
- **Danceability** (Ballabilità): [0, 100], indica quanto la traccia è adatta al ballo;
- **Loudness** (Sonorità)(dB): [-60, 0], intensità acustica della traccia;
- **Liveness** (Vivacità): [0, 100], indica la presenza di spettatori durante l'esecuzione della traccia;
- **Valence** (Positività): [0, 100], misura la positività trasmessa dalla canzone;
- **Length** (Durata): durata della traccia;
- **Acousticness** (Acustica): [0, 100], misura la naturalezza del suono, ovvero quanto il suono della canzone è modificato elettronicamente;
- **Speechiness** (Loquacità): [0, 100], misura la componente vocale o parlata della traccia;
- **Popularity**: [0, 100], indica la popolarità della canzone.

Dataset - Standardizzazione

Necessaria per utilizzare la metodologia PCA (Principal Component Analysis).

Per garantire che ogni variabile contribuisca in modo equo è necessario che abbiano la stessa deviazione standard e di conseguenza, lo stesso peso. Ciò è stato fatto utilizzando lo **z-score** calcolato con la formula mostrata, che utilizza:

$$Z = \frac{X - \mu}{\sigma}$$

- X è il valore originale della variabile;
- μ (mi) è la media della variabile;
- σ (sigma) è la deviazione standard della variabile.

Dataset - PCA

PCA (Principal Component Analysis) consente di ridurre il numero di variabili necessarie a descrivere un insieme di dati, minimizzando la perdita di informazioni.

Vantaggi:

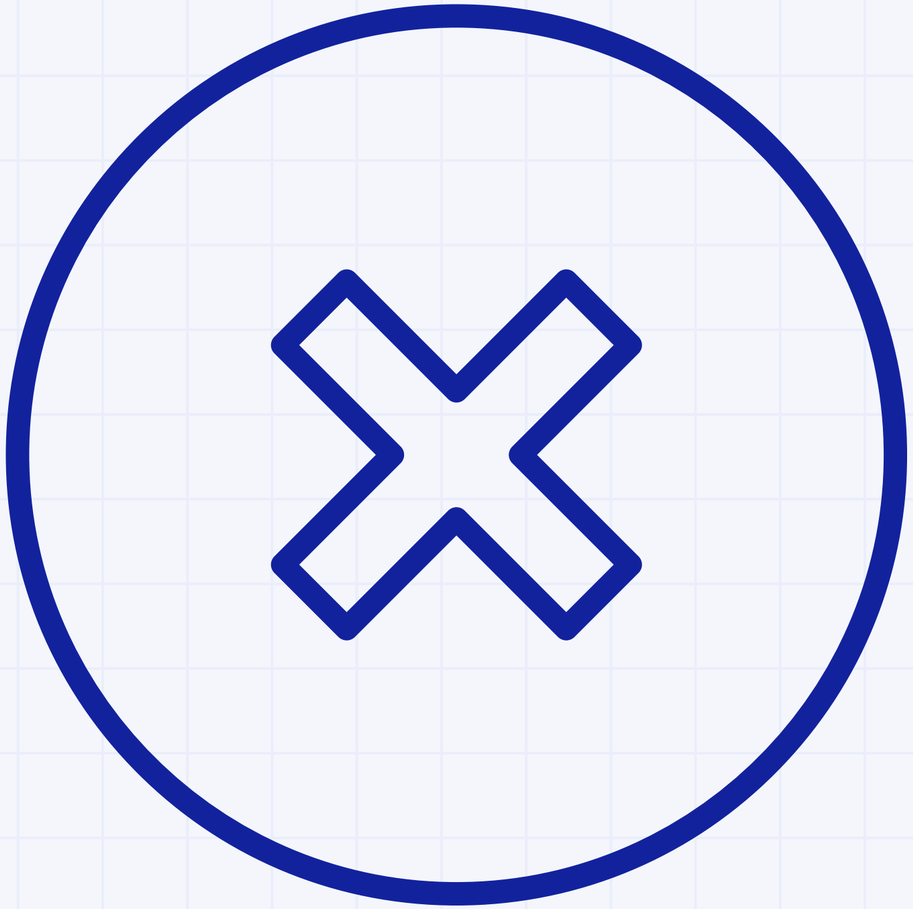
- **Riduzione del costo computazionale:** la riduzione delle dimensioni diminuisce la complessità;
- **Visualizzazione grafica:** consente di rappresentare i dati nello spazio tridimensionale;
- **Riduzione del rumore:** consente di neutralizzare l'impatto dei valori anomali.

Implementato usando la libreria **pca-js.**

Dataset - PCA

Escluse le features:

- **popularity** che si basa sul numero di riproduzioni effettuate nel breve periodo è stata esclusa perchè le canzoni più recenti in media registrano un numero maggiore di ascolti rispetto a quelle del passato;
- **length** che indica la durata del brano. Dato che la lunghezza media delle canzoni è cambiata significativamente dal 1950 a oggi con una tendenza alla riduzione, si è deciso di escluderla.



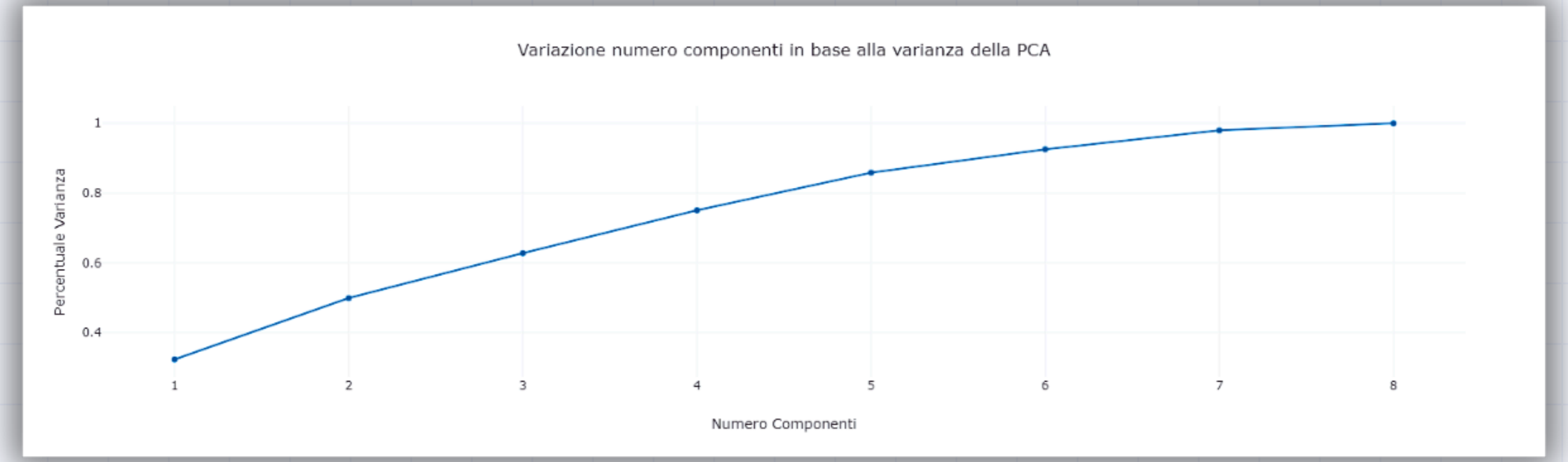
Dataset - PCA

Passi:

- Calcolo degli **autovalori e degli autovettori** (cercando un compromesso tra il numero di componenti da utilizzare e la perdita di varianza);
- L’algoritmo di clusterizzazione opera sul numero di componenti principali sufficiente a **preservare almeno il 70%** della varianza originale del dataset;
- I risultati della clusterizzazione vengono proiettati in uno **spazio tridimensionale**, utilizzando sempre le prime tre componenti principali.

Dataset - PCA Risultato

Mostra la percentuale di varianza conservata dopo l'applicazione della PCA, in funzione del numero di componenti principali selezionate. Il numero di componenti principali (PC) necessario per raggiungere almeno il 70% della varianza originale è pari a **4 PC**, ottenendo circa il **77% della varianza**. Il numero di features è stato ridotto da n a m, con $m < n$.

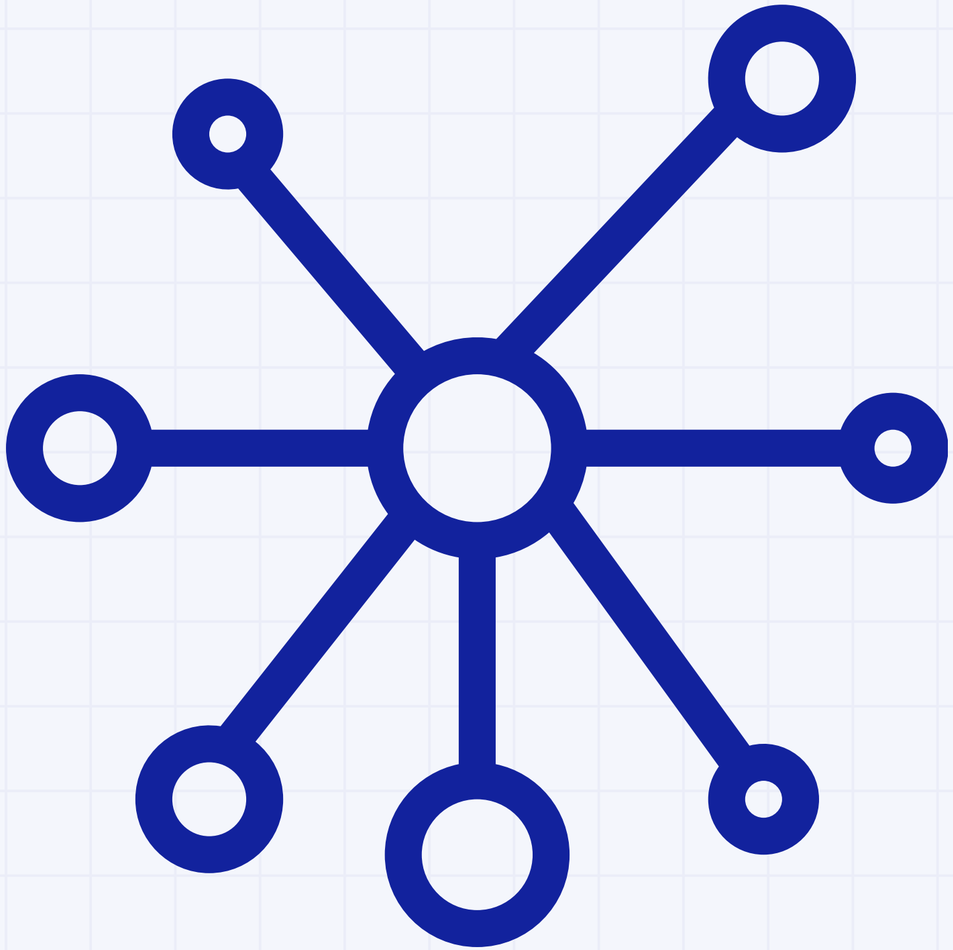


Clustering

Un insieme di metodologie utilizzate per raggruppare oggetti in classi omogenee dette cluster.

Gli algoritmi utilizzati sono stati:

- **DBScan**
- **K-means**



Clustering - DBScan

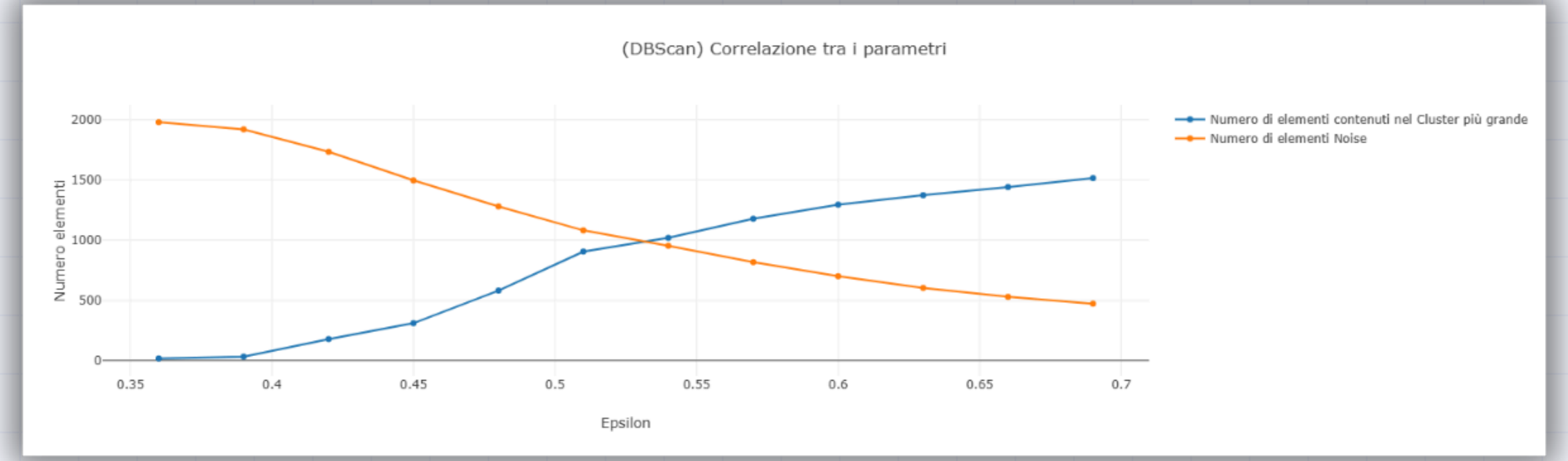
Il DBScan è un algoritmo di clustering basato sulla densità.

Utilizza due parametri:

- **Epsilon(ϵ)**: la massima distanza tra due punti che consente di determinare se un punto fa parte di un cluster o meno, che nel nostro caso varia da 0.35 a 0.70 .
- **minPoints:** il numero minimo di punti necessari per formare un cluster, che nel nostro caso è stato è stato fissato a 10.

Clustering - DBScan

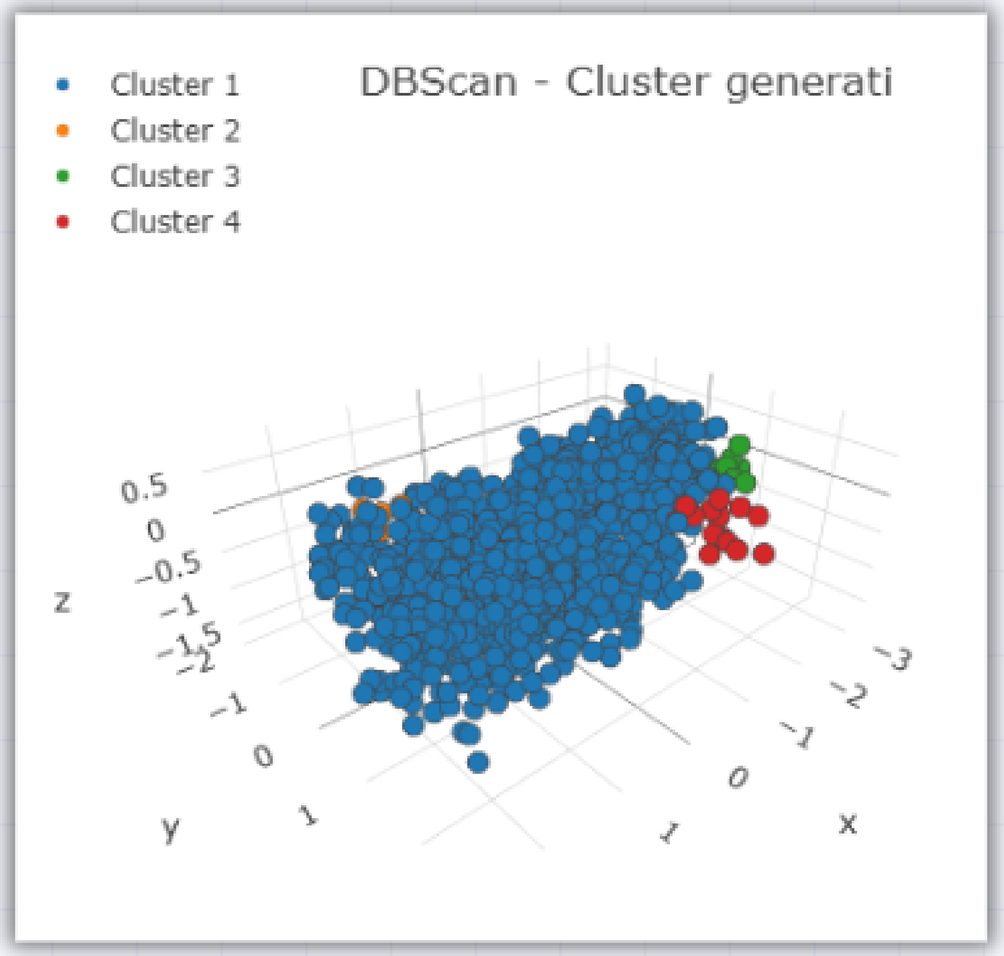
Il grafico mostra che il numero di elementi esclusi dall'analisi (noise) era troppo alto o che erano stati creati pochi cluster con molti elementi, risultando inefficiente. Per questo è stato scelto un compromesso con ϵ (Epsilon) pari a 0.54.



Clustering - DBScan Risultato

- Ha portato alla formazione di quattro cluster.
- Il più grande contenente 1032 elementi.
- Numero di elementi esclusi (noise) pari a 941.

Non va bene: Un noise così elevato compromette la qualità del raggruppamento.



Clustering - K-means

Il K-means è un algoritmo partizionale che permette di suddividere un insieme di punti in K gruppi, ciascuno identificato da un centroide.

Passi:

- Crea k partizioni assegnando casualmente i punti a ciascuna partizione;
- Calcola il centroide di ogni gruppo;
- Ogni punto viene assegnato al gruppo il cui centroide risulta più vicino;
- Vengono ricalcolati i centroidi per i nuovi gruppi.

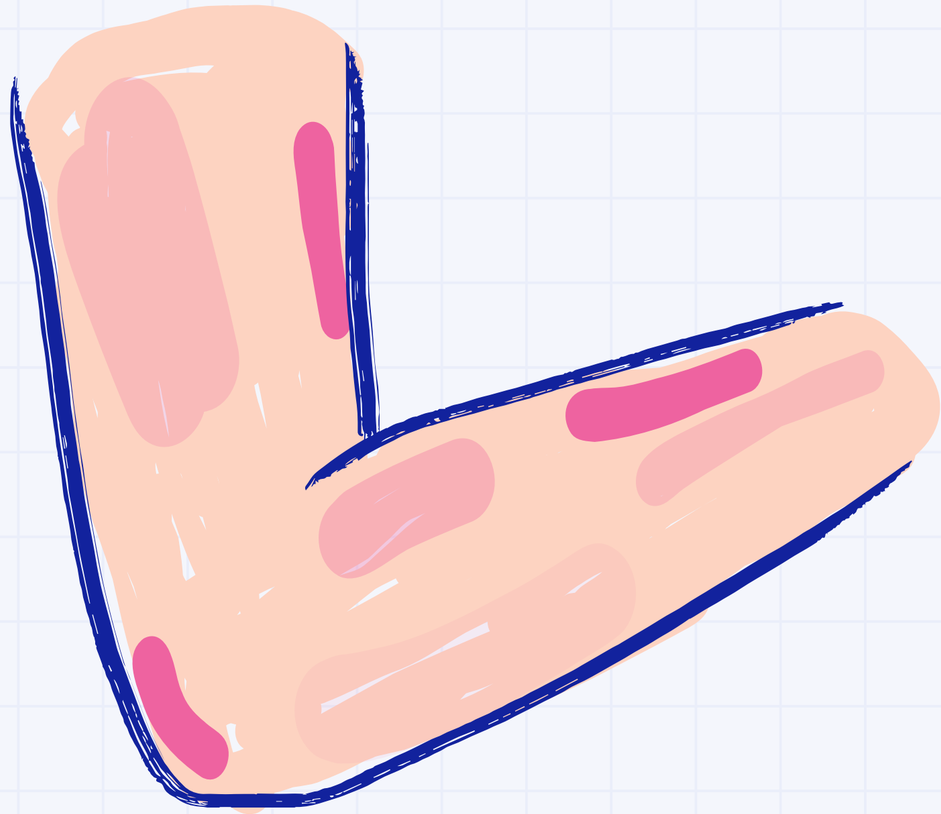
Il processo si ripete fino a quando l'algoritmo non converge o fino al raggiungimento del numero massimo di iterazioni predefinito.

Clustering - K-means Problema

Problema: determinazione del valore ottimale di K per il clustering.

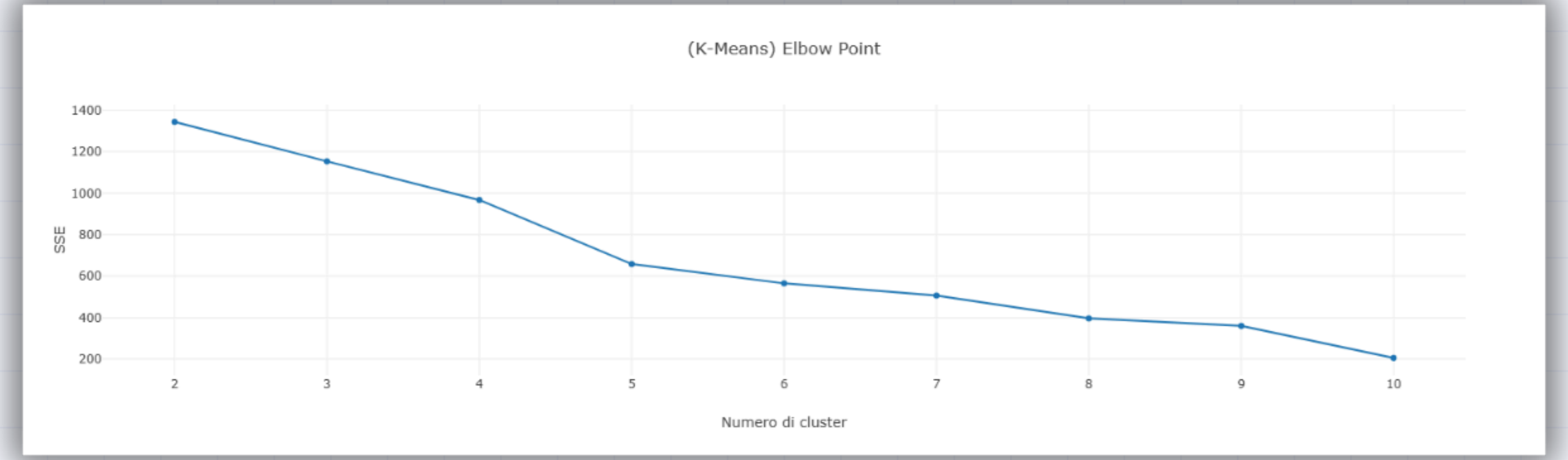
Soluzione: **Elbow method**

Consiste nel graficare la Somma degli Errori Quadrati (SSE) in funzione del numero di cluster.



Clustering - K-means Elbow point

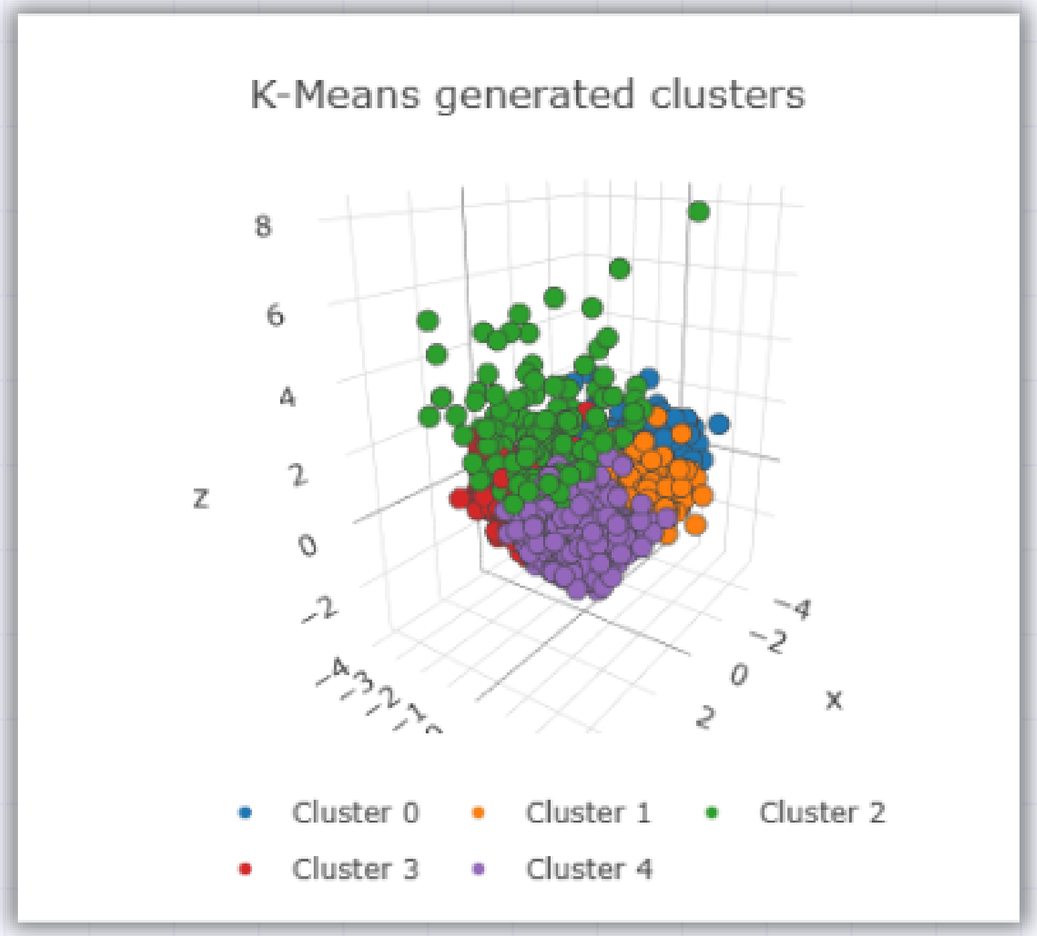
Il grafico mi ha permesso di determinare il valore ottimale di K, il quale è scelto come il punto di "gomito" della curva risultante, dove si osserva una significativa riduzione della SSE (somma delle distanze al quadrato di ciascun punto dal proprio centroide all'interno di ogni cluster).



Clustering - K-means Risultato

- I punti sono stati raggruppati in cinque cluster distinti.
- Nonostante la distribuzione dei punti nello spazio tridimensionale sia piuttosto concentrata, l’algoritmo k-means ha effettuato un buon raggruppamento.
- Ogni cluster rappresenta un insieme coerente di punti con caratteristiche simili.
- Raggruppamento risulta efficiente.

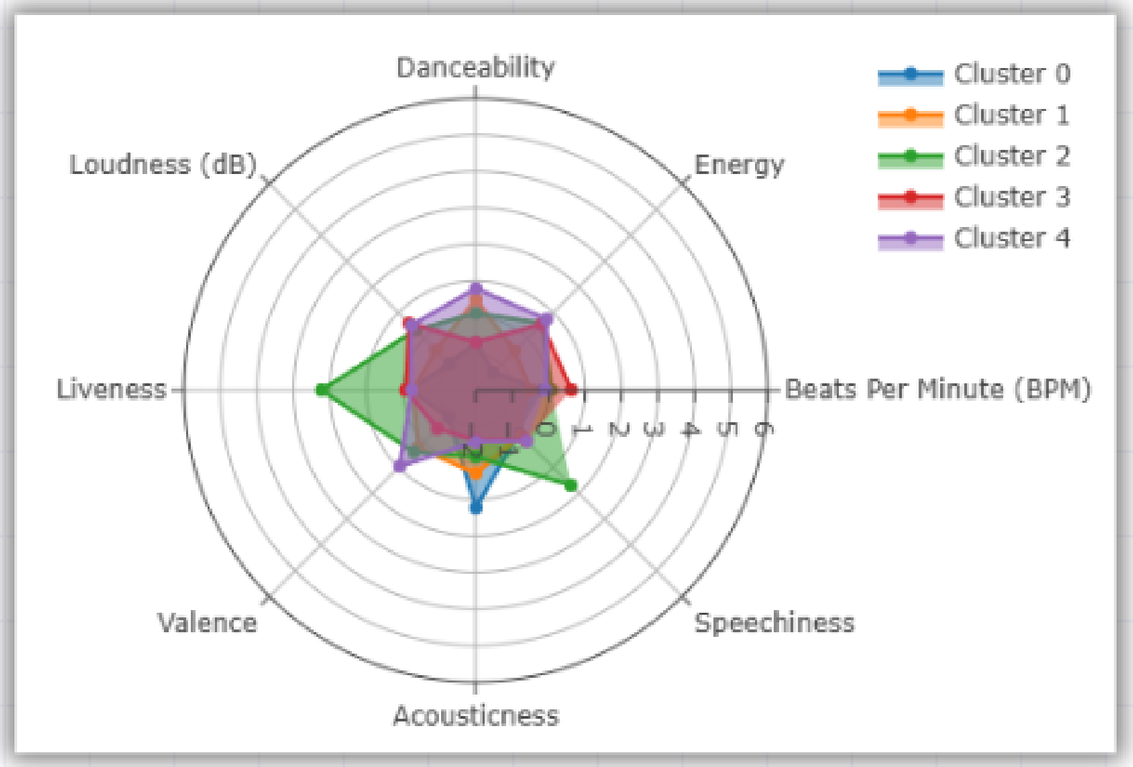
K-means viene scelto come algoritmo di clusering.



Soluzione - Valutazione

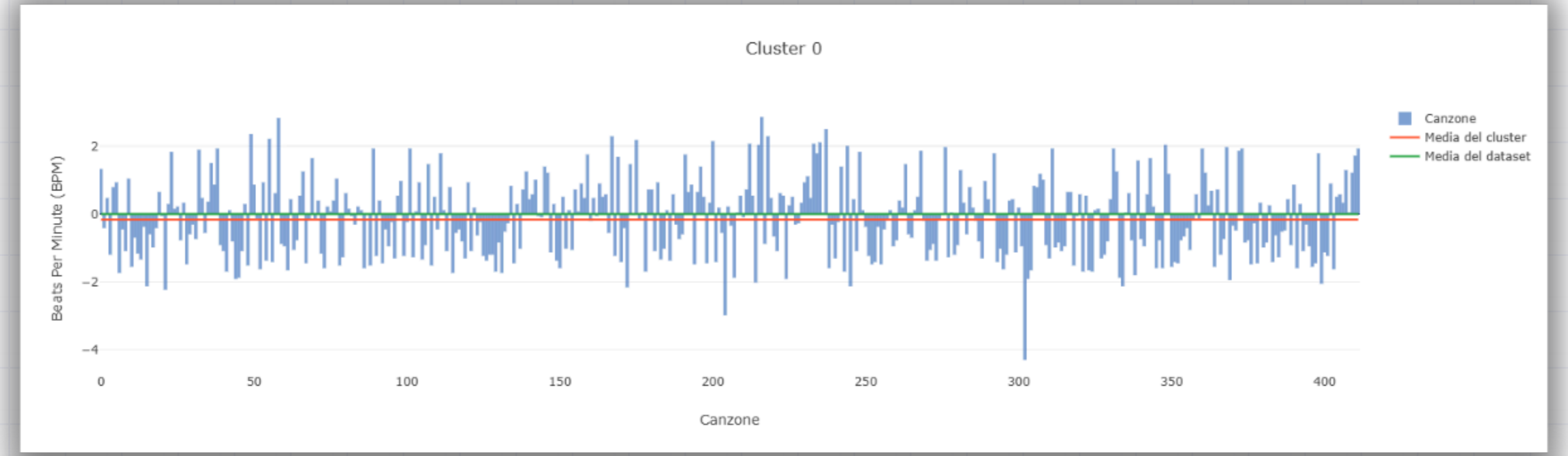
Soluzione finale utilizza: **PCA e K-means.**
Sul grafico radar sono mostrati i valori medi delle features per ciascun cluster.
Osserviamo:

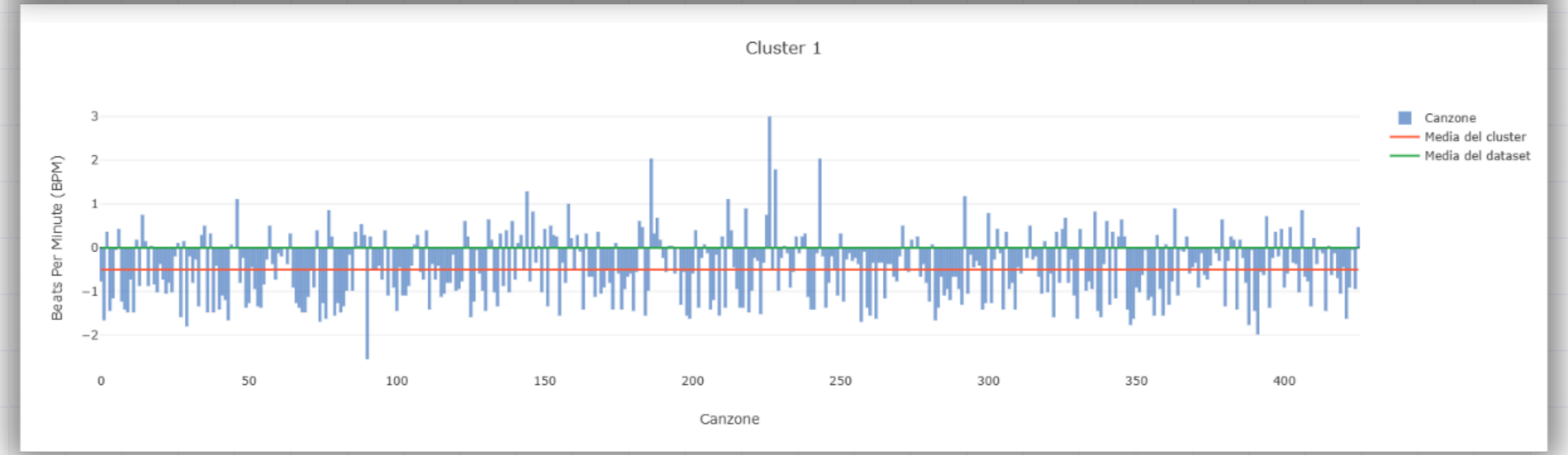
- **Cluster 0** possiede dei valori di Acousticness (Acustica) più alti ciò indica che contiene canzoni che sono meno distorte da effetti elettronici;
- **Cluster 2** possiede dei valori elevati di Speechiness (Loquacità) che di Liveness (Vivacità), il che suggerisce una predominanza di canzoni con una forte componente parlata con anche la presenza di spettatori.



Soluzione - Valutazione

- Sono stati rappresentati i valori medi delle features dei cluster attraverso dei grafici a barre
- Obiettivo di ottenere un’ulteriore rappresentazione bidimensionale dell’andamento delle features in ciascun cluster.
- Alcuni esempi per la feature Beats Per Minute (BPM).


Cluster 0
Beats Per Minute (BPM)
2
0
−2
−4
0
50
100
150
200
250
300
350
400
Canzone
Canzone
Media del cluster
Media del dataset


Cluster 1
Beats Per Minute (BPM)
3
2
1
0
−1
−2
0
50
100
150
200
250
300
350
400
Canzone
Canzone
Media del cluster
Media del dataset

Risultati

Dopo aver applicato l’algoritmo di clustering k-means e osservato che produce la suddivisione più efficace, si è ottenuto un totale di 5 cluster di canzoni. Ogni cluster è stato trasformato in una playlist, contenente le canzoni associate. Un esempio parziale di playlist ottenuta:

```
Playlist 0:
0   Titolo: Sunrise, Artista: Norah Jones, Top Genere: adult standards
1   Titolo: Love Me Tender, Artista: Elvis Presley, Top Genere: adult standards
2   Titolo: Zonder Jou, Artista: Paul de Leeuw, Top Genere: dutch cabaret
3   Titolo: Music, Artista: John Miles, Top Genere: classic uk pop
4   Titolo: Der Weg, Artista: Herbert Gr�nemeyer, Top Genere: german pop
5   Titolo: The Scientist, Artista: Coldplay, Top Genere: permanent wave
6   Titolo: De Weg, Artista: Guus Meeuwis, Top Genere: dutch pop
7   Titolo: Just Breathe, Artista: Pearl Jam, Top Genere: alternative rock
```

Sviluppi futuri

- Sviluppo di un modulo aggiuntivo che consenta l’integrazione diretta con **Spotify;**
- Estrarre automaticamente le canzoni dalla sezione preferiti;
- Salvarle in un file;
- File sarà elaborato dall’agente.

Vantaggi:

- Creazione di playlist dinamiche basate sulle preferenze musicali dell’utente;
- Migliorerebbe l’esperienza dell’utente.

Laurea Triennale in Informatica - Università di Salerno
Corso di Fondamenti di Intelligenza Artificiale - Prof. Fabio Palomba

PROGETTO FIA 2024-2025

Grazie per l'attenzione

Vincenzo Medica
(matr. 0512116808)