Report di progetto - MelodyMind

VINCENZO MEDICA, Università degli Studi di Salerno, Italia

In questo report si descrive il processo volto a risolvere il problema di creare delle playlist che contengano canzoni che hanno caratteristiche simili, ciò attraverso la creazione dell'agente intelligente MelodyMind. Per raggiungere questo obiettivo, ho adottato un approccio di apprendimento non supervisionato basato sul clustering. Fornendo a diversi algoritmi di clustering un set di canzoni, ciascuna descritta non solo da informazioni di base come il titolo e il nome dell'artista, ma anche da diverse features musicali, con l'intento di ottenere playlist che contengano canzoni con caratteristiche simili.

Si riporta, dunque, il link al repository GitHub che contiene il codice del progetto per la demo: GitHub Repository.

CCS Concepts: • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: Fundamentals of Artificial Intelligence, Clustering, K-Means, DBScan, PCA, Principal Component Analysis

ACM Reference Format:
Vincenzo Medica. 2025. Report di progetto - MelodyMind. 1, 1 (January 2025), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Author's Contact Information: Vincenzo Medica, Università degli Studi di Salerno, Fisciano, Italia, v.medica@studenti.unisa.it.

ACM XXXX-XXXX/2025/1-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Contents

1 Introduzione

1.1 Descrizione del problema

Il problema affrontato in questo progetto riguarda la necessità di raggruppare un insieme di canzoni che hanno caratteristiche simili all'interno di playlist. Le canzoni devono essere selezionate da un dataset ottenuto da Spotify, una delle principali piattaforme di streaming musicale, in modo tale che in futuro questa soluzione possa essere facilmente integrata con la piattaforma.

1.2 Obiettivo

L'obiettivo di questo progetto è la realizzazione di un agente intelligente in grado di analizzare un insieme di canzoni provenienti da un dataset ottenuto tramite le Spotify API, identificarne le caratteristiche comuni e raggrupparle in playlist omogenee. Per raggiungere questo scopo verranno utilizzati diversi algoritmi di clustering, valutandone le performance al fine di ottenere una suddivisione ottimale delle canzoni in gruppi significativi.

1.3 Strumenti e tecnologie impiegate

Il progetto è stato interamente sviluppato in JavaScript, sfruttando il runtime environment Node.js per garantire una gestione efficiente del codice e delle risorse. Node.js ha permesso di integrare agevolmente i dati del dataset, di eseguire elaborazioni complesse e, grazie all'impiego di svariate librerie JavaScript, di generare una vasta gamma di rappresentazioni grafiche. Questi strumenti hanno facilitato l'interpretazione dei risultati, offrendo una panoramica chiara e interattiva degli output ottenuti. Per l'analisi e la visualizzazione dei risultati sono state impiegate svariate librerie JavaScript, che hanno consentito di generare una vasta gamma di rappresentazioni grafiche. Questi strumenti hanno facilitato l'interpretazione, offrendo una panoramica chiara e interattiva dei risultati ottenuti.

2 Descrizione dell'ambiente

2.1 Specifica PEAS dell'ambiente

Di seguito si delinea la specifica PEAS dell'ambiente in cui opererà l'agente:

- **Performance:** sono adottate per valutare l'operato dell'agente:
  - Massimizzare la similarità intra-cluster (distanza media minima tra canzoni dello stesso cluster).
  - Minimizzare il numero di outlier generati dagli algoritmi di clustering.
- **Environment:** sono i vari elementi che lo formano:
  - Dataset: Un set di dati contenente canzoni con caratteristiche estratte tramite le API di Spotify, come BPM, Acousticness, Danceability, Energy, ecc. .
  - Playlists: vuote che andranno ad accogliere le canzoni raggruppate per caratteristiche comuni.
- **Actuators:** sono gli strumenti che l'agente utilizza per generare le playlist:
  - Algoritmi di clustering (DBScan, K-means).
  - Funzione per generare playlist sulla base dei cluster.
- **Sensors:** sono i metodi tramite cui l'agente raccoglie i dati dalle canzoni:
  - L'agente raccoglie le caratteristiche delle canzoni da un dataset preesistente, ottenuto tramite le API di Spotify. I dati includono vari parametri musicali come BPM, energia, acusticità, e altri attributi che descrivono ogni canzone.

2.2 Caratteristiche

L'agente si interfaccia in un ambiente con le seguenti caratteristiche:

- **Completamente osservabile:** l'agente è a conoscenza in ogni istante della configurazione del dataset e delle playlist.
- **Deterministico:** la configurazione e il numero delle playlist è il risultato solo e soltanto dell'agente
- **Statico:** l'ambiente non cambia mentre l'agente sta deliberando.
- **Discreto:** le azioni dell'agente (eseguire il clustering su un insieme di canzoni) sono chiaramente definite e non vi è una continua variazione nel tempo. Le canzoni vengono raggruppate in playlist definite, e ogni playlist è un insieme discreto di canzoni.
- **Singolo:** l'agente lavora in modo autonomo, senza interazione con altri agenti.

Altre caratteristiche:

- **Non supervisionato:** non ci sono etichette predefinite che indicano a quale playlist musicale appartiene una canzone. Le feature delle canzoni vengono utilizzate come input, ma l'agente deve individuare a quale playlist appartengono.

3 Analisi sul Dataset

3.1 Selezione del dataset

Il dataset selezionato per il progetto è stato scelto in base alla necessità di lavorare con campioni descritti da un insieme uniforme di feature, in linea con quelle utilizzate dalla piattaforma musicale più utilizzata al giorno d'oggi, ovvero Spotify. Questa scelta è stata fatta anche con l'intenzione di lasciare aperta la possibilità, in futuro, di integrare l'agente con Spotify.
Per soddisfare questa esigenza, si è deciso di cercare un dataset su Kaggle estratto direttamente da Spotify. Sono stati individuati diversi dataset pertinenti, ciascuno con un numero di campioni e una varietà musicale differenti. Alcuni di questi includevano principalmente canzoni degli ultimi 10 o 20 anni, in prevalenza appartenenti al genere pop. Tuttavia, tali dataset sono stati considerati non ottimali, poiché non rappresentavano in modo equilibrato tutte le principali caratteristiche dei macro-generi musicali, come rock, rap, musica classica, ecc. che consentissero di dare una visione globale del mondo musicale.

Successivamente, è stato analizzato un dataset distribuito in modo più equilibrato tra le varietà musicali, ma contenente un numero elevato di campioni. La grande dimensione del dataset, combinata con le tecnologie utilizzate, ha comportato prestazioni insufficienti e difficoltà computazionali in diverse fasi del progetto. Perciò, si è optato per un dataset più contenuto ma comunque rappresentativo di tutte le varietà musicali, composto da circa 2.000 tracce rilasciate tra il 1956 e il 2019. Questo dataset, selezionato per la sua qualità e adeguatezza, è disponibile al seguente link.
Il dataset include, oltre alle informazioni di base come il nome, l'artista e l'anno di uscita, le seguenti features per ogni canzone:

- **Top Genre:** genere principale associato alla canzone in base al genere predominante dell'artista.
- **Beats Per Minute (BPM):** unità di misura del numero di battiti per minuto;
- **Energy:** [0, 100], misura l'intensità della canzone e l'energia trasmessa;
- **Danceability:** [0, 100], indica quanto la traccia è adatta al ballo;
- **Loudness (dB):** [-60, 0], intensità acustica della traccia;
- **Liveness:** [0, 100], indica la presenza di spettatori durante l'esecuzione della traccia;
- **Valence:** [0, 100], misura la positività trasmessa dalla canzone;
- **Length (Duration):** durata della traccia;

- **Acousticness:** [0, 100], misura la naturalezza del suono, ovvero quanto il suono della canzone è modificato elettronicamente;
- **Speechiness:** [0, 100], misura la componente vocale o parlata della traccia;
- **Popularity:** [0, 100], indica la popolarità della canzone;

Questo dataset, per la sua varietà e completezza, è risultato ideale per lo sviluppo dell'agente, fornendo una base solida per l'analisi e la creazione di playlist musicali coerenti.
La Figura 1 mostra parte del dataset utilizzato con le relative features per ogni canzone.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Title	Artist	Top Genre	Year	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Length (Duration)	Acousticness	Speechiness	Popularity
2	Sunrise	Norah Jones	adult standards	2004	157	30	53	-14	11	68	201	94	3	71
3	Black Night	Deep Purple	album rock	2000	135	79	50	-11	17	81	207	17	7	39
4	Clint Eastwood	Gorillaz	alternative hip hop	2001	168	69	66	-9	7	52	341	2	17	69
5	The Pretender	Foo Fighters	alternative metal	2007	173	96	43	-4	3	37	269	0	4	76
6	Waitin' On A Sunny Day	Bruce Springsteen	classic rock	2002	106	82	58	-5	10	87	256	1	3	59
7	The Road Ahead (Miles Of The Unknown)	City To City	alternative pop rock	2004	99	46	54	-9	14	14	247	0	2	45
8	She Will Be Loved	Maroon 5	pop	2002	102	71	71	-6	13	54	257	6	3	74
9	Knights of Cydonia	Muse	modern rock	2006	137	96	37	-5	12	21	366	0	14	69
10	Mr. Brightside	The Killers	modern rock	2004	148	92	36	-4	10	23	223	0	8	77
11	Without Me	Eminem	detroit hip hop	2002	112	67	91	-3	24	66	290	0	7	82
12	Love Me Tender	Elvis Presley	adult standards	2002	109	5	44	-16	11	31	162	88	4	49
13	Seven Nation Army	The White Stripes	alternative rock	2003	124	46	74	-8	26	32	232	1	8	74
14	Als Het Golft	De Dijk	dutch indie	2000	102	88	54	-6	53	59	214	2	3	34
15	I'm going home	Ten Years After	album rock	2005	117	93	38	-2	81	40	639	18	10	26
16	Fluorescent Adolescent	Arctic Monkeys	garage rock	2007	112	81	65	-5	14	82	173	0	3	66
17	Zonder Jou	Paul de Leeuw	dutch cabaret	2006	133	42	42	-10	16	25	236	84	4	48
18	Speed of Sound	Coldplay	permanent wave	2005	123	90	52	-7	7	36	288	0	6	69
19	Uninvited	Alanis Morissette	alternative rock	2005	127	54	38	-5	9	19	276	2	3	57
20	Music	John Miles	classic uk pop	2004	87	31	27	-13	63	12	352	1	3	46
21	Cry Me a River	Justin Timberlake	dance pop	2002	74	65	62	-7	10	56	288	57	18	74
22	Fix You	Coldplay	permanent wave	2005	138	42	21	-9	11	12	296	16	3	81
23	The Cave	Mumford & Sons	modern folk rock	2009	142	51	60	-10	11	35	218	5	4	67
24	Als De Morgen Is Gekomen	Jan Smit	dutch pop	2006	96	89	63	-6	9	81	176	5	3	55
25	Somebody Told Me	The Killers	modern rock	2004	138	99	51	-3	12	65	197	0	9	69
26	Dichterbij Dan Ooit	BLØF	dutch pop	2002	112	74	65	-7	23	52	261	18	3	16
27	Miracle	Ilse DeLange	dutch americana	2008	130	48	55	-8	10	18	270	48	3	50
28	Smokers Outside the Hospital Doors	Editors	alternative dance	2007	123	68	53	-4	12	55	298	0	4	56
29	Cleanin' Out My Closet	Eminem	detroit hip hop	2002	148	76	91	-5	8	87	298	7	17	71
30	Der Weg	Herbert Grönemeyer	german pop	2008	142	24	34	-11	12	19	259	92	4	48

Fig. 1. Tabella che mostra parte del dataset contenente per ogni canzone le relative features.

3.2 Standardizzazione

Dopo la selezione del dataset, il passo successivo è stato quello della standardizzazione dei dati. Questo processo è stato necessario poiché, utilizzando la metodologia PCA (Principal Component Analysis), il nuovo dataset viene ottenuto proiettando i campioni originali su un sistema di riferimento diverso. I nuovi assi generati dipendono dalla deviazione standard delle variabili, il che implica che una variabile con una deviazione standard più elevata avrà un impatto maggiore sugli assi principali rispetto a una variabile con una deviazione standard più bassa. Per garantire che ogni variabile contribuisca in modo equo alla determinazione dei nuovi assi, è stato effettuato un processo di standardizzazione, in modo che tutte le variabili avessero la stessa deviazione standard e, di conseguenza, lo stesso peso nei calcoli derivati dalla PCA.
La tecnica adottata per la standardizzazione è stata lo z-score, che viene calcolato con la seguente formula:

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

dove:

- X è il valore originale della variabile,
- μ è la media della variabile,
- σ è la deviazione standard della variabile.

A seguito di questa trasformazione, tutti i valori sono stati normalizzati, garantendo così che ogni variabile avesse la stessa importanza nel processo di analisi principale.
Sono riportate in Figura 2, Figura 3, Figura 4 gli snippet di codice relativi a questa fase.

```
function standardize(array) : void {
    const mean = getMean(array);
    const standardDeviation : number  = getStandardDeviation(array, mean);
    array.forEach((value, index, array) : void =>{
        zscore = (value - mean)/standardDeviation;
        array[index] = zscore;
    });
}
```

Fig. 2. Snippet di codice che implementa la funzione di standardizzazione tramite z-score.

```
function getMean(array){
    let n = array.length;
    let sum : number  = 0;
    array.forEach((value, index, array) : void =>{
        if(value!=""||value){
            sum += parseInt(value);
        } else {
            n--;
        }
    });
    return sum/n;
}
```

Fig. 3. Snippet di codice che calcola la media di una variabile.

```
function getStandardDeviation(numbersArr, meanVal) {

    var SDprep : number  = 0;
    for(var key in numbersArr) {
        if(numbersArr[key]!=""||numbersArr[key]){
            SDprep += Math.pow((parseFloat(numbersArr[key]) - meanVal), y: 2);
        }
    }
    var SDresult : number  = Math.sqrt( x: SDprep/numbersArr.length);
    return SDresult;

}
```

Fig. 4. Snippet di codice che calcola la deviazione standard di una variabile.

3.3 PCA

Dopo aver standardizzato i dati del dataset, è stato ottenuto un insieme di campioni, ciascuno caratterizzato da 10 variabili (features). È stata presa la decisione di escludere la variabile (popularity) relativa alla popolarità di una canzone, poiché tale metrica, calcolata da Spotify, si basa sul numero di riproduzioni effettuate nel breve periodo. Questo criterio è stato ritenuto poco espressivo e affidabile, dato che il dataset include canzoni pubblicate fino a 70 anni fa. Si ritiene che le canzoni

più recenti tendano, in media, a registrare un numero maggiore di ascolti rispetto a quelle del passato.

Inoltre, è stata esclusa la variabile relativa alla durata del brano (length). Questa scelta si basa sul fatto che la lunghezza media delle canzoni è cambiata significativamente dal 1950 a oggi (con una tendenza alla riduzione) e, al contempo, non si è ritenuto che tale variabile fosse indicativa del genere musicale.

Delle 8 variabili rimanenti, si è scelto di non utilizzarle singolarmente ma di sottoporle a un processo di analisi delle componenti principali (Principal Component Analysis, PCA). La PCA è una tecnica che consente di ridurre il numero di variabili necessarie a descrivere un insieme di dati, minimizzando la perdita di informazioni. Nel contesto del progetto, ciò significa descrivere ogni canzone non più attraverso tutte le 8 variabili originali, ma tramite una combinazione delle stesse, generata dalla PCA.

L'adozione della PCA offre diversi benefici:

- **Riduzione del costo computazionale:** la riduzione delle dimensioni diminuisce la complessità del calcolo delle distanze tra i punti del dataset, con conseguente diminuzione dei tempi di elaborazione e delle risorse necessarie;
- **Visualizzazione grafica:** la possibilità di rappresentare i dati nello spazio tridimensionale utilizzando le prime tre componenti principali;
- **Riduzione del rumore:** la PCA consente di neutralizzare l'impatto dei valori anomali, proiettandoli sulle componenti principali più significative e riducendo l'influenza delle componenti minori.

Per quanto riguarda l'implementazione, è stata utilizzata la libreria **pca-js**, che adotta un approccio basato sulla risoluzione matriciale. La prima fase del processo prevede il calcolo degli autovalori e degli autovettori attraverso il seguente metodo:

```
const vectors = pca.getEigenVectors(data);
```

Una volta calcolati gli autovettori, si è reso necessario determinare quali e quanti utilizzare per raggiungere un livello di precisione accettabile. Riducendo il numero di componenti principali rispetto al numero iniziale di variabili, si verifica inevitabilmente una perdita di informazione, poiché non tutta la varianza originaria viene preservata. È quindi emersa la necessità di trovare un compromesso accettabile tra il numero di componenti da utilizzare e la perdita di varianza.

In un primo momento, avevo considerato di utilizzare esclusivamente le prime tre componenti principali, in modo da facilitare la rappresentazione grafica dei dati e rendere evidente l'efficacia della PCA. Tuttavia, durante lo sviluppo del progetto, si è osservato che questo approccio comportava una significativa perdita di varianza.
Di conseguenza, è stata adottata la seguente strategia:

- L'algoritmo di clusterizzazione opera sul numero di componenti principali sufficiente a preservare almeno il 70% della varianza originale del dataset;
- Per la rappresentazione grafica dei risultati della clusterizzazione, i dati vengono proiettati in uno spazio tridimensionale, utilizzando sempre le prime tre componenti principali, che risultano comunque le più significative.

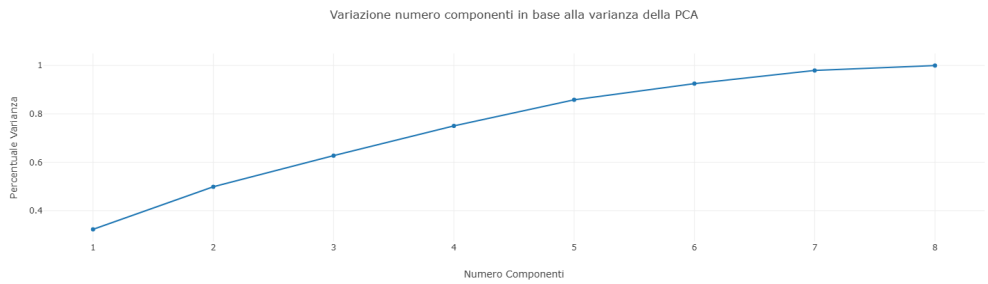

Fig. 5. Grafico che riporta la variazione del numero di componenti in base alla varianza della PCA

La Figura 5 illustra la percentuale di varianza conservata dopo l'applicazione della PCA, in funzione del numero di componenti principali selezionate.
Va sottolineato che, nel caso del dataset utilizzato, il numero di componenti principali (PC) necessario per raggiungere almeno il 70% della varianza originale è pari a 4, ottenendo circa il 77% della varianza.

Dopo aver completato la fase di preparazione, l'ultimo passo da eseguire è stata l'applicazione della PCA. Utilizzando la funzione:

pca.computeAdjustedData(data, ... vectors)

(dove in `vectors` sono calcolate e conservate dinamicamente le componenti principali necessarie per raggiungere il 70% della varianza originale), è stato possibile ridurre il numero di features da n a m, con $m < n$.

4 Clustering

Il passo successivo si è concretizzato finalmente nel clustering. Per clustering si intende un insieme di metodologie utilizzate per raggruppare oggetti in classi omogenee. Ogni cluster (classe) è un insieme di oggetti con caratteristiche simili agli altri oggetti dello stesso cluster ma che si differenziano più o meno notevolmente dagli oggetti presenti negli altri cluster. Il clustering rientra nel ramo dell'apprendimento non supervisionato. Gli algoritmi di clustering presi in considerazione sono stati il **DBScan** e il **K-means**. Entrambi sono stati implementati, utilizzati e valutati al fine di determinare quale tra i due ottenesse i migliori risultati.

4.1 DBScan

Il DBScan è un algoritmo di clustering basato sulla densità: dato un set di punti in uno spazio, raggruppa insieme (e di conseguenza proietta nello stesso cluster) punti che sono sufficientemente vicini e segna come valori anomali (noise) quei punti che giacciono più o meno "da soli" in regioni poco dense (dove per ogni punto il suo prossimo vicino è comunque troppo lontano per essere considerato parte dello stesso cluster). L'algoritmo presenta due parametri:

- **Epsilon(ϵ):** la massima distanza tra due punti che consente di determinare se un punto fa parte di un cluster o meno.
- **minPoints:** il numero minimo di punti necessari per formare un cluster.

Nel caso specifico, l'algoritmo è stato eseguito con il parametro `minPoints` fissato a 10 e con valori di `epsilon` che variano da 0.35 a 0.7. Come evidenziato nella Figura 6 che riporta un grafico nel

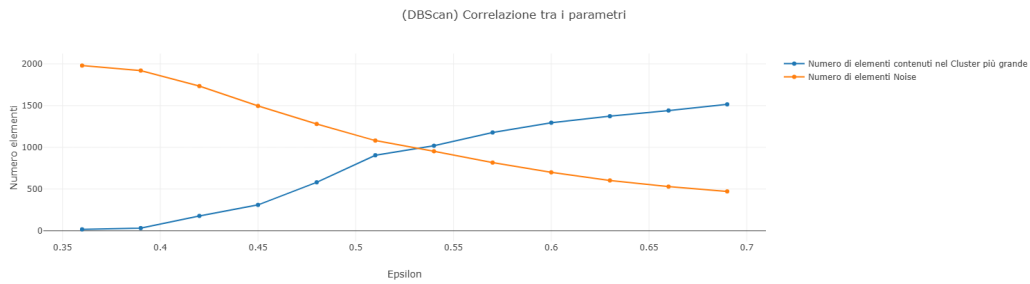

Fig. 6. Grafico che riporta la relazione tra Epsilon e il Numero di elementi

quale, il numero di elementi esclusi dall'analisi (noise) si è rivelato essere troppo elevato, oppure sono stati creati pochi cluster, ciascuno comprendente un numero elevato di elementi, risultando così poco efficace.

Dall'analisi del grafico mostrato in Figura 6, è stato identificato come valore di compromesso per ϵ (Epsilon) 0.54. Questa scelta ha portato alla formazione di quattro cluster, il più grande dei quali contiene 1032 elementi, con un numero di elementi esclusi (noise) pari a 941. Tuttavia, questo risultato non è stato ritenuto accettabile poiché un noise così elevato compromette la qualità del raggruppamento.

Il grafico riportato in Figura 7 illustra la correlazione tra il valore di ϵ (Epsilon) e il numero di cluster generati dall'algoritmo. Tale analisi, è stata fondamentale per individuare un valore di ϵ ottimale.

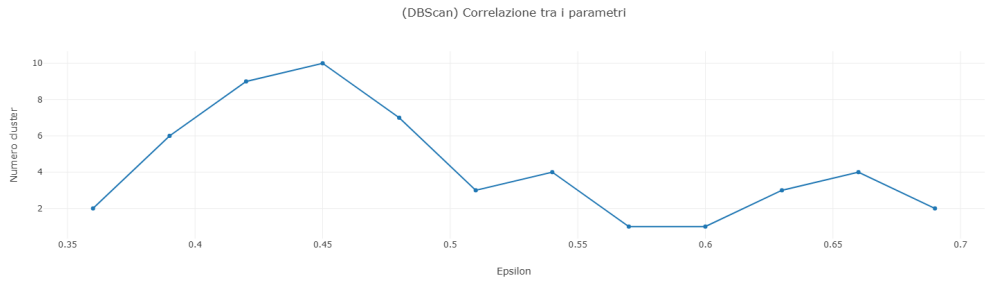

Fig. 7. Grafico che riporta la relazione tra Epsilon e il Numero di Cluster

Nella Figura 8, viene rappresentato in uno spazio tridimensionale il risultato ottenuto dall'algoritmo DBScan con $\epsilon = 0.54$. Dall'analisi del grafico emerge l'esistenza di un cluster predominante (indicato in blu), il quale evidenzia un risultato del processo di clustering non ottimale, suggerendo la necessità di ulteriori perfezionamenti nei parametri utilizzati.

4.2 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. L'algoritmo segue una procedura iterativa: inizialmente crea

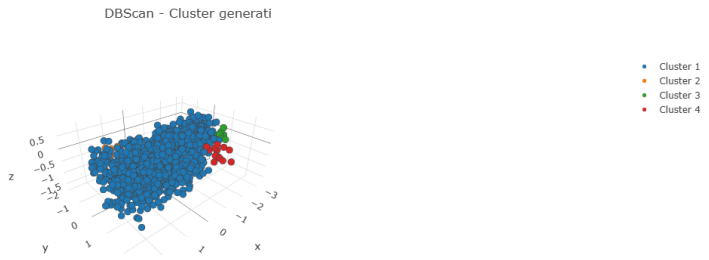

Fig. 8. Rappresentazione tridimensionale del risultato del DBScan con $\epsilon = 0.54$.

k partizioni, assegnando casualmente i punti a ciascuna partizione, quindi calcola il centroide di ogni gruppo. Successivamente, ogni punto viene assegnato al gruppo il cui centroide risulta più vicino. A questo punto, vengono ricalcolati i centroidi per i nuovi gruppi e il processo si ripete fino a quando l'algoritmo non converge o fino al raggiungimento del numero massimo di iterazioni predefinito.

Durante l'applicazione dell'algoritmo K-means, è emerso un nuovo problema: la determinazione del valore ottimale di K per il clustering. Per risolvere questa difficoltà, ho utilizzato l'Elbow method, che consiste nel graficare la Somma degli Errori Quadrati (SSE) in funzione del numero di cluster. Il valore di K ottimale è scelto come il punto di "gomito" della curva risultante, dove si osserva una significativa riduzione della SSE. La SSE rappresenta la somma delle distanze al quadrato di ciascun punto dal proprio centroide all'interno di ogni cluster. Il grafico ottenuto, mostrato in Figura 9, mi ha permesso di determinare il valore ottimale di K per il problema.

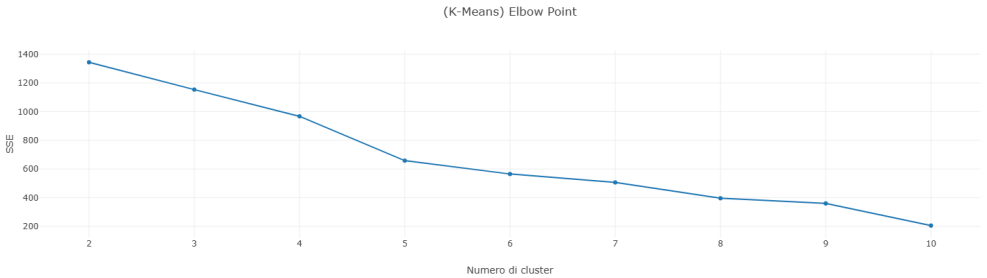

Fig. 9. Grafico che mostra la somma degli errori quadrati (SSE) in funzione del numero di cluster (K) per il metodo Elbow.

A questo punto, ho sviluppato una funzione mostrata in Figura 10 per l'individuazione automatica del K ottimale. L'Elbow point è stato individuato confrontando la variazione di SSE tra i valori $K-1, K$ e $K, K+1$ per ogni K compreso tra 3 e 9. Il valore di K per cui questa variazione è massima è stato considerato come l'Elbow point, che viene restituito dalla funzione.

I risultati ottenuti dall'applicazione dell'algoritmo k-means sono rappresentati in un grafico tridimensionale mostrato in Figura 11.

```
function elbowPoint(dataset,min,max){
    let kmin=min; //valore minimo di k
    let kmax=max; //valore max a cui puo arrivare k
    let sse : any[] =[]; //squared sum estimate
    for(k=kmin;k<=kmax;k++) { //Calcolo l'sse per ogni k
        clusterMaker.k(k);
        clusterMaker.iterations(100);
        clusterMaker.data(dataset);
        let cluster : ({...})[]  = clusterMaker.clusters();
        var distortions : number  = 0;
        for (i = 0; i < k; i++)
            distortions = distortions + sommaDistanze(cluster[i].centroid, cluster[i].points);
        sse.push(distortions);
    }
    // Calcolo elbow point
    deltas = [];
    for (i = 1; i < sse.length - 1; i++){
        delta1 = Math.abs( x: sse[i] - sse[i-1]);
        delta2 = Math.abs( x: sse[i+1] - sse[i]);
        deltas.push(Math.abs( x: delta2-delta1));
    }
    const maximumDelta : number  = Math.max(...deltas);
    const elbowPoint = deltas.indexOf(maximumDelta) + 1 + kmin;
    return elbowPoint;
}
```

Fig. 10. Codice che mostra le istruzioni utilizzate per l'individuazione automatica del K ottimale.

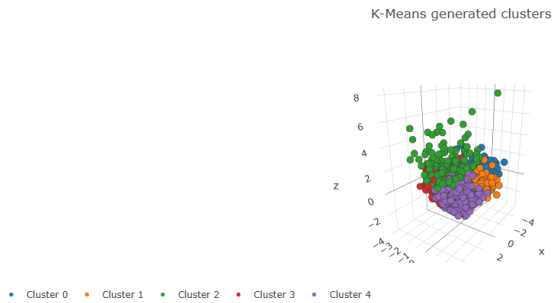

Fig. 11. Rappresentazione tridimensionale del risultato del K-means.

I punti nel grafico sono stati raggruppati in cinque cluster distinti, ciascuno rappresentato con un colore specifico. Nonostante la distribuzione dei punti nello spazio tridimensionale sia piuttosto concentrata, l'algoritmo k-means ha identificato con successo ben cinque regioni. Ogni cluster rappresenta un insieme coerente di punti con caratteristiche simili, mettendo in evidenza la capacità dell'algoritmo di suddividere i dati in modo efficace.

4.3 Valutazione e scelta dell'algoritmo di clustering

Dai risultati e dalle valutazioni eseguite, si è deciso di adottare il K-means come algoritmo principale per il clustering. Il DBScan, nonostante il suo approccio basato sulla densità, ha generato un elevato

numero di elementi esclusi (noise) e ha prodotto pochi cluster, di cui uno predominante con un numero eccessivo di elementi, compromettendo la qualità del raggruppamento. Inoltre, l'algoritmo si è dimostrato particolarmente sensibile alla scelta del parametro Epsilon, che non ha portato a risultati soddisfacenti anche dopo diversi tentativi di ottimizzazione.

Al contrario, il K-means ha permesso di identificare un numero ottimale di cluster utilizzando l'Elbow method, il quale ha garantito una suddivisione più equilibrata e accurata dei dati. Grazie alla riduzione significativa della somma degli errori quadrati (SSE) e al supporto di un processo iterativo ben definito, il K-means ha mostrato una maggiore capacità di creare cluster omogenei e rappresentativi rispetto al DBScan. Questi vantaggi hanno reso il K-means la scelta più adeguata per il problema affrontato.

5 Soluzione

In base alle valutazioni precedenti, si è deciso di combinare PCA e K-means. Questa scelta ha consentito una riduzione della dimensionalità e un raggruppamento più efficace dei dati.

5.1 Valutazione

I vari cluster non sono più stati rappresentati su un grafico tridimensionale, ma su un grafico radar per ottenere una lettura più immediata e accurata dei valori medi delle features per ciascun cluster.

Fig. 12. Grafico radar che rappresenta i valori medi delle features per ciascun cluster.

Il grafico risultante è mostrato in Figura 12. Osservandolo possiamo notare ad esempio che per il cluster 0, è evidente che i valori di Acousticness (Acustica) sono significativamente più alti rispetto agli altri cluster. Questo dato conferma che al suo interno sono presenti canzoni caratterizzate dalla presenza di strumenti musicali il cui suono è meno distorto da effetti elettronici. In modo simile, esaminando il cluster 2, si nota un valore elevato 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 durante l'esecuzione della traccia. Questi risultati sono in linea con le aspettative e confermano che il clustering ha raggruppato in modo appropriato i brani.

Infine, ho deciso di rappresentare i valori medi delle features dei cluster attraverso dei grafici a barre, al fine di ottenere un'ulteriore rappresentazione bidimensionale dell'andamento delle features in ciascun cluster. Questa visualizzazione mostra in modo chiaro e dettagliato le differenze tra i cluster. Nelle Figure 13, 14, 15, 16 e 17, sono riportati a titolo di esempio solo i grafici a barre per la feature Beats Per Minute (BPM). Tuttavia, grafici analoghi sono stati generati anche per le altre features; per consultarli, si rimanda al repository GitHub indicato all'inizio della documentazione.

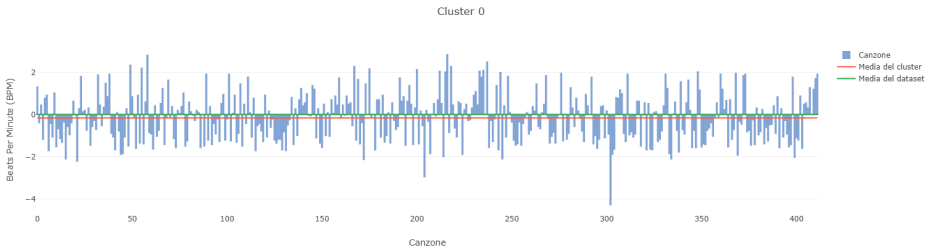

Fig. 13. Grafico a barre dei valori medi delle features BPM per il cluster 0.

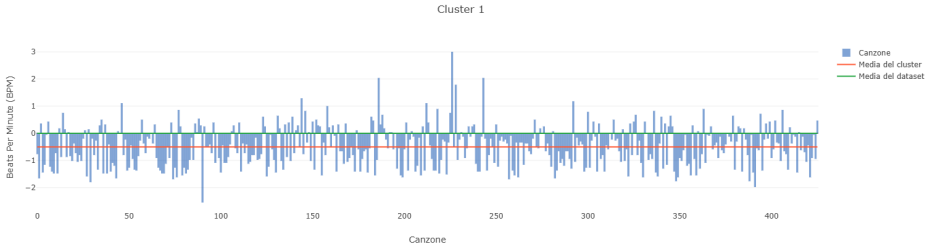

Fig. 14. Grafico a barre dei valori medi delle features BPM per il cluster 1.

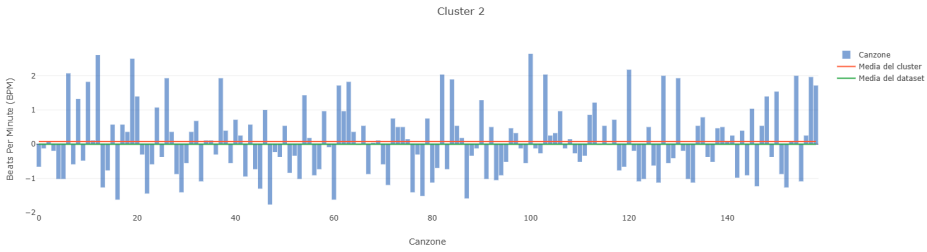

Fig. 15. Grafico a barre dei valori medi delle features BPM per il cluster 2.

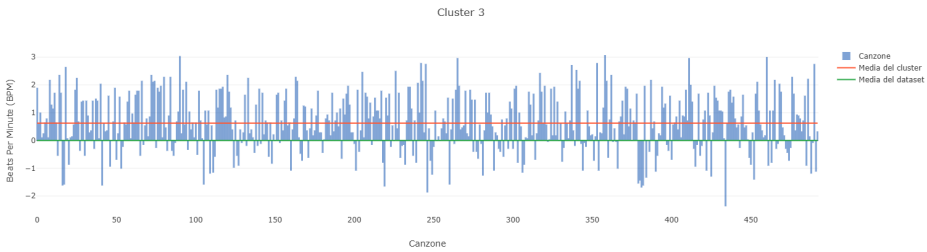

Fig. 16. Grafico a barre dei valori medi delle features BPM per il cluster 3.

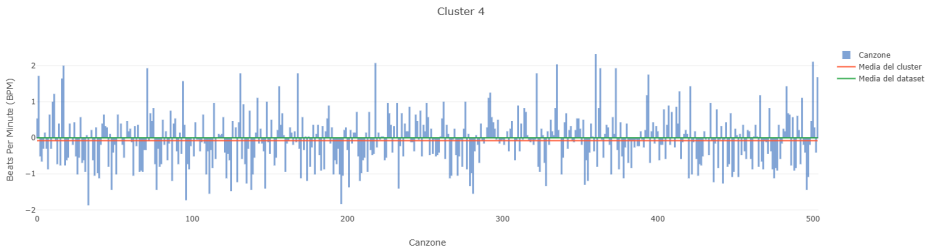

Fig. 17. Grafico a barre dei valori medi delle features BPM per il cluster 4.

6 Risultati

Il processo di clustering ha generato un totale di 5 cluster di canzoni, ciascuno trasformato in una playlist contenente i relativi brani. Per motivi di spazio, le Figure 18, 19, 20, 21 e 22 presentano una selezione parziale delle playlist. L'elenco completo delle canzoni associate a ciascuna playlist è disponibile nel repository GitHub indicato all'inizio della documentazione. Le figure includono i titoli delle canzoni e i rispettivi artisti.

```
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
8   Titolo: Goodbye My Lover, Artista: James Blunt, Top Genere: neo mellow
9   Titolo: Hurt, Artista: Christina Aguilera, Top Genere: dance pop
10   Titolo: Make You Feel My Love, Artista: Adele, Top Genere: british soul
11   Titolo: Ne me quitte pas, Artista: Jacques Brel, Top Genere: chanson
12   Titolo: W�r Bisto, Artista: Twarres, Top Genere: dutch pop
13   Titolo: Only Time, Artista: Enya, Top Genere: celtic
14   Titolo: De Vondeling Van Ameland, Artista: Boudewijn de Groot, Top Genere: dutch indie
15   Titolo: Blaosmuziek, Artista: G� Reinders, Top Genere: carnaval limburg
16   Titolo: Halt mich, Artista: Herbert Gr�nemeyer, Top Genere: german pop
17   Titolo: Empire State of Mind (Part II) Broken Down, Artista: Alicia Keys, Top Genere: hip pop
18   Titolo: Feeling Good, Artista: Muse, Top Genere: modern rock
19   Titolo: Voltooid Verleden Tijd, Artista: IOS, Top Genere: dutch pop
20   Titolo: Red Mij Niet - Single Edit, Artista: Maarten Van Roozendaal, Top Genere: dutch indie
```

Fig. 18. Playlist corrispondente al cluster 0: elenco di canzoni con i rispettivi artisti.

7 Conclusioni

Il progetto presentato è stato sviluppato facendo leva sulle conoscenze acquisite durante il corso di Fondamenti di Intelligenza Artificiale. In particolare, l'approccio di clustering applicato alle canzoni è stato guidato dalla volontà di esplorare nuove modalità di organizzazione dei dati attraverso metodi analitici avanzati. La scelta di utilizzare PCA e K-means, in particolare, è stata dettata dalla necessità di ottenere una rappresentazione più chiara e strutturata dei brani musicali, sfruttando le caratteristiche condivise tra le canzoni.

```
Playlist 1:
412   Titolo: The Road Ahead (Miles Of The Unknown), Artista: City To City, Top Genere: alternative pop rock
413   Titolo: Cry Me a River, Artista: Justin Timberlake, Top Genere: dance pop
414   Titolo: Miracle, Artista: Ilse DeLange, Top Genere: dutch americana
415   Titolo: Chasing Pavements, Artista: Adele, Top Genere: british soul
416   Titolo: Don't Know Why, Artista: Norah Jones, Top Genere: adult standards
417   Titolo: The Eyes of Jenny, Artista: Sandy Coast, Top Genere: classic uk pop
418   Titolo: De Neus Umhoeg, Artista: Rowwen H♦ze, Top Genere: carnaval limburg
419   Titolo: Solitary Man, Artista: Johnny Cash, Top Genere: arkansas country
420   Titolo: Maybe Tomorrow, Artista: Stereophonics, Top Genere: britpop
421   Titolo: Sorry Seems To Be The Hardest Word, Artista: Blue, Top Genere: boy band
422   Titolo: Rosie, Artista: Claw Boys Claw, Top Genere: dutch indie
423   Titolo: Eternal Flame, Artista: The Bangles, Top Genere: album rock
424   Titolo: Daughters, Artista: John Mayer, Top Genere: neo mellow
425   Titolo: Sadeness, Artista: Enigma, Top Genere: downtempo
426   Titolo: Old Man, Artista: Neil Young, Top Genere: album rock
427   Titolo: Black Magic Woman, Artista: Santana, Top Genere: blues rock
428   Titolo: Set The Fire To The Third Bar, Artista: Snow Patrol, Top Genere: irish rock
429   Titolo: Is Dit Alles, Artista: Doe Maar, Top Genere: dutch cabaret
430   Titolo: Heartbreak Warfare, Artista: John Mayer, Top Genere: neo mellow
431   Titolo: One Word, Artista: Anouk, Top Genere: dutch indie
432   Titolo: Sweet Goodbyes, Artista: Krezip, Top Genere: dutch indie
```

Fig. 19. Playlist corrispondente al cluster 1: elenco di canzoni con i rispettivi artisti.

```
Playlist 2:
838   Titolo: Als Het Golft, Artista: De Dijk, Top Genere: dutch indie
839   Titolo: I'm going home, Artista: Ten Years After, Top Genere: album rock
840   Titolo: Tears Dry On Their Own, Artista: Amy Winehouse, Top Genere: british soul
841   Titolo: Sweet Jane, Artista: Lou Reed, Top Genere: art rock
842   Titolo: Grounds for Divorce, Artista: Elbow, Top Genere: britpop
843   Titolo: Family Portrait, Artista: P!nk, Top Genere: dance pop
844   Titolo: The Wild Rover, Artista: The Dubliners, Top Genere: celtic
845   Titolo: Wat Zou Je Doen - Live, Artista: Marco Borsato, Top Genere: dutch cabaret
846   Titolo: Limburg - Live, Artista: Rowwen H♦ze, Top Genere: carnaval limburg
847   Titolo: Numb / Encore, Artista: JAY-Z, Top Genere: east coast hip hop
848   Titolo: Lose Yourself - From "8 Mile" Soundtrack, Artista: Eminem, Top Genere: detroit hip hop
849   Titolo: Harder, Better, Faster, Stronger, Artista: Daft Punk, Top Genere: electro
850   Titolo: Single Ladies (Put a Ring on It), Artista: Beyonc♦, Top Genere: dance pop
851   Titolo: You're the Voice, Artista: John Farnham, Top Genere: australian pop
852   Titolo: Crazy In Love (feat. Jay-Z), Artista: Beyonc♦, Top Genere: dance pop
853   Titolo: De wedstrijd, Artista: Bram Vermeulen, Top Genere: belgian rock
854   Titolo: Sonne, Artista: Rammstein, Top Genere: alternative metal
855   Titolo: It's Raining Men, Artista: The Weather Girls, Top Genere: disco
856   Titolo: Zing - Vecht - Huil - Bid - Lach - Werk En Bewonder, Artista: Ramses Shaffy, Top Genere: dutch cabaret
857   Titolo: Hollereer, Artista: De Jeugd Van Tegenwoordig, Top Genere: dutch hip hop
858   Titolo: Go With The Flow, Artista: Queens of the Stone Age, Top Genere: alternative metal
```

Fig. 20. Playlist corrispondente al cluster 2: elenco di canzoni con i rispettivi artisti.

Il processo di clustering ha permesso di creare playlist ben definite, ognuna delle quali rispecchia un gruppo di brani con caratteristiche comuni.

L'approccio analitico e iterativo adottato ha portato a un risultato che, oltre a raggruppare le canzoni in modo efficace, offre una solida base per futuri sviluppi. Le numerose sperimentazioni e ottimizzazioni eseguite durante il processo hanno confermato l'efficacia del clustering nel risolvere il problema, evidenziando però alcune aree che potrebbero essere ulteriormente perfezionate in fasi future.

```
Playlist 3:
997   Titolo: The Pretender, Artista: Foo Fighters, Top Genere: alternative metal
998   Titolo: Knights of Cydonia, Artista: Muse, Top Genere: modern rock
999   Titolo: Mr. Brightside, Artista: The Killers, Top Genere: modern rock
1000   Titolo: Speed of Sound, Artista: Coldplay, Top Genere: permanent wave
1001   Titolo: Uninvited, Artista: Alanis Morissette, Top Genere: alternative rock
1002   Titolo: Fix You, Artista: Coldplay, Top Genere: permanent wave
1003   Titolo: The Cave, Artista: Mumford & Sons, Top Genere: modern folk rock
1004   Titolo: Smokers Outside the Hospital Doors, Artista: Editors, Top Genere: alternative dance
1005   Titolo: Big Log - 2006 Remaster, Artista: Robert Plant, Top Genere: album rock
1006   Titolo: Iris, Artista: The Goo Goo Dolls, Top Genere: alternative rock
1007   Titolo: The Saints Are Coming, Artista: U2, Top Genere: irish rock
1008   Titolo: All My Life, Artista: Foo Fighters, Top Genere: alternative metal
1009   Titolo: Traffic - Radio Edit, Artista: Ti◆sto, Top Genere: big room
1010   Titolo: Take Me Out, Artista: Franz Ferdinand, Top Genere: alternative rock
1011   Titolo: American Idiot, Artista: Green Day, Top Genere: modern rock
1012   Titolo: Come Undone, Artista: Robbie Williams, Top Genere: dance pop
1013   Titolo: Run, Artista: Snow Patrol, Top Genere: irish rock
1014   Titolo: The Day After Tomorrow, Artista: Saybia, Top Genere: danish pop rock
1015   Titolo: Dansen Aan Zee, Artista: BL◆F, Top Genere: dutch pop
1016   Titolo: The Unforgiven III, Artista: Metallica, Top Genere: alternative metal
1017   Titolo: Sometimes You Can't Make It On Your Own, Artista: U2, Top Genere: irish rock
1018   Titolo: Listen (From the Motion Picture "Dreamgirls"), Artista: Beyonc◆, Top Genere: dance pop
```

Fig. 21. Playlist corrispondente al cluster 3: elenco di canzoni con i rispettivi artisti.

```
Playlist 4:
1490   Titolo: Black Night, Artista: Deep Purple, Top Genere: album rock
1491   Titolo: Clint Eastwood, Artista: Gorillaz, Top Genere: alternative hip hop
1492   Titolo: Waitin' On A Sunny Day, Artista: Bruce Springsteen, Top Genere: classic rock
1493   Titolo: She Will Be Loved, Artista: Maroon 5, Top Genere: pop
1494   Titolo: Without Me, Artista: Eminem, Top Genere: detroit hip hop
1495   Titolo: Seven Nation Army, Artista: The White Stripes, Top Genere: alternative rock
1496   Titolo: Fluorescent Adolescent, Artista: Arctic Monkeys, Top Genere: garage rock
1497   Titolo: Als De Morgen Is Gekomen, Artista: Jan Smit, Top Genere: dutch pop
1498   Titolo: Somebody Told Me, Artista: The Killers, Top Genere: modern rock
1499   Titolo: Dichterbij Dan Ooit, Artista: BL◆F, Top Genere: dutch pop
1500   Titolo: Cleanin' Out My Closet, Artista: Eminem, Top Genere: detroit hip hop
1501   Titolo: 7 Seconds (feat. Neneh Cherry), Artista: Youssou N'Dour, Top Genere: afropop
1502   Titolo: Don't Let Me Be Misunderstood, Artista: Santa Esmeralda, Top Genere: disco
1503   Titolo: Breaking the Habit, Artista: Linkin Park, Top Genere: alternative metal
1504   Titolo: You're The First, The Last, My Everything, Artista: Barry White, Top Genere: adult standards
1505   Titolo: I Want It That Way, Artista: Backstreet Boys, Top Genere: boy band
1506   Titolo: Smoorverliefd, Artista: Doe Maar, Top Genere: dutch cabaret
1507   Titolo: Me Gustas Tu, Artista: Manu Chao, Top Genere: latin alternative
1508   Titolo: Kryptonite, Artista: 3 Doors Down, Top Genere: alternative metal
1509   Titolo: Crazy, Artista: Seal, Top Genere: british soul
1510   Titolo: In The Army Now, Artista: Status Quo, Top Genere: album rock
```

Fig. 22. Playlist corrispondente al cluster 4: elenco di canzoni con i rispettivi artisti.

7.1 Sviluppi futuri

Un miglioramento significativo potrebbe consistere nello sviluppo di un modulo aggiuntivo che consenta l'integrazione diretta con Spotify. Questo modulo potrebbe estrarre automaticamente le canzoni dalla sezione preferiti dell'account Spotify dell'utente e salvarle in un file. Tale file poi sarà elaborato dall'agente. In questo modo, l'agente potrebbe interagire con l'utente in maniera più fluida e semplice, creando playlist dinamiche basate sulle preferenze musicali dell'utente. Tale

approccio migliorerebbe l'esperienza dell'utente, rendendo il sistema più facile da usare e capace di adattarsi alle preferenze dell'utente.

7.2 Considerazioni finali

In conclusione, sono pienamente soddisfatto dei risultati ottenuti, che non solo hanno dimostrato la validità dell'approccio scelto, ma hanno anche arricchito la mia esperienza pratica nel campo dell'intelligenza artificiale. Questo progetto ha offerto l'opportunità di applicare concetti teorici in un contesto concreto, facendo emergere soluzioni pratiche ed efficienti.