Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione
Corso di Laurea Magistrale in Computer Science Engineering (Information Systems)

AIML/BDA Final Writeup

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Students:

Pietro Urso
Ernesto Petroni

Professors:

Tommaso Di Noia
Vito Walter Anelli
Antonio Ferrara

A.A. 2021/2022

Summary

1. Abstract

Il topic scelto per il Progetto è *"Tell Me How to Continue My Playlist: Sequential Recommendation Techniques".* L'obiettivo del project work è l'ideazione di un sistema di raccomandazione di Machine Learning che, data in ingresso una canzone, produca in uscita una playlist di canzoni consigliate. Il sistema di raccomandazione fornisce il risultato richiesto mediante l'utilizzo delle tecniche di apprendimento non supervisionato.

2. Introduction

Quando si parla di sistemi di raccomandazione, si fa riferimento a sistemi che vengono usati nella vita di ogni giorno. L'obiettivo di questi sistemi è aiutare l'utente nella ricerca delle informazioni di interesse all'interno della immensa mole di dati generati, che si hanno a disposizione. Questo avviene considerando le preferenze/comportamenti passati degli utenti.

Nel sistema di raccomandazione in oggetto, il problema che si vuole risolvere, è legato alla predizione di un set di items (canzoni) generato a partire dalla similarità calcolata tra la canzone in input e tutte le rimanenti canzoni (items) a disposizione.

La similarità tra due oggetti può essere calcolata in vari modi; tra le principali soluzioni abbiamo la distanza euclidea, la Manhattan distance e la cosine similarity.

Durante lo sviluppo del progetto sono state utilizzate delle tecniche di unsupervised learning per il clustering dei datapoints. Nello specifico gli algoritmi *k-means*, *k-medoids*, *DBSCAN* e *Gaussian Mixture Model* hanno consentito la costruzione della playlist delle canzoni consigliate a partire dalla canzone in riproduzione (input).

4. Dataset and Features

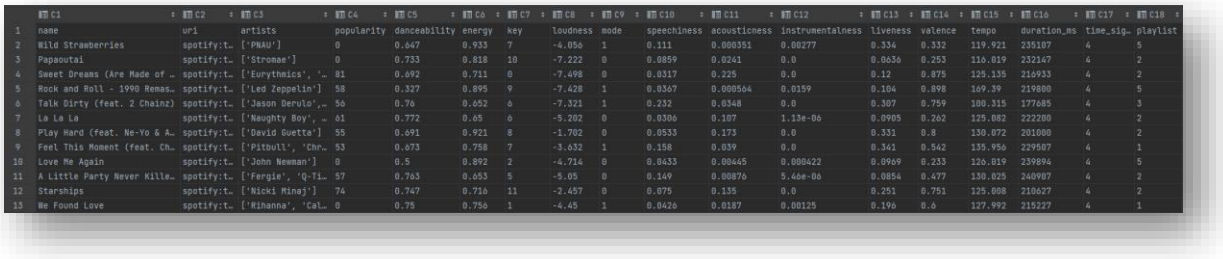
Il **dataset** utilizzato per questo progetto contiene le informazioni inerenti ad un insieme di canzoni pubblicate nel panorama musicale degli ultimi anni. Il numero di sample ovvero di canzoni, è pari a circa 56.000 e il numero di features è pari a 18.

Per quanto riguarda le **features** di ciascuna canzone, esse sono sia numeriche che categoriche. Esse vengono riportare nel dettaglio di seguito.

- **Name**: indica il nome della canzone pubblicata;
- **URI**: indica il collegamento alla traccia di Spotify della canzone;
- **Artists**: indica l'elenco degli autori della canzone;
- **Popularity**: indica, tramite un valore numerico compreso tra 0 e 100, la popolarità della canzone;
- **Danceability**: indica, tramite un valore compreso tra 0.0 e 0.99, la danzabilità della canzone;
- **Energy**: indica, tramite un valore compreso tra 0.0 e 1.0, l'energia trasmessa dalla canzone;
- **Key**: è un valore compreso tra 0 e 11;
- **Loudness:** valore compreso tra -9.99 e -0.1;
- **Mode:** è un valore binario (0 o 1);
- **Speechiness:** valore compreso tra 0.0 e 0.99;
- **Acusticness:** valore compreso tra 0.0 e 0.99;
- **Instrumentalness:** valore compreso tra 0.0 e 0.99;
- **Liveness:** valore compreso tra 0.0 e 1.0;

- **Valence:** valore compreso tra 0.0 e 1.0;
- **Tempo:** indica la durata in secondi della canzone;
- **Duration_ms:** indica la durata in millisecondi della canzone;
- **Time signature:** valore compreso tra 0 e 5;
- **Playlist:** indica la playlist di appartenenza teorica della canzone (valore compreso tra 1 e 7).

Si riporta di seguito una porzione esemplificativa del dataset:



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
1	name	uri	artists	popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_sig…	playlist
2	Wild Strawberries	spotify:t…	['PNAU']	0	0.647	0.933	7	-4.056	1	0.111	0.000351	0.00277	0.334	0.332	119.921	235107	4	5
3	Papaoutai	spotify:t…	['Stromae']	0	0.733	0.818	10	-7.222	0	0.0859	0.0241	0.0	0.0636	0.253	116.019	232147	4	2
4	Sweet Dreams (Are Made of …	spotify:t…	['Eurythmics', '…	81	0.692	0.711	0	-7.498	0	0.0317	0.225	0.0	0.12	0.875	125.135	216933	4	2
5	Rock and Roll - 1990 Remas…	spotify:t…	['Led Zeppelin']	58	0.327	0.895	9	-7.428	1	0.0367	0.000564	0.0159	0.104	0.898	169.39	219800	4	5
6	Talk Dirty (feat. 2 Chainz)	spotify:t…	['Jason Derulo',…	56	0.76	0.652	6	-7.321	1	0.232	0.0348	0.0	0.307	0.759	100.315	177685	4	3
7	La La La	spotify:t…	['Naughty Boy', …	61	0.772	0.65	6	-5.202	0	0.0306	0.107	1.13e-06	0.0905	0.262	125.082	222200	4	2
8	Play Hard (feat. Ne-Yo & A…	spotify:t…	['David Guetta']	55	0.691	0.921	8	-1.702	0	0.0533	0.173	0.0	0.331	0.8	130.072	201080	4	2
9	Feel This Moment (feat. Ch…	spotify:t…	['Pitbull', 'Chr…	53	0.673	0.758	7	-3.632	1	0.158	0.039	0.0	0.341	0.542	135.956	229507	4	1
10	Love Me Again	spotify:t…	['John Newman']	0	0.5	0.892	2	-4.714	0	0.0433	0.00445	0.000422	0.0969	0.233	126.019	239894	4	5
11	A Little Party Never Kille…	spotify:t…	['Fergie', 'Q-Ti…	57	0.763	0.653	5	-5.05	0	0.149	0.00876	5.46e-06	0.0854	0.477	130.025	240907	4	2
12	Starships	spotify:t…	['Nicki Minaj']	74	0.747	0.716	11	-2.457	0	0.075	0.135	0.0	0.251	0.751	125.008	210627	4	2
13	We Found Love	spotify:t…	['Rihanna', 'Cal…	0	0.75	0.756	1	-4.45	1	0.0426	0.0187	0.00125	0.196	0.6	127.992	215227	4	1

Immagine 1 – Porzione del dataset utilizzato

Il dataset completo è reperibile al seguente link:
https://drive.google.com/file/d/1Q9CRt4i2pRBMMGOGjpGdnrPKoTHxqHqw/view?usp=sharing

5. Methods

Le tecniche di Machine Learning che si intendono utilizzare per implementare il sistema di raccomandazione sono tecniche di **Unsupervised Learning**. Si è scelto di adoperare questo approccio in quanto l'obiettivo dello studio è clusterizzare, ovvero identificare gruppi di possibili playlists dal dataset originale. Questi clusters sono composti da canzoni con caratteristiche simili. Pertanto, trovare delle similarità tra i datapoints, in modo tale da ricercare caratteristiche comuni tra loro, motiva la scelta fatta rispetto agli algoritmi utilizzati.
A partire dalla canzone in ingresso, quindi, si identifica il cluster di appartenenza e lo si utilizza successivamente per scegliere le n canzoni più vicine per caratteristiche, rispetto alla canzone in riproduzione (ovvero la canzone in input). La selezione delle n canzoni più vicine viene effettuata calcolando la similarità tra la canzone in input e tutte le altre canzoni presenti nel cluster. Nello specifico la misura di similarità adoperata è la distanza Euclidea:

$$d(x,y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Per quanto concerne gli algoritmi di clustering implementati, sono stati utilizzati il **K-means**, il **K-medoids**, il **DBSCAN** e il **Gaussian Mixture Model**.

5.1 K-Means

L'idea del k-means è trovare le k centroidi, i punti centrali dei k clusters, ovvero i punti che hanno in media una migliore distanza tra tutti gli altri datapoints all'interno dei clusters. I datapoints che sono più vicini ad una specifica centroide appartengono al cluster rappresentato da quella centroide. In questo modo clusterizziamo tutti i datapoints. Per la scelta del miglior k si è utilizzato l'*elbow method*, che fornisce il giusto compromesso tra il numero di cluster e il costo del clustering.

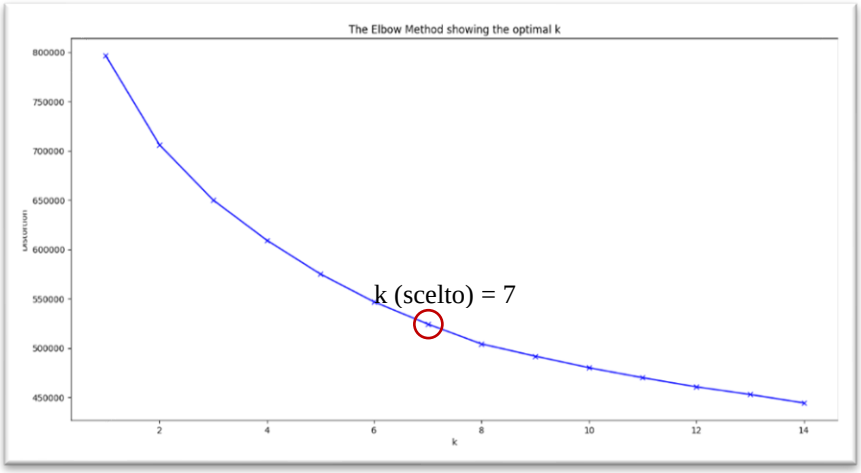


Immagine 2 – Elbow Method (k-means)

5.2 K-Medoids

Alcuni ricercatori hanno deciso di proporre un nuovo algoritmo di clustering chiamato K-medoids. L'obiettivo è eliminare l'idea di generare le centroidi non legate ai datapoints, come avviene nel K-means. Il K-medoids propone di usare i datapoints per rappresentare le centroidi dei clusters. Per fare ciò ci sono diversi algoritmi, il più usato è il PAM (Partitioning Around Medoids) che partiziona i datapoints tra le medoids trovando il miglior clustering a partire da queste medoids. Questo algoritmo candida ciascun datapoint come possibile medoide, calcolandone il costo. La configurazione con il costo minore rappresenterà il clustering finale. Anche in questo caso, per la scelta del numero di clusters k, si è applicato l'*elbow method.*

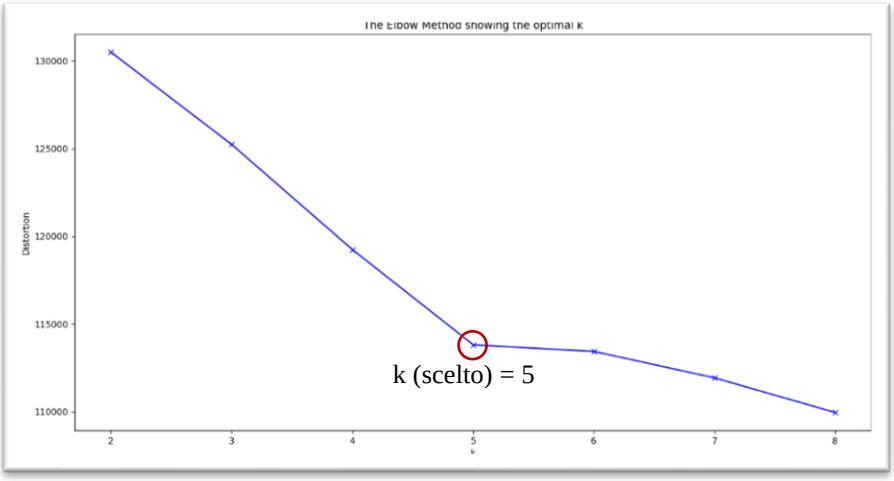


Immagine 2 – Elbow Method (k-medoids)

5.3 DBSCAN

Il DBSCAN è un algoritmo agglomerativo che si basa sul concetto di densità. L'obiettivo di questo algoritmo è quello di trovare le regioni, nello spazio dei dati, con alta densità di datapoints e separarle dalle regioni con una bassa densità. La densità viene definita attraverso un valore, chiamato *eps*, in cui sulla base del valore che si definisce, si specifica il raggio che consente di tracciare virtualmente una sfera o cerchio (spazio bidimensionale). All'interno di questa sfera si può calcolare il numero di datapoints che sono al suo interno.

Nel DBSCAN, un ulteriore concetto importante è legato al numero minimo di punti (MinPts) che si vogliono considerare per formare un cluster. Pertanto, è necessario impostare due parametri:

- **Epsilon**: distanza fisica o raggio;
- **MinPts:** numero minimo di punti necessari per formare un cluster.

La scelta del parametro Epsilon può essere fatta utilizzando il k-distance graph mentre il valore del MinPts può essere scelto attraverso la seguente regola empirica:

$$minPts \leq numero\_featres + 1$$

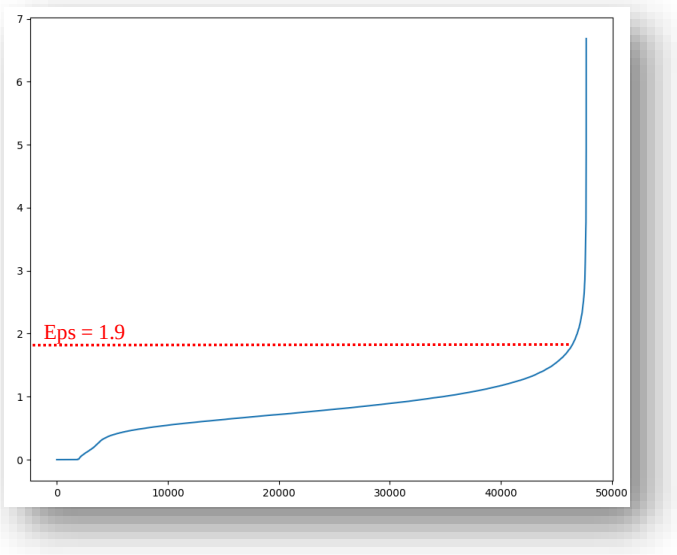


Immagine 3 – k-distance graph

5.4 Gaussian Mixture Model

Alcune distribuzioni complesse richiedono sistemi più sofisticati per il calcolo dei nuovi centri dei cluster. Il Gaussian Mixture Models (GMMs) hanno quindi una flessibilità superiore al K-Means, K-Medoids e DBSCAN consentendo di elaborare correttamente anche le distribuzioni più insidiose.

In questo metodo si assume che la distribuzione delle osservazioni sia Gaussiana, un'assunzione meno restrittiva rispetto a considerarli circolari e impiegare la media per il calcolo dei clusters.

In questo modo i parametri che descrivono la forma dei cluster sono due:

- media
- deviazione standard

La forma che i cluster possono assumere si estende ora a qualsiasi ellisse.
Se il numero delle componenti k è noto, la tecnica più comunemente usata per stimare i parametri del Mixture Model è l'Expectation-Maximization (EM). L'EM è una tecnica numerica, basata su un algoritmo iterativo, per massimizzare la stima della likelihood, pertanto minimizzando il costo del clustering.

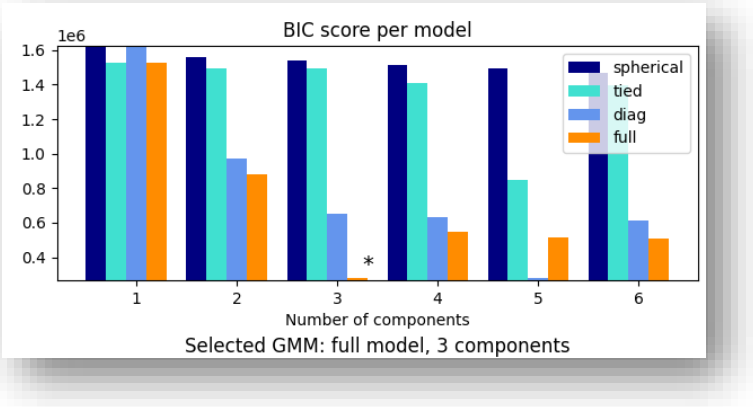


Immagine 4 – BIC score per la scelta del n° di clusters

6. Experiments and Results

6.1 Preprocessing e Normalizzazione

La prima fase dello studio si è concentrata sulla lettura del dataset in formato csv. Successivamente è stata condotta un'analisi del dataset (gestione missing values), seguita da una fase di preprocessing (rimozione di specifici caratteri), da una fase di normalizzazione dei dati (z-score normalizzation) e dalla rimozione dei samples duplicati. Infine, la fase di preprocessing è stata completata attraverso la features selection, valutando l'IG (Information Gain) associato a ciascuna feature.

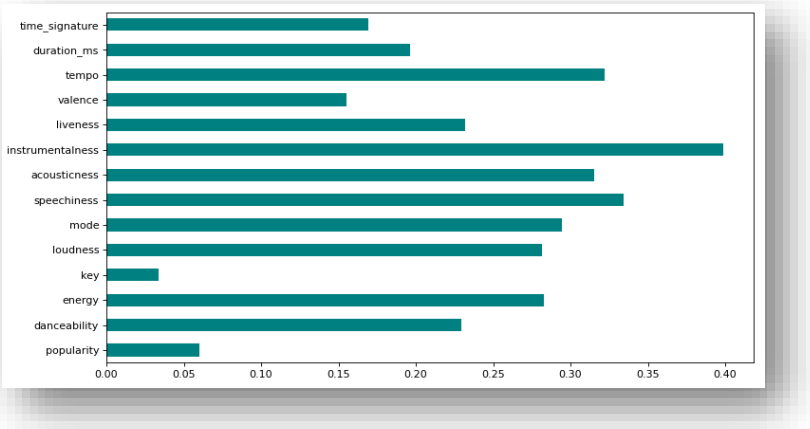


Immagine 5 – Feature importance (Information Gain)

6.2 Calcolo parametri ottimali degli algoritmi di Clustering

Prima dell'implementazione degli algoritmi di clustering, sono state sviluppate delle procedure per il calcolo ottimale dei parametri necessari per l'implementazione dei vari algoritmi. Nello specifico:

- Calcolo del numero k di clusters ottimale, mediante l'utilizzo dell'*elbow method*, per gli algoritmi K-Means e K-Medoids.

- Calcolo del valore ottimale di *eps*, attraverso la costruzione del k-distance graph, per l'algoritmo DBSCAN.

- Utilizzo della regola empirica: $minPts \leq numero\_featres + 1$ per il calcolo del valore ottimale del parametro *minPts* per l'algoritmo DBSCAN.

- Valutazione mediante il Bayesian Information Criterion (*BIC*) per la scelta del numero di componenti ottimali (numero di clusters) per l'algoritmo di Gaussian Mixture Model.

6.3 Implementazione degli algoritmi di Clustering

Sono state sviluppate quattro funzioni, una per ogni algoritmo di clustering, che implementano la procedura di clusterizzazione dei datapoints.
Ciascuna di queste funzioni, a seguito dell'applicazione dell'algoritmo, si occupa dell'aggiunta di un'etichetta (feature aggiuntiva), che identifica il cluster di appartenenza per ciascun sample.
Queste funzioni ritorneranno in uscita il dataset originale, con l'aggiunta dell'informazione frutto della clusterizzazione.

6.4 La classe SongRecommender

La classe SongRecommender() ha l'obiettivo di restituire la playlist consigliata sulla base della canzone in input.
Nello specifico si occupa inizialmente di estrarre tutte le canzoni che appartengono allo stesso cluster della canzone fornita in input. Successivamente vi è il calcolo del valore di similarità, espressa mediante la distanza euclidea (indicatore di correlazione), tra la canzone in riproduzione (input) e ciascuna delle altre canzoni precedentemente estratte.

Questo valore calcolato viene aggiunto come feature all'interno della lista delle canzoni che appartengono allo stesso cluster della canzone fornita in input.
A seguito del calcolo delle distanze, vi è un ordinamento decrescete della lista, seguita dalla selezione delle prime n canzoni che formeranno la playlist consigliata, restituita come output.

6.5 Valutazione dei risultati

All'interno del dataset originale è presente la feature playlist; essa indica la playlist di riferimento di appartenenza di ogni sample ed ha un valore compreso tra 1 e 7. Questa informazione viene utilizzata per valutare l'error rate e l'accuracy sulla base del risultato ottenuto a partire dalla canzone in riproduzione.

- L'**error rate** è calcolato come:

$$error\ rate = \frac{n°\ miss\ classified\ samples}{total\ number\ of\ playlist\ samples}$$

- L'**accuracy** è calcolata come:

$$accuracy = 1 - error\ rate$$

6.6 Data visualizzation

A causa dell'elevato numero di features, per la visualizzazione dei sample clusterizzati, si è scelto di utilizzare l'algoritmo di dimensionality reduction t-SNE.
Il problema iniziale è stato trasformato nell'equivalente bidimensionale, passando quindi da un numero elevato di features (dimensioni) a solo due dimensioni.
Tale trasformazione ha consentito una visualizzazione grafica dei datasamples opportunamente clusterizzati secondo la label definita dall'algoritmo di clustering applicato.

6.7 Results

La canzone in riproduzione (input) scelta per la playlist recommendation è 'Call me maybe', mentre, i diversi algoritmi sono stati valutati sulla base dell'***accuracy*** e dell'**error rate**, calcolati a partire da una playlist consigliata contenente 25 canzoni.
Di seguito sono riportati i risultati dei vari esperimenti, che includono:

- *Scelta dei parametri ottimali*;
- *Visualizzazione dei clusters*;
- *Playlist consigliata*;
- *Parametri di valutazione*.

6.7.1 K-Means

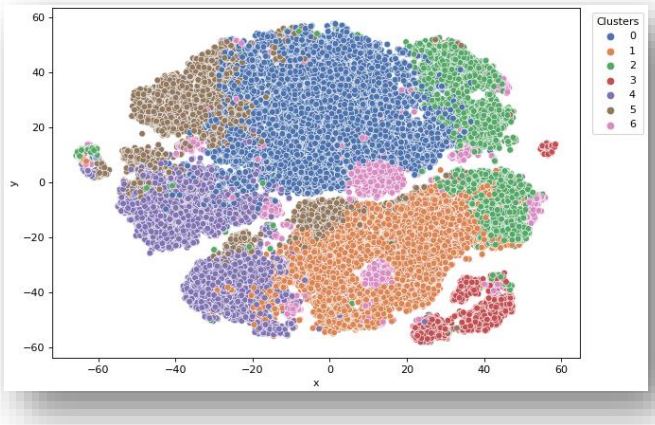


Immagine 6 – Visualizzazione (t-SNE) clustering k-means

```
La canzone in riproduzione (input) è:  Call me maybe
                                                  name                                                artists
5870                                       Pay My Rent                                                DNCE
6635                                              Sexy                                            JoeVille
9335                                         Supernova                                  Mr Hudson, Kanye West
13339                        I Saw Her Again Last Night                                The Mamas & The Papas
29645                                        Polaroid                  Jonas Blue, Liam Payne, Lennon Stella
1561                                         Last Nite                                         The Strokes
169                                Crazy World - Radio Edit                                DJ Antoine, Mad Mark
9866          The Steady Song (feat. Isabel Reyes Feeney)                                   Republic Of Loose
43539                   NO MERCY (feat. Lil Wayne, Ph4de)  "Its Different", Forever M.C., Lil Wayne, PH4DE
1451   One Night In Bangkok - From "Chess" / Remaster...                                       Murray Head
```

Immagine 7 – Playlist consigliata contenente le prime 10 canzoni.

```
Il numero di miss classified sample è:  6
L'error rate sulle prime  25 canzoni è pari al  24.0 %
L'accuracy sulle prime  25 canzoni è pari al  76.0 %
```

Immagine 8 – Parametri di valutazione k-means

6.7.2 K-Medoids

Per quanto riguarda l'algoritmo k-medoids, a causa dell'elevata capacità computazionale richiesta, in termini di memoria necessaria per l'applicazione dell'algoritmo, è stata effettuata una riduzione del numero di samples presi in esame, da circa 47.000 iniziali a

40.000 utilizzati. Di conseguenza, i risultati sono stati ottenuti a partire da questo numero ridotto di canzoni.

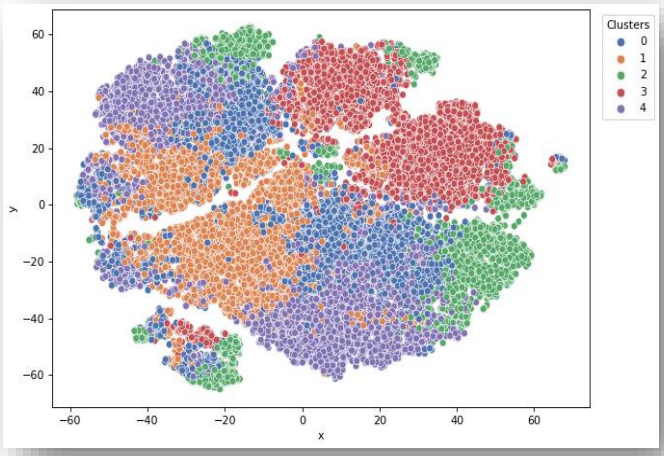


Immagine 9 – Visualizzazione (t-SNE) clustering k-medoids

```
La canzone in riproduzione (input) è:  Call me maybe
                                                  name                          artists
5870                                       Pay My Rent                            DNCE
6635                                               Sexy                        JoeVille
9335                                          Supernova            Mr Hudson, Kanye West
1561                                          Last Nite                      The Strokes
9066          The Steady Song (feat. Isabel Reyes Feeney)                Republic Of Loose
1451   One Night In Bangkok - From "Chess" / Remaster...                      Murray Head
26854                          Don't Stop - 2004 Remaster                  Fleetwood Mac
12970                              Does Your Mother Know                             ABBA
12024                                    Beautiful Noise                      Neil Diamond
27640                               Kiss Me Thru The Phone                Soulja Boy, Sammie
```

Immagine 10 – Playlist consigliata contenente le prime 10 canzoni.

```
Il numero di miss classified sample è:  8
L'error rate sulle prime  25 canzoni è pari al  32.0 %
L'accuracy sulle prime  25 canzoni è pari al  68.0 %
```

Immagine 11 – Parametri di valutazione k-Medoids

6.7.3 DBSCAN

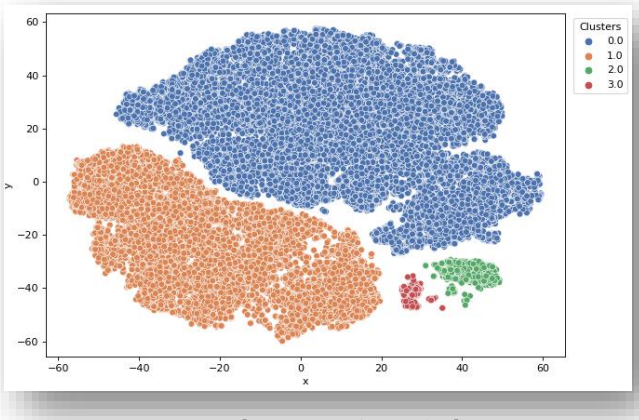


Immagine 12 – Visualizzazione (t-SNE) clustering DBSCAN

```
La canzone in riproduzione (input) è:  Call me maybe
                                                  name                          artists
5870                                       Pay My Rent                            DNCE
23150  Rnw@y (Backyard Bangers Reanimation) [feat. Ph...  Linkin Park, Phoenix Orion
9396                                              Zero                         Alpines
19121                        Aretha, Sing One For Me                        Cat Power
9335                                          Supernova            Mr Hudson, Kanye West
22531                                  How Peculiar                    Robbie Williams
40217                                        Scream                  Tiësto, John Christian
1561                                          Last Nite                      The Strokes
169                          Crazy World - Radio Edit                 DJ Antoine, Mad Mark
9342                                     Cold Shoulder                           N-Dubz
```

Immagine 13 – Playlist consigliata contenente le prime 10 canzoni.

```
Il numero di miss classified sample è:  17
L'error rate sulle prime  25 canzoni è pari al  68.0 %
L'accuracy sulle prime  25 canzoni è pari al  31.999999999999996 %
```

Immagine 14 – Parametri di valutazione DBSCAN

6.7.4 Gaussian Mixture Model

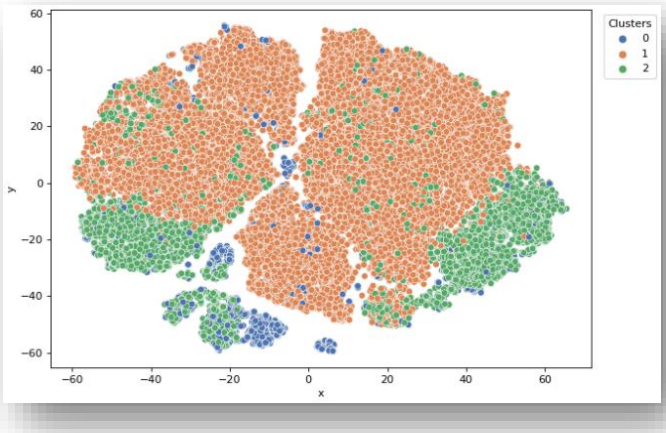


Immagine 15 – Visualizzazione (t-SNE) clustering GMM

```
La canzone in riproduzione (input) è:  Call me maybe
                                                  name                          artists
5870                                       Pay My Rent                            DNCE
23150  Rnw@y (Backyard Bangers Reanimation) [feat. Ph...  Linkin Park, Phoenix Orion
9396                                              Zero                         Alpines
19121                        Aretha, Sing One For Me                        Cat Power
9335                                          Supernova            Mr Hudson, Kanye West
13339                      I Saw Her Again Last Night                The Mamas & The Papas
29645                                        Polaroid  Jonas Blue, Liam Payne, Lennon Stella
22531                                  How Peculiar                    Robbie Williams
40217                                        Scream                  Tiësto, John Christian
```

Immagine 16 – Playlist consigliata contenente le prime 10 canzoni.

```
Il numero di miss classified sample è:  14
L'error rate sulle prime  25 canzoni è pari al  56.00000000000001 %
L'accuracy sulle prime  25 canzoni è pari al  43.99999999999999 %
```

Immagine 17 – Parametri di valutazione GMM

7. Conclusion

Il progetto realizzato aveva come obiettivo lo sviluppo di un sistema di raccomandazione in grado di predire, a partire da una canzone in riproduzione, una playlist con cui continuare l'ascolto musicale. All'interno del lavoro si è dimostrato come gli algoritmi di unsupervised learning risultano essere utili allo sviluppo del sistema di raccomandazione richiesto in fase di progettazione. Nello specifico gli algoritmi k-means, k-medoids, DBASCAN e Gaussian Mixture Model hanno consentito di clusterizzare i samples presenti nel dataset e, mediante la valutazione delle similarità tra i datasamples e la canzone in input, predire la playlist consigliata a partire dai cluster ottenuti.
Analizzando i risultati riportati nel paragrafo precedente, si può dedurre come l'algoritmo k-means consenta una clusterizzazione dei samples tale da produrre una playlist in uscita con un'accuracy superiore rispetto agli altri algoritmi implementati. Di conseguenza, anche il numero di miss classified samples ottenuti con l'algoritmo k-means risulta essere il più basso. La valutazione delle performance dei vari algoritmi è strettamente legata alla canzone in input e al numero di canzoni all'interno della playlist consigliata.
Concludendo, nel caso specifico analizzato, il progetto ha evidenziamo come il k-means risulta essere l'algoritmo più accurato per la realizzazione del sistema di raccomandazione.