Polytechnic University of Bari

Department of Electrical and Information Engineering
Master's Degree in Computer Science Engineering (Information Systems)

AIML/BDA Final Writeup

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Students:	**Professors:**
Pietro Urso	Tommaso Di Noia
Ernesto Petroni	Vito Walter Anelli
	Antonio Ferrara

A.Y. 2021/2022

Summary

1. Abstract

The topic chosen for the Project is *"Tell Me How to Continue My Playlist: Sequential Recommendation Techniques".* The goal of the project work is the creation of a Machine Learning recommendation system that, given in a song, produces a playlist of recommended songs. The recommendation system provides the required result using unsupervised learning techniques.

2. Introduction

When we talk about recommendation systems, we refer to systems that are used in everyday life. The goal of these systems is to help the user in the search for information of interest within the immense amount of data generated, which is available. This is done considering the past preferences/behaviors of users.

In our recommendation system, the problem to be solved is related to the prediction of a set of items (songs) generated starting from the similarity calculated between the input song and all the remaining songs (items) available.

The similarity between two objects can be calculated in various ways; among the main solutions we have the Euclidean distance, the Manhattan distance and the cosine similarity.

During the development of the project, unsupervised learning techniques were used for the clustering of datapoints. Specifically, *k-means*, *k-medoids*, *DBSCAN* and *Gaussian Mixture Model algorithms* have enabled the construction of the recommended song playlist from the song being played (input).

4. Dataset and Features

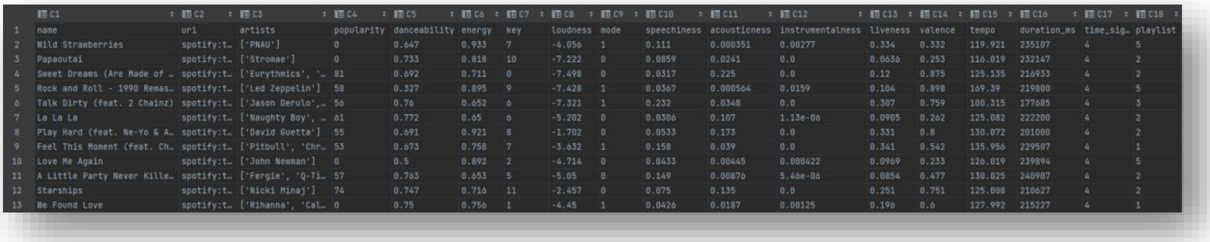
The **dataset** used for this project contains information about a set of songs published in the music scene in recent years. The number of samples or songs is equal to about 56,000 and the number of features is equal to 18.

As for the **features** of each song, they are both numerical and categorical. They are detailed below:

- **Name**: indicates the name of the published song;
- **URI**: indicates the link to the Spotify track of the song;
- **Artists**: indicates the list of songwriters;
- **Popularity**: indicates, through a numerical value between 0 and 100, the popularity of the song;
- **Danceability**: indicates, through a value between 0.0 and 0.99, the danceability of the song;
- **Energy**: indicates, through a value between 0.0 and 1.0, the energy transmitted by the song;
- **Key**: is a value between 0 and 11;
- **Loudness:** value between -9. 99 and -0.1;
- **Mode:** is a binary value (0 or 1);
- **Speechiness:** value between 0.0 and 0.99;
- **Acousticness:** value between 0.0 and 0.99;
- **Instrumentalness:** value between 0.0 and 0.99;
- **Liveness:** value between 0.0 and 1.0;

- **Valence:** value between 0.0 and 1. 0;
- **Time:** indicates the duration in seconds of the song;
- **Duration_ms:** indicates the duration in milliseconds of the song;
- **Time signature:** value between 0 and 5;
- **Playlist:** indicates the playlist of theoretical belonging to the song (value between 1 and 7).

Below is reported a portion of the 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	201000	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

Image 1 – Portion of the dataset used

The complete dataset can be found at the following link: https://drive.google.com/file/d/1Q9CRt4i2pRBMMGOGjpGdnrPKoTHxqHqw/view?usp=sharing

5. Methods

The Machine Learning techniques used to implement the recommendation system are **Unsupervised Learning techniques**. We choose to use this approach because the objective of the study is clustering to identify groups of possible playlists from the original dataset. These clusters are composed of songs with similar characteristics. Therefore, finding similarities between the datapoints, in such a way as to search for common characteristics between them, motivates the choice made with respect to the algorithms used.

Starting from the incoming song, therefore, the cluster of belonging is identified, and it is used later to choose the n songs closest by characteristics, compared to the song being played (the input song). The selection of the nearest n songs is made by calculating the similarity between the input song and all the other songs in the cluster. Specifically, the measure of similarity used is the **Euclidean distance**:

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

As for the clustering algorithms implemented, **K-means**, **K-medoids**, **DBSCAN** and the **Gaussian Mixture Model** were used.

5.1 K-Means

The idea of k-means is to find the k centroids, the central points of k clusters, that are the points having on average a better distance between all the other datapoints within their clusters. The datapoints which are closest to a specific centroid belong to the cluster represented by that centroid. In this way it clusters all the datapoints.

For the choice of the best K, is used the *elbow method*, which provides the right balance between the number of clusters and the cost of clustering.

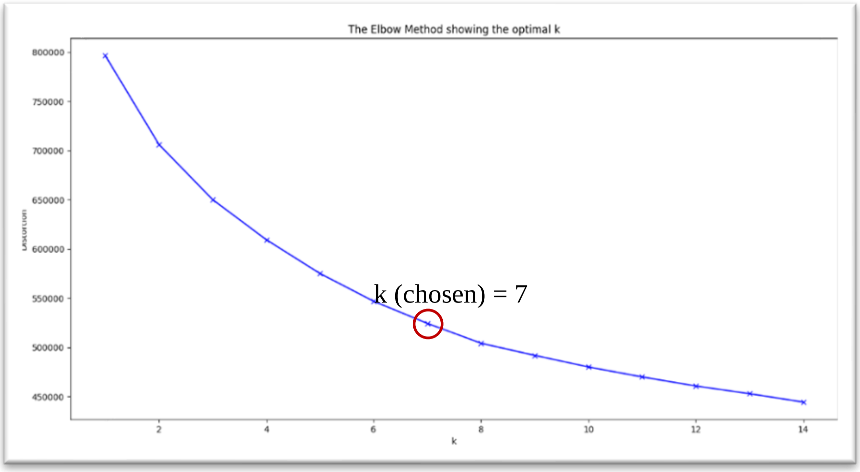


Image 2 – Elbow Method (k-means)

5.2 K-Medoids

Some researchers have decided to propose a new clustering algorithm called K-medoids. The goal is about the idea of generating the centroids not related to datapoints, as it happens in K-Means algorithm. The K-medoids proposes is to use the datapoints to represent the centroids of the clusters. To do that there are several algorithms, the most used is the PAM (Partitioning Around Medoids) that partitions datapoints among the medoids, finding the best clustering from this medoids. This algorithm nominates each datapoint as possible medoid, calculating the cost. The lowest cost configuration will represent the final clustering. Also in this case, for the choice of the number of clusters k, has been applied the *elbow method.*

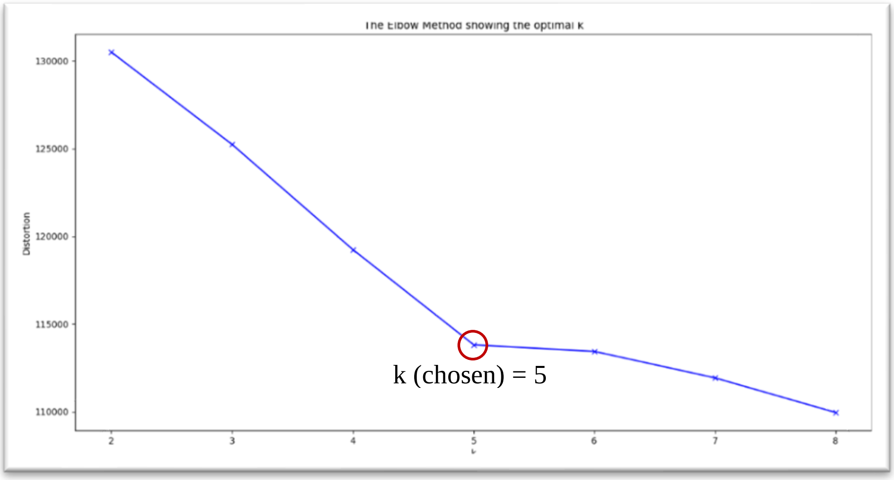


Image 2 – Elbow Method (k-medoids)

5.3 DBSCAN

DBSCAN is an agglomerative algorithm that is based on the concept of density. The goal of this algorithm is to find regions, in the dataspace, with high density of datapoints and separate them from regions with a low density. The density is defined through a value, called *eps*, in which based on the value that is defined, you specify the radius that allows you to virtually draw a sphere or circle (two-dimensional space). Within this sphere you can calculate the number of datapoints that are inside it.

In DBSCAN, another important concept is related to the minimum number of points (MinPts) that you want to consider forming a cluster. Therefore, it is necessary to set two parameters:

- **Epsilon**: physical distance or radius;
- **MinPts:** Minimum number of points needed to form a cluster.

The choice of the Epsilon parameter can be made using the k-distance graph while the value of the MinPts can be chosen through the following rule of thumb:

$$minPts \leq number\_of\_features + 1$$

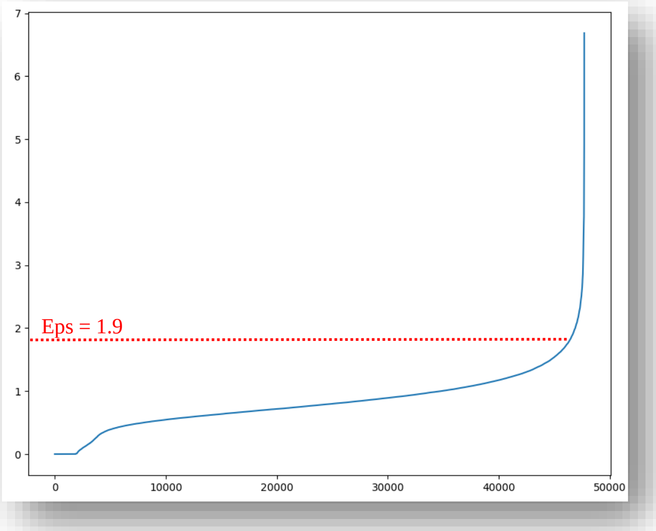


Image 3 – k-distance graph

5.4 Gaussian Mixture Model

Some complex deployments require more sophisticated systems for computing new cluster centers. The Gaussian Mixture Models (GMMs) therefore have a greater flexibility than the K-Means, K-Medoids and DBSCAN allowing even the most insidious distributions to be processed correctly.

In this method it is assumed that the distribution of observations is Gaussian, a less restrictive assumption than considering them circular and employing the average for the calculation of clusters.

In this way the parameters that describe the shape of the clusters are two:

- average
- standard deviation

The shape that clusters can take now extends to any ellipse.
If the number of components k is known, the most common used technique for estimating the parameters of the Mixture Model is Expectation-Maximization (EM). EM is a numerical technique, based on an iterative algorithm, to maximize the estimation of likelihood, thus minimizing the cost of clustering.

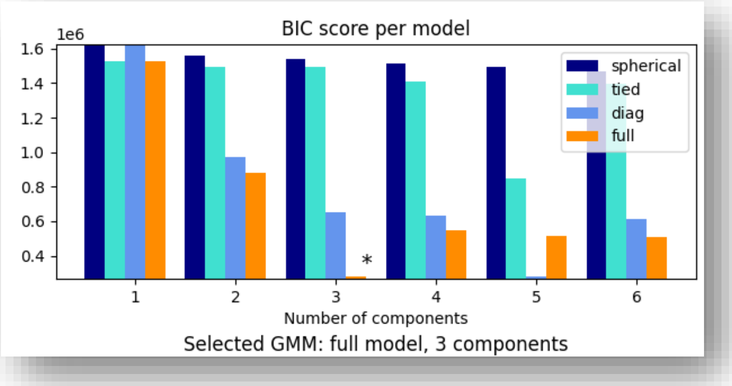


Image 4 – BIC score for the choice of the number of clusters

6. Experiments and Results

6.1 Preprocessing and Normalization

The first phase of the study focused on reading the dataset in csv format. Subsequently, an analysis of the dataset (missing values management) was conducted, followed by a preprocessing phase (removal of specific characters), a data normalization phase (z-score normalizzation) and the removal of duplicate samples. Finally, the preprocessing phase was completed through features selection, evaluating the IG (Information Gain) associated with each feature.

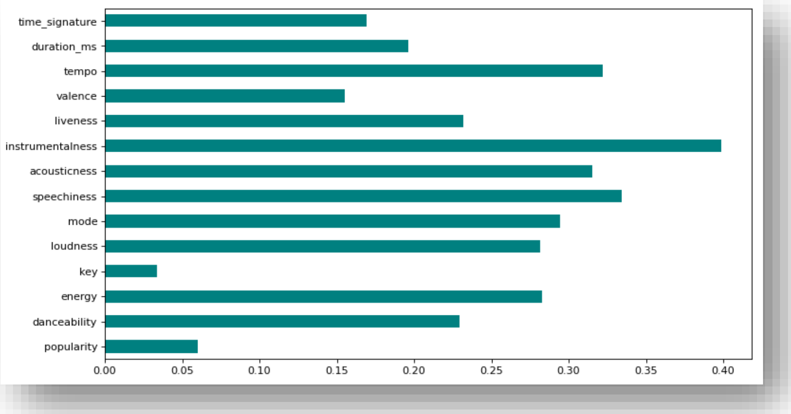


Image 5 – Feature importance (Information Gain)

6.2 Calculation of optimal parameters of Clustering algorithms

Before the implementation of clustering algorithms, procedures were developed for the optimal calculation of the parameters necessary for the implementation of the various algorithms. Specifically:

- Calculation of the optimal number k of clusters, using the *elbow method*, for the K-Means and K-Medoids algorithms.

- Calculation of the optimal value of *eps*, through the construction of the k-distance graph, for the DBSCAN algorithm.

- Use of the rule of thumb: $minPts \leq number\_of\_features + 1$ for calculating the optimal value of the *minPts* parameter for the DBSCAN algorithm.

- Evaluation using the Bayesian Information Criterion (*BIC*) for choosing the number of optimal components (number of clusters) for the Gaussian Mixture Model algorithm.

6.3 Implementation of Clustering algorithms

Four functions have been developed, one for each clustering algorithm, which implement the datapoints clustering procedure.
Each of these functions, following the application of the algorithm, takes care of adding a label (additional feature), which identifies the cluster to which each sample belongs.
These functions will return to the original dataset, with the addition of the information resulting from clustering.

6.4 The SongRecommender class

The SongRecommender() class aims to return the recommended playlist based on the input song.
Specifically, it initially deals with extracting all the songs that belong to the same cluster of the song provided in input. Then, there is the calculation of the similarity value, expressed by the *Euclidean distance* (correlation indicator), between the song being played (input) and each of the other songs previously extracted.

This calculated value is added as a feature within the list of songs that belong to the same cluster as the input song.
After the calculation of the distances, there is a decreased sorting of the list, followed by the selection of the first n songs that will form the recommended playlist, returned as output.

6.5 Evaluation of results

Within the original dataset there is the playlist feature; it indicates the reference playlist to which each sample belongs and has a value between 1 and 7. This information is used to evaluate the error rate and accuracy based on the results obtained from the song being played.

- The **error rate** is calculated as:

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

- The **accuracy** is calculated as:

$$accuracy = 1 - error\ rate$$

6.6 Data visualization

Due to the large number of features, for the display of clustered samples, it was decided to use the t-SNE dimensionality reduction algorithm.

The initial problem was transformed into the two-dimensional equivalent, thus moving from many features (dimensions) to only two dimensions.

This transformation has allowed a graphic display of the datasamples suitably clustered according to the label defined by the clustering algorithm applied.

6.7 Results

The song being played (input) chosen for the playlist recommendation is 'Call me maybe', while the different algorithms have been evaluated based on ***accuracy*** and **error rate**, calculated from a recommended playlist containing 25 songs.

Below are the results of the various experiments, which include:

- *Choice of optimal parameters*;
- *Visualization of clusters*;
- *Recommended playlist*;
- *Evaluation parameters*.

6.7.1 K-Means

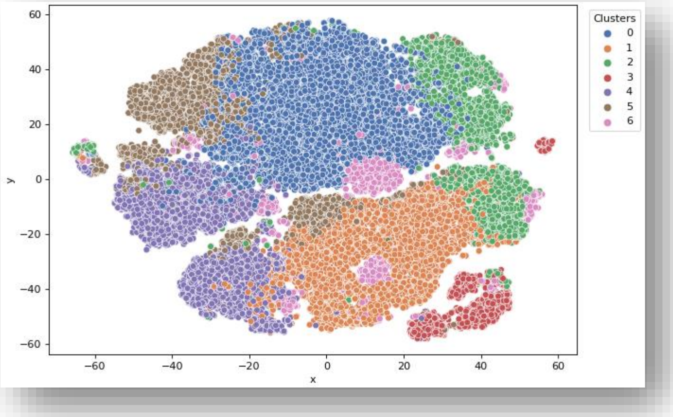
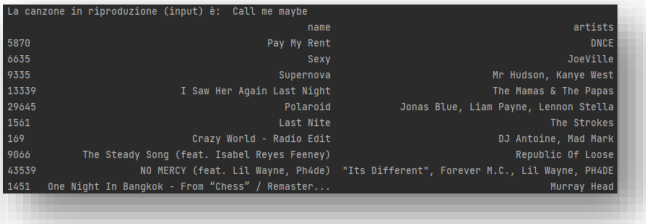
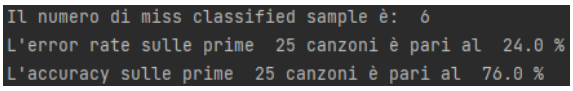


Image 6 – Visualization (t-SNE) clustering k-means



```
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
9066        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
```

Image 7 – Recommended playlist containing the first 10 songs.



```
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 %
```

Image 8 – K-means evaluation parameters

6.7.2 K-Medoids

As for the k-medoids algorithm, due to the high computational capacity required, in terms of the memory required for the application of the algorithm, was made a reduction of the number of samples examined, from about 47,000 initials to 40.000 used. The results have been obtained from this reduced set of songs.

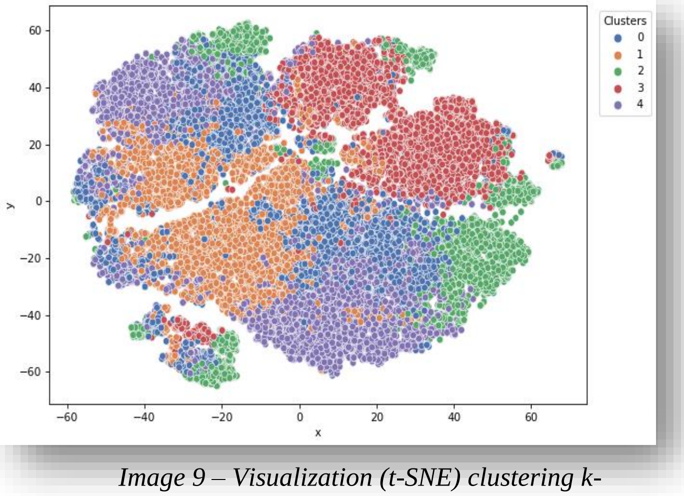
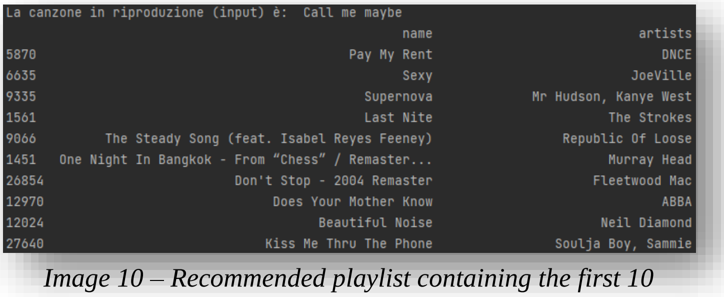
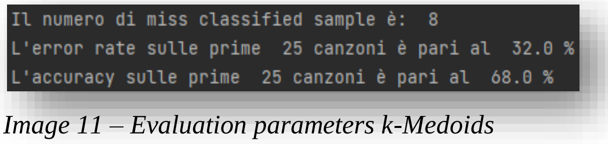


Image 9 – Visualization (t-SNE) clustering k-medoids



```
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
```

Image 10 – Recommended playlist containing the first 10 songs.



```
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 %
```

Image 11 – Evaluation parameters k-Medoids

6.7.3 DBSCAN

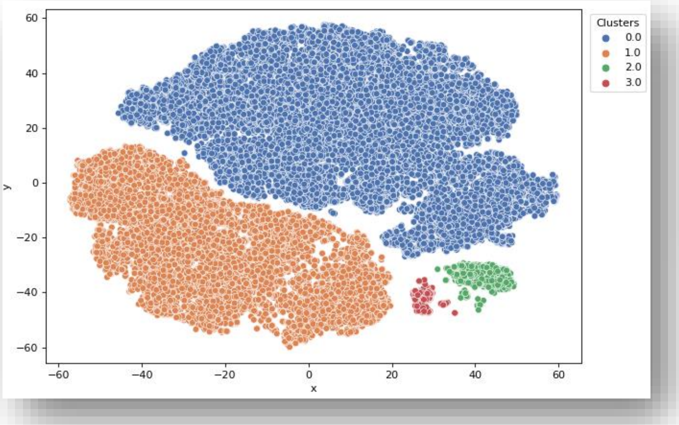
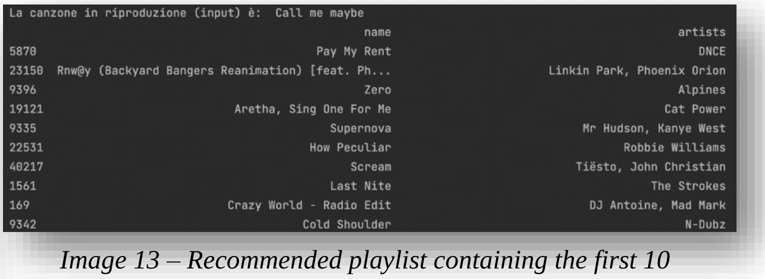
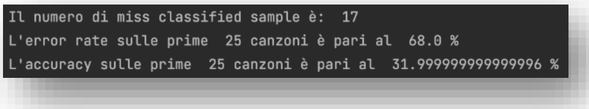


Image 12 – Visualization (t-SNE) DBSCAN clustering



```
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
```

Image 13 – Recommended playlist containing the first 10 songs.



```
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 %
```

Image 14 — DBSCAN evaluation parameters

6.7.4 Gaussian Mixture Model

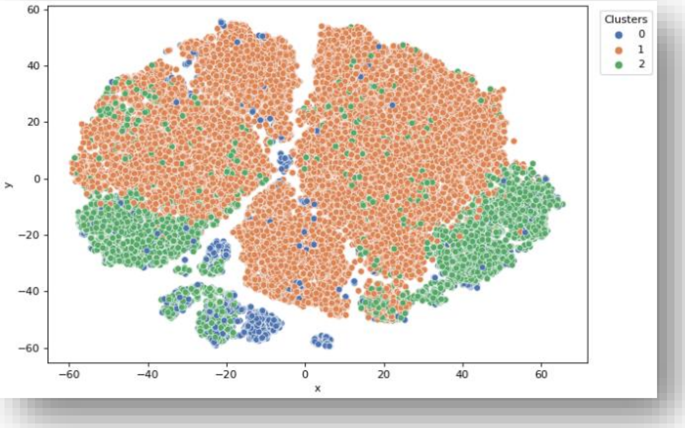
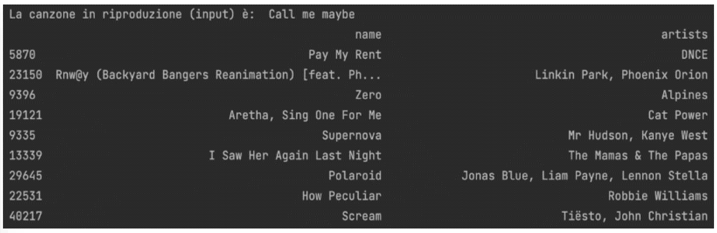
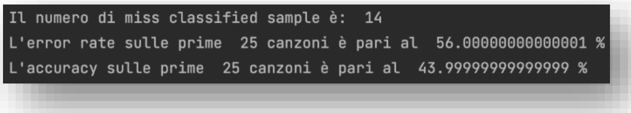


Image 15 – Visualization (t-SNE) GMM clustering



```
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
```

Image 16 – Recommended playlist containing the first 10 songs.



```
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 %
```

Image 17 – GMM evaluation parameters

7. Conclusion

The project carried out had as its objective the development of a recommendation system able to predict, starting from a song being played, a playlist with which to continue listening to music. The work demonstrated how unsupervised learning algorithms are useful for the development of the recommendation system required at the design stage. Specifically, the k-means, k-medoids, DBASCAN and Gaussian Mixture Model algorithms allowed to cluster the samples present in the dataset and, by evaluating the similarities between the datasamples and the input song, predict the recommended playlist starting from the clusters obtained.

Analyzing the results reported in the previous paragraph, it can be deduced that the k-means algorithm allows a clustering of the samples such as to produce an output playlist with a higher accuracy than the other algorithms implemented. As a result, the number of mis classified samples obtained with the k-means algorithm is also the lowest. The evaluation of the performance of the various algorithms is closely linked to the input song and the number of songs within the recommended playlist.

In conclusion, in the specific case analyzed, the project highlighted how the k-means turns out to be the most accurate algorithm for the realization of the recommendation system.