PLAYLIST RECOMMENDER

Caso di studio Ingegneria della Conoscenza A.A. 2021-2022

Matteo Esposito – 718240

Giuseppe Galgano - 717510

INFORMAZIONI

- Per il caso di studio abbiamo optato per la realizzazione di un recommender system di canzoni basato su un dataset di Spotify acquisito dal sito Kaggle. Il dataset in questione contiene circa 230000 canzoni ed un totale di 26 generi. Sono implementati inoltre 4 algoritmi di classificazione quali Decision Tree Classifier, Random Forest Classifier, Logistic Regression e K-Nearest Neighbour Classifier per classificare la popolarità di una canzone.

STRUMENTI & LIBRERIE UTILIZZATI

- Linguaggio: Python
- IDE: PyCharm
- Hosting: GitHub

- Sklearn
- Scikit-learn e una libreria open source di apprendimento automatico per il linguaggio di programmazione Python.

- Pandas
- Pandas è una libreria software scritta per il linguaggio di programmazione Python per la manipolazione e l'analisi dei dati.

- Matplotlib
- Matplotlib è una libreria per la creazione di grafici per il linguaggio di programmazione Python e la libreria matematica NumPy.

PREPROCESSING PER IL CLUSTERING

Per l’attività di clustering abbiamo apportato le seguenti modifiche al dataset:

1. Creazione di un indice utilizzando le features ‘track_name’ e ‘artist_name’.
2. Creazione di una prima tabella chiamata attributes , partendo dal dataset iniziale, rimuovendo 'track_id', 'track_name', 'time_signature', 'track_name', 'artist_name', 'key’.
3. Creazione di una seconda tabella chiamata genres dove per ogni tipologia di genere abbiamo aggiunto una feature di tipo binario cosi che ogni canzone abbia 1 al proprio genere. Questa modifica permette di effettuare la similaritá del coseno.
4. Unione delle due tabelle attributes e genres grazie all’indice in un'unica tabella chiamata songs.
5. Eliminazione di eventuali duplicati in songs.

CLUSTERING

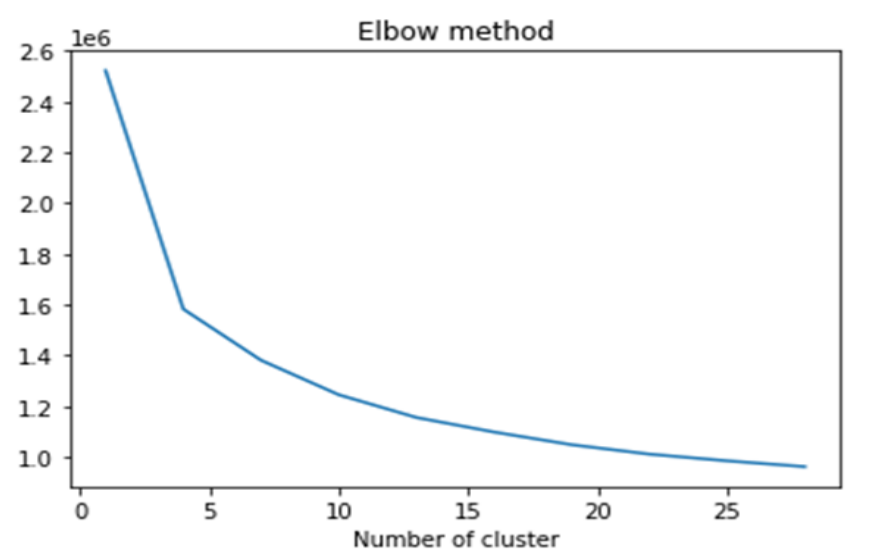
- Il clustering consiste in un insieme di metodi per raggruppare oggetti in classi omogenee. Un cluster è un insieme di oggetti che presentano tra loro delle similarità, ma che, per contro, presentano dissimilarità con oggetti in altri cluster.
- Nel nostro progetto abbiamo applicato l'algoritmo K-Means. Quest'ultimo ha lo scopo di suddividere un insieme di oggetti in k gruppi sulla base dei loro attributi.

ELBOW METHOD

Applicando l'Elbow Method, "Metodo del Gomito", abbiamo scoperto il numero di cluster più adatto per il dataset.

Questo grafico deve essere letto da destra verso sinistra. Si deve trovare il punto in cui la curva tende a salire in modo più consistente.

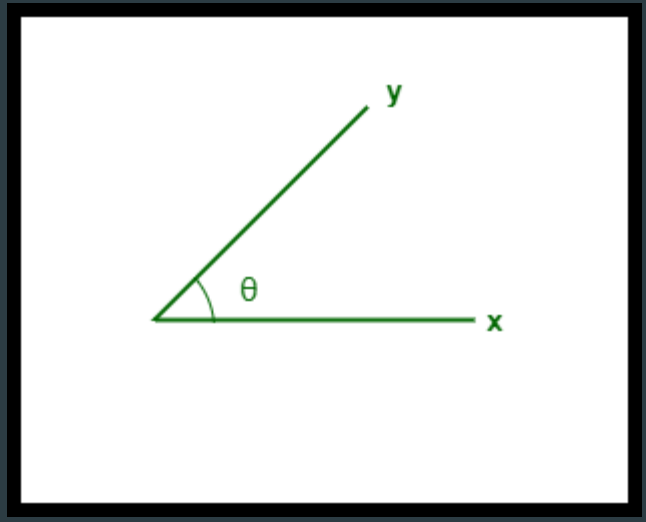
Nel nostro caso abbiamo scelto 5.



RECOMMENDER SYSTEM

Una volta effettuato il cluster sul dataset si è proceduto ad applicare la raccomandazione basata sulla similarità delle canzoni.

In particolare si è scelto di utilizzare la similarità del coseno.



PREPROCESSING PER LA CLASSIFICAZIONE

Per l’attività di classificazione abbiamo apportato le seguenti modifiche al dataset:

1. Per la feature «key» convertiamo le 12 chiavi in numeri, utilizzando l’indice.
2. Per la feature «mode» convertiamo le Major in 1 e Minor in 0.
3. Per la feature «time_signature» convertiamo i battiti in numeri, utilizzando l'indice.
4. Rendiamo la feature «popularity» binaria, una canzone é popolare se ha uno score maggiore o uguale a 75. Non é popolare altrimenti.

CLASSIFICAZIONE

Uno degli scopi principali del Machine Learning è la classificazione, cioè il problema di indentificare la classe di un nuovo obiettivo sulla base di conoscenza estratta da un training set.

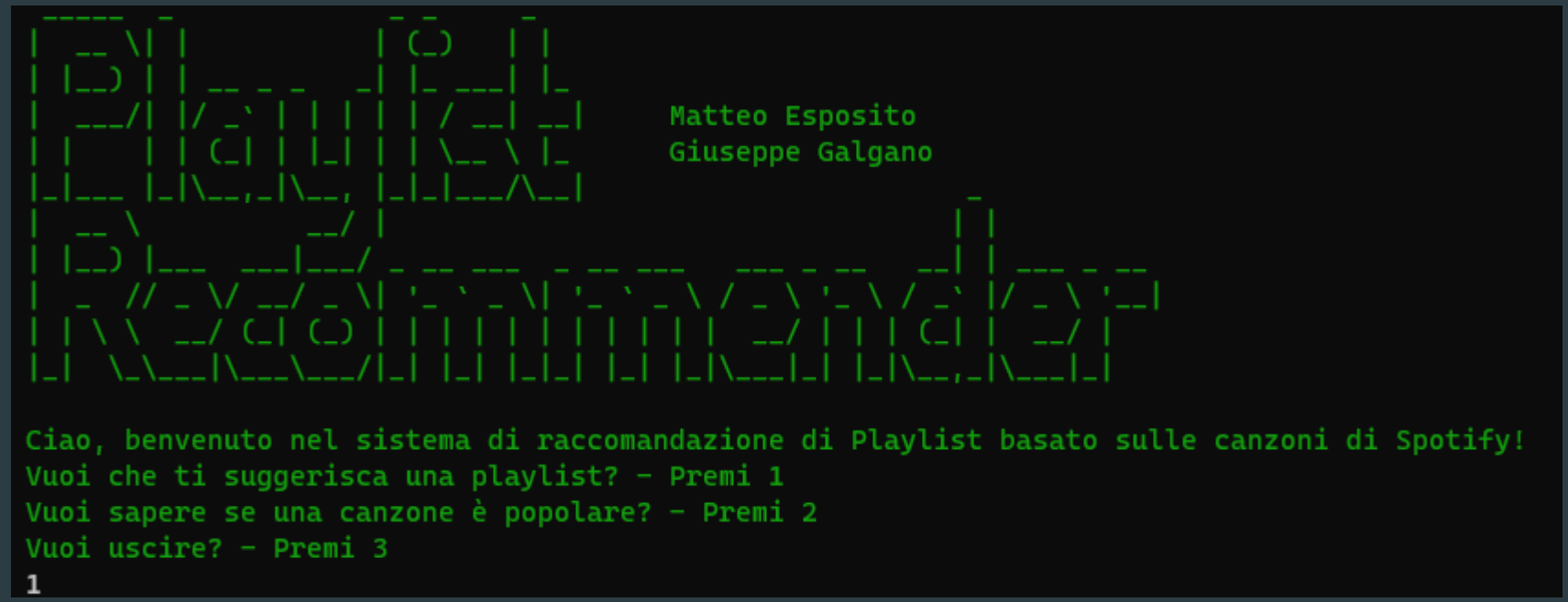
Per lo scopo del nostro progetto abbiamo deciso di suddividere i dati in un insieme di training e un insieme di test fissando quest'ultimo al 20%. La variabile target sulla quale effettuare la predizione sarà «popularity»".

MODELLI A CONFRONTO ACCURATEZZA

Dopo aver calcolato l'accuratezza dei modelli, abbiamo scoperto che il Random Forest Classifier performa meglio degli altri.

- Random Forest Classifier 0.994
- Decision Tree Classifier 0.985
- Logistic Regression 0.983
- K-Nearest Neighbout Classifier 0.981

GUIDA ALL'USO: MENÚ PRINCIPALE



GUIDA ALL'USO:
DOMANDE CREAZIONE PLAYLIST

```
Adesso dovrai suggerirmi su quale canzone basare la tua playlist!
Qual'é il nome di una traccia che hai apprezzato?
DNA
Adesso dimmi il nome dell'artista che ha scritto la traccia.
BTS
Quante canzoni vuoi inserire nella tua playlist?
15
```

GUIDA ALL'USO: PLAYLIST CREATA

```
Playlist basata sulla canzone "DNA" di BTS

crushcrushcrush - Paramore
Good to Me - SEVENTEEN
Siren - SUNMI
Run - BTS
Treasure - ATEEZ
Outro: Wings - BTS
La Vie en Rose - IZ*ONE
Summer - Calvin Harris
Hard Times - Paramore
BBoom BBoom - MOMOLAND
No - CLC
DDD - EXID
Maps - Maroon 5
Trivia 轉 : Seesaw - BTS
Valkyrie - ONEUS
```

GUIDA ALL'USO: DOMANDE CLASSIFICAZIONE

```
Adesso dovrai suggerirmi la canzone su cui predire la popolaritá!
Qual'é il nome della traccia?
Without Me
Adesso dimmi il nome dell'artista che ha scritto la traccia.
Eminem
```

GUIDA ALL'USO:
SCELTA ALGORITMO DI CLASSIFICAZIONE E RISULTATO

```
Canzone trovata nel dataset!
Quale classificatore vuoi utilizzare?
Random Forest Classifier - Premi 1
K-Nearest Neighbors Classifier - Premi 2
Decision Tree Classifier - Premi 3
Logistic Regression - Premi 4

1
La canzone é popolare!
```