Università degli Studi di Salerno

Liking Songs

Progetto di Fondamenti di Intelligenza Artificiale

Autore: Giovanni Citro
Matricola: 0512115103
Anno Accademico 2023-2024

Il Mio GitHub

Gennaio 2024

Indice

1 Introduzione **2**

2 Definizione del Problema **3**

3 Specifica PEAS **4**
3.1 Performance 4
3.2 Environment 4
3.3 Actuators 4
3.4 Sensors 4
3.5 Specifiche dell'Ambiente 4
3.5.1 Singolo Agente 4
3.5.2 Completamente Osservabile 4
3.5.3 Deterministico 4
3.5.4 Episodico 4
3.5.5 Statico 4
3.5.6 Discreto 4

4 Modello CRISP-DM **5**
4.1 Scelta del dataset 5
4.2 Data Understanding 5
4.2.1 Descrizione dei dati 5
4.3 Data Preparation 6
4.3.1 Data Cleaning 6
4.3.2 Feature Engineering 8

5 Modello di Classificazione **9**
5.1 Decision Tree Classifier 9
5.2 One Hot Encoding 10

6 Conclusioni e Sviluppi Futuri **11**

1 Introduzione

Il sistema proposto è un modello di classificazione che si basa sull'interazione attiva degli utenti per predire le loro preferenze musicali. Gli utenti inseriscono le canzoni preferite, assegnando feedback positivi (1) o negativi (0). Questi dati addestrano un modello avanzato che apprende in modo personalizzato le preferenze di ciascun utente.

2 Definizione del Problema

Il sistema implementato si basa su un sofisticato modello di classificazione che permette agli utenti di inserire le proprie canzoni preferite all'interno di una lista apposita. La fase successiva coinvolge l'utente nel processo decisionale, chiedendogli di assegnare un feedback positivo (1) o negativo (0) per ciascuna canzone, indicando se la stessa gli piaccia o meno. Questo feedback utente diventa fondamentale per addestrare il modello, che utilizza i dati raccolti per apprendere e comprendere le preferenze musicali specifiche di ciascun utente.

Il sistema, sfruttando le informazioni fornite dall'utente durante la fase di classificazione delle canzoni preferite, è in grado di creare un modello personalizzato di predizione delle preferenze musicali. In altre parole, il sistema apprende i tratti distintivi delle canzoni gradite e non gradite da parte dell'utente in questione, consentendo predizioni future sul gradimento di nuove canzoni che non sono state precedentemente valutate.

3 Specifica PEAS

3.1 Performance

La misura di performance dell'agente è valutata in base alla sua abilità di avvicinarsi il più possibile a una situazione ideale, in cui agli utenti vengono mostrati esattamente i valori che indicano se una canzone piace (1) o non piace (0). Questa valutazione è effettuata utilizzando il dataset come riferimento, con l'aggiunta della colonna "Liking" che rappresenta il valore ideale di gradimento associato a ciascuna canzone.

3.2 Environment

L'ambiente in cui opera l'agente è costituito dall'insieme di canzoni con i relativi dati di gradimento. Le canzoni presenti nell'ambiente sono caratterizzate dai feedback degli utenti, rappresentati attraverso i valori binari di gradimento (piace o non piace), e sono descritte nel dataset utilizzato per l'addestramento e la valutazione del modello.

3.3 Actuators

Gli attuatori dell'agente consistono nella lista di canzoni con la predizione del parametro che ci dice quale canzone può piacere e quale no

3.4 Sensors

I sensori rilevano il feedback degli utenti, espressi attraverso i valori binari di gradimento (piace o non piace), per tutte le canzoni presenti nell'ambiente. Questi dati sensoriali sono fondamentali per l'addestramento e la valutazione del modello di classificazione, permettendo all'agente di apprendere e migliorare la sua capacità di predire le preferenze musicali degli utenti.

3.5 Specifiche dell'Ambiente

3.5.1 Singolo Agente

3.5.2 Completamente Osservabile

Il sistema ha accesso a tutte le informazioni riguardanti le caratteristiche e le preferenze delle canzoni.

3.5.3 Deterministico

La predizione del gradimento di una canzone è basata su regole fisse e non varia, indipendentemente da altri fattori o da eventuali ripetizioni delle azioni.

3.5.4 Episodico

Le esperienze dell'agente sono divise in episodi distinti, e le decisioni possono essere prese basandosi solo sulle informazioni raccolte durante l'episodio corrente. In questo caso, ogni interazione dell'utente con il sistema potrebbe significare un episodio separato.

3.5.5 Statico

Le preferenze musicali degli utenti o le caratteristiche delle canzoni rimangono costanti e invariate durante le interazioni

3.5.6 Discreto

Le variabili sono definite in modi discreti(numeri finiti)

4 Modello CRISP-DM

4.1 Scelta del dataset

Per i dataset del modello di classificazione ho usato la strada più semplice, ovvero quella di cercare un dataset con le canzoni di Spotify e i suoi attributi. Infine ho creato un dataset inserendo la colonna che mi serviva per far apprendere al modello i valori di gradimento da predire

4.2 Data Understanding

Data Understanding, vengono acquisiti ed esplorati i dati

I dataset si caricano nello spazio di archiviazione della sessione

```
#Dataset che contiene le canzoni con le informazioni

datasetSongs = pd.read_csv("Spotify-Songs.csv")

datasetSongs
```

1 to 10 of 1003 entries Filter

index	title	artist	top genre	year released	added	bpm	nrgy	dnce	dB	live	val	dur	acous	spch	pop	top year	artist type
0	STARSTRUKK (feat. Katy Perry)	3OH!3	dance pop	2009.0	2022-02-17	140.0	81.0	61.0	-6.0	23.0	23.0	203.0	0.0	6.0	70.0	2010.0	Duo
1	My First Kiss (feat. Ke$ha)	3OH!3	dance pop	2010.0	2022-02-17	138.0	89.0	68.0	-4.0	36.0	83.0	192.0	1.0	8.0	68.0	2010.0	Duo
2	I Need A Dollar	Aloe Blacc	pop soul	2010.0	2022-02-17	95.0	48.0	84.0	-7.0	9.0	96.0	243.0	20.0	3.0	72.0	2010.0	Solo
3	Airplanes (feat. Hayley Williams of Paramore)	B.o.B	atl hip hop	2010.0	2022-02-17	93.0	87.0	66.0	-4.0	4.0	38.0	180.0	11.0	12.0	80.0	2010.0	Solo
4	Nothin' on You (feat. Bruno Mars)	B.o.B	atl hip hop	2010.0	2022-02-17	104.0	85.0	69.0	-6.0	9.0	74.0	268.0	39.0	5.0	79.0	2010.0	Solo
5	Magic (feat. Rivers Cuomo)	B.o.B	atl hip hop	2010.0	2022-02-17	82.0	93.0	55.0	-4.0	35.0	79.0	196.0	1.0	34.0	71.0	2010.0	Solo
6	The Time (Dirty Bit)	Black Eyed Peas	dance pop	2010.0	2022-02-17	128.0	81.0	82.0	-8.0	60.0	44.0	308.0	7.0	7.0	75.0	2010.0	Band/Group
7	Imma Be	Black Eyed Peas	dance pop	2009.0	2022-02-17	92.0	52.0	60.0	-7.0	31.0	41.0	258.0	18.0	37.0	71.0	2010.0	Band/Group
8	Talking to the Moon	Bruno Mars	dance pop	2010.0	2022-02-17	146.0	59.0	50.0	-5.0	11.0	8.0	218.0	51.0	3.0	87.0	2010.0	Solo
9	Just the Way You Are	Bruno Mars	dance pop	2010.0	2022-02-17	109.0	84.0	64.0	-5.0	6.0	42.0	221.0	1.0	4.0	86.0	2010.0	Solo

Show 10 per page 1 2 10 100 101

```
#Dataset che contiene le canzoni ascoltate con i valori di gradimento(0,1)

datasetSongsMe = pd.read_csv("LikingDef.csv", delimiter=";")

datasetSongsMe
```

1 to 10 of 249 entries Filter

index	title	artist	year released	bpm	dur	artist type	top genre	liking
0	STARSTRUKK (feat. Katy Perry)	3OH!3	2009	140	203	Duo	dance pop	0
1	My First Kiss (feat. Ke$ha)	3OH!3	2010	138	192	Duo	dance pop	0
2	I Need A Dollar	Aloe Blacc	2010	95	243	Solo	pop soul	0
3	Airplanes (feat. Hayley Williams of Paramore)	B.o.B	2010	93	180	Solo	atl hip hop	1
4	Nothin' on You (feat. Bruno Mars)	B.o.B	2010	104	268	Solo	atl hip hop	1
5	Magic (feat. Rivers Cuomo)	B.o.B	2010	82	196	Solo	atl hip hop	1
6	The Time (Dirty Bit)	Black Eyed Peas	2010	128	308	Band	dance pop	0
7	Imma Be	Black Eyed Peas	2009	92	258	Band	dance pop	0
8	Talking to the Moon	Bruno Mars	2010	146	218	Solo	dance pop	0
9	Just the Way You Are	Bruno Mars	2010	109	221	Solo	dance pop	0

Show 10 per page 1 2 10 20 25

4.2.1 Descrizione dei dati

- **Title**: Titolo della Canzone
- **Artist**: Cantante
- **Top Genre**: Genere della Canzone
- **Year Released**: Anno di Rilascio
- **Dur**: Durata della canzone in secondi
- **Artist Type** : Tipo di Artista(Solista o Band)

4.3 Data Preparation

Guardando bene il dataset ho visto che alcune colonne erano inutili al mio intento e quindi sono andato ad eliminarle(added, top year, acous, spch, pop, dB, live, nrgy, dnce, bpm), poi ho formattato alcuni valori all'interno di alcune colonne come: Anno di Rilascio, Durata e Tipo di Artista.

4.3.1 Data Cleaning

Il Data Cleaning (pulizia dei dati) è un passaggio essenziale nella preparazione dei dati, durante il quale vengono affrontati vari problemi presenti nei dati grezzi per assicurarsi che siano adatti all'analisi o all'addestramento del modello.Nel mio caso sono andato ad eseguire i seguenti passaggi:

-Eliminare alcune colonne che non erano correlate con il problema in esame

Rimozione di alcune colonne inutili

```
[ ] #Rimozione colonne inutili
    datasetSongs.drop(['added','top year', 'acous', 'spch', 'pop', 'dB','live','nrgy', 'dnce','val','bpm'], axis=1, inplace=True)

    #Formattazione Nome delle Colonne
    datasetSongs.columns = datasetSongs.columns.str.title()

    datasetSongs
```

1 to 10 of 1003 entries Filter

index	Title	Artist	Top Genre	Year Released	Dur	Artist Type
0	STARSTRUKK (feat. Katy Perry)	3OH!3	dance pop	2009.0	203.0	Duo
1	My First Kiss (feat. Ke$ha)	3OH!3	dance pop	2010.0	192.0	Duo
2	I Need A Dollar	Aloe Blacc	pop soul	2010.0	243.0	Solo
3	Airplanes (feat. Hayley Williams of Paramore)	B.o.B	atl hip hop	2010.0	180.0	Solo
4	Nothin' on You (feat. Bruno Mars)	B.o.B	atl hip hop	2010.0	268.0	Solo
5	Magic (feat. Rivers Cuomo)	B.o.B	atl hip hop	2010.0	196.0	Solo
6	The Time (Dirty Bit)	Black Eyed Peas	dance pop	2010.0	308.0	Band/Group
7	Imma Be	Black Eyed Peas	dance pop	2009.0	258.0	Band/Group
8	Talking to the Moon	Bruno Mars	dance pop	2010.0	218.0	Solo
9	Just the Way You Are	Bruno Mars	dance pop	2010.0	221.0	Solo

Show 10 per page 1 2 10 100 101

-Formattare alcuni dati che non erano rappresentati correttamente

Formattazione Anno di Rilascio

```
datasetSongs[['Year Released']] = datasetSongs[['Year Released']].apply(lambda x: x.astype(str).str.split('.').str[0])

datasetSongs
```

1 to 10 of 1003 entries Filter

index	Title	Artist	Top Genre	Year Released	Dur	Artist Type
0	STARSTRUKK (feat. Katy Perry)	3OH!3	dance pop	2009	203.0	Duo
1	My First Kiss (feat. Ke$ha)	3OH!3	dance pop	2010	192.0	Duo
2	I Need A Dollar	Aloe Blacc	pop soul	2010	243.0	Solo
3	Airplanes (feat. Hayley Williams of Paramore)	B.o.B	atl hip hop	2010	180.0	Solo
4	Nothin' on You (feat. Bruno Mars)	B.o.B	atl hip hop	2010	268.0	Solo
5	Magic (feat. Rivers Cuomo)	B.o.B	atl hip hop	2010	196.0	Solo
6	The Time (Dirty Bit)	Black Eyed Peas	dance pop	2010	308.0	Band/Group
7	Imma Be	Black Eyed Peas	dance pop	2009	258.0	Band/Group
8	Talking to the Moon	Bruno Mars	dance pop	2010	218.0	Solo
9	Just the Way You Are	Bruno Mars	dance pop	2010	221.0	Solo

Show 10 per page 1 2 10 100 101

Formattazione Tipo di Artista

```
datasetSongs['Artist Type'] = datasetSongs['Artist Type'].apply(lambda x: str(x).split('/')[0])

datasetSongs
```

1 to 10 of 1003 entries Filter

index	Title	Artist	Top Genre	Year Released	Dur	Artist Type
0	STARSTRUKK (feat. Katy Perry)	3OH!3	dance pop	2009	3:23	Duo
1	My First Kiss (feat. Ke$ha)	3OH!3	dance pop	2010	3:12	Duo
2	I Need A Dollar	Aloe Blacc	pop soul	2010	4:03	Solo
3	Airplanes (feat. Hayley Williams of Paramore)	B.o.B	atl hip hop	2010	3:00	Solo
4	Nothin' on You (feat. Bruno Mars)	B.o.B	atl hip hop	2010	4:28	Solo
5	Magic (feat. Rivers Cuomo)	B.o.B	atl hip hop	2010	3:16	Solo
6	The Time (Dirty Bit)	Black Eyed Peas	dance pop	2010	5:08	Band
7	Imma Be	Black Eyed Peas	dance pop	2009	4:18	Band
8	Talking to the Moon	Bruno Mars	dance pop	2010	3:38	Solo
9	Just the Way You Are	Bruno Mars	dance pop	2010	3:41	Solo

Show 10 per page 1 2 10 100 101

-Eliminare i dati nulli

```
print(df.head())

df.dropna(inplace=True)
```

```
                                      Title      Artist  Year Released  \
0                 STARSTRUKK (feat. Katy Perry)       3OH!3           2009
1                   My First Kiss (feat. Ke$ha)       3OH!3           2010
2                               I Need A Dollar  Aloe Blacc           2010
3  Airplanes (feat. Hayley Williams of Paramore)       B.o.B           2010
4              Nothin' on You (feat. Bruno Mars)       B.o.B           2010

   Dur Artist Type  Liking        Genre
0  203         Duo       0    Dance Pop
1  192         Duo       0    Dance Pop
2  243        Solo       0     Pop Soul
3  180        Solo       1  Atl Hip Hop
4  268        Solo       1  Atl Hip Hop
```

4.3.2 Feature Engineering

Il feature engineering è il processo di creazione, trasformazione o selezione di nuove feature (variabili) a partire dai dati esistenti per migliorare le prestazioni del modello o facilitare l'interpretazione dei dati. Nel mio caso sono andato a trasformare le feature della colonna Durata e della colonna Liking andandole a categorizzare rispettivamente come Corta/Media/Lunga e MiPiace/NonMiPiace

```
for i, row in df.iterrows():
  dur = row['Dur']
  if dur < 180 :
    df.loc[i, 'Dur'] = 'Corta'
  elif dur <= 240 :
    df.loc[i, 'Dur'] = 'Media'
  else:
    df.loc[i, 'Dur'] = "Lunga"

print(df['Dur'].value_counts())
df
```

```
Media    163
Lunga     69
Corta      9
Name: Dur, dtype: int64
```

1 to 10 of 241 entries Filter

index	Title	Artist	Year Released	Dur	Artist Type	Liking	Genre
0	STARSTRUKK (feat. Katy Perry)	3OH!3	2009	Media	Duo	0	Dance Pop
1	My First Kiss (feat. Ke$ha)	3OH!3	2010	Media	Duo	0	Dance Pop
2	I Need A Dollar	Aloe Blacc	2010	Lunga	Solo	0	Pop Soul
3	Airplanes (feat. Hayley Williams of Paramore)	B.o.B	2010	Media	Solo	1	Atl Hip Hop
4	Nothin' on You (feat. Bruno Mars)	B.o.B	2010	Lunga	Solo	1	Atl Hip Hop
5	Magic (feat. Rivers Cuomo)	B.o.B	2010	Media	Solo	1	Atl Hip Hop
6	The Time (Dirty Bit)	Black Eyed Peas	2010	Lunga	Band	0	Dance Pop
7	Imma Be	Black Eyed Peas	2009	Lunga	Band	0	Dance Pop
8	Talking to the Moon	Bruno Mars	2010	Media	Solo	0	Dance Pop
9	Just the Way You Are	Bruno Mars	2010	Media	Solo	0	Dance Pop

```
df['Liking'] = pd.to_numeric(df['Liking'], errors='coerce')
for i, row in df.iterrows():
  liking = row['Liking']
  if liking == 1 :
    df.loc[i, 'Liking'] = 'Mi Piace'
  else:
    df.loc[i, 'Liking'] = 'Non Mi Piace'

#Mostra il conteggio di Mi Piace e Non Mi Piace
print(df['Liking'].value_counts())
df
```

```
Non Mi Piace    165
Mi Piace         79
Name: Liking, dtype: int64
```

1 to 10 of 244 entries Filter

index	Title	Artist	Year Released	Dur	Artist Type	Liking	Genre
0	STARSTRUKK (feat. Katy Perry)	3OH!3	2009	Media	Duo	Non Mi Piace	Dance Pop
1	My First Kiss (feat. Ke$ha)	3OH!3	2010	Media	Duo	Non Mi Piace	Dance Pop
2	I Need A Dollar	Aloe Blacc	2010	Lunga	Solo	Non Mi Piace	Pop Soul
3	Airplanes (feat. Hayley Williams of Paramore)	B.o.B	2010	Media	Solo	Mi Piace	Atl Hip Hop
4	Nothin' on You (feat. Bruno Mars)	B.o.B	2010	Lunga	Solo	Mi Piace	Atl Hip Hop
5	Magic (feat. Rivers Cuomo)	B.o.B	2010	Media	Solo	Mi Piace	Atl Hip Hop
6	The Time (Dirty Bit)	Black Eyed Peas	2010	Lunga	Band	Non Mi Piace	Dance Pop
7	Imma Be	Black Eyed Peas	2009	Lunga	Band	Non Mi Piace	Dance Pop
8	Talking to the Moon	Bruno Mars	2010	Media	Solo	Non Mi Piace	Dance Pop
9	Just the Way You Are	Bruno Mars	2010	Media	Solo	Non Mi Piace	Dance Pop

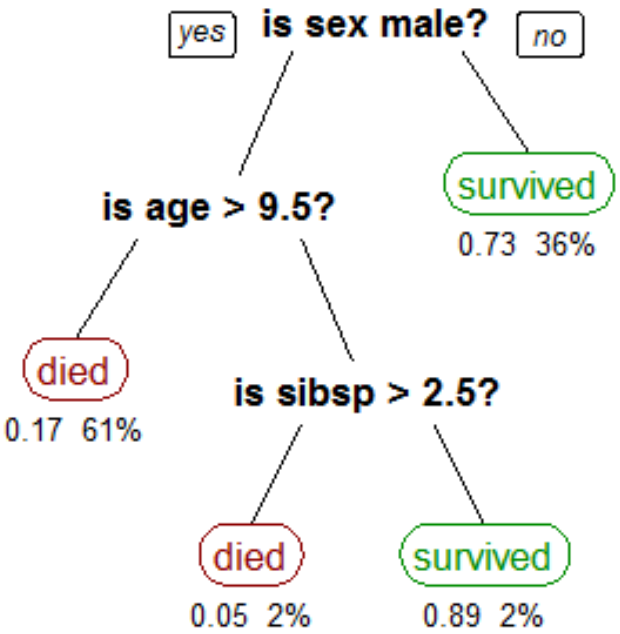
5 Modello di Classificazione

Questo modello utilizza un DecisionTreeClassifier per classificare se una persona gradisce o meno una canzone basandosi sulle caratteristiche fornite nel DataFrame. La preparazione dei dati, la codifica delle variabili categoriche, la divisione dei dati e la valutazione delle prestazioni sono passaggi cruciali in questo processo. Inoltre uso il One Hot Encoding che è una codifica per convertire le variabili categoriche in array di soli 0 ed 1 poichè alcuni modelli potrebbero non accettare valori diversi da quelli numerici, in quanto le stringhe per loro non hanno significato.

Come ultimo passaggio verifichiamo i risultati che ha ottenuto grazie ai dati che avevamo messo da parte con il test set. Come accuratezza abbiamo una percentuale che va dallo 0.76 all'1.00 e come precisione dallo 0.75 al 1.00

5.1 Decision Tree Classifier

La scelta del Decision Tree Classifier è stata fatta perchè offre una chiara interpretazione delle decisioni prese dal modello, in quanto può essere visualizzato sotto forma di un albero decisionale. La struttura a ramificazione dell'albero inoltre consente di modellare in modo più diretto le decisioni basate su caratteristiche musicali rilevanti per gli utenti. Quello che vanno a fare gli alberi è di mettere come radice la migliore caratteristica del training set e va a dividere questa in altri sotto-insiemi ripetendo questi passaggi finchè non si ottengono i nodi foglia, nell'albero decisionale i nodi rappresentano le caratteristiche mentre gli archi le decisioni. Inoltre sono molto facili per ottenere una predizione poichè basta navigare all'interno di questo albero fino ad arrivare ad un conclusione(nodo foglia).



5.2 One Hot Encoding

La scelta di utilizzare il One Hot Encoding per gestire le variabili categoriche nel dataset è stata presa perchè è particolarmente efficace nel trattare queste variabili, come il genere musicale, l'artista o altri attributi categorici presenti nelle canzoni. Trasformare queste variabili in rappresentazioni binarie facilita l'addestramento di modelli di machine learning. Inoltre, visto che alcuni algoritmi di machine learning potrebbero erroneamente attribuire un ordine o una gerarchia alle variabili categoriche,il one-hot encoding è stato scelto anche per evitare questo problema, poiché con questo ogni categoria ottiene una colonna separata, andando ad eliminare l'interpretazione di un ordine implicito.

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

6 Conclusioni e Sviluppi Futuri

In conclusione, il sistema implementato costituisce un modello di classificazione che offre agli utenti la possibilità di arricchire una lista personalizzata di canzoni preferite. Attraverso l'interazione attiva dell'utente, il sistema raccoglie preziosi feedback positivi e negativi, creando una solida base di dati per l'addestramento del modello di machine learning. Questo modello, alimentato dalle preferenze musicali espresse dagli utenti, è in grado di apprendere e identificare i tratti distintivi delle canzoni gradite e non gradite. Gli obiettivi che mi ero preposto all'inizio sono stati soddisfatti, ma non del tutto, perchè al momento con questo modello posso semplicemente fare una predizione del gradimento di una canzone già presente all'interno del mio dataset.

Per sviluppi futuri, è possibile ampliare e perfezionare il sistema attraverso l'introduzione di nuove feature, l'utilizzo di algoritmi di machine learning più complessi, l'esplorazione di tecniche avanzate di embedding per rappresentare meglio le caratteristiche musicali oppure integrare tecniche per farsi che l'utente tramite la propria playlist possa cercare delle canzoni che non sono presenti in quest'ultima che possano piacergli.