Hit or flop?

**NOMINATIVO + MATRICOLA COMPONENTI GRUPPO**

Calvano Miriana - 699645

Curci Antonio - 697498

Lomonte Nunzia – 697175

**LINK REPOSITORY:** https://github.com/NunziaL/ICon20-21\_Cal-Cur-Lom/tree/main

Questo progetto si basa su **quattro** argomenti principali: le ontologie, i modelli supervisionati, il classificatore bayesiano e l’apprendimento non supervisionato con clustering.

Istruzioni di esecuzione

Comandi di esecuzione

Per eseguire il file.py occorre scaricare la cartella del progetto da GitHub e successivamente digitare sul terminale i seguenti comandi:

1. cd *percorso\_cartella*/ICon20-21\_Cal-Cur-Lom
2. python main.py

Dopo aver lanciato i due comandi sopra citati, si aprirà la seguente schermata:

Immagine che contiene testo

Descrizione generata automaticamente

L’applicazione fornisce **3 funzionalità** principali, ognuna delle quali include degli argomenti diversi del programma:

Per poter garantire un’esecuzione corretta, è necessario inserire il nome di una canzone nella barra di ricerca, selezionare un’opzione tra le tre indicate e, infine, premere “*Invio*”

1. Selezionare “*Decennio*” per **predire** il **decennio** della canzone inserita dall’utente

Immagine che contiene testo

Descrizione generata automaticamente

1. Selezionare “*Previsione*” per **predire** se la canzone inserita dall’utente è stata un **flop oppure una hit:**

Immagine che contiene testo

Descrizione generata automaticamente

1. Selezionare “*Suggerimenti*” per fornire all’utente dei **suggerimenti** con delle **canzoni simili** a quella presa in esame

Immagine che contiene testo

Descrizione generata automaticamente

***NB***: Se viene premuto il pulsante “*Invio*”, ma non viene selezionata nessuna opzione o se la barra di ricerca è vuota, verrà visualizzato il seguente messaggio:

Immagine che contiene testo

Descrizione generata automaticamente

**Informazioni su implementazione**

Il dataset utilizzato è stato preso da <https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset>; consiste in 6 file con estensione .csv in cui sono conservate le informazioni riguardanti le canzoni presenti nel database di Spotify. Ogni canzone presenta i seguenti attributi: *track, artist, uri, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration\_ms, time\_signature, chorus\_hit, sections, target*. In totale, sommando tutte le entrate per ogni file, le canzoni prese in esame sono circa 41000.

**Acquisizione del dataset**

Il dataset viene acquisito dai file .csv. Durante questa fase, viene aggiunto un attributo “*decade*”, che indica il decennio di appartenenza delle canzoni in base al nome del file in cui esse si trovano (ES: *“Misty Roses, Astrud Gilberto”* avrà in corrispondenza dell’attributo “*decade*” il valore *1990* perché si trova nel file dataset-of-90’s).

**Identificazione canzone tramite URI**

Successivamente, viene chiesto all’utente di fornire in input una stringa che dovrà contenere almeno il titolo della canzone che si vuole analizzare (ES: “*your song elton john*” oppure “*your song*”), simulando una semplice ricerca che si farebbe nell’applicazione di Spotify poiché sono state sfruttate le API di Spotify attraverso il modulo *spotipy*.

È stato possibile risalire all’**URI** (*Uniform Resource Identifier*) per poter identificare il brano in maniera univoca. Grazie alla funzione spotipy.audio\_features() è stato possibile ottenere le features della canzone (danceability, energy, key, ecc…).

**Classificatore bayesiano**

Nella fase successiva, attraverso il classificatore bayesiano viene eseguita la previsione della decade di appartenenza della canzone inserita in input dall’utente, in quanto questa informazione non è presente nelle features fornite da Spotify. Il **classificatore *bayesiano*** è basato sull'applicazione del [teorema di Bayes](https://it.wikipedia.org/wiki/Teorema_di_Bayes) e richiede la conoscenza delle probabilità a priori e condizionali relative al problema al fine di assegnare la classe di appartenenza all’oggetto preso in considerazione.

**Apprendimento supervisionato**

Nel modulo “supervisionato.py” sono state definite le funzioni necessarie per la fase di predizione. Durante questa fase dell’esecuzione vengono rimossi dal dataset gli attributi “track, artist e uri” in quanto non contengono informazioni utili all’apprendimento del mondo. Successivamente, il dataset viene diviso in due ulteriori dataframe, X e y, che, rispettivamente rappresentano un insieme di training e uno di test. Questi due insiemi saranno poi “splittati” in modo tale che le feature di appredimento siano separate da quelle di target. Dunque l’insieme y conterrà solo la feature “target”, mentre X tutte le altre. Affinché venga costruito un modello sulla base di dati confrontabili e ben pesati, viene svolta la standardizzazione dei valori.

Alla conclusione di questo step, vengono definiti i 6 modelli che verranno sfruttati per effettuare le predizioni: Regressione logistica, KNN, Alberi di decisione, Rete neurale, random forest e gradient boosting.

* **Regressione logistica:** è un modello di [regressione non lineare](https://it.wikipedia.org/wiki/Regressione_nonlineare) in cui vengono determinati i pesi di una funzione lineare appiattita dal sigmoide, minimizzando un errore sull’ insieme di esempi. Inoltre, esso viene utilizzato quando la [variabile dipendente](https://it.wikipedia.org/wiki/Variabile_dipendente) è di tipo [dicotomico](https://it.wikipedia.org/wiki/Variabile_dicotomica). Pertanto, l'obiettivo del modello è di stabilire la probabilità con cui un'osservazione può generare uno o l'altro valore della variabile dipendente; può inoltre essere utilizzato per classificare le osservazioni, in base alle proprie caratteristiche
* **KNN:** Il classificatore knn è usato per classificare gli oggetti basandosi sulle caratteristiche degli oggetti vicini a quello considerato. Il tutto viene implementato attraverso l’algoritmo *k-nearest neighbors* (k-NN) il cui input è costituito dai k esempi di addestramento più vicini nello spazio delle funzionalità, mentre l’output è un'appartenenza a una classe.
* **Alberi di decisione:** L’albero di decisione è un modello predittivo costituito da nodi interni, archi e nodi foglia. Esso viene costruito utilizzando tecniche di apprendimento a partire dall'insieme dei dati iniziali (data set), il quale può essere diviso in due sottoinsiemi: il training set sulla base del quale si crea la struttura dell'albero e il test set che viene utilizzato per testare l'accuratezza del modello predittivo così creato.
* **Reti neurali**: Sono state usate le reti neurali di tipo *MLPC*, quindi un classificatore a più strati che allena il modello secondo la back-propagation (*MLPC*) tramite la funzione *MLPClassifier*().
* **Random Forest**: Classifica i dati in base all’addestramento fatto su più alberi in cui vengono fatte le varie predizioni che poi saranno aggregate. Le feature su cui vengono fatte le predizioni sono permutate a random.
* **Gradient Boosting**: Con questo algoritmo di classificazione viene restituita una predizione che viene migliorata ad ogni iterata, in modo tale da poter minimizzarne l’errore.

A questo punto, ci sarà la possibilità di poter predire la classe tramite dei modelli allenati precedentemente salvati su file .sav oppure allenarli nuovamente in caso in cui questi file non dovessero essere presenti nel percorso del progetto. Verrà dunque definita la classe predetta per ogni modello.

**Apprendimento non supervisionato: clustering**

Infine, il sistema restituisce all’utente una lista di canzoni simili a quella inserita. I risultati sono elaborati durante ad una fase di **clustering** basato sull’algoritmo di **KMeans**, in cui sono calcolati 4000 cluster. Si tratta di apprendimento non supervisionato in cui i dati sono divisi in gruppi di dati che hanno caratteristiche simili attraverso la definizione di un centroide. Grazie a questa fase è possibile risalire al cluster di appartenenza della canzone inserita dall’utente per restituire all’utente una lista di canzoni che appartengono allo stesso cluster presenti nel dataset.