CLASSIFICAZIONE DI GENERI CINEMATOGRAFICI

Progetto di Ingegneria della Conoscenza

A.A. 2020/2021

Componenti del gruppo:

- *Antony Campana*
  - Matricola: 698341 E-mail: a.campana9@studenti.uniba.it
- *Alessio Cervino*
  - Matricola: 697884 E-mail: a.cervino@studenti.uniba.it

Link progetto su GitHub

1. INTRODUZIONE

Il progetto realizzato è un classificatore di generi cinematografici di tipo multi-label (variante del problema della classificazione che ammette per ogni film l'assegnazione di uno o più generi). La classificazione si basa sull'analisi della trama di ogni film presente in un dataset che contiene più di 40000 film con metadati collezionati da IMDB (Internet Movie Database), sito web che gestisce informazioni su film, attori, registi, personale di produzione e programmi televisivi. Il dataset è contenuto nella repository con il nome "movies_metadata.csv". Il task consiste nel confronto e nella valutazione di alcuni dei principali modelli di classificazione basati su apprendimento supervisionato e nella predizione di generi di film scelti dall'utente in input con la descrizione della trama, oppure preimpostati in un set di prova (è stato utilizzato un file xlsx, foglio di calcolo Excel). L'apprendimento supervisionato è una tecnica di machine learning che mira a istruire un sistema informatico in modo da consentirgli di elaborare automaticamente previsioni sui valori di uscita di un sistema rispetto ad un input sulla base di una serie di esempi ideali, costituiti da coppie di input e di output, che vengono inizialmente forniti al modello.

GENERI PRESENTI:

- Drammatico
- Commedia
- Thriller
- Romantico
- Azione
- Horror
- Crimine
- Documentario
- Avventura
- Fantascienza
- Famiglia
- Mistero
- Fantasy
- Animazione
- Straniero
- Musicale
- Storia
- Guerra
- Western
- Serie TV

I modelli di classificazione utilizzati sono:

- Decision Tree Classifier
- Random Forest Classifier
- Extra Trees Classifier
- Multi-layer Perceptron (MLP) Classifier

Al termine dell'addestramento, il modello più performante in termini di accuratezza, precisione e richiamo è stato il Multi-layer Perceptron (MLP).

2. GESTIONE DEL DATASET

Per realizzare questo sistema, sono state isolate dal dataset le features *title*, *overview* e *genres* e sono stati eliminati i film aventi valore N/A nella feature *overview* e *genres*, ossia senza una trama e senza genere. Successivamente, ad ogni genere è stato associato un valore binario con l'utilizzo del

MultiLabelBinarizer e sono stati selezionati soltanto i primi 20 generi con più occorrenze, in quanto gli altri sono case cinematografiche o canali tv.

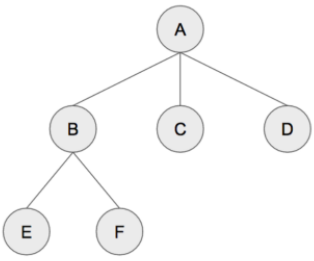
Il dataset è stato, quindi, suddiviso in una parte di training che si occupa dell'addestramento del modello (70% della dimensione del dataset) e in una parte di validation (30% della dimensione del dataset) che valuta le prestazioni del sistema addestrato.

Per addestrare il modello, è stato necessario convertire ogni trama in valori TF-IDF, usati per misurare l'importanza di un termine rispetto ad ogni trama. L'idea alla base di questo comportamento è di dare più importanza ai termini che compaiono nella trama, ma che in generale sono poco frequenti.

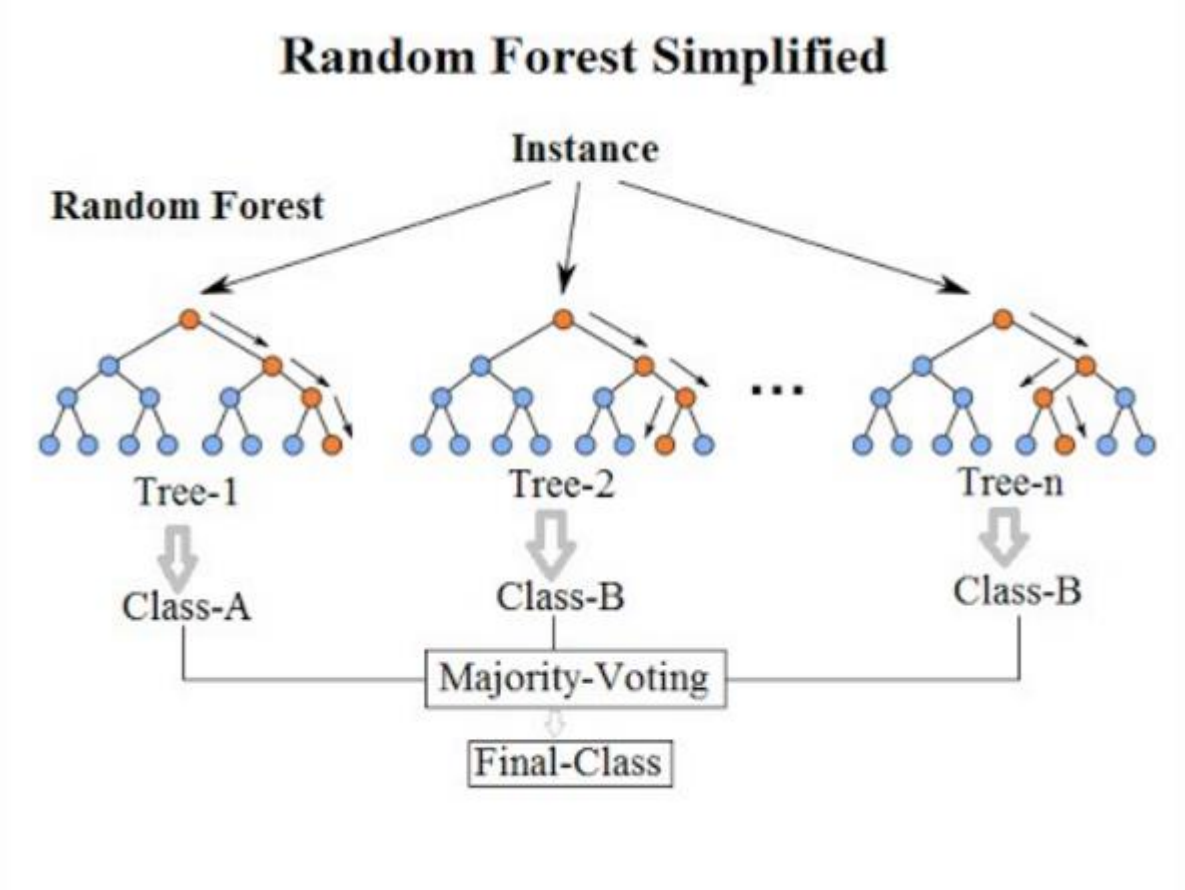
3. MODELLI DI CLASSIFICAZIONE

Al fine di ottenere una predizione sui nuovi esempi, sono stati applicati modelli di classificazione basati su apprendimento supervisionato, derivati dalla libreria di Python *sklearn*. L'idea di utilizzare più modelli ha avuto lo scopo di valutare l'accuratezza di ogni singolo modello in fase di test.

1. Decision Tree Classifier:
   Un albero di decisione è un modello predittivo, dove ogni nodo interno rappresenta una variabile, un arco verso un nodo figlio rappresenta un possibile valore per quella proprietà e una foglia il valore predetto per la variabile obiettivo a partire dai valori delle altre proprietà, che nell'albero è rappresentato dal cammino (path) dal nodo radice (root) al nodo foglia. Normalmente un albero di decisione 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 validation set che viene utilizzato per testare l'accuratezza del modello predittivo così creato.

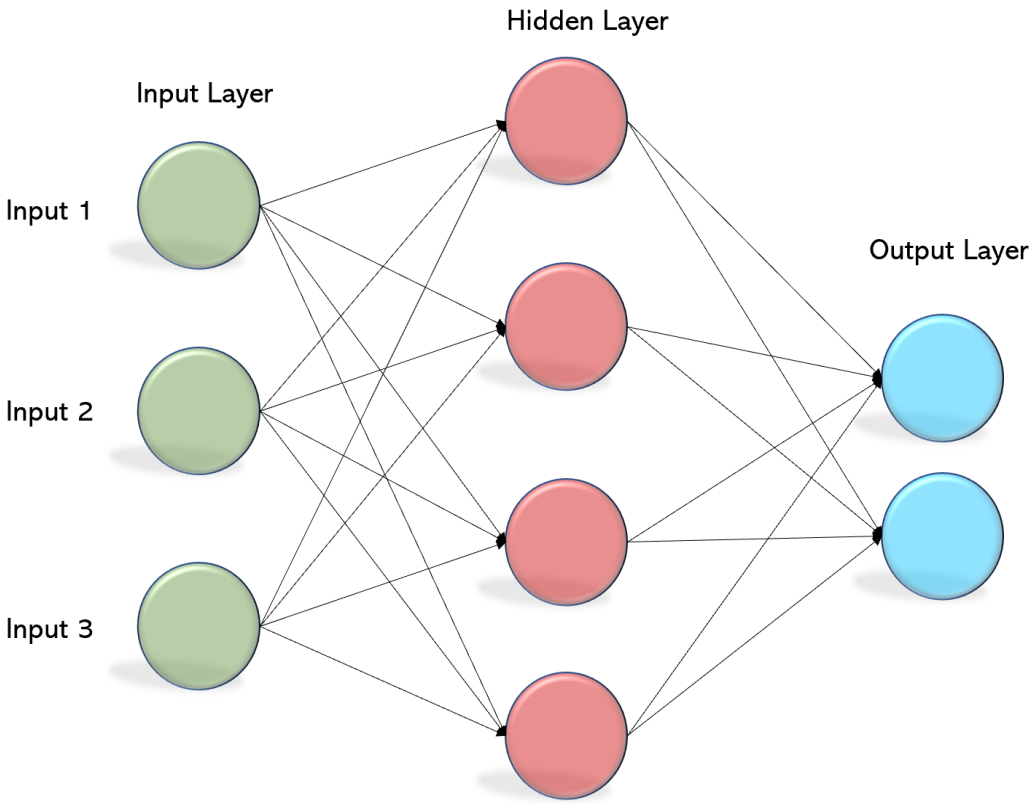


2. Random Forest:
   Una foresta casuale (random forest) è un classificatore d'insieme ottenuto dall'aggregazione di alberi di decisione. Le foreste casuali si pongono come soluzione che minimizza l'overfitting del training set rispetto agli alberi di decisione.



3. Extra Trees Classifier:
   L' Extremely Randomized Trees utilizza un numero di alberi decisionali casuali (noti anche come "extra trees") su vari campioni del dataset e utilizza la media per migliorare l'accuratezza predittiva e controllare l'overfitting. Rispetto al Random Forest Classifier, l'Extra Trees presenta un algoritmo computazionalmente meno costoso e, quindi, più veloce, in quanto non calcola il set ottimale ma lo seleziona casualmente.

4. Multi-Layer Perceptron:
   Il Percettrone multistrato (in acronimo MLP dall'inglese Multi-Layer Perceptron) è un modello di rete neurale artificiale che mappa insiemi di dati in ingresso in un insieme di dati in uscita appropriati.
   È fatta di strati multipli di nodi in un grafo diretto, con ogni strato completamente connesso al successivo. Il Percettrone multistrato usa una tecnica di apprendimento supervisionato chiamata backpropagation, un algoritmo che confronta il valore in uscita del sistema con il valore desiderato e, sulla base della differenza così calcolata (errore), l'algoritmo modifica i pesi sinaptici della rete neurale, facendo convergere progressivamente il set dei valori di uscita verso quelli desiderati.



ACCURATEZZA DEI CLASSIFICATORI UTILIZZATI:

	Training Set Accuracy	Validation Set Accuracy
Decision Tree	~ 0.992304	~ 0.100725
Random Forest	~ 0.991967	~ 0.155221
Extra Trees	~ 0.992304	~ 0.152386
MLP	~ 0.366131	~ 0.157346

Il Decision Tree è il classificatore meno efficiente in quanto tende ad overfittare il set.

Il Random Forest e l'Extra Trees danno risultati di accuratezza simili e migliori rispetto a quelli del Decision Tree.

L'MLP anche se ha una bassa accuratezza nel training set, raggiunge i risultati migliori di accuratezza nel validation set rispetto agli altri tre classificatori.

L'accuratezza dei vari modelli non raggiunge percentuali elevate perché la predizione è di tipo multi-label. Ad esempio, se si addestra il modello analizzando la trama di un film contenente il termine "arma", il modello assegnerà al suddetto film i generi <thriller, azione, guerra> presenti nel dataset. Tuttavia, può accadere che film contenenti il termine "arma" possano appartenere ai generi <western, azione>, facendo diminuire l'accuratezza della predizione.

4. FASE DI PREDIZIONE DA INPUT

Dopo aver ricevuto i risultati delle valutazioni delle precisioni di tutti i classificatori, l'utente decide quale di questi quattro modelli utilizzare. Successivamente, l'utente sceglie se importare un set di prova preesistente da file contenente i titoli e le trame di alcuni film oppure se inserire manualmente i titoli e le trame di film, in modo tale da predirne i generi. L'output sarà una tabella contenente per ogni film i generi predetti.

```
The Shawshank Redemption                                  [Drama]
The Godfather                          [Drama, Thriller, Action, Crime]
The Godfather: Part II                                    [Drama]
Schindler's List                                        [Drama, War]
The Lord of the Rings: The Return of the King             [Drama]
Pulp Fiction                           [Drama, Thriller, Action, Crime]
Forrest Gump                                            [Drama, War]
The Lord of the Rings: The Two Towers                     [Comedy]
Star Wars: Episode V - The Empire Strikes Back    [Science Fiction, Animation]
Goodfellas                                        [Thriller, Crime]
One Flew Over the Cuckoo's Nest                     [Drama, Crime]
Parasite                                                  [Comedy]
```

In questo test, è stato utilizzato il classificatore MLP.

5. CONCLUSIONI

L'MLP Classifier restituisce risultati più accurati, in quanto a differenza degli altri tre classificatori assegna più generi precisi ai film che l'utente ha dato in input al sistema. Ad esempio, utilizzando il Random Forest Classifier come modello predittivo, può capitare che non venga assegnato alcun genere ad un film.

Durante le fasi di codifica dei modelli predittivi, è stata provata l'ottimizzazione degli iperparametri tramite gli algoritmi di CV (cross validation) **Grid Search** e **Randomized Search**. Quest'ultimo, rispetto al primo, è più veloce e meno costoso a livello computazionale nella ricerca di parametri migliori per i modelli di classificazione. Nel nostro caso, entrambi gli algoritmi con iperparametri settati non hanno restituito risultati migliori rispetto all'utilizzo dei parametri standard.