Capitolo 1

Implementazione

1.1 Data Preprocessing

La fase di preprocessing dei dati consiste nel preparare i dati per l'addestramento del modello. Una volta acqusiti i dati e averne studiato le peculiaritá e le caratteristiche, si procede con la pulizia dei dati. Abbiamo usato il metodo *pivot_table* di Pandas per creare una tabella pivot che contiene i genome-score degli utenti per ogni film che lo possiede(andando cosi a ridurre la cardialita' da 60k a 13.816).
In seguito a ció verrá effettuata una merge con un i generi dei film applicando una One-Hot Encoding, attraverso la funzione *get_dummies()*. Ci siamo inoltre andati a focalizzare sui voti da parte degli utenti, raggruppandoli per film ci abbiamo calcolato la media dei voti. Quest'ultimo capo sará il nostro *target* per l'addestramento del modello.

1.2 Modeling

La nostra fase di modellazione include uno studio di regressione basato sulla predizione del voto medio di un film dati i suoi genome-score(ed in seguito anche i generi). Abbiamo inoltre applicato una tecnica di *PCA* per ridurre la cardinalitá dei dati, ma non abbiamo ottenuto risultati soddisfacenti, quindi abbiamo deciso di non applicarla.

Eseguiamo il train-test split con un rapporto 80-20 ed in seguito addestreremo il nostro modello.

1.2.1 Tecniche di ML Supervisionate con Approccio non-Deep

In questa sezione verranno descritte le tecniche di ML supervisionate con approccio non-Deep utilizzate per la predizione del voto medio di un film.

- **Linear Regression**: La regressione lineare é una tecnica di ML supervisionata che permette di predire il valore di una variabile dipendente a partire da una o piú variabili indipendenti.
  - –
  - **Lasso**: Lasso é un algoritmo di ML supervisionato che permette di effettuare la regressione di un dato mediante la creazione di un modello lineare.
- **Random Forest**: Random Forest é un algoritmo di ML supervisionato che permette di effettuare la classificazione o la regressione di un dato mediante la creazione di piú alberi (Metodo di Emsable Learning).
- **SVR**: Support Vector Regression é un algoritmo che cerca di trovare una funzione che minimizzi la distanza tra i dati di training e una fascia di tolleranza definita dall'utente.
- **KNN**: K-Nearest Neighbors é un algoritmo di ML supervisionato che permette di effettuare la classificazione o la regressione di un dato cerca di stimare il valore di una variabile dipendente su nuovi dati in base alla loro vicinanza ad altri dati di training.
- **NB Gaussian**: Naive Bayes é un algoritmo di ML supervisionato che permette di effettuare la classificazione o la regressione di un dato mediante la creazione di un modello probabilistico.

Per ogni modello precedentemente citato é stato applicato il Tuning dei parametri per cercare di ottimizzare il modello. Attraverso la funzione *RandomizedSearchCV* di Scikit-Learn abbiamo cercato di trovare i migliori parametri per ogni modello.

1.2.2 Tecniche di ML Supervisionate con Reti Neurali

Tabular Data