Introduzione alla Data Science
Progetto

Per ogni passaggio, commentare opportunamente e fornire giustificazioni delle scelte operate.

Dataset

Ti sono stati assegnati i dataset

- netflix_titles_1.csv
- netflix_titles_2.csv
- amazon_titles_1.csv
- amazon_titles_2.csv

1. Integrazione

Per ciascun gruppo di dati hai 2 tabelle ***titles***. Procedi con i seguenti passi:

- Fai join tra quest 2 tabelle per ottenere una tabella integrata
- Crea una nuova tabella <piattaforma>_titles_combinata.csv ottenuta a partire da <piattaforma>_titles_1.csv ed aggiungendo le colonne *date_added* e *country* che trovi in <piattaforma>_titles_2.csv
- Procedi con la pulizia dei dati: elimina colonne inutili o ripetute, righe non significative, gestisci la presenza di eventuali valori nulli o mancanti, ecc...

Dopo aver preparato i due dataset, puoi decidere di concatenarli per ottenerne uno unico, oppure mantenerli separati.

2. Trasformazione

- Sostituisci la colonna *date_added* con 2 colonne *year_added* e *month_added*
- Sostituisci la colonna *genres* con una colonna *genres_number* che contiene il numero di generi associati a quel dato

3. Esplorazione

Rappresentare la distribuzione dell'*imdb_score* suddividendo i programmi per *age_certification*

Rappresentare il numero di programmi (FILM + TV SHOW) prodotti negli anni. Rappresentare la stessa informazione rispetto all'anno di caricamento sulla piattaforma.

4. Test Statistico

Verificare con un test statistico se ci siano differenze significative tra la distribuzione di FILM suddivisi per anno di produzione o per anno di caricamento sulla piattaforma tenendo separate i dati delle due piattaforme. Ripetere lo stesso sul tipo TV SHOW. Infine, eseguire un unico test sull'intero insieme di dati.

5. **OLAP**

   Costruire una rappresentazione OLAP che conteggi i dati nelle due piattaforme raggruppando per

   - Anno di caricamento sulla piattaforma
   - Tipologia (TV SHOW o FILM)
   - Paese di produzione

   Proporre e discutere 1 visualizzazione

6. **Metodi predittivi**

   Creare un descrittore composto da

   - imdb_score
   - tmdbs_score
   - tmdbs_popularity
   - runtime

   ed utilizzarlo come input per un metodo predittivo supervisionato (es. regressione logistica) per predire type (ossia la categoria che può essere FILM o TV SHOW).

   Usare come training i dati di una piattaforma, e come test quelli della seconda.