

## Laboratorio 7

### Sistema de recomendaciones

**Instrucciones:** Usted usará Python a través de un Jupyter Notebook para llevar a cabo este laboratorio. Recuerde utilizar comentarios para describir lo que está haciendo en cada parte del proceso y así dejar constancia del entendimiento de su trabajo. Estará usando el dataset proporcionado dentro del portal. Al finalizar recuerde subir al portal un link de su repositorio en el que se pueda correr su notebook (no enviarlo por correo).

El juego de datos proporcionado es de 3 archivos para el sistema de recomendación de películas:

- El archivo `movies_metadata.csv` comprimido que contiene la información de varias películas:
  - `Adult`
  - `belongs_to_collection`
  - `Budget`
  - `Genres` (género (característica))
  - `Homepage`
  - `Id` (identificador de película TMDb)
  - `imdb_id` (identificador de película IMDb)
  - `original_language`
  - `original_title`
  - `overview`
  - `popularity`
  - `poster_path`
  - `production_companies`
  - `production_countries`
  - `release_date`
  - `revenue`
  - `runtime`
  - `spoken_languages`
  - `status`
  - `tagline`
  - `title`
  - `video`
  - `vote_average`
  - `vote_count`.
- El archivo `links_small.csv` que contiene el identificador de película (TMDb) y el identificador de película (IMDb) de un pequeño subconjunto de películas del juego de datos completo:
  - `movieId` (identificador de la película)

- imdbId (identificador de película IMDB)
- tmdbId (identificador de película TMDB).
- El archivo ratings\_small.csv comprimido que contiene un subconjunto de evaluaciones de varios usuarios:
  - userId (identificador del usuario)
  - movieId (identificador de la película)
  - rating (evaluación de la película)
  - timestamp (fecha de registro).

El detalle de las películas ha sido obtenido del TMDB Open API. Los enlaces y evaluaciones se han obtenido del sitio web oficial de GroupLens.

Para este ejercicio se le pide que utilice el algoritmo de contenido basado en recomendaciones y filtrado colaborativo, usando las características (géneros):

- {'id': 10402, 'name': 'Music'}
- {'id': 10749, 'name': 'Romance'}
- {'id': 10751, 'name': 'Family'}
- {'id': 10752, 'name': 'War'}
- {'id': 10769, 'name': 'Foreign'}
- {'id': 10770, 'name': 'TV Movie'}
- {'id': 12, 'name': 'Adventure'}
- {'id': 14, 'name': 'Fantasy'}
- {'id': 16, 'name': 'Animation'}
- {'id': 18, 'name': 'Drama'}
- {'id': 27, 'name': 'Horror'}
- {'id': 28, 'name': 'Action'}
- {'id': 35, 'name': 'Comedy'}
- {'id': 36, 'name': 'History'}
- {'id': 37, 'name': 'Western'}
- {'id': 53, 'name': 'Thriller'}
- {'id': 80, 'name': 'Crime'}
- {'id': 878, 'name': 'Science Fiction'}
- {'id': 9648, 'name': 'Mystery'}
- {'id': 99, 'name': 'Documentary'}.

A continuación se mencionan las generalidades de los pasos sugeridos a realizar:

1. Leer los archivos CSV proporcionados y almacenarlos en un **np.array** correspondiente para ser trabajado en el notebook.
2. Ajustar un modelo de regresión múltiple en base al juego de datos cargado que relacione cualquier subconjunto de características.

3. Asegúrese de hacer los ajustes y trabajos necesarios al modelo que ha implementado para describir y predecir la información sobre la nube de datos. No olvide utilizar el algoritmo de contenido basado en recomendaciones y filtrado colaborativo, usando las características que considere necesarias.
4. Haga un análisis sobre sus hallazgos, donde mencione claramente las razones por las que considera que su modelo es bueno justificando adecuadamente.