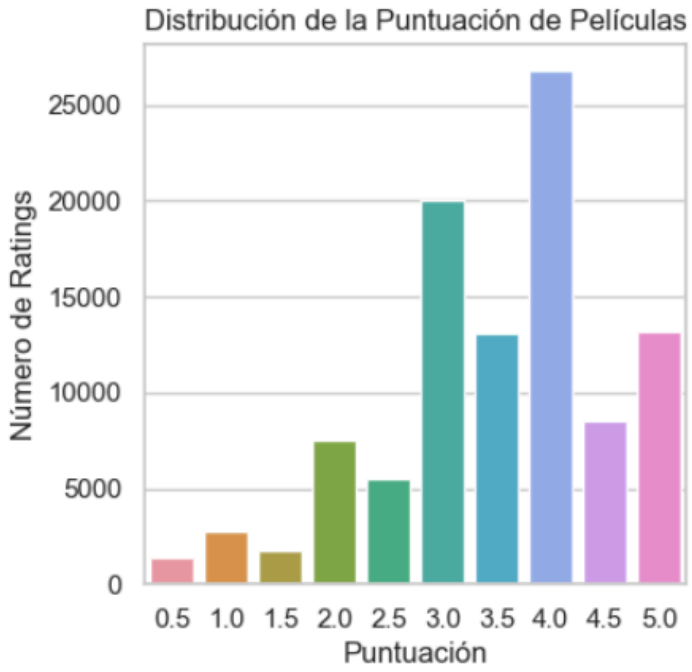
Sistema de Recomendación por Distancias usando el dataset Movielens

Universidad Nacional de San Agustin
Escuela Profesional de Ciencias de la Computación
EDWIN FREDY CHAMBI MAMANI
echambimam@unsa.edu.pe
repositorio: https://github.com/

Arquitectura para el Análisis de Datos

Analizar un dataset de 1M millón de registros utilizando diferentes medidas de distancia y similitud (Manhattan, Euclidiana, Pearson y Coseno) implica realizar operaciones sobre datos para comprender relaciones y patrones, para:

Métrica	Valor
Número de Ratings	100836
Número de ID de Películas Únicos	9724
Número de ID de Usuarios Únicos	610
Promedio de Ratings por Usuario	165.3
Promedio de Ratings por Película	10.4



1. Distancia de Manhattan

La distancia de Manhattan (también conocida como distancia L1) entre dos puntos se calcula como la suma de las diferencias absolutas de sus componentes. En el contexto de un dataset de películas y ratings:

Aplicación: Se puede usar para medir la distancia entre dos vectores de ratings de películas por usuarios.
Proceso: Se calcula la suma de las diferencias absolutas entre las calificaciones de dos usuarios para todas las películas comunes.
Inferencias: Menor distancia implica mayor similitud en términos de patrones de rating entre usuarios. Se utiliza comúnmente en sistemas de recomendación basados en contenido para encontrar usuarios o ítems similares.

```
Ingrese el ID del usuario a recomendar:  5
Los 5 géneros favoritos del usuario 5 son: Drama, Children, Comedy, Crime, Animation
Ingrese el número de películas a recomendar:  5

Tenemos 5 películas recomendadas para el usuario 5 usando distancia Manhattan:
```

No.	Título	Géneros	Rating Promedio	Cantidad de Puntuaciones
1	Ace Ventura: When Nature Calls (1995)	Comedy	2.73	88
2	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery|Sci-Fi|Thriller	3.98	177
3	Seven (a.k.a. Se7en) (1995)	Mystery|Thriller	3.98	203
4	Die Hard: With a Vengeance (1995)	Action|Crime|Thriller	3.56	144
5	Dumb & Dumber (Dumb and Dumber) (1994)	Adventure|Comedy	3.06	133

2. Distancia Euclidiana

La distancia Euclidiana (distancia L2) entre dos puntos en un espacio n-dimensional se calcula como la raíz cuadrada de la suma de las diferencias al cuadrado de sus componentes.

Aplicación: Se utiliza para medir la similitud entre vectores de características, como los perfiles de usuario o ítem.
Proceso: Se calcula la raíz cuadrada de la suma de las diferencias al cuadrado entre las calificaciones de dos usuarios para todas las películas comunes.
Inferencias: Se utiliza en sistemas de recomendación para encontrar usuarios o ítems que son similares en términos de sus perfiles de calificación. Es sensible a valores extremos.

```
Ingrese el ID del usuario a recomendar:  5
Los 5 géneros favoritos del usuario 5 son: Drama, Children, Comedy, Crime, Animation
Ingrese el número de películas a recomendar:  5

Tenemos 5 películas recomendadas para el usuario 5 usando distancia Euclidiana:
```

No.	Título	Géneros	Rating Promedio	Cantidad de Puntuaciones
1	Ace Ventura: When Nature Calls (1995)	Comedy	2.73	88
2	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery|Sci-Fi|Thriller	3.98	177
3	Seven (a.k.a. Se7en) (1995)	Mystery|Thriller	3.98	203
4	Die Hard: With a Vengeance (1995)	Action|Crime|Thriller	3.56	144
5	Dumb & Dumber (Dumb and Dumber) (1994)	Adventure|Comedy	3.06	133

3. Similitud Coseno

La similitud coseno mide el coseno del ángulo entre dos vectores en un espacio n-dimensional, proporcionando un valor entre -1 y 1, donde 1 significa que los vectores son idénticos en dirección, 0 indica independencia y -1 significa que los vectores son opuestos en dirección.

Aplicación: Se utiliza para medir la similitud entre vectores de características, como los perfiles de calificación de usuarios o ítems.

Proceso: Se calcula como el producto punto entre los vectores dividido por el producto de sus normas.

Inferencias: Una similitud coseno cercana a 1 indica una alta similitud entre los patrones de calificación de usuarios, independientemente de la magnitud de sus calificaciones.

```
Ingrese el ID del usuario a recomendar:  5
Los 5 géneros favoritos del usuario 5 son: Drama, Children, Comedy, Crime, Animation
Ingrese el número de películas a recomendar:  5

Tenemos 5 películas recomendadas para el usuario 5 usando similitud de coseno:
```

No.	Título	Géneros	Rating Promedio	Cantidad de Puntuaciones
1	Jumanji (1995)	Adventure|Children|Fantasy	3.43	110
2	Grumpier Old Men (1995)	Comedy|Romance	3.26	52
3	Father of the Bride Part II (1995)	Comedy	3.07	49
4	Heat (1995)	Action|Crime|Thriller	3.95	102
5	Sabrina (1995)	Comedy|Romance	3.19	54

4. Correlación de Pearson

La correlación de Pearson mide la relación lineal entre dos conjuntos de datos, proporcionando un valor que va desde -1 (correlación negativa perfecta) hasta +1 (correlación positiva perfecta), con 0 indicando ausencia de correlación lineal.

Aplicación: Se aplica para medir la relación lineal entre los ratings dados por diferentes usuarios a películas.

Proceso: Se calcula utilizando la covarianza y las desviaciones estándar de los conjuntos de datos.

Inferencias: Un valor alto de correlación de Pearson (cercano a 1) indica una relación fuerte entre los patrones de calificación de los usuarios, mientras que valores cercanos a 0 indican falta de relación lineal.

```
Ingrese el ID del usuario a recomendar:  5
Los 5 géneros favoritos del usuario 5 son: Drama, Children, Comedy, Crime, Animation
Ingrese el número de películas a recomendar:  5

Tenemos 5 películas recomendadas para el usuario 5 usando correlación de Pearson:
```

No.	Título	Géneros	Rating Promedio	Cantidad de Puntuaciones
1	Jumanji (1995)	Adventure|Children|Fantasy	3.43	110
2	Grumpier Old Men (1995)	Comedy|Romance	3.26	52
3	Father of the Bride Part II (1995)	Comedy	3.07	49
4	Heat (1995)	Action|Crime|Thriller	3.95	102
5	Sabrina (1995)	Comedy|Romance	3.19	54

Resumen del Análisis

Dataset: Un dataset de 1M y 20M de registros de ratings de películas por usuarios.

Métodos de Análisis: Se aplicaron las distancias de Manhattan y Euclidiana junto con las medidas de correlación de Pearson y similitud coseno, donde el 1M si procesa y el de 20M muestra un overflow

```
C:\Users\CHAMBI\AppData\Roaming\Python\Python311\site-packages\pandas\core\reshape\reshape.py:125: RuntimeWarning:

overflow encountered in scalar multiply
```

Objetivo: Comprender la similitud entre usuarios o películas basándose en sus patrones de calificación.

Inferencias Generales: Cada medida proporciona una perspectiva diferente de similitud o relación entre los datos, lo que permite diferentes enfoques en sistemas de recomendación, agrupamiento de usuarios o análisis de tendencias.

Este tipo de análisis es crucial en sistemas de recomendación y análisis de datos para entender la estructura subyacente y las relaciones entre los datos, facilitando la toma de decisiones informadas en diferentes dominios como el comercio electrónico, entretenimiento digital y más.