Universidad Peruana de Ciencias Aplicadas

Trabajo Parcial

Curso: Aplicaciones de Data Science

Sección: 279

Docente: Carlos Fernando Montoya Cubas

Grupo: Grupo 2

Integrantes:

Rafael Tomas Chui Sanchez - U201925837

Rodrigo Alejandro Meza Polo - U202224016

Axel Yamir Pariona Rojas - U202222148

Liam Mikael Quino Neff - U20221e167

Mayo 2025-01

ÍNDICE

1. Introducción

En la era del contenido digital, los usuarios se enfrentan a una sobrecarga informativa al intentar seleccionar películas o series de extensos catálogos en plataformas como Netflix, Amazon Prime Video, Disney+ y HBO Max. Esta saturación, conocida como la "paradoja de la elección" (Iyengar & Lepper, 2000), genera frustración y complica la toma de decisiones, afectando la experiencia del usuario.

Ante esta situación, los sistemas de recomendación se han convertido en herramientas clave para personalizar el consumo de contenido, anticipando las preferencias del usuario y optimizando su navegación. Existen múltiples enfoques para ello, como el filtrado colaborativo, los modelos basados en contenido, y los sistemas híbridos, que combinan lo mejor de ambos mundos (Ricci et al., 2015).

En este proyecto se propone el desarrollo de un sistema de recomendación personalizado de películas, empleando técnicas de aprendizaje automático, procesamiento de lenguaje natural (NLP) y principios de ciencia de datos. Para ello, se construyó un dataset actualizado a partir de la API pública de The Movie Database (TMDb), que permitió recopilar información estructurada y semiestructurada sobre películas recientes.

A través de un enfoque completo, que abarca desde el análisis exploratorio de datos (EDA), preprocesamiento, visualización, hasta la modelización y validación de algoritmos, se desarrolló una solución que permite recomendar películas similares a las preferencias del usuario, predecir calificaciones promedio futuras, y clasificar títulos según su probabilidad de éxito comercial. Además, se integró una interfaz amigable que facilita la interacción del usuario final con el sistema.

2. Descripción del caso de uso

El caso de uso desarrollado en este proyecto tiene como objetivo construir un sistema de recomendación de películas personalizado, adaptado al perfil de un usuario hipotético. La propuesta se basa en aplicar técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático, utilizando datos reales obtenidos mediante la API pública de The Movie Database (TMDb). Esto nos permitió construir un dataset actualizado con información detallada sobre películas recientes.

El sistema considera múltiples características técnicas de cada título, como géneros, duración, presupuesto, popularidad, elenco, director, sinopsis y número de votos. Esta riqueza de atributos permite crear modelos que anticipen el interés del usuario por ciertas producciones, sin requerir un historial de visualización.

Este proyecto aborda uno de los principales retos en la industria del entretenimiento digital: la personalización de contenido en tiempo real. Según un informe de McKinsey (2013), un sistema de recomendaciones efectivo puede representar hasta un 35 % del consumo total en plataformas como

Netflix. Por tanto, diseñar soluciones inteligentes que mejoren la curaduría de contenido audiovisual resulta clave para mejorar la experiencia del usuario.

En función del objetivo planteado, se abordan tres preguntas clave:

- ¿Qué películas se pueden recomendar a un usuario basándose en atributos como género, elenco o palabras clave de películas que le hayan gustado previamente?
- ¿Se puede predecir el nivel de aprobación (voto promedio) que tendrá una nueva película, con base en sus características antes de su estreno?
- ¿Podemos clasificar si una película será un éxito comercial o no, considerando variables como presupuesto, popularidad, y equipo de producción?

Para responder estas preguntas se utilizan modelos tanto supervisados como no supervisados, incluyendo técnicas de regresión, clasificación binaria, y sistemas de recomendación basados en contenido, junto con la vectorización textual usando TF-IDF y la medida de similitud coseno. Todo el sistema está integrado en una interfaz gráfica amigable, que permite al usuario final acceder fácilmente a las recomendaciones, predicciones y clasificaciones generadas por el modelo.

3. Descripción del conjunto de datos (dataset):

El conjunto de datos utilizado en este proyecto fue construido a partir de la API pública de The Movie Database (TMDb), una base de datos colaborativa y abierta ampliamente utilizada para análisis de contenido audiovisual.

Para su elaboración, se desarrolló un script en Python que automatiza la recolección de información sobre películas próximas a estrenarse y recientemente lanzadas. Mediante peticiones a la API, se recopilaron características técnicas y descriptivas de cada película, generando así un archivo .csv estructurado con los siguientes campos:

- id: Identificador único de la película
- title: Título
- release_date: Fecha de estreno
- genres: Lista de géneros asociado, por ejemplo "Action" o "Thriller".
- overview: Sinopsis o descripción textual
- popularity: Métrica de popularidad en TMDb
- runtime: Duración en minutos
- production_companies: Productoras responsables
- cast: Lista de actores principales
- director: Director principal
- vote_average: Calificación promedio (si ya disponible)
- vote_count: Número de votos

- budget: Presupuesto estimado

El dataset final incluye películas previstas para el año 2025, lo cual permite trabajar con una muestra actualizada y representativa del mercado cinematográfico contemporáneo. Algunas variables, como genres, cast y production_companies, presentan formato de lista (semiestructurado), mientras que overview contiene texto libre (no estructurado), y los campos como popularity, runtime, vote_average o budget son datos numéricos.

Esta combinación de datos estructurados, semiestructurados y no estructurados permite abordar tareas tanto de regresión, como la estimación de la calificación promedio de una película; de clasificación, como la predicción del éxito comercial; y de recomendación personalizada, mediante el análisis del contenido textual y metadatos de cada título.

4. Análisis exploratorio de los datos (EDA)

El análisis exploratorio de datos (EDA) constituye una fase fundamental en los proyectos de ciencia de datos, ya que permite comprender la estructura, calidad y comportamiento de las variables antes de aplicar modelos predictivos o prescriptivos.

En este proyecto, se trabajó con un conjunto de datos generado a partir de la API pública de The Movie Database (TMDb), que reúne información actualizada de películas previstas para el año 2025. El dataset combina variables estructuradas, semiestructuradas y no estructuradas, lo cual presenta tanto oportunidades como desafíos en términos de preprocesamiento y análisis.

4.1. Carga e inspección de datos

El archivo movies_tmdb_2025.csv fue generado mediante un script en Python que automatiza la recolección de información desde la API pública de TMDb, consolidando los datos en formato estructurado para su posterior análisis.

Este archivo fue cargado utilizando la biblioteca Pandas, generando un dataframe con más de 40 películas (la cantidad puede variar según el momento de extracción de datos) y 13 atributos principales por registro.

- Variables semiestructuradas: genres, production_companies, cast
- Texto no estructurado: overview (sinopsis), title
- Variables numéricas: budget, popularity, runtime, vote_average, vote_count
- Variables categóricas estructuradas: director, release_date, release_year

Durante la inspección inicial, se identificaron valores nulos en columnas clave como budget, runtime y vote_average. Asimismo, las columnas tipo lista (genres, cast, production_companies) estaban codificadas como strings, por lo que fue necesario

convertirlas en estructuras manejables para análisis textual. Estos hallazgos motivaron un proceso riguroso de limpieza y preprocesamiento, que garantizó la integridad y utilidad del conjunto de datos para las etapas posteriores del proyecto.

4.2. Preprocesamiento de datos

Con el objetivo de disponer de un conjunto de datos limpio y adecuado para el análisis, se realizaron diversas tareas de preprocesamiento. Estas acciones permitieron mejorar la calidad de los datos, corregir inconsistencias y preparar la información para su posterior uso en los modelos de recomendación, regresión y clasificación. A continuación, se detallan los pasos aplicados:

- Conversión de tipos de datos: Se ajustaron los tipos de datos de cada columna según su naturaleza. Las variables numéricas como budget, popularity, runtime, vote_average y vote_count fueron convertidas explícitamente a tipo float. La columna release_date fue transformada al tipo datetime, y a partir de esta se generó una nueva columna release_year para permitir análisis temporales.
- Tratamiento de datos faltantes: Se identificaron valores nulos principalmente en las columnas budget, runtime y vote_average. Para no eliminar registros valiosos, se optó por técnicas de imputación:
    - En budget y runtime, los valores faltantes fueron reemplazados por la media aritmética de cada variable.
    - En vote_average, cuando fue necesario, se imputó con una calificación promedio neutral (por ejemplo, 5.0).

  Las columnas con datos faltantes irrelevantes para los modelos fueron descartadas durante la generación del dataset, por lo que no se requirió eliminar campos como homepage o tagline.

- Transformación de variables semiestructuradas: Las columnas genres, production_companies y cast, obtenidas desde la API en formato de listas codificadas como strings, fueron transformadas en texto plano mediante concatenación de los elementos. Esto permitió su posterior vectorización usando técnicas de procesamiento de lenguaje natural (NLP), como TF-IDF. Esta transformación también fue clave para la generación de un "perfil textual" de cada película, que se usa en el sistema de recomendación.
- Limpieza de texto: La columna overview, que contiene la sinopsis textual de cada película, fue sometida a un preprocesamiento básico que incluyó:
    - Conversión a minúsculas
    - Eliminación de caracteres especiales
    - Eliminación de espacios redundantes

Esto preparó el texto para ser representado mediante vectores numéricos (embeddings) en el modelo de recomendación basado en contenido.

- Eliminación de duplicados: Se verificó que no existieran registros duplicados mediante el uso de combinaciones únicas de id y title. No se identificaron duplicados en la extracción actual, por lo que no fue necesario realizar esta limpieza.
- Detección de outliers: Se identificaron valores atípicos en las variables budget, popularity y vote_count, especialmente aquellos con órdenes de magnitud muy superiores a la media. Para los campos numéricos se aplicaron métodos estadísticos (como el rango intercuartílico, IQR) y análisis visual mediante boxplots, a fin de detectar y mitigar el efecto de estos outliers. En algunos casos, se aplicó capping para limitar los valores extremos sin perder la integridad del conjunto de datos.

Este proceso de preprocesamiento aseguró la preparación adecuada del dataset para su uso en los algoritmos de aprendizaje automático, mejorando la calidad de los resultados y facilitando la interpretabilidad de los modelos construidos.

4.3. Estadísticas descriptivas

Se realizó un análisis estadístico descriptivo de las variables numéricas más relevantes del conjunto de datos, con el objetivo de identificar patrones generales, rangos de valores y posibles sesgos. Este análisis incluyó variables como presupuesto (budget), duración (runtime), calificación promedio (vote_average), número de votos (vote_count) y popularidad (popularity).

A continuación, se presenta un resumen estadístico generado con el método .describe() de Pandas:

	id	popularity	runtime	vote_average	vote_count	budget
count	9.999000e+03	9999.000000	9999.000000	9999.000000	9999.000000	9.999000e+03
mean	3.492977e+05	5.333389	103.243324	6.423010	1820.710371	2.158891e+07
std	4.191676e+05	17.441715	28.089014	1.305506	3286.366836	4.084619e+07
min	5.000000e+00	0.746000	0.000000	0.000000	0.000000	0.000000e+00
25%	1.183950e+04	2.291450	91.000000	5.993000	101.000000	0.000000e+00
50%	9.909900e+04	3.036000	102.000000	6.600000	654.000000	2.500000e+06
75%	6.048470e+05	4.940750	117.000000	7.200000	1965.000000	2.500000e+07
max	1.506450e+06	759.568400	367.000000	10.000000	37631.000000	4.600000e+08

- Presupuesto (budget): El presupuesto promedio de las películas es de aproximadamente 21.6 millones de dólares, pero la mediana es solo de 2.5 millones.

  Este contraste tan grande sugiere una distribución altamente asimétrica positiva, es decir, unas pocas películas con presupuestos extremadamente altos (hasta 460 millones) elevan el promedio general.

  El 25% de las películas tienen presupuesto 0, lo cual puede indicar valores no registrados o películas independientes con información incompleta desde la API.

- Duración (runtime): La duración media es de 103.2 minutos, con una mediana de 102 minutos, lo que indica una distribución relativamente simétrica.

  El rango de duración va desde 0 minutos (posiblemente valores faltantes o mal ingresados) hasta un máximo de 367 minutos, lo cual representa un outlier claro.

  El 25% de las películas duran menos de 91 minutos, mientras que el 75% no supera los 117. Esto está en línea con los estándares comerciales actuales.

- Calificación promedio (vote_average): Las calificaciones oscilan entre 0.0 y 10.0, con una media de 6.42 y una mediana de 6.6.

  La mayoría de películas se sitúan entre 6 y 7.2 puntos, lo que sugiere una tendencia a evaluaciones moderadamente positivas.

- La existencia de valores tan bajos como 0.0 podría indicar registros incompletos o títulos aún no valorados.
- Número de votos (vote_count): El número promedio de votos es de 1,821, pero la mediana es 654, lo que implica una distribución muy sesgada positivamente.

  Esto significa que unas pocas películas han sido votadas masivamente (hasta 37,631 votos), mientras que muchas otras han recibido poca atención.

  Nuevamente, el valor mínimo es 0, lo que puede indicar películas recién lanzadas o sin visibilidad.

- Popularidad (popularity): El valor promedio de popularidad es de 5.33, con una mediana de 3.03 y un máximo de 759.57, lo que indica una alta concentración de popularidad en unos pocos títulos.

  El mínimo es 0.746, lo cual también confirma que hay películas con muy baja tracción.

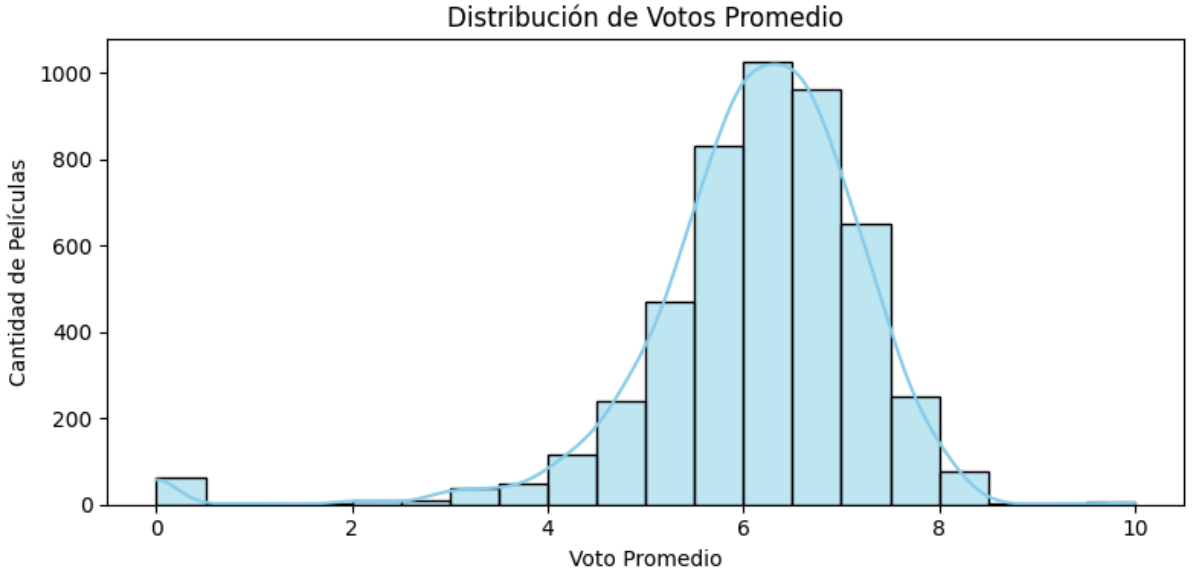
  La desviación estándar de 17.44 es muy superior a la media, confirmando que la variable tiene una alta dispersión.

Estas estadísticas permiten identificar que la mayoría de variables numéricas presentan distribuciones sesgadas, muchas veces con outliers extremos. Este comportamiento es común en datasets del mundo real donde los productos (en este caso, películas) tienen visibilidad y presupuestos muy diversos. Conocer estas características es fundamental para aplicar correctamente técnicas de normalización, imputación, o transformación de variables durante la etapa de modelado.

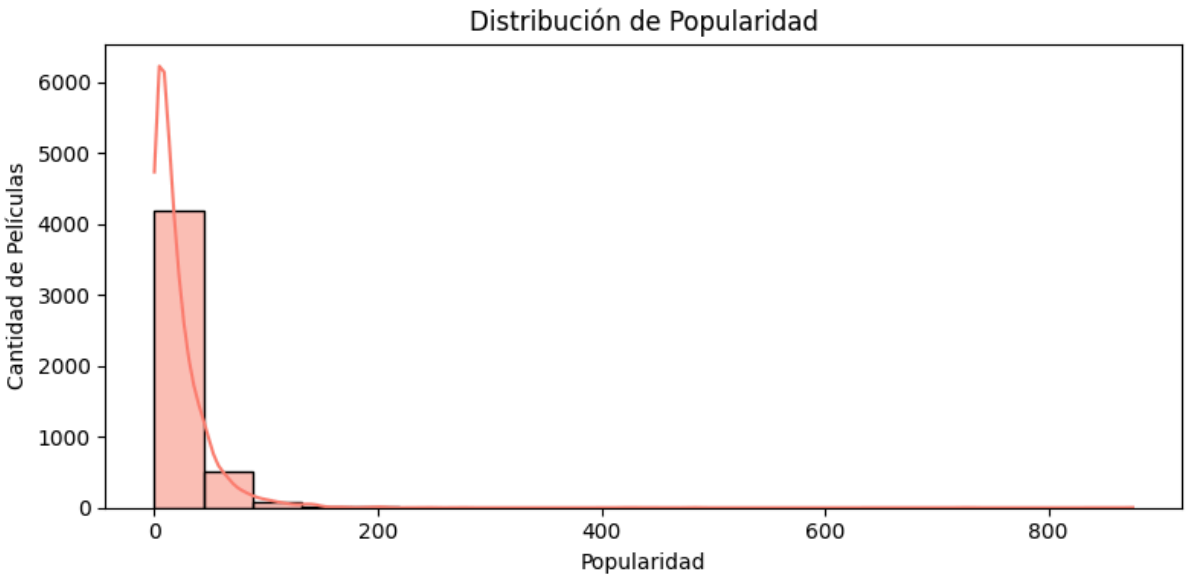
4.4. Visualización de datos

Con el objetivo de complementar el análisis estadístico y explorar visualmente las principales características del conjunto de datos, se realizaron diversas visualizaciones que permitieron identificar tendencias, relaciones y patrones relevantes. Estas representaciones gráficas facilitan la interpretación de los datos y ofrecen una base sólida para el desarrollo de modelos predictivos. A continuación, se describen las visualizaciones más representativas:

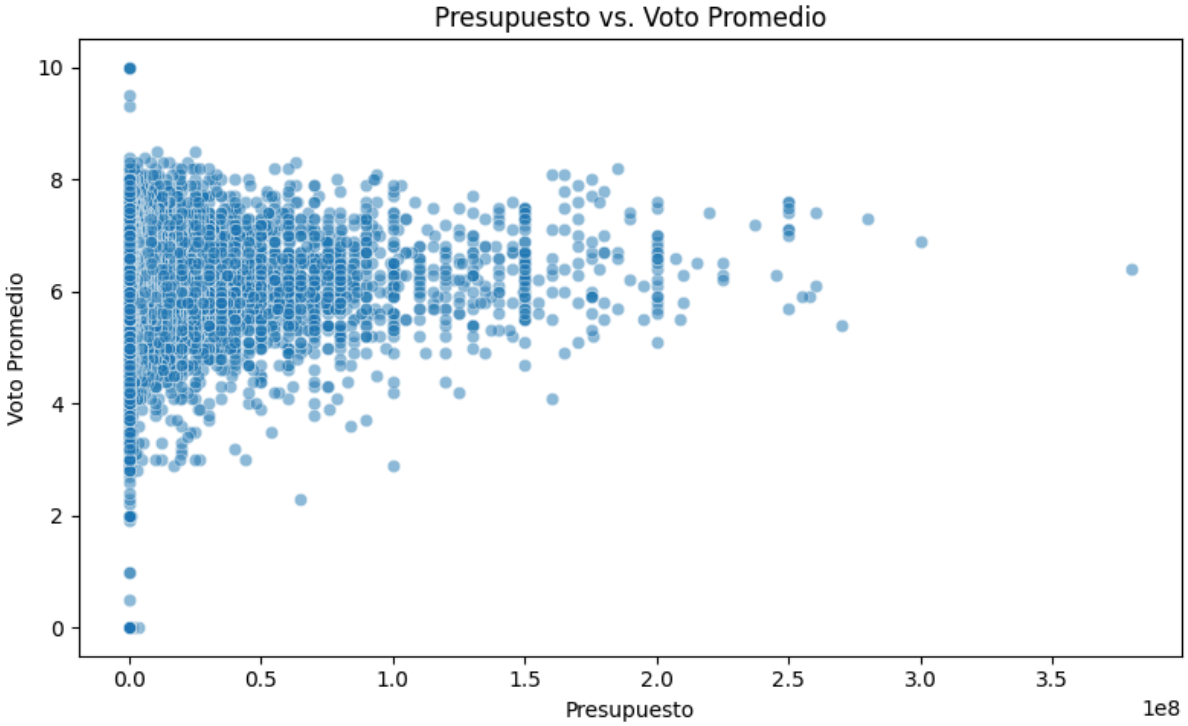
Distribución de calificaciones promedio: Se elaboró un histograma sobre la variable vote_average, con el fin de analizar cómo se distribuyen las calificaciones otorgadas a las películas. La mayoría se concentra entre los 5.5 y 7.5 puntos, mostrando una tendencia hacia valoraciones moderadamente positivas. La distribución presenta una leve asimetría negativa, lo que indica que las puntuaciones altas son más comunes que las bajas.



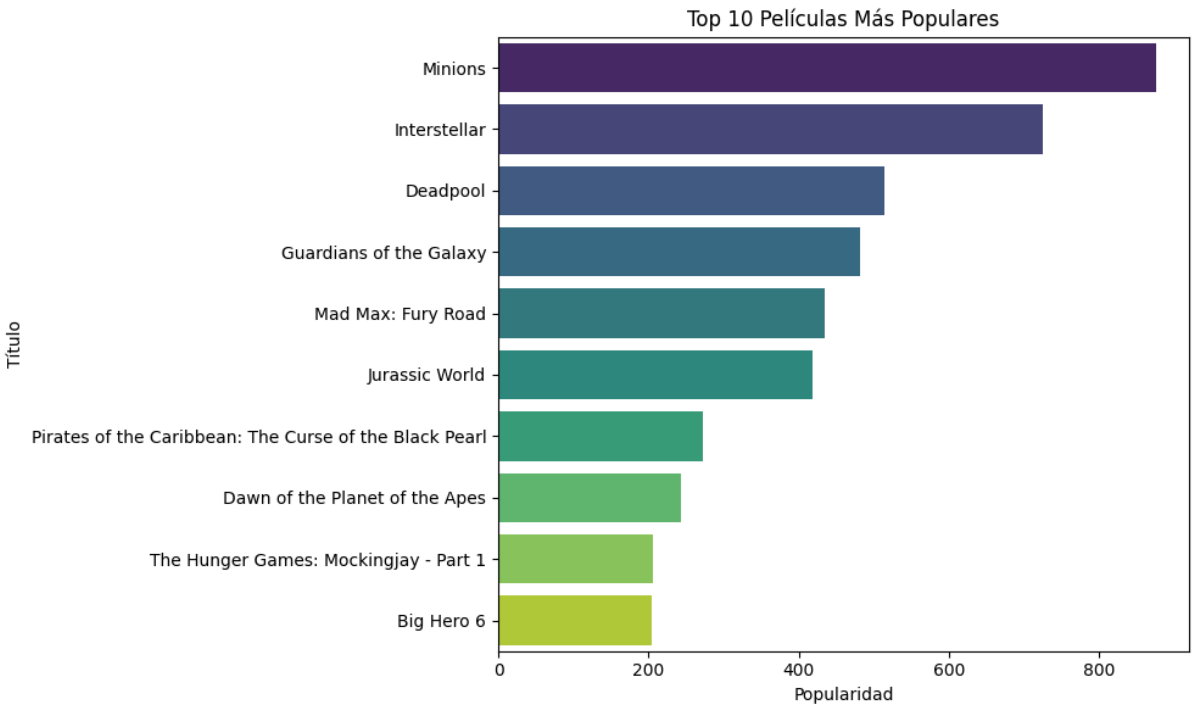
Distribución de popularidad: El histograma de popularity revela una distribución altamente sesgada hacia valores bajos, con unas pocas películas acumulando una gran popularidad. Esto es coherente con el comportamiento del mercado: mientras algunas producciones generan altas expectativas y visibilidad, muchas otras mantienen un perfil más discreto.



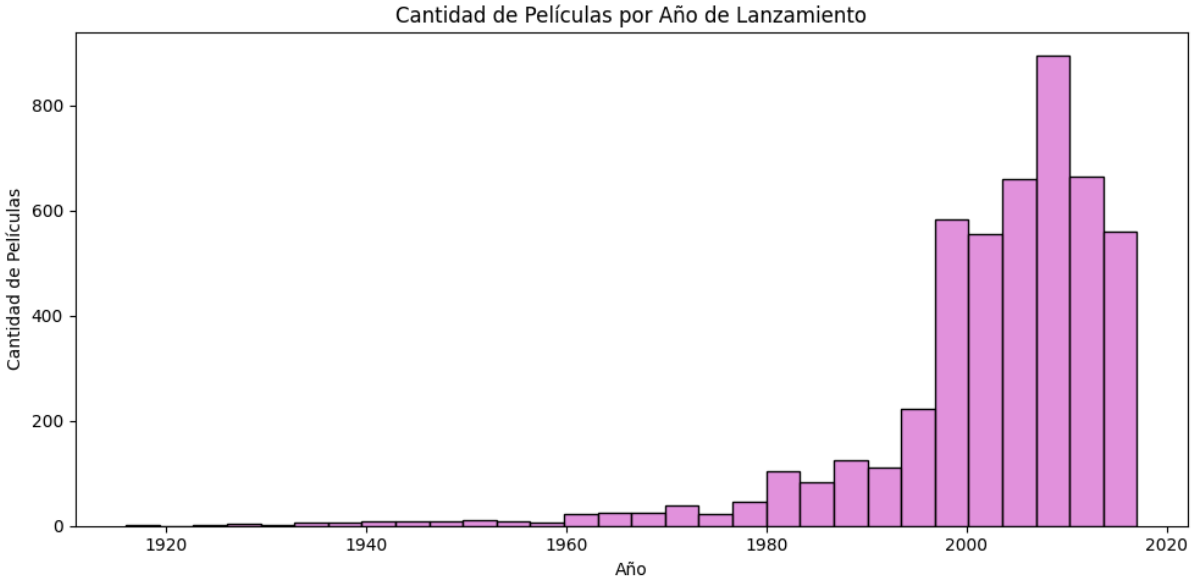
Relación entre presupuesto y calificación: Se utilizó un gráfico de dispersión para analizar la relación entre budget y vote_average. Los resultados muestran una alta dispersión y ausencia de una correlación clara. Esto sugiere que un mayor presupuesto no garantiza una mejor valoración por parte del público, confirmando que otros factores como el guion, la dirección o la actuación influyen significativamente en la percepción del espectador.



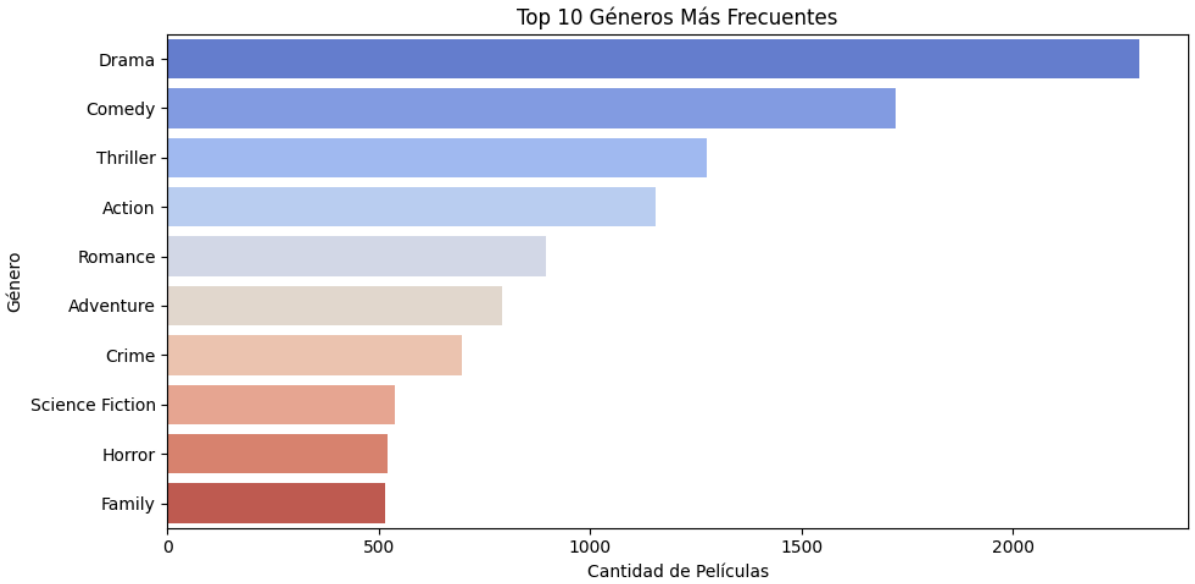
Top 10 películas más populares: Se identificaron las 10 películas con mayor puntuación de popularidad. Estas producciones destacan por su alto presupuesto, reparto reconocido y estrategias de marketing. Títulos como Ballerina o Thunderbolts encabezan la lista. Esta visualización es clave para entrenar modelos de recomendación basados en tendencias actuales.



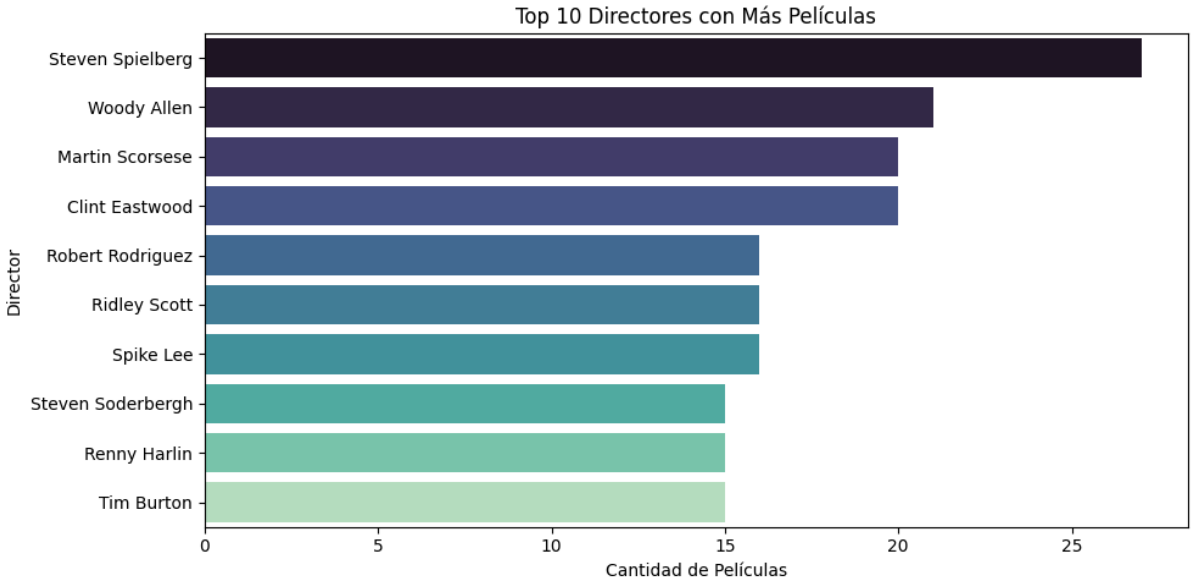
Evolución de estrenos por año: Aunque el dataset contiene principalmente películas programadas para estrenarse en 2025, se incorporaron algunos registros de años anteriores. El histograma de release_year permite observar que el mayor volumen de estrenos corresponde al año actual, lo que refleja el enfoque del dataset en contenidos recientes.



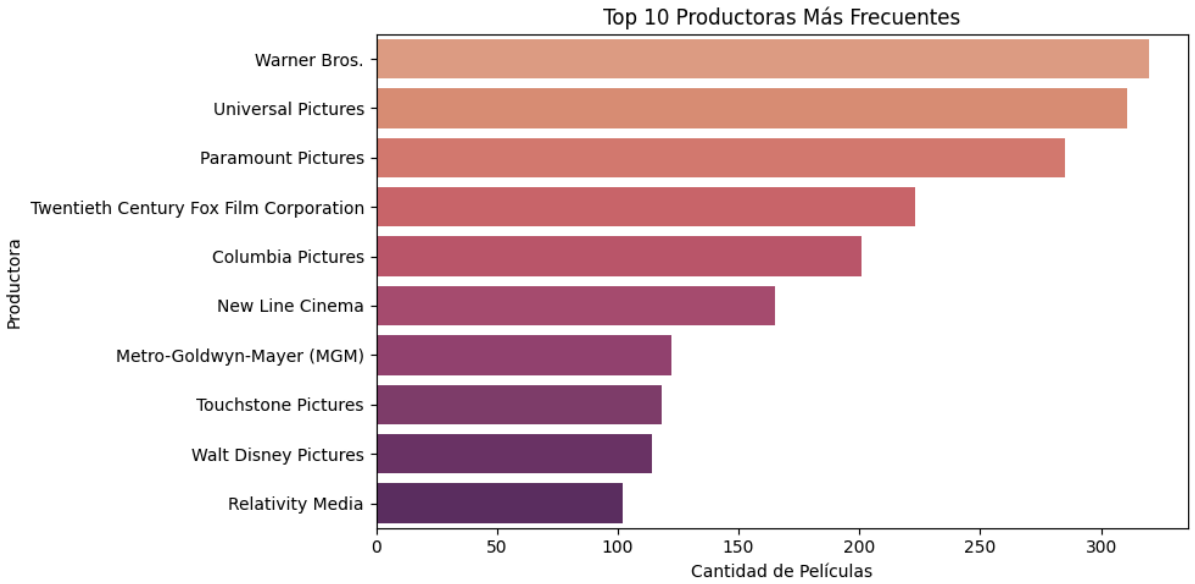
Géneros más frecuentes: A partir del campo genres, se generó un gráfico de barras con los géneros más comunes. En el conjunto de datos predominan los géneros de Action, Thriller, Science Fiction, Fantasy y Adventure. Esto indica una clara inclinación del mercado hacia producciones de alto impacto visual y narrativas épicas, lo cual también influye en las preferencias del público objetivo.



Directores con más producciones: Se identificaron los directores con mayor presencia en el dataset. Figuras como Jake Schreier y Victoria Mahoney aparecen repetidamente en producciones recientes. Esta métrica es relevante si se desea evaluar el impacto de ciertos directores en la recepción crítica o el éxito comercial de las películas.



Productoras más frecuentes: Finalmente, se analizó el campo production_companies, encontrándose que estudios como Marvel Studios, Lionsgate, y Skydance Media lideran la presencia en el dataset. Estas productoras están detrás de las películas más populares y de mayor presupuesto, lo cual puede relacionarse directamente con su capacidad de producción y distribución a gran escala.



Estas visualizaciones no solo complementan el análisis descriptivo, sino que también aportan información clave para la selección de variables en los modelos de recomendación, regresión y clasificación empleados en este proyecto.

5. Propuesta de Modelización

Con base en los objetivos del caso de uso, se propone el uso de modelos de aprendizaje automático y técnicas de procesamiento de lenguaje natural (NLP) para resolver tres tareas clave:

- Recomendar películas similares a las seleccionadas por un usuario.
- Predecir la calificación promedio esperada de una película aún no valorada.
- Clasificar si una película será un "éxito comercial" o no, a partir de sus características antes del estreno.

Estas tareas cubren distintos enfoques de la ciencia de datos: sistemas de recomendación, regresión y clasificación. A continuación, se detalla la propuesta técnica para cada una.

5.1. Modelo basado en contenido para recomendación

Para responder a la pregunta "¿Qué películas se pueden recomendar a un usuario basándose en atributos como género, elenco o palabras clave de películas que le hayan gustado previamente?", se propone la implementación de un sistema de recomendación basado en contenido (Content-Based Filtering).

Este enfoque analiza las características propias de cada película y recomienda otras similares sin necesidad de tener historial de valoraciones de otros usuarios. Es ideal para escenarios en los que no se cuenta con feedback colaborativo, como ratings individuales o interacciones históricas.

Técnica utilizada:

- Enriquecimiento textual de cada película combinando: genres, overview, cast, director, production_companies
- Vectorización mediante TF-IDF (Term Frequency - Inverse Document Frequency)
- Cálculo de similitud mediante cosine_similarity

De este modo, se genera un perfil vectorial para cada película, y se calcula la similitud con otras para generar recomendaciones.

Ejemplo de aplicación:

Si un usuario selecciona la película Inception, el sistema buscará películas con sinopsis, reparto, director o géneros similares y podría sugerir Interstellar o Tenet, que comparten estilo narrativo, director (Christopher Nolan), y elementos de ciencia ficción y suspenso.

Este enfoque permite recomendaciones altamente relevantes y personalizadas sin necesidad de interacción previa del usuario.

5.2. Modelo de regresión para predecir la calificación promedio

Para responder a la pregunta "¿Se puede predecir el nivel de aprobación (voto promedio) que tendrá una nueva película, con base en sus características antes de su estreno?", se propone el uso de modelos de regresión supervisada.

Variables independientes:

- budget, runtime, popularity, vote_count, release_year
- Codificaciones one-hot para variables categóricas como genres, director, production_companies

Modelos propuestos:

- Regresión lineal múltiple
- Random Forest Regressor (modelo principal)

Métricas esperadas:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 (Coeficiente de determinación)

Este modelo permitirá anticipar el rendimiento crítico de una película antes de su lanzamiento, lo que puede ser útil para productoras o servicios de streaming que evalúan decisiones de inversión y promoción.

5.3. Modelo de clasificación de éxito comercial

La tercera pregunta planteada es: "¿Podemos clasificar si una película será un éxito comercial o no, considerando variables como presupuesto, popularidad y características técnicas antes del estreno?"

Para ello, se propone un modelo de clasificación supervisada, donde la variable objetivo éxito es binaria:

- 1 = éxito comercial, por ejemplo, si popularity > 75 y vote_count > 1000.
- 0 = no éxito

Variables consideradas:

- budget, runtime, popularity, vote_count, genres, director

Modelos propuestos:

- Logistic Regression
- Random Forest Classifier (modelo principal)

Métricas a evaluar:

- Accuracy
- Precision / Recall
- F1 Score
- Matriz de confusión

Este modelo permite evaluar anticipadamente el potencial comercial de una película y puede integrarse en dashboards o motores de análisis de inversión de estudios.

Cada modelo ha sido seleccionado en función de la naturaleza de los datos disponibles, la estructura del problema y la facilidad de interpretación de los resultados. Se utilizarán técnicas de validación cruzada y comparación de métricas para seleccionar la mejor opción en cada caso.

6. Publicación de los resultados

En esta sección se presentan los resultados obtenidos tras el entrenamiento y evaluación de los tres modelos desarrollados: un sistema de recomendación basado en contenido, un modelo de regresión para predecir la calificación promedio (vote_average), y un modelo de clasificación para anticipar el éxito comercial de una película.

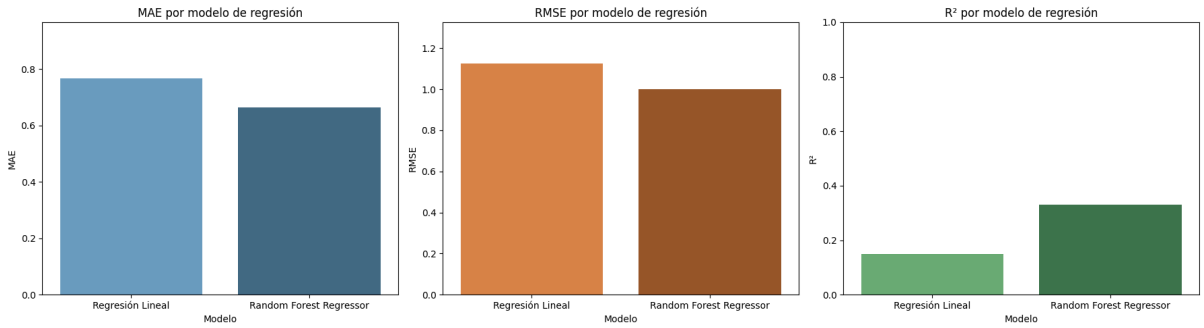
A continuación, se muestran las métricas clave para los modelos de regresión y clasificación. Se incluye una captura de consola con las métricas obtenidas y un gráfico de barras comparativo para facilitar la visualización.

6.1. Resultados del modelo de regresión

El objetivo del modelo de regresión fue predecir la calificación promedio esperada (vote_average) de una película en base a atributos como presupuesto (budget), popularidad (popularity) y duración (runtime). Se entrenaron dos modelos: una Regresión Lineal como punto de partida, y un Random Forest Regressor como modelo más flexible y no lineal.

```
========== Regresión de calificación promedio ==========
                    Modelo   MAE  RMSE    R²
         Regresión Lineal 0.767 1.127 0.149
Random Forest Regressor 0.665 1.000 0.330
```

A continuación, se presentan los resultados obtenidos para cada métrica evaluada, junto con su interpretación:



MAE (Error Absoluto Medio)

- Regresión Lineal: 0.77
- Random Forest: 0.67

El modelo de Random Forest tiene un MAE menor, lo que significa que, en promedio, sus predicciones se desvían menos del valor real. Un MAE más bajo es mejor, ya que representa menor error absoluto por predicción.

RMSE (Raíz del Error Cuadrático Medio)

- Regresión Lineal: 1.13
- Random Forest: 1.00

El modelo de Random Forest también tiene un RMSE menor, lo cual es deseable ya que esta métrica penaliza más los errores grandes. Un valor más bajo indica mejor rendimiento global frente a predicciones con errores significativos.

R^2 (Coeficiente de Determinación)

- Regresión Lineal: 0.15
- Random Forest: 0.33

El R^2 mide cuánta varianza de los datos es explicada por el modelo. En este caso, el Random Forest tiene un valor significativamente más alto, lo que sugiere que captura mejor las relaciones no lineales presentes en los datos. Un R^2 más cercano a 1 indica un mejor ajuste.

En las tres métricas evaluadas (MAE, RMSE y R^2), el modelo de Random Forest Regressor superó claramente a la Regresión Lineal:

- Presentó errores más bajos en promedio (menor MAE y RMSE).
- Explicó más varianza de los datos (mayor R^2).

Sin embargo, ambos modelos obtuvieron un R^2 relativamente bajo (por debajo de 0.4), lo que indica que aún queda una gran proporción de variabilidad sin explicar. Esto puede deberse a factores externos no considerados en el modelo, como reseñas profesionales, tendencias sociales, o campañas de marketing.

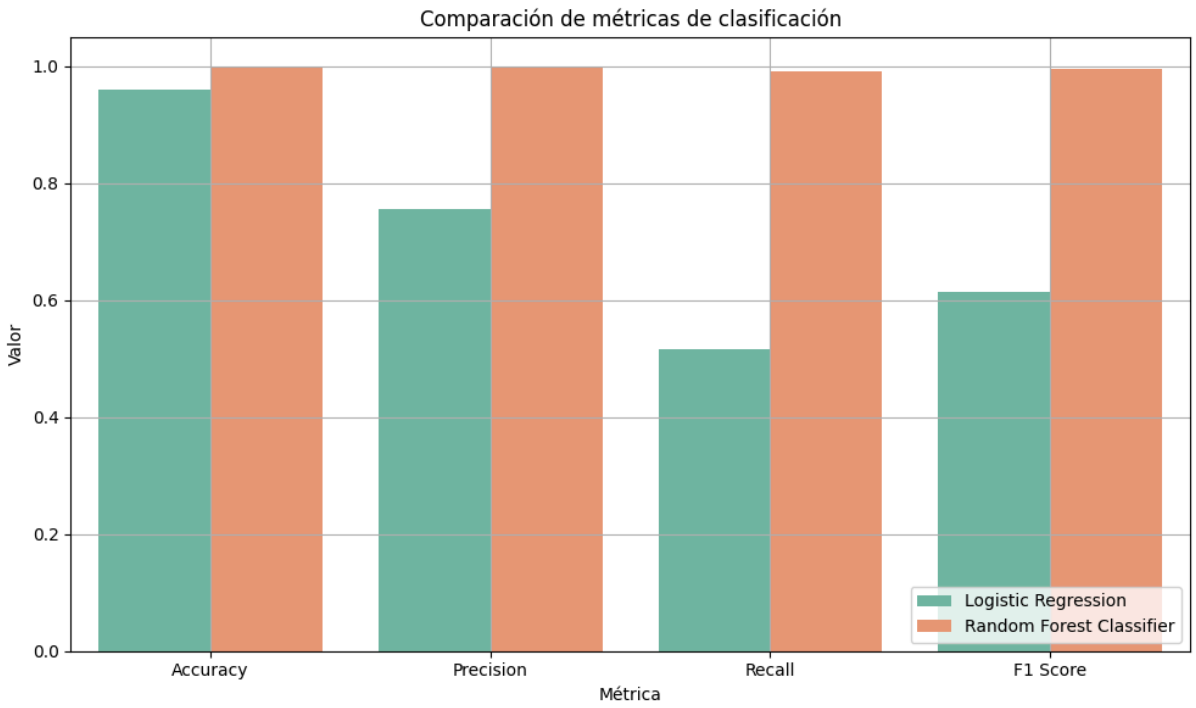
Se concluye que el Random Forest Regressor es el modelo más adecuado entre los evaluados para esta tarea, aunque podrían explorarse otras alternativas (como regresión por boosting, redes neuronales o ingeniería de características más avanzada, para mejorar el rendimiento.

6.2. Resultados del modelo de regresión

Para anticipar si una película será un "éxito comercial", se definió como criterio que esta supere los 75 puntos de popularidad y cuente con más de 1000 votos. A partir de esta definición, se construyó una variable objetivo binaria (éxito) y se entrenaron dos modelos supervisados: Regresión Logística y Random Forest Classifier.

```
========== Clasificación de éxito comercial ==========
                  Modelo  Accuracy  Precision  Recall  F1 Score
     Logistic Regression      0.96       0.76    0.52      0.61
Random Forest Classifier      1.00       1.00    0.99      1.00
```

A continuación, se presentan los resultados obtenidos en el conjunto de prueba, junto con el análisis detallado de cada métrica:



Accuracy (Exactitud)

- Regresión Logística: 0.96
- Random Forest: 1.00

Ambos modelos presentan una exactitud muy alta, lo que indica que clasificaron correctamente la mayoría de los casos. Sin embargo, el Random Forest logra una exactitud casi perfecta, sin errores en las predicciones del conjunto de prueba.

Precision (Precisión)

- Regresión Logística: 0.76
- Random Forest: 1.00

La precisión mide cuántas de las predicciones positivas realmente lo eran. El Random Forest tiene una precisión perfecta, mientras que la Regresión Logística genera más falsos positivos. Esto implica que el Random Forest es más confiable cuando predice un éxito comercial.

Recall (Sensibilidad)

- Regresión Logística: 0.52
- Random Forest: 0.99

Esta es la métrica donde se observa la mayor diferencia. El Random Forest identifica prácticamente todos los éxitos reales, mientras que la Regresión Logística solo detecta aproximadamente la mitad, lo que podría llevar a subestimar oportunidades comerciales.

F1 Score

- Regresión Logística: 0.61
- Random Forest: 0.99

El F1 Score, al combinar precisión y recall, muestra que el Random Forest tiene un equilibrio casi perfecto entre ambas métricas, mientras que la Regresión Logística ofrece un desempeño aceptable, pero claramente inferior.

El modelo Random Forest demostró ser claramente superior a la regresión logística en la tarea de predecir si una película será un "éxito comercial". Obtuvo métricas cercanas a 1.00 en accuracy, precision, recall y F1 Score, superando ampliamente a la regresión logística, que aunque presentó una exactitud alta (0.96), mostró un recall bajo (0.52), indicando que no logró identificar muchos éxitos reales.

Esta diferencia se debe a que el Random Forest, al ser un modelo no lineal, captura mejor relaciones complejas entre variables como presupuesto, duración y popularidad. Además, maneja mejor el desbalance de clases, siendo capaz de detectar casi todos los casos positivos sin comprometer la precisión.

El excelente equilibrio entre precisión y sensibilidad del Random Forest lo convierte en una solución robusta y confiable para este tipo de problema. Como trabajo futuro, se sugiere probar técnicas de balanceo de clases, como SMOTE, y explorar modelos de ensamble más avanzados, como XGBoost, para seguir mejorando la capacidad predictiva.

En conclusión, el Random Forest es el modelo más adecuado para predecir el éxito comercial de una película en este contexto.

6.3. Evaluación del modelo de recomendación

Este modelo no se evalúa con métricas cuantitativas clásicas, sino de forma cualitativa. Las recomendaciones generadas fueron coherentes con las características de las películas base, mostrando similitudes temáticas y estilísticas. Por ejemplo, al seleccionar Inception, el sistema sugirió Interstellar, Tenet y The Prestige, todas películas con elementos narrativos comunes, misma dirección y géneros similares.

7. Conclusiones:

Al terminar el proyecto logramos desarrollar un sistema de recomendación de películas utilizando técnicas de ciencia de datos, aprendizaje automático y procesamiento de lenguaje natural (NLP), aplicadas sobre un conjunto de datos actualizado obtenido desde la API de The Movie Database (TMDb). El enfoque abordó tres tareas clave: recomendación personalizada, predicción de calificación promedio, y clasificación del éxito comercial de una película.

En la primera tarea, se implementó un modelo basado en contenido mediante la vectorización TF-IDF y la similitud coseno, que generó recomendaciones coherentes y relevantes a partir de las características textuales y técnicas de cada película. Aunque su evaluación es cualitativa, los resultados demostraron una alta capacidad del sistema para identificar similitudes significativas entre títulos, incluso sin requerir historial del usuario.

En la segunda tarea, se entrenaron modelos de regresión para anticipar la calificación promedio (vote_average). El modelo de Random Forest superó a la regresión lineal en todas las métricas (MAE, RMSE, R^2), mostrando mejor capacidad para capturar relaciones no lineales entre variables como presupuesto, popularidad y duración. A pesar de ello, el valor de R^2 fue moderado, lo que sugiere oportunidades de mejora mediante técnicas avanzadas o incorporación de nuevas variables.

En la tercera tarea, se aplicaron modelos de clasificación para predecir si una película será un éxito comercial. El modelo de Random Forest obtuvo resultados sobresalientes con accuracy, precision, recall y F1 Score cercanos a 1.00, siendo claramente superior a la regresión logística. Su desempeño equilibrado y robusto lo convierte en la mejor opción para esta tarea. No obstante, se reconoce que el dataset presenta cierto desbalance de clases, por lo que a futuro podrían evaluarse técnicas como SMOTE o modelos de ensamble como XGBoost.

En conjunto, el sistema desarrollado representa una solución integral y funcional que puede ser integrada en plataformas de recomendación o análisis de contenido audiovisual. El uso combinado de modelos supervisados y no supervisados, junto con una interfaz amigable, permite explorar múltiples escenarios de personalización, evaluación y predicción. Como futuras actualizaciones, sugerimos ampliar la base de datos, incorporar feedback del usuario y explorar técnicas de aprendizaje profundo para mejorar aún más la precisión y relevancia de las recomendaciones.

8. Bibliografía:

Aggarwal, C. C. (2016). Recommender systems: The textbook. Springer. https://link.springer.com/book/10.1007/978-3-319-29659-3

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing?. Journal of Personality and Social Psychology, 79(6), 995–1006. https://psycnet.apa.org/doiLanding?doi=10.1037%2F0022-3514.79.6.995

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems handbook (2nd ed.). Springer. https://link.springer.com/book/10.1007/978-1-4899-7637-6

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Robxburgh, C, & Hung Byers, A. (1 de mayo de 2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Digital. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

The Movie Database (TMDb). (s.f.). https://www.themoviedb.org/

9. Anexos:

Link del repositorio en Github: https://github.com/LiamQuinoNeff/CC219-TP-TF-2025-1