



TECNICATURA SUPERIOR EN

# Ciencia de Datos e Inteligencia Artificial

***Módulo: Científico de Datos***

***Año: 2024***

***Trabajo Práctico Integrador***

---

**Integrantes del Grupo 13:**

- Ingaramo, Ma. Eugenia (eugenia.ingaramo@gmail.com)
  - Lonardi, Pablo (lonardipablo@gmail.com)
  - Margheim, Carolina (caro08.m@gmail.com)
  - Sosa, Rodrigo (sosarodrigox@gmail.com)
  - Zenere, Mauricio (zeneremauricio@gmail.com)
-

## ÍNDICE

<b>INTRODUCCIÓN.....</b>	<b>3</b>
<b>COMPRENSIÓN DEL NEGOCIO.....</b>	<b>4</b>
Situación Actual.....	4
Objetivo del Negocio.....	4
Objetivos del Proyecto de Datos.....	5
<b>COMPRENSIÓN DE LOS DATOS.....</b>	<b>6</b>
Recolección de Datos.....	6
Exploración Inicial de los Datos.....	6
Descripción de los Datos.....	12
<b>PREPARACIÓN DE LOS DATOS.....</b>	<b>15</b>
<b>MODELADO.....</b>	<b>17</b>
<b>EVALUACIÓN.....</b>	<b>19</b>
<b>DESPLIEGUE.....</b>	<b>22</b>
<b>CONCLUSIÓN.....</b>	<b>23</b>
<b>LECCIONES APRENDIDAS.....</b>	<b>25</b>

## INTRODUCCIÓN

Para llevar a cabo este proyecto, se ha seguido la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), un modelo de proceso estándar ampliamente utilizado en el ámbito de la minería de datos. CRISP-DM proporciona una estructura organizada y repetible para el desarrollo de proyectos de minería de datos, dividiéndose en seis fases:

- Comprensión del Negocio
- Comprensión de los Datos
- Preparación de los Datos
- Modelado
- Evaluación
- Despliegue

La elección de CRISP-DM se debe a su flexibilidad y aplicabilidad en diferentes industrias, lo que nos permite adaptarnos a las necesidades específicas de este proyecto. Además, esta metodología proporciona un marco de trabajo claro y estructurado, facilitando la colaboración entre los diferentes miembros del equipo y la comunicación de los resultados obtenidos.

En el presente proyecto integrador, se ha desarrollado una aplicación de recomendación que ahora no solo permite la entrada de texto, sino que también admite entradas por voz y análisis de imágenes de portada. Esta ampliación de funcionalidades responde a la necesidad de ofrecer una experiencia de usuario más integral y accesible, alineada con los avances en tecnologías de Procesamiento del Habla, Procesamiento de Imágenes y Sistemas de Inteligencia Artificial.

El proyecto ha sido desarrollado siguiendo la metodología CRISP-DM, lo cual ha permitido una estructuración robusta y sistemática en cada etapa del proceso. Además, hemos implementado técnicas avanzadas de minería de datos y aprendizaje automático para mejorar la precisión y relevancia de las recomendaciones.

## COMPRENSIÓN DEL NEGOCIO

### Situación Actual

Actualmente, una “Biblioteca Popular” lleva a cabo la gestión de libros de manera manual; esto incluye:

- La adquisición, catalogación y clasificación de libros.
- El etiquetado y la digitalización.
- El préstamo y devolución de libros, así como la gestión de multas y renovaciones.
- La asistencia al usuario.
- La programación de eventos y actividades para la promoción de la lectura, la alfabetización y el compromiso comunitario.
- La gestión de colecciones para la rotación de libros obsoletos o poco utilizados y la identificación de aquellos de interés de los usuarios.
- La preservación y conservación de libros para garantizar su durabilidad y accesibilidad a largo plazo.
- La gestión de sistemas de gestión de bibliotecas, bases de datos en línea, recursos electrónicos y otros aspectos tecnológicos relacionados con la operación de la biblioteca.

La recomendación de libros, por su parte, se ve limitada y condicionada por el personal de la biblioteca (sus conocimientos de materiales de lectura e intereses propios). La consulta de disponibilidad y estado de préstamo de libros requiere que los usuarios visiten físicamente la biblioteca o realicen consultas telefónicas, generando demoras y mayor carga administrativa.

Por otro lado, debido a la falta de un registro exhaustivo de la demanda, la adquisición de libros no está completamente alineada con las necesidades e intereses de los usuarios. Otro aspecto a destacar es que la disposición de recursos y servicios en la biblioteca se basa en suposiciones generales y no en datos concretos sobre el uso y las preferencias de los usuarios.

### Objetivo del Negocio

La biblioteca busca mejorar la eficiencia operativa así como la experiencia de usuario, reduciendo la carga de trabajo manual del personal de la biblioteca. En base a esto, se identifican los siguientes objetivos del negocio:

- Optimizar la experiencia del usuario permitiendo la consulta online de existencia de textos y la recomendación de libros afines y disponibles en la biblioteca en base a dichas consultas.

- Optimizar el proceso de préstamos y devolución de textos llevando un mejor registro y seguimiento de los mismos.
- Liberar al personal de las tareas repetitivas como la recomendación de libros.
- Mejorar la accesibilidad a usuarios utilizando diferentes medios de entrada de consulta, incluyendo texto, voz e imágenes.

Lograr estos objetivos tendría un impacto positivo tanto en los usuarios de la biblioteca como en la eficiencia operativa de la institución, lo que contribuiría a su sostenibilidad a largo plazo.

### Objetivos del Proyecto de Datos

Mediante este proyecto se pretende implementar un sistema automatizado de consulta de disponibilidad y recomendación de libros afines mediante técnicas de procesamiento del lenguaje natural, procesamiento de imágenes y procesamiento de voz. De esta manera, se espera mejorar la experiencia de los usuarios de la biblioteca, facilitando el acceso a los materiales y reduciendo el tiempo dedicado a la búsqueda, lo que podría traducirse en una mayor fidelidad y participación de los mismos.

No obstante, en una próxima etapa del proyecto se prevé ampliar el alcance del proyecto para dar respuesta a otros objetivos del negocio:

- Utilizar análisis de datos para predecir la demanda de libros y ajustar las adquisiciones de manera más precisa, optimizando así la gestión de recursos y evitando la escasez o el exceso de inventario.
- Analizar patrones de uso de los libros y servicios de la biblioteca para identificar oportunidades de mejora en la disposición de recursos y servicios, aumentando así su utilización y maximizando el valor para los usuarios.

En conjunto, estas medidas permitirán reducir la carga de trabajo manual del personal de la biblioteca en tareas como la consulta de disponibilidad y el seguimiento de préstamos, permitiendo que se centren en actividades de valor añadido, como la asistencia al usuario y la programación de eventos.

Asimismo, con la implementación de estas medidas, se espera:

- Reducir los costos asociados con el exceso de inventario o la necesidad de adquirir libros de manera urgente para satisfacer la demanda inesperada.
- Aumentar la utilización de los recursos y servicios de la biblioteca, lo que maximizará su impacto y valor para la comunidad.
- Liberar tiempo y recursos que podrían ser reasignados a actividades de mayor valor añadido, como la asistencia al usuario y la planificación de programas y eventos.

## COMPRENSIÓN DE LOS DATOS

### Recolección de Datos

Para obtener los datos necesarios para el proyecto, se recabaron datos de diversas fuentes. En esta primera etapa del proyecto, se conformó una base de datos con un catálogo de textos. Dado que la biblioteca no cuenta con la totalidad de sus textos digitalizados, se trabajó con datos recolectados de la web para completar el archivo CSV proporcionado por la biblioteca.

Inicialmente, y considerando el tiempo limitado disponible para el desarrollo de la solución, se utilizó un conjunto de datos de Kaggle ([Books Dataset \(kaggle.com\)](https://kaggle.com/datasets/BooksDataset)). Paralelamente, se trabajó en la recolección de datos de la web para ampliar el banco de datos que alimentaría el modelo. Para ello, se realizó un web scraping utilizando la herramienta en línea Octoparse (<http://octoparse.es/>), la cual permitió automatizar la recopilación de datos de libros publicados en la página [Cúspide – Libros \(cuspide.com\)](http://cuspide.com).

### Próximas Etapas del Proyecto

En una próxima etapa del proyecto, se prevé completar esta base de datos con la siguiente información:

- **Datos de Préstamos:** Información sobre los préstamos de libros, incluyendo el número de veces que un libro ha sido prestado, fechas de préstamo y devolución, y el historial de préstamos por usuario.
- **Datos de Disponibilidad:** Información sobre la disponibilidad de los libros en la biblioteca, como el número de copias disponibles, ubicación en la biblioteca y estado de préstamo.
- **Datos de Usuarios:** Información sobre los usuarios de la biblioteca, incluyendo registros de membresía, historial de préstamos y preferencias de lectura.

Estas actividades adicionales son esenciales para enriquecer la base de datos y mejorar la precisión y relevancia del sistema de recomendación de libros.

### Exploración Inicial de los Datos

Para comprender la naturaleza y la calidad de los datos disponibles, llevamos a cabo una serie de tareas que nos ayudan a orientar las decisiones sobre el preprocesamiento, la selección de variables y las técnicas de análisis de datos adecuadas para abordar los objetivos del proyecto.

- **Examinación de la Estructura de los Conjuntos de Datos:** Se examinó la estructura de los conjuntos de datos disponibles para comprender cómo están organizados, qué variables están incluidas y cómo se relacionan entre sí.

- **Cálculo de Estadísticas Descriptivas:** Se calcularon estadísticas descriptivas básicas para cada variable en los conjuntos de datos con el fin de comprender la distribución y la variabilidad de los datos.
- **Creación de Visualizaciones Gráficas:** Se crearon visualizaciones gráficas para explorar las relaciones entre las variables e identificar posibles patrones, tendencias o correlaciones que puedan ser relevantes para los objetivos del proyecto.
- **Identificación de Problemas en los Datos:** Se identificaron valores faltantes, atípicos, anomalías, inconsistencias y errores de entrada que pudieran afectar la validez y la confiabilidad de los resultados del análisis.

A continuación el detalle del análisis:

## Estructura de los datos

```
# Obtener información de las variables dataset recibido:
descripcion = data.describe(include="all")

print("\nDescripción de las variables:")
print(descripcion)
```

Descripción de las variables:

	isbn13	isbn10	title	subtitle	\
count	6.810000e+03	6810	6810	2381	
unique	NaN	6810	6398	2009	
top	NaN	0002005883	The Lord of the Rings	A Novel	
freq	NaN	1	11	226	
mean	9.780677e+12	NaN	NaN	NaN	
...	...	...	...	...	
min	9.780002e+12	NaN	NaN	NaN	
25%	9.780330e+12	NaN	NaN	NaN	
50%	9.780553e+12	NaN	NaN	NaN	
75%	9.780810e+12	NaN	NaN	NaN	
max	9.789042e+12	NaN	NaN	NaN	

	authors	categories	\
count	6738	6711	
unique	3780	567	
top	Agatha Christie	Fiction	
freq	37	2588	
mean	NaN	NaN	
...	...	...	
min	NaN	NaN	
25%	NaN	NaN	
50%	NaN	NaN	
75%	NaN	NaN	
max	NaN	NaN	

```

count 6481
unique 6481
top http://books.google.com/books/content?id=KQZCPgAACAAJ&printsec=frontcover&img=1&zoom=1&source=gbs_api
freq 1
mean NaN
...
min NaN
25% NaN
50% NaN
75% NaN
max NaN

description \
count 6548
unique 6474
top This is a reproduction of the original artefact. Generally these books are created from careful scans of the original. This allows us to preserve
the book accurately and present it in the way the author intended. Since the original versions are generally quite old, there may occasionally be certain
imperfections within these reproductions. We're happy to make these classics available again for future generations to enjoy!
freq 6
mean NaN
...
min NaN
25% NaN
50% NaN
75% NaN
max NaN

published_year average_rating num_pages ratings_count
count 6804.000000 6767.000000 6767.000000 6.767000e+03
unique NaN NaN NaN NaN
top NaN NaN NaN NaN
freq NaN NaN NaN NaN
mean 1998.630364 3.933284 348.181026 2.106910e+04
...
min 1853.000000 0.000000 0.000000 0.000000e+00
25% 1996.000000 3.770000 208.000000 1.590000e+02
50% 2002.000000 3.960000 304.000000 1.018000e+03
75% 2005.000000 4.130000 420.000000 5.992500e+03
max 2019.000000 5.000000 3342.000000 5.629932e+06

[11 rows x 12 columns]

```



```
print(data.head())
```

	isbn13	isbn10	title	subtitle	
0	9780002005883	0002005883	Gilead	NaN	
1	9780002261982	0002261987	Spider's Web	A Novel	
2	9780006163831	0006163831	The One Tree	NaN	
3	9780006178736	0006178731	Rage of angels	NaN	
4	9780006280897	0006280897	The Four Loves	NaN	

	authors	categories
0	Marilynne Robinson	Fiction
1	Charles Osborne;Agatha Christie	Detective and mystery stories
2	Stephen R. Donaldson	American fiction
3	Sidney Sheldon	Fiction
4	Clive Staples Lewis	Christian life

	thumbnail
0	<a href="http://books.google.com/books/content?id=KQZCPgAACAAJ&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api">http://books.google.com/books/content?id=KQZCPgAACAAJ&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api</a>
1	<a href="http://books.google.com/books/content?id=gA5GPgAACAAJ&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api">http://books.google.com/books/content?id=gA5GPgAACAAJ&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api</a>
2	<a href="http://books.google.com/books/content?id=OmQawEACAAJ&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api">http://books.google.com/books/content?id=OmQawEACAAJ&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api</a>
3	<a href="http://books.google.com/books/content?id=FKo2TgANz74C&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api">http://books.google.com/books/content?id=FKo2TgANz74C&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api</a>
4	<a href="http://books.google.com/books/content?id=XhQ5XsFcpGIC&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api">http://books.google.com/books/content?id=XhQ5XsFcpGIC&amp;printsec=frontcover&amp;img=1&amp;zoom=1&amp;source=gbs_api</a>

```
description \
```

0 A NOVEL THAT READERS and critics have been eagerly anticipating for over a decade, Gilead is an astonishingly imagined story of remarkable lives. John Ames is a preacher, the son of a preacher and the grandson (both maternal and paternal) of preachers. It's 1956 in Gilead, Iowa, towards the end of the Reverend Ames's life, and he is absorbed in recording his family's story, a legacy for the young son he will never see grow up. Haunted by his grandfather's presence, John tells of the rift between his grandfather and his father: the elder, an angry visionary who fought for the abolitionist cause, and his son, an ardent pacifist. He is troubled, too, by his prodigal namesake, Jack (John Ames) Boughton, his best friend's lost son who returns to Gilead searching for forgiveness and redemption. Told in John Ames's joyous, rambling voice that finds beauty, humour and truth in the smallest of life's details, Gilead is a song of celebration and acceptance of the best and the worst the world has to offer. At its heart is a tale of the sacred bonds between fathers and sons, pitch-perfect in style and story, set to dazzle critics and readers alike.

1 A new 'Christie for Christmas' -- a full-length novel adapted from her acclaimed play by Charles Osborne Following BLACK COFFEE and THE UNEXPECTED GUEST comes the final Agatha Christie play novelisation, bringing her superb storytelling to a new legion of fans. Clarissa, the wife of a Foreign Office diplomat, is given to daydreaming. 'Supposing I were to come down one morning and find a dead body in the library, what should I do?' she muses. Clarissa has as her chance to find out when she discovers a body in the drawing-room of her house in Kent. Desperate to dispose of the body before her husband comes home with an important foreign politician, Clarissa persuades her three house guests to become accessories and accomplices. It seems that the murdered man was not unknown to certain members of the house party (but which ones?), and the search begins for the murderer and the motive, while at the same time trying to persuade a police inspector that there has been no murder at all... SPIDER'S WEB was written in 1954 specifically for Margaret Lockwood and opened first at the Theatre Royal Nottingham before moving to the Savoy Theatre in London on 14 December 1954. With THE MOUSETRAP and WI

2 Volume Two of Stephen Donaldson's acclaimed second trilogy featuring the compelling anti-hero Thomas Covenant.

3 A memorable, mesmerizing heroine Jennifer -- brilliant, beautiful, an attorney on the way up until the Mafia's schemes win her the hatred of an implacable enemy -- and a love more destructive than hate. A dangerous, dramatic world The Dark Arena of organized crime and flashbulb lit courtrooms where ambitious prosecutors begin their climb to political power.

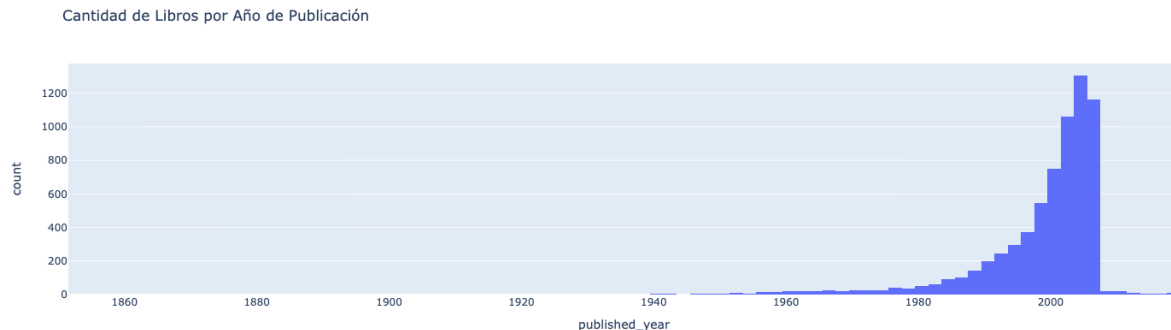
4 Lewis' work on the nature of love divides love into four categories; Affection, Friendship, Eros and Charity. The first three come naturally to humanity. Charity, however, the Gift-love of God, is divine, and without this supernatural love, the natural loves become distorted and even dangerous.

	published_year	average_rating	num_pages	ratings_count
0	2004.0	3.85	247.0	361.0
1	2000.0	3.83	241.0	5164.0
2	1982.0	3.97	479.0	172.0
3	1993.0	3.93	512.0	29532.0
4	2002.0	4.15	170.0	33684.0

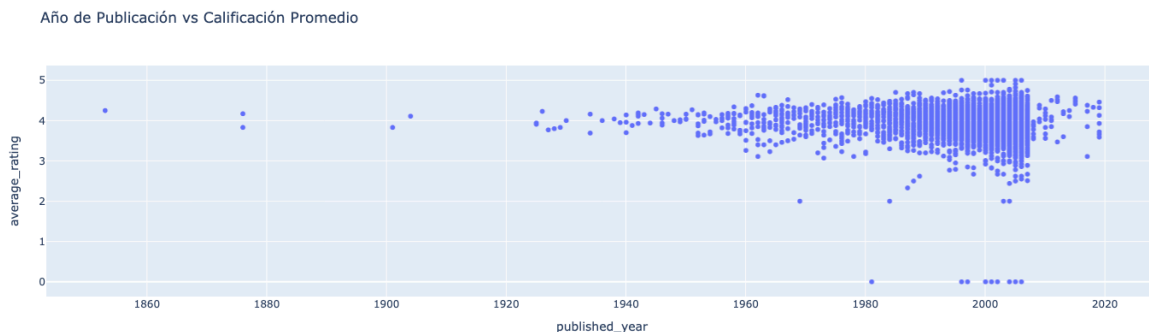
## Distribución de datos por año de publicación

Este gráfico muestra la cantidad de libros publicados según el año. Se observa un aumento significativo en la cantidad de libros publicados a partir de la década de 1990, con un pico notable alrededor de 2004-2005. Esto sugiere un incremento en la producción editorial en los últimos años, lo cual puede estar relacionado con la creciente accesibilidad a herramientas de autopublicación y el auge del mercado digital de libros. Por otro lado, puede significar un aumento en la diversidad de títulos y géneros disponibles para los usuarios de la biblioteca.



### Calificación Promedio según año de publicación

Este gráfico de dispersión muestra la relación entre el año de publicación y la calificación promedio de los libros. No se observa una tendencia clara que indique que los libros publicados en un año específico reciben mejores calificaciones que otros. Sin embargo, se puede notar que la mayoría de las calificaciones se agrupan alrededor de una calificación promedio de 4.0, independientemente del año de publicación la cual indica que los usuarios de la biblioteca están satisfechos en general con los libros ofrecidos.



### Cantidad de Calificaciones según Año de Publicación

El siguiente gráfico de dispersión muestra la relación entre el año de publicación y el conteo de calificaciones. Se observa una gran variabilidad en la cantidad de calificaciones, con algunos libros publicados recientemente (después del año 2000) obteniendo un alto número de calificaciones. Esto podría reflejar la popularidad de ciertos libros modernos o el impacto de las plataformas de calificación en línea. Sin embargo, la mayoría de los libros tienen un conteo de calificaciones relativamente bajo, lo cual es típico en grandes conjuntos de datos de libros.



## Descripción de los Datos

Se llevó a cabo un análisis exhaustivo del conjunto de datos para comprender su utilidad en el desarrollo del sistema de recomendación de textos, así como para identificar necesidades de preprocesamiento o la posible necesidad de datos adicionales.

El conjunto de datos analizado consta de 6,810 registros, cada uno representando un libro único. Estos registros están descritos mediante 12 atributos, que incluyen tanto identificadores numéricos como descripciones.

A continuación se detalla la estructura y las características principales del conjunto de datos:

```
# Obtener informacion del dataset recibido:  
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 6810 entries, 0 to 6809  
Data columns (total 12 columns):  
#   Column          Non-Null Count  Dtype    
---  ---             
0   isbn13          6810 non-null   int64    
1   isbn10          6810 non-null   object   
2   title           6810 non-null   object   
3   subtitle        2381 non-null   object   
4   authors         6738 non-null   object   
5   categories      6711 non-null   object   
6   thumbnail       6481 non-null   object   
7   description     6548 non-null   object   
8   published_year  6804 non-null   float64   
9   average_rating  6767 non-null   float64   
10  num_pages       6767 non-null   float64   
11  ratings_count   6767 non-null   float64   
dtypes: float64(4), int64(1), object(7)  
memory usage: 638.6+ KB
```

- **isbn13**: Número ISBN de 13 dígitos, identificador único para libros.
- **isbn10**: Número ISBN de 10 dígitos, otra forma de identificador único.
- **title**: Título del libro.
- **subtitle**: Subtítulo del libro.
- **authors**: Autor(es) del libro.
- **categories**: Categorías o géneros del libro.
- **thumbnail**: URL de la imagen de portada del libro.
- **description**: Descripción del libro.
- **published\_year**: Año de publicación del libro.
- **average\_rating**: Calificación promedio recibida por el libro.
- **num\_pages**: Número de páginas del libro.
- **ratings\_count**: Número de calificaciones recibidas por el libro.

## Variables Categóricas

### Identificadores de Libro

- **isbn13**: Cada libro tiene un identificador ISBN de 13 dígitos, proporcionado para todos los libros (6810 entradas no nulas).
- **isbn10**: Un identificador ISBN alternativo de 10 dígitos para cada libro, también completamente poblado.

### Títulos y Autores

- **title**: El título del libro, sin entradas nulas en el dataset.
- **subtitle**: Subtítulos proporcionados para 2,381 libros, sugiriendo que no todos los libros tienen un subtítulo explícito.

- **authors:** Nombres de los autores, presentes para 6,738 libros, indicando una pequeña proporción de entradas sin esta información. Esto podría representar un problema a futuro.

#### Categorización y Descripción

- **categories:** Categorías o géneros bajo los cuales se clasifican los libros, con 6,711 libros categorizados. Hay casi 100 libros no categorizados lo cual podría generar conflictos a la hora de realizar recomendaciones.
- **description:** Descripciones textuales detalladas disponibles para 6,548 libros, proporcionando un contexto valioso sobre el contenido y el tema de cada libro y que podría utilizarse para las recomendaciones a clientes de la biblioteca.

#### Aspectos Visuales y Editoriales

- **thumbnail:** URLs de imágenes de portada para 6,481 libros, lo que ayuda en la visualización y el marketing digital del libro. Se testeó valores al azar y se verificó que algunos links estaban caídos. Evaluar la solución para estos casos.
- **published\_year:** Año de publicación registrado para 6,804 libros, ofreciendo una perspectiva sobre la temporalidad de las publicaciones.

#### Calificaciones y Popularidad

- **average\_rating:** La calificación promedio asignada por los usuarios, disponible para 6,767 libros, lo que refleja la recepción del libro por parte de los lectores.
- **num\_pages:** Número de páginas, también disponible para 6,767 libros, indicando la extensión de los textos.
- **ratings\_count:** El número de calificaciones que ha recibido cada libro, también notado para 6,767 libros, que puede ser utilizado para medir la popularidad o el engagement de los lectores.

#### **Variables Numéricas**

- **published\_year:** Año de publicación registrado para 6,804 libros, ofreciendo una perspectiva sobre la temporalidad de las publicaciones. Los años de publicación varían entre 95 años distintos, con el año más común siendo 2006 (894 libros publicados en ese año).

#### Calificaciones y Popularidad

- **average\_rating:** La calificación promedio asignada por los usuarios, disponible para 6,767 libros, lo que refleja la recepción del libro por parte de los lectores. Los valores de calificación varían entre 1 y 5, en promedio se distribuyen en 201 valores únicos, siendo 4.0 la calificación más común con 126 ocurrencias.
- **num\_pages:** Número de páginas, también disponible para 6,767 libros, indicando la extensión de los textos. El número de páginas varía en 916 valores únicos, con el valor más frecuente siendo 288 páginas (141 libros).
- **ratings\_count:** El número de calificaciones que ha recibido cada libro, también notado para 6,767 libros, que puede ser utilizado para medir la popularidad o el engagement de

los lectores. Hay 3882 valores únicos en el conteo de calificaciones, con el valor más común apareciendo 43 veces.

El conjunto de datos analizado está mayormente completo, aunque presenta algunos valores faltantes en los campos "subtitle" y "thumbnail", lo cual es esperable en contextos de recopilación de datos de textos. Además, se han identificado valores atípicos en el campo "num\_pages" que pueden requerir un análisis adicional para determinar su validez.

En términos de calidad, el conjunto de datos proporciona una base sólida para varios componentes esenciales del proyecto. Es adecuado para el desarrollo de un sistema automatizado de recomendación de libros, tanto basado en contenido como en colaboración, alineándose con el objetivo de optimizar la experiencia del usuario y aumentar su satisfacción.

A pesar de algunas limitaciones, el conjunto de datos actual es suficiente y presenta datos de calidad para una primera iteración en la construcción de un sistema de recomendación.

### **Próximas Etapas del Proyecto**

Para respaldar completamente la iniciativa de mejorar la eficiencia en la gestión de recursos y optimizar su utilización, se recomienda enriquecer el conjunto de datos con información adicional sobre las interacciones de los usuarios con la biblioteca, tales como:

- Registros de Préstamos: Información sobre el historial de préstamos de los usuarios, incluyendo fechas y frecuencia de préstamos.
- Frecuencia de Uso de los Textos: Datos que indiquen cuántas veces y por cuánto tiempo se han utilizado los textos.
- Retroalimentación Directa de los Usuarios: Comentarios y calificaciones que los usuarios proporcionen sobre los libros.

Esta información adicional permitirá realizar análisis predictivos más precisos sobre tendencias de publicación y demanda de libros, facilitando una mejor alineación con las necesidades e intereses de los stakeholders. Además, la integración y automatización de los registros de disponibilidad y estado de préstamo en el conjunto de datos podrían reducir significativamente la carga administrativa del personal, apoyando la eficiencia operativa. Esto proporcionará una base de datos más completa sobre la cual tomar decisiones informadas para la gestión de colecciones y la programación de actividades de la biblioteca.

## PREPARACIÓN DE LOS DATOS

Con el objetivo de garantizar que los datos estén en condiciones óptimas para su análisis y uso en el desarrollo del presente proyecto, se llevaron a cabo las siguientes actividades:

- **Selección de Datos Relevantes:** Se seleccionaron datos relevantes para el objetivo del proyecto, incluyendo título, autor, categoría, descripción y año de publicación.

```
# Obtenemos las variables de interés para el análisis:

variables_seleccionadas = [
    "title",
    "authors",
    "categories",
    "description",
    "published_year",
]

print(data[variables_seleccionadas].info())
```

---

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6810 entries, 0 to 6809
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype  
---  ---            -
0   title           6810 non-null  object 
1   authors         6738 non-null  object 
2   categories      6711 non-null  object 
3   description     6548 non-null  object 
4   published_year  6804 non-null  float64
dtypes: float64(1), object(4)
memory usage: 266.1+ KB
None
```

- **Limpieza y Transformación de los Datos:**
  - Detección y Corrección de Valores Atípicos: Se identificaron y corrigieron valores atípicos para asegurar la precisión de los análisis.
  - Manejo de Valores Faltantes: Se gestionaron valores faltantes para garantizar que los datos estén libres de errores y anomalías que puedan afectar la calidad de los análisis.



```
# Revisar si hay valores nulos en las variables seleccionadas y reemplazarlos por un valor por defecto
for feature in variables_seleccionadas:
    data[feature] = data[feature].fillna("")

# Revisar si hay valores nulos en las variables seleccionadas
print(data[variables_seleccionadas].info())
```


---

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6810 entries, 0 to 6809
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   title            6810 non-null   object
1   authors          6810 non-null   object
2   categories        6810 non-null   object
3   description       6810 non-null   object
4   published_year    6810 non-null   object
dtypes: object(5)
memory usage: 266.1+ KB
None
```

Por su parte, en el módulo de procesamiento de imágenes se generó una función que realiza la limpieza de caracteres especiales y espacios adicionales del texto.

#### Función `limpiar\_texto`

python

 Copiar código

```
def limpiar_texto(texto):
    return re.sub(r'^a-zA-Z\s', '', texto).strip()
```

- **Resolución de Inconsistencias:**

- Inconsistencias entre Portadas y Metadatos: Se trabajó en la inconsistencia entre la información de las portadas y los metadatos de los libros. Se eliminaron registros inconsistentes y se completaron campos faltantes utilizando la información disponible en las portadas.
- Nomenclatura y Estandarización de Categorías: Se identificaron inconsistencias en la nomenclatura de las categorías y la falta de estandarización en esta columna. Estas inconsistencias se abordaron mediante técnicas de limpieza y transformación de datos, corrigiendo errores tipográficos, agrupando categorías similares y estableciendo una taxonomía coherente.



Función `encontrar\_coincidencias`

```
python Copiar código
```

```
def encontrar_coincidencias(lista1, lista2, umbral=0.25):  
    coincidencias = []  
    for item in lista1:  
        coincidencias_cercanas = get_close_matches(item, lista2, n=1, cutoff=umbral)  
        if coincidencias_cercanas:  
            coincidencias.append((item, coincidencias_cercanas[0]))  
    return coincidencias
```

- **Objetivo:** Encontrar coincidencias aproximadas entre dos listas de texto usando un umbral para la similitud.

- **Creación de una Columna Integrada:** Se combinaron las columnas de interés en una nueva columna llamada "libros", la cual será utilizada por el sistema de recomendación de textos para encontrar similitudes entre los libros de la biblioteca.

```
# Combinar las columnas de interés en una sola columna  
libros = (  
    data["title"]  
    + " "  
    + data["categories"]  
    + " "  
    + data["authors"]  
    + " "  
    + f"{data['published_year']}"  
)  
pd.set_option("display.max_colwidth", None)
```

## MODELADO

Una vez comprendidos los datos disponibles, procedimos a su preparación para el modelado.

En este proyecto, hemos recurrido a la técnica de TF-IDF (Term Frequency-Inverse Document Frequency) para crear un sistema de recomendación basado en contenido.

### Técnica TF-IDF

TF-IDF es una técnica fundamental en el procesamiento del lenguaje natural que convierte texto en vectores numéricos. Esta técnica se emplea para evaluar la importancia de un término dentro de un documento específico en relación con un corpus de documentos. A continuación, se detallan los componentes de TF-IDF:

- Frecuencia de Término (TF): Mide cuántas veces aparece un término en un documento determinado. Se calcula como el número de veces que un término aparece en un documento dividido por el total de términos en ese documento.
- Frecuencia Inversa de Documento (IDF): Evalúa la importancia de un término en el conjunto de documentos. Los términos comunes que aparecen en muchos documentos reciben un peso menor. Se calcula como el logaritmo del número total de documentos dividido por el número de documentos que contienen el término.

El producto de TF e IDF otorga más peso a los términos significativos en un documento específico, mientras reduce el peso de los términos comunes.

```
from sklearn.feature_extraction.text import TfidfVectorizer

# Crear un vectorizador TF-IDF
vector_tfidf = TfidfVectorizer()

vector_caracteristicas = vector_tfidf.fit_transform(libros)

print(vector_caracteristicas.shape)
```

### Similitud del Coseno (Cosine Similarity)

Para medir la similitud entre dos vectores obtenidos a partir de TF-IDF, se ha empleado la técnica de Similitud del Coseno. Esta métrica se define como el coseno del ángulo entre dos vectores en un espacio multidimensional y se utiliza para evaluar la similitud entre textos.

Matemáticamente, la Similitud del Coseno se calcula tomando el coseno del ángulo entre dos vectores no nulos, y su valor oscila entre -1 y 1, donde:

- Un valor de 1 indica que los vectores son idénticos.
- Un valor de 0 indica que los vectores son ortogonales (no tienen similitudes).
- Un valor de -1 indica que los vectores son diametralmente opuestos.

```
from sklearn.metrics.pairwise import cosine_similarity

# Obtener la similitud coseno entre los vectores de características:
similitud = cosine_similarity(vector_caracteristicas, vector_caracteristicas)

print(similitud)
```

Además de tomar datos de entrada en formato texto para la recomendación de libros trabajamos en el escaneo de portadas como datos de entrada para el sistema. Para ello se trabajó con un lector OCR (Optical Character Recognition, o Reconocimiento Óptico de Caracteres) a fin de convertir el texto de las portadas en datos editables y buscables. A partir de ello buscamos integrar esta funcionalidad al sistema de recomendación de libros.

#### Cargar la imagen con Streamlit

```
python
imagen = st.file_uploader(label='Sube tu foto aquí:')

if imagen:
    with open("uploaded_image.jpg", "wb") as f:
        f.write(imagen.getbuffer())
    st.success('Imagen subida correctamente!')

    portada = cv2.imread("uploaded_image.jpg")
    st.image(portada, width=200, channels='BGR')
```

1. **Subir imagen:** Permite al usuario subir una imagen.
2. **Guardar imagen:** Guarda la imagen subida en el servidor.
3. **Mostrar imagen:** Muestra la imagen subida en la aplicación.


#### Realizar OCR en la imagen

```
python
ocr = easyocr.Reader(['es'])
lectura = ocr.readtext(portada)
lectura_texto = " ".join([item[1] for item in lectura]).upper()
libro = [lectura_texto]
```

1. **Inicializar EasyOCR:** Configura EasyOCR para leer texto en español.
2. **Leer texto:** Extrae texto de la imagen.
3. **Convertir a mayúsculas:** Convierte el texto leído a mayúsculas y lo almacena en una lista.

#### Procesar el texto

python

 Copiar código


```
libro_str = libro[0][1:-1]
pattern = re.compile(r'\[[^\]]*\]|\d+\.\d+|\d+|[A-Z][A-Z ]*[A-Z]')
matches = pattern.findall(libro_str)
result = []
for match in matches:
    if re.match(r'\d+\.\d+', match):
        result.append(float(match))
    elif re.match(r'\d+', match):
        result.append(int(match))
    else:
        result.append(match.strip())

libro_limpio = [limpiar_texto(str(item)) for item in result if limpiar_texto(str(item))]
libro_usuario = [item.upper() for item in libro_limpio]
```

1. **Eliminar caracteres no deseados:** Usa una expresión regular para encontrar y extraer patrones de texto específicos.
2. **Convertir y limpiar:** Convierte números a sus respectivos tipos (flotantes o enteros) y limpia el texto restante.
3. **Formatear el texto:** Limpia y convierte a mayúsculas.

#### Cargar el dataset y buscar coincidencias

python


 Copiar código

```
df = pd.read_csv('grupo13_pp2\data\Datos_Integrados.csv', header=0, encoding='latin-1')
titles = [title.upper().strip() for title in df['titulo'].tolist()]
```

1. **Cargar datos:** Lee un archivo CSV que contiene datos, incluyendo títulos de libros.
2. **Preparar títulos:** Convierte los títulos a mayúsculas y elimina espacios adicionales.

### Encontrar coincidencias

python

 Copiar código

```
coincidencias = encontrar_coincidencias(libro_usuario, titles, umbral=0.75)

if coincidencias:
    coincidencia_index = max(coincidencias, key=lambda tupla: len(max(tupla, key=len)))
    coincidencia_mas_relevante = coincidencia_index[1]

    st.write("Se encontraron las siguientes coincidencias:")
    df_resultado = df[df['titulo'].str.upper().str.strip() == coincidencia_mas_relevante]
    df_resultado = df_resultado.drop(columns=['portada'])
    df_resultado = df_resultado.reset_index(drop=True)
    df_resultado.columns = [col.upper() for col in df_resultado.columns]

    st.write("<style>div[data-testid='column-description'] {width: 100%;}</style>", unsafe
    st.dataframe(df_resultado, height=600)
else:
    st.warning("No se encontraron coincidencias.")
```

1. **Buscar coincidencias:** Encuentra coincidencias aproximadas entre el texto extraído y los títulos del dataset.
2. **Mostrar resultados:** Si se encuentran coincidencias, muestra los resultados en una tabla. Si no, muestra una advertencia.

## EVALUACIÓN

Una vez concluido el modelado, procedimos a evaluar el sistema de recomendación utilizando un conjunto de datos de prueba. Este paso es crucial para verificar la precisión y relevancia de las recomendaciones, asegurando que el sistema cumpla con los objetivos del negocio.

Para evaluar si la lista de libros recomendados es similar al libro solicitado y mostrar un porcentaje de similitud en cada ítem, realizamos los siguientes pasos:

- **Cálculo de la Similitud del Coseno:** Se calcula la Similitud del Coseno entre el libro ingresado y los libros recomendados. Esta métrica se utiliza para medir cuán similares son dos vectores de características, obteniendo un valor que oscila entre -1 y 1.
- **Asignación del Puntaje de Similitud:** Se agrega a la lista el porcentaje de similitud de cada libro recomendado en relación con el libro ingresado. Esto permite que el usuario visualice qué tan similar es cada recomendación respecto al libro inicial.
- **Ordenamiento por Similitud:** Los libros recomendados se muestran ordenados por similitud, de mayor a menor, facilitando al usuario identificar las recomendaciones más relevantes.

### Implementación Técnica

Para realizar la comparación de títulos y encontrar coincidencias cercanas, utilizamos la librería `difflib`. Esta librería proporciona clases y funciones para comparar secuencias, siendo particularmente útil para tareas de procesamiento de texto y análisis de diferencias. En este caso, empleamos la función `get_close_matches()` para encontrar las coincidencias más cercanas al título del libro proporcionado por el usuario.

```
import difflib
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import pandas as pd
```

```
# Carga
url = "https://raw.githubusercontent.com/sosarodrigox/grupo13_pp2/main/data/data.csv"
data = pd.read_csv(url)

# Preprocesamiento
data.fillna("", inplace=True)
libros = (
    data["title"]
    + " "
    + data["subtitle"]
    + " "
    + data["categories"]
    + " "
    + data["authors"]
    + " "
    + data["published_year"].astype(str)
)

# Vectorización TF-IDF
vector_tfidf = TfidfVectorizer()
vector_caracteristicas = vector_tfidf.fit_transform(libros)

# Cálculo similitud de coseno
similitud = cosine_similarity(vector_caracteristicas, vector_caracteristicas)
```

```
def recomendar_libros(nombre_libro):
    print(f"Libro seleccionado por el usuario: {nombre_libro}")
    lista_titulos_completa = data["title"].tolist()
    encontrar_cercanos = difflib.get_close_matches(
        nombre_libro, lista_titulos_completa)

    if encontrar_cercanos:
        cercanos = encontrar_cercanos[0]
        indice_de_libro = data[data.title == cercanos].index[0]
        puntaje_similitud = list(enumerate(similitud[indice_de_libro]))
        libros_similares_ordenados = sorted(
            puntaje_similitud, key=lambda x: x[1], reverse=True
        )

        # Lista de diccionarios con los libros recomendados y su similitud
        recomendaciones = []
        # Muestra los 10 primeros
        for i, (index, score) in enumerate(libros_similares_ordenados[1:11], 1):
            libro = {
                "Indice": i,
                "Titulo": data.iloc[index]["title"],
                "Autor": data.iloc[index]["authors"],
                "Año": data.iloc[index]["published_year"],
                "Categoria": data.iloc[index]["categories"],
                "Similitud": f"{score * 100:.2f}%",
            }
            recomendaciones.append(libro)

        # Convertir a un DataFrame
        df_recomendaciones = pd.DataFrame(recomendaciones)
        return df_recomendaciones

    else:
        print("No se encontró ninguna coincidencia para el libro ingresado.")
        return pd.DataFrame() # DataFrame vacío si no hay coincidencias

# Prueba
libro_usuario = input("Ingrese el nombre de su libro favorito: ")
df_recomendaciones = recomendar_libros(libro_usuario)
display(df_recomendaciones)
```



Libro seleccionado por el usuario: frankenstein						
Indice	Titulo		Autor	Año	Categoria	Similitud
0	1	Frankenstein, Or, The Modern Prometheus	Mary Wollstonecraft Shelley	2002.0	Fiction	48.01%
1	2	Mary Wollstonecraft Shelley's Frankenstein, or, The modern Prometheus	Mary Wollstonecraft Shelley;Susan J. Wolfson	2007.0	Fiction	47.79%
2	3	Frankenstein	Mary Shelley	2002.0	LITERARY CRITICISM	42.28%
3	4	Frankenstein	Mary Shelley	2004.0	Fiction	41.80%
4	5	Frankenstein	Mary Wollstonecraft Shelley	2003.0	Fiction	33.24%
5	6	Sense and Sensibility	Jane Austen	2002.0	Fiction	22.39%
6	7	The Crucible	Arthur Miller	1996.0	Drama	20.45%
7	8	Robinson Crusoe	Daniel Defoe;Michael Shinagel	1994.0	Fiction	18.72%
8	9	The Turn of the Screw	Henry James;Deborah Esch;Jonathan Warren	1999.0	Fiction	18.55%
9	10	1 Henry IV	William Shakespeare	2003.0	Drama	18.38%

De los resultados obtenidos se observa que los mismos se alinean con los objetivos del negocio y del presente proyecto de datos, permitiendo realizar la recomendación de libros requerida por los stakeholders.

## DESPLIEGUE

A fin de integrar el modelo realizado con los procesos de negocio y mejorar la experiencia de los usuarios de la biblioteca, se ha creado una interfaz mediante Streamlit, un marco de trabajo de Python de código abierto que permite crear aplicaciones web personalizadas para el aprendizaje automático y la ciencia de datos.

La aplicación web admite múltiples formatos de entrada:

- **TEXTO:** El usuario puede ingresar el título de un libro para obtener una lista de libros recomendados ordenados por similitud de contenido, facilitando la exploración y la selección de nuevos materiales de lectura.
- **VOZ:** Utilizando la biblioteca de procesamiento de voz `speechrecognition`, los usuarios pueden proporcionar comandos de voz que son transcritos y analizados para generar recomendaciones de libros.
- **IMÁGENES:** Empleando `pytesseract` para el reconocimiento óptico de caracteres (OCR) y `Pillow` para el manejo de imágenes, los usuarios pueden cargar imágenes de portadas de libros. La aplicación extrae texto de las imágenes para sugerir libros similares.

Para el desarrollo de estas funcionalidades, se han utilizado las siguientes librerías de Inteligencia Artificial:

- **scikit-learn:** Para la implementación de algoritmos de aprendizaje automático que generan las recomendaciones utilizando la técnica de vectorización TF-IDF y similitud de coseno.
- **pytesseract:** Para la extracción de texto de las imágenes de las portadas de los libros mediante Reconocimiento Óptico de Caracteres (OCR).
- **speechrecognition:** Para para convertir audio en texto permitiendo la transcripción y procesamiento de comandos de voz.
- **Pillow:** Para la manipulación y procesamiento de imágenes.

Una vez validado con los responsables de la Biblioteca, planificaremos y ejecutaremos la implementación de esta aplicación en el entorno de producción del portal de la institución, junto con gestionar los accesos correspondientes y proporcionar la documentación de usuario final para su correcto uso.

Imágenes de la aplicación:

## Sistema de Recomendación de Libros

### Entrada de Texto

Ingrese el nombre de su libro favorito:

Buscar Similares

### Entrada de Voz

Start Recording Stop Reset Download

0:00 / 0:00

### Entrada de Imagen

Sube tu foto aquí:

Drag and drop file here  
Limit 200MB per file

Browse files

Por favor, carga una imagen.

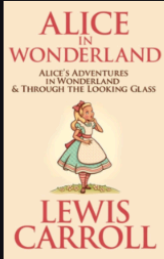
Sube tu foto aquí:

Drag and drop file here  
Limit 200MB per file

Browse files

alice.png 0.7MB

Imagen subida correctamente!



He reconocido el siguiente texto: Alice Wonderland Alice S

Libro seleccionado por el usuario: Alice Wonderland Alice S

Libros similares encontrados:

	Índice	Título	Autor	Año	Categoría	Similitud
0	1	Alice's Adventures in Wonderland	Lewis Carroll	2,000	Juvenile Fiction	87.23%
1	2	Dreaming in Pictures	Douglas Robert Nic	2,002	Photography	37.34%
2	3	Wonderland	Joyce Carol Oates	2,006	Fiction	32.47%
3	4	Sleeping in Flame	Jonathan Carroll	2,004	Fiction	27.32%
4	5	Lewis Carroll's Jabberwocky	Graeme Base	1,996	Juvenile Nonfiction	25.35%
5	6	The Complete C. S. Lewis Signature Classics	C. S. Lewis	2,007	Religion	23.95%
6	7	The Courtesan	Susan Carroll	2,005	Fiction	23.67%
7	8	Glass Soup	Jonathan Carroll	2,006	Fiction	22.70%
8	9	The Bride Finder	Susan Carroll	1,999	Fiction	21.80%
9	10	C.S. Lewis	C. S. Lewis	1,996	Religion	21.74%

### Entrada de Texto

Ingrese el nombre de su libro favorito:

Dracula

Buscar Similares

Libro seleccionado por el usuario: Dracula

Libros similares encontrados:

	Índice	Título	Autor	Año	Categoría
0	1	Dracula	Bram Stoker	2,003	Fiction
1	2	Bram Stoker's Dracula	Harold Bloom	2,003	Juvenile Nonfic
2	3	The Bram Stoker Bedside Companion	Bram Stoker	1,973	Fiction
3	4	In Search of Dracula	Raymond T. McNally;Radu Florescu	1,994	Biography & Au
4	5	Dracula	Bram Stoker	1,997	Fiction
5	6	A Dracula Handbook	Elizabeth Miller	2,005	Education
6	7	Happy Hour at Casa Dracula	Marta Acosta	2,006	Fiction
7	8	Best Ghost and Horror Stories	Bram Stoker;Richard Dalby;Stefan R.	1,997	Fiction
8	9	Count Zero	William Gibson	2,006	Fiction
9	10	The Lord of the Rings		2,002	Baggins, Frodo

### Entrada de Voz

Start Recording Stop Reset Download

0:00 / 0:01

El libro detectado es: Frankenstein

Libro seleccionado por el usuario: Frankenstein

Libros similares encontrados:

	Índice	Título	Autor	Año
0	1	Frankenstein, Or, The Modern Prometheus	Mary Wollstonecraft Shelley	2,002
1	2	Mary Wollstonecraft Shelley's Frankenstein,	Mary Wollstonecraft Shelley;Susan J. Wolfson	2,007
2	3	Frankenstein	Mary Shelley	2,002
3	4	Frankenstein	Mary Shelley	2,004
4	5	Frankenstein	Mary Wollstonecraft Shelley	2,003
5	6	Sense and Sensibility	Jane Austen	2,002
6	7	The Crucible	Arthur Miller	1,996
7	8	Robinson Crusoe	Daniel Defoe;Michael Shinagel	1,994
8	9	The Turn of the Screw	Henry James;Deborah Esch;Jonathan Warren	1,999
9	10	1 Henry IV	William Shakespeare	2,003

## CONCLUSIÓN

El presente proyecto de implementación de un sistema de recomendación de libros basado en técnicas de minería de datos y procesamiento del lenguaje natural ha demostrado ser una solución eficaz para mejorar la gestión y la experiencia del usuario en la Biblioteca Popular. A través de la metodología CRISP-DM, hemos logrado estructurar el proyecto en fases claras y organizadas, permitiendo una ejecución coherente y bien documentada de cada etapa.

El uso de técnicas como TF-IDF y la Similitud del Coseno ha proporcionado una base sólida para la recomendación de libros, facilitando la conversión de texto en vectores numéricos y la medición de similitud entre documentos. Estas herramientas no solo mejoraron la precisión de las recomendaciones, sino que también optimizaron el tiempo de búsqueda para los usuarios, incrementando su satisfacción y fidelidad.

Luego, la integración de nuevas tecnologías, como el reconocimiento de voz y el reconocimiento óptico de caracteres (OCR), ha ampliado significativamente las capacidades del sistema. Estas mejoras permiten a los usuarios interactuar con la aplicación a través de múltiples medios, incluyendo texto, voz e imágenes de portadas de libros, proporcionando una experiencia de usuario más rica e inclusiva.

A pesar de los desafíos encontrados, como la digitalización incompleta de los datos y las dificultades iniciales en el procesamiento de voz e imágenes, las soluciones implementadas han permitido superar estos obstáculos y fortalecer el sistema. La flexibilidad y robustez de las tecnologías utilizadas han sido clave para adaptar y mejorar continuamente el sistema de recomendaciones.

Es por esto que consideramos que el proyecto no solo ha cumplido con los objetivos iniciales de mejorar la eficiencia operativa y la experiencia del usuario, sino que también ha sentado las bases para futuras mejoras y expansiones del sistema. La metodología CRISP-DM y las técnicas empleadas en este proyecto pueden servir de modelo para otras iniciativas similares, demostrando la aplicabilidad y efectividad de la minería de datos y las tecnologías de inteligencia artificial en diversos contextos.

## ANEXO: LECCIONES APRENDIDAS

En el desarrollo del presente proyecto, se ha reafirmado que la fase de recopilación y preparación de datos es la que demanda mayor tiempo y esfuerzo en un proyecto de minería de datos. Esta observación es consistente con nuestra experiencia, ya que los datos necesarios no estaban completamente digitalizados ni sistematizados. Para superar este desafío, inicialmente utilizamos un conjunto de datos de Kaggle, lo que nos permitió avanzar en el modelado mientras paralelamente se realizaba la recolección de una base de datos más robusta mediante técnicas de web scraping.

Uno de los principales desafíos fue la falta de una base de datos de usuarios y transacciones, lo cual impidió alcanzar uno de los objetivos iniciales del proyecto: predecir la demanda de libros y ajustar las adquisiciones de nuevos textos en consecuencia. Este obstáculo resaltó la importancia de contar con datos transaccionales detallados para la implementación de modelos predictivos precisos en el contexto de gestión de inventarios.

Asimismo, al abordar el procesamiento de imágenes de portadas de libros para la búsqueda y recomendación de textos similares, encontramos que estas imágenes presentan composiciones complejas, dificultando su interpretación automática. Inicialmente, el modelo propuesto logró una precisión del 20%, lo cual era insuficiente para nuestras necesidades. Como solución, se implementó la tecnología OCR (Optical Character Recognition) utilizando `pytesseract`, que permite convertir documentos escaneados, archivos PDF y fotografías de texto en datos editables y buscables. La integración de esta tecnología mejoró significativamente la precisión y funcionalidad del sistema de recomendación de libros.

Además, la incorporación de la entrada por voz, utilizando la biblioteca `speechrecognition`, presentó desafíos en la calidad de la transcripción y la interpretación del lenguaje natural. Sin embargo, estos desafíos fueron superados con ajustes en la limpieza de datos y la normalización de texto, mejorando la precisión de las recomendaciones basadas en voz.

Estas experiencias subrayan la necesidad de una preparación de datos meticulosa y de considerar múltiples enfoques tecnológicos para superar las limitaciones inherentes en los proyectos de minería de datos. También hemos aprendido que algunos objetivos iniciales de la biblioteca eran imposibles de cumplir con el dataset disponible debido a la falta de datos sistematizados y transaccionales. No obstante, el proyecto final ha logrado cumplir muy bien con la integración de la inteligencia artificial para abordar problemáticas reales de la biblioteca, mejorando la experiencia del usuario y optimizando los procesos de recomendación de libros. Este proyecto establece una base sólida para futuras iteraciones y mejoras, demostrando la aplicabilidad y efectividad de la inteligencia artificial en la gestión de bibliotecas.