UNIVERSIDAD
Panamericana

Alumnos:

Adriana Alejandra Pérez Echeverría

Luis Rodrigo Consuelos Figueroa

Claudio Diego Vázquez Nava

Trabajo:

Proyecto Final: Books

Materia:

Fundamentos de Ciencia de Datos

Profesor:

Carlos Sebastian Loredo Gómez

Fecha de entrega:

30 de septiembre de 2025

1. Entendimiento del negocio y enfoque analítico

1.1 Determinar objetivos del negocio

Books to Scrape es un negocio en línea que lleva operando en el sector editorial durante años. El negocio está enfocado en ofrecer un catálogo amplio y variado a buen precio, sin embargo, con el crecimiento de internet, la entrada de nuevos competidores ha significado una alerta para identificar varias áreas de mejora donde el análisis de datos es un factor clave.

Objetivos principales

Predecir el precio de un libro con base en sus características sin una alta tasa de pérdidas.

A lo largo de los años, la definición de los precios se ha realizado de manera simple, manual y estática, tomando en cuenta costos bases, impuestos y un margen estándar. Este enfoque es sencillo de aplicar mientras haya un volumen bajo de libros, pero en este nuevo entorno competitivo, puede significar problemas porque se puede caer en una pérdida de ventas por sobreprecio o una pérdida de rentabilidad por sub-precio.

Clasificar libros dentro de una categoría (género) de forma automática

Es una tarea que se ha realizado manualmente dentro del negocio, basándose en la descripción del producto y en criterios del equipo editorial. Al incrementar el catálogo, este proceso significa una pérdida de dinero y tiempo, aparte de ser propenso a errores. Un libro mal catalogado puede afectar su visibilidad en la plataforma, reducir sus ventas y ser una mala experiencia para el usuario.

Criterios de éxito (Eficiencia en el negocio)

- Incrementar conversiones en el sitio web de Book to Scrape.
- Mejorar la competitividad en cuestión de precios, aumentar el margen a un 5% con el modelo de precios.
- Reducir a la mitad (a un 50%) el tiempo invertido en la clasificación de libros.
- Disminuir errores de categorización.

1.2 Estado actual del negocio

Inventario y recursos

- Equipo: científicos de datos, analistas de precios, equipo de marketing.
- Información: dataset con 1000 registros de libros, los datos que contiene son: título, descripción, categoría, rating, reseñas, disponibilidad, impuestos y precios.
- Arquitectura computacional y software: Python, Colab, librerías de análisis de datos y ML (scikit-learn, pandas, etc).

Requerimientos y restricciones

- Tiempo estimado: 1 mes de desarrollo y validación.
- Temas legales: No se cuenta con datos sensibles de usuarios, solo información pública de los libros.
- Restricciones: Se cuenta con pocos datos (1000 datos) y vairables, y el dataset no está balanceado en categorías.

Riesgos y contingencias

- Overfitting: Hay poca cantidad de datos en diferentes categorías.
- Predicción sesgada: Si no se encuentran todas las categorías necesarias y se introduce uno nuevo sin antes de enseñarle al modelo, no podrá catalogar bien.
- Contingencia: Recolección de datos

Glosario de terminología

- Precio (price_incl_tax): Valor de venta final de un libro.
- Categoría: Género literario.
- Rating: Calificación dada al libro por usuario (1-5)
- n_reviews: número de reseñas que tiene un libro.

Análisis costo-beneficio

Costos: Horas del equipo de analítica, cómputo y tiempo de integración.

Beneficios: Precios competitivos, organización automática de catálogo e insights para marketing.

1.3 Crear un plan de trabajo

1. Recolección de datos: Recolectar datos a través del sitio web
2. Limpieza de datos: Quitar valores innecesarios, estructurar los datos.
3. EDA (Exploración de datos): Análisis general de los datos en el dataset, como análisis de precios, ratings y distribución de categorías.
4. Modelado: Generar modelos de regresión para evaluar el precio y generar modelos de clasificación para las categorías.
5. Evaluación y validación: Ajustar parámetros, ver desempeño.
6. Comunicar resultados.
7. Despliegue.

2. Requerimiento de datos y extracción de datos

Los datos necesarios para el estudio se encuentran en el catálogo de nuestro sitio web Books to Scrape. Cada libro cuenta con la siguiente información:

- Nombre del libro
- Categoría
- UPC
- Product type
- Precio (excl. tax)
- Precio (incl. tax)
- Tax
- Disponibilidad
- Número de reseñas
- Rating (en estrellas)
- Descripción
- URL de la imagen

Los datos fueron extraídos de la página mediante web scraping utilizando Selenium con Python en Google Colab. Extraer los datos de esta manera nos permitió visualizar la información del sitio de la manera que un usuario lo haría.

La extracción de los datos consistió en obtener los url de cada título recorriendo las categorías, con estos, se pudo acceder a la página de cada título y recolectar la información completa.

El dataset que se obtuvo cuenta con los 1000 títulos del sitio web y las 12 columnas, que hacen referencia a las características de los títulos mencionados anteriormente.

3. Preparamiento de datos

Una vez recolectada la información fue necesario depurar, transformar y estructurar los datos con el fin de proceder con un análisis adecuado.

Limpieza

- Se eliminaron características que no aportan información relevante
    - **UPC:** Los ID del producto son únicos para todos los productos
    - **product_type:** indica que el artículo es un libro, siendo irrelevante porque todos los artículos analizados son libros.
    - **Price (incl. tax) y Tax**: El impuesto de todos los artículos es de 0.
    - **Descripción:** El 99.8% de los títulos tenían descripción, por lo que se prefirió la creación y uso de la variable description_length (número de palabras).
    - **URL de la imagen**: Las url de las imágenes son únicas.
- Se estandarizaron formatos:
    - **Precio, Disponibilidad**, **Rating**: Se pasó de texto a números.
- Manejo de valores atípicos:
    - **Description_length:** se calculó un valor mínimo y máximo, de forma que los títulos que pasarán estos sustituyeran su valor por uno de los límites.

Adición de características

- Se generaron variables derivadas para capturar más información del catálogo:
    - description_length: número de palabras en la descripción.
    - sentiment_label: análisis de sentimiento de la descripción (positivo, neutro, negativo) con procesamiento de lenguaje natural a través de un modelo ya entrenado encontrado en nltk (SentimentIntensityAnalyzer).

- Se revaluó categorías ambiguas:
    - Libros con categorías “Default” y “Add a comment” fueron reclasificados con apoyo de un modelo de lenguaje (Gemini) hacia categorías existentes.
- Se creó una columna consolidada categoría_agrupada para agrupar categorías muy específicas con pocos títulos, mejorando la representatividad en los modelos y se eliminó la columna de categoría.

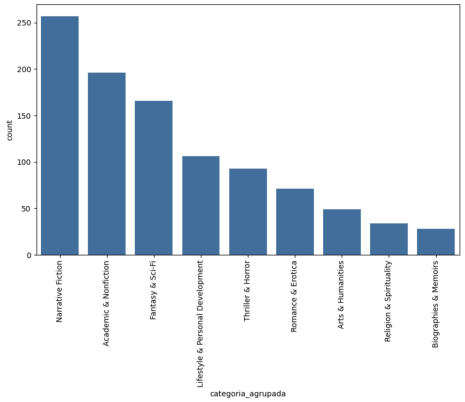
Resultado final

Tras la preparación, el dataset quedó conformado por las siguientes variables clave para el análisis:

- availability (disponibilidad en unidades)
- categoría_agrupada
- rating (1 a 5)
- description_length (número de palabras)
- price (valor numérico)
- sentiment_label (positivo, neutro, negativo)

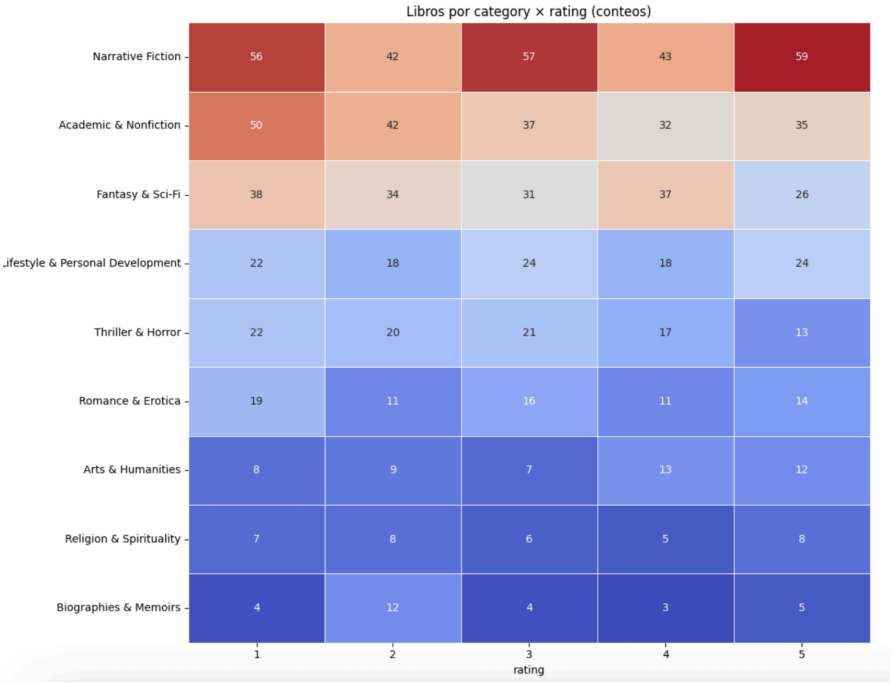
4. Entendimiento de los datos

- Hay una diferencia en la cantidad de títulos entre las categorías agrupadas.

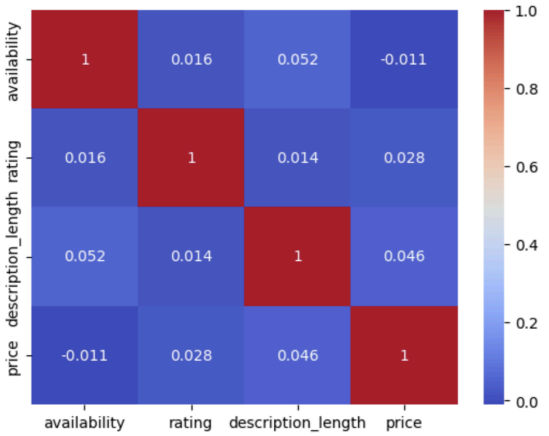


- Si se tuvieran más títulos por categorías podría analizarse mejor el rating que reciben. No obstante, en general parece haber una cantidad relativamente

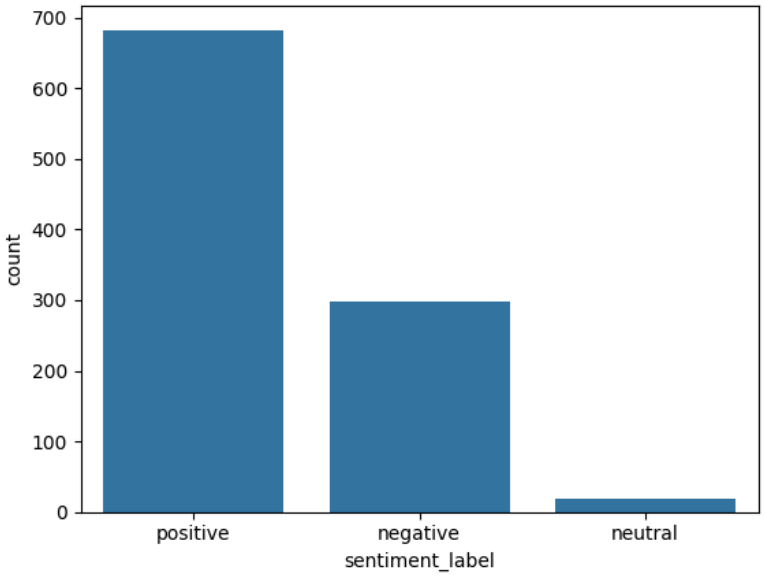
similar en rating, es decir, que no hay una categoría que califique totalmente bien o totalmente mal.



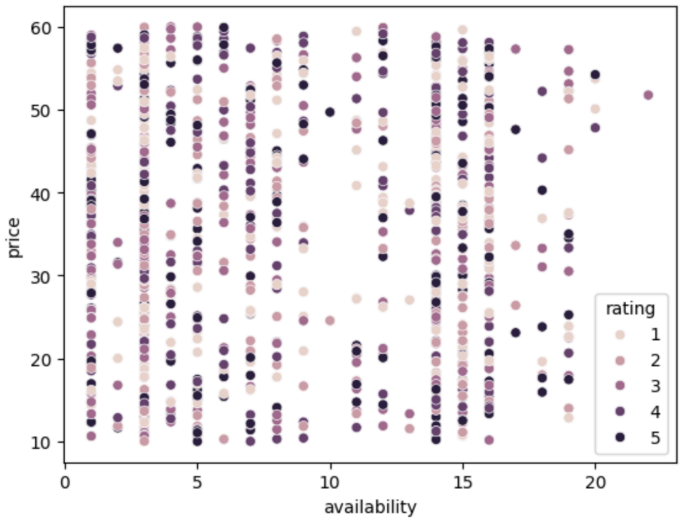
- No hay relación lineal entre las variables numéricas, ya que las correlaciones son cero.



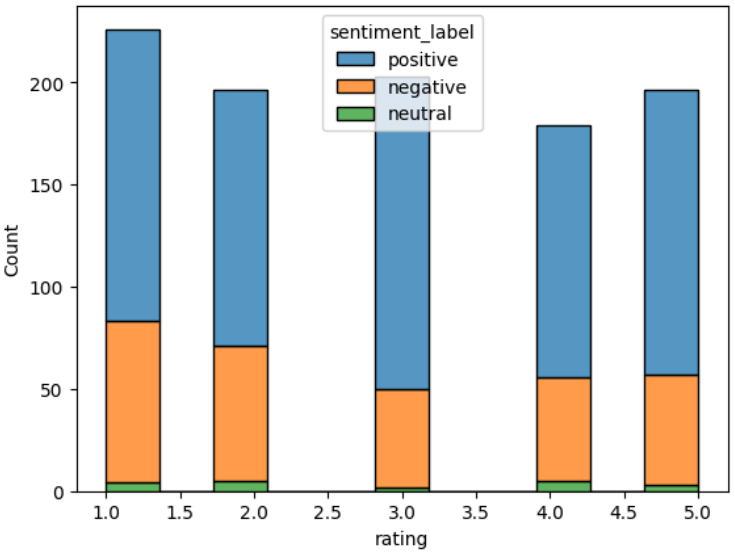
- La mayoría de los libros cuentan con un sentimiento positivo. Es muy raro que haya libros con sentimientos neutros.



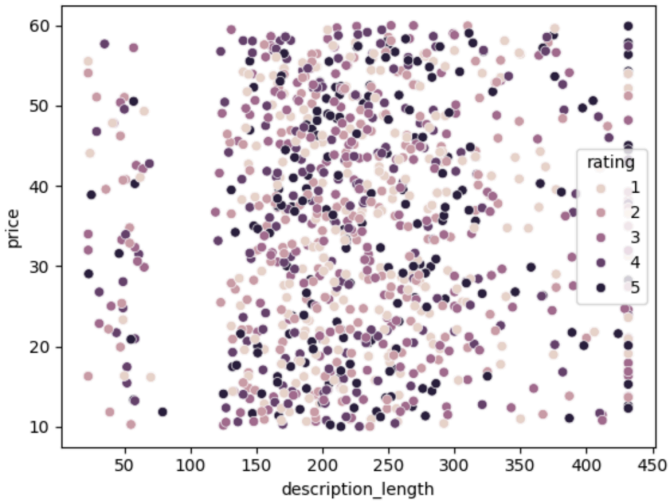
- No hay evidencia de que los libros con menor disponibilidad tengan mejor calificación y sean más costosos. Tampoco se muestra que los libros con mayor calificación sean los más caros.



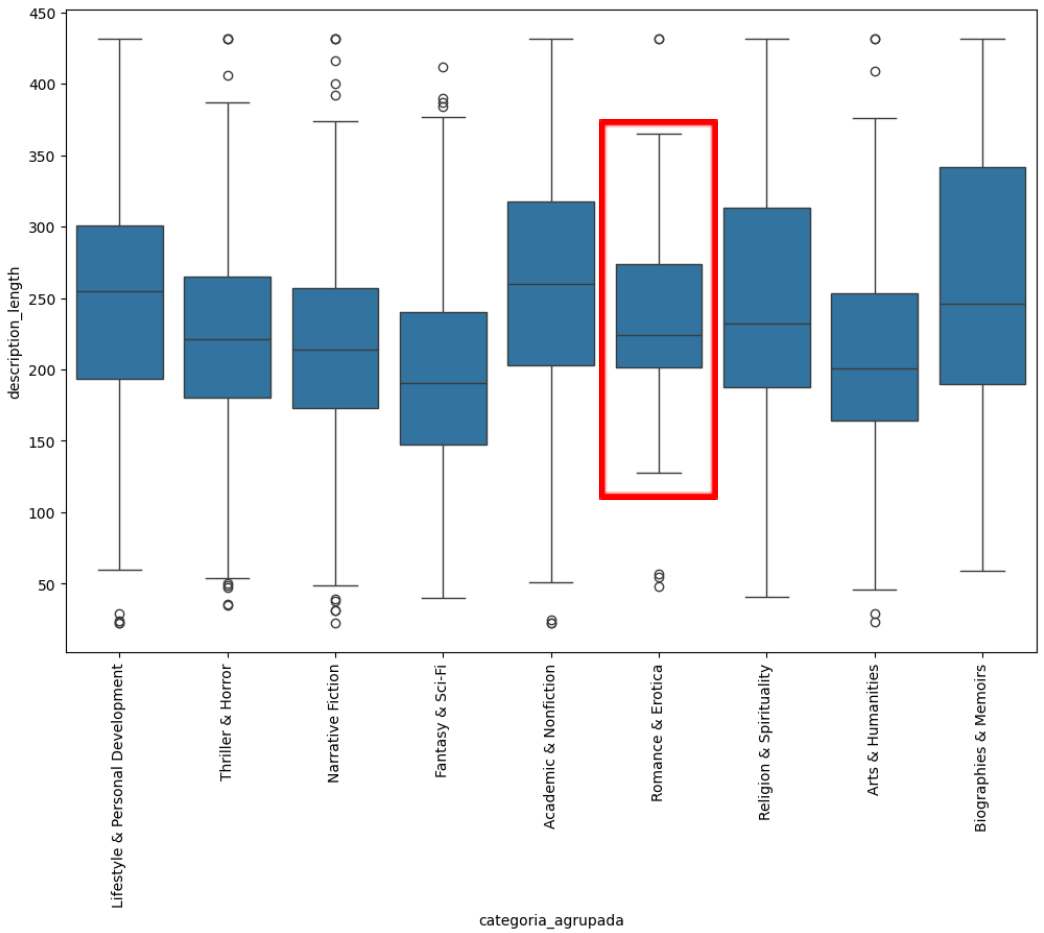
- Los libros con un sentimiento negativo, porcentualmente tienen menor rating comparados con los libros de sentimiento positivo.



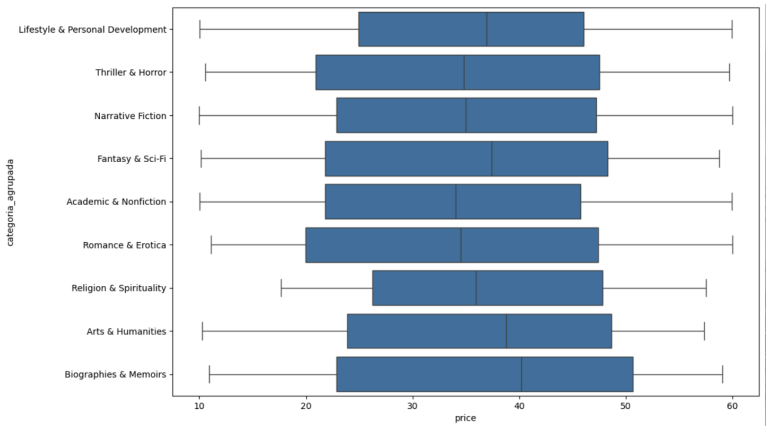
- La longitud de la descripción de los libros no influye en los libros. La longitud de descripción de los libros se concentra en el rango de 140 a 330 palabras.



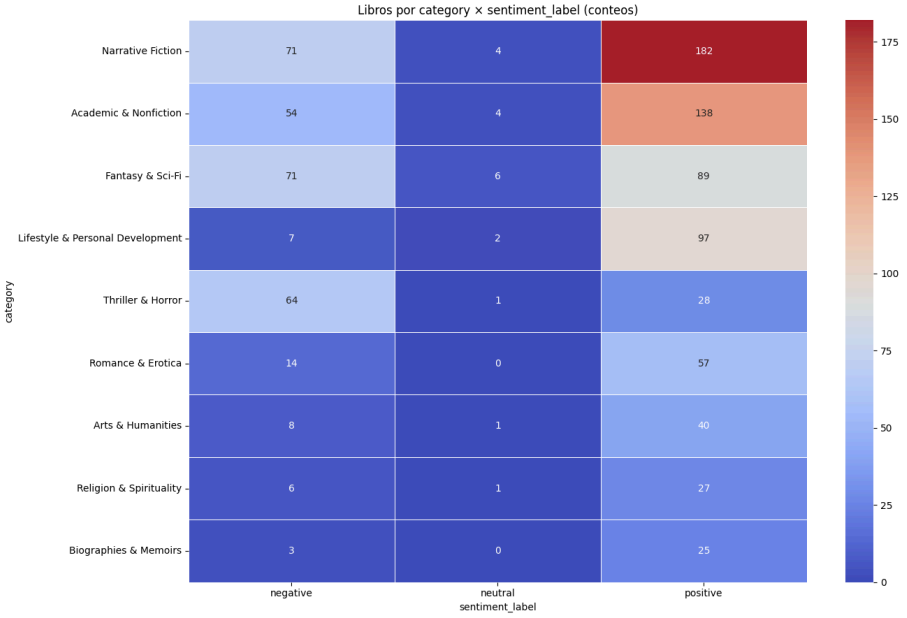
- La cantidad de palabras de la descripción para todas las categorías tienen un rango similar, aunque su distribución es diferente. La categoría con menor rango, que tiende a ser más consistente en su cantidad de palabras, es Romance y Erótica.



- La categoría de libros sí parece tener algo de influencia en el precio de los libros, pero no tan significante. Aproximadamente el 50% de libros de las categorías se encuentran entre 22 y 50 dólares.



- La mayoría de las categorías tienen más libros con descripciones positivas. Aunque sí se aprecia que la categoría influye en el sentimiento, ya que la mayoría de los libros de la categoría Thriller & Horror cuentan con una descripción que genera sentimiento negativo.



- Tabla de medidas de tendencia central y de desviación.

	mean	std	median	min	Q1	Q3	max
availability	8.585000	5.654622	7.00	1.000	3.0000	14.0000	22.000
rating	2.923000	1.434967	3.00	1.000	2.0000	4.0000	5.000
description_length	230.490125	86.167001	221.00	22.625	176.0000	278.2500	431.625
price	35.070350	14.446690	35.98	10.000	22.1075	47.4575	59.990

6. Prueba de Hipótesis

6.1 ¿El sentimiento del libro (positivo o negativo) influye en el precio?

Se realizó una prueba t de Student para muestras independientes con el fin de comparar el precio promedio de los libros entre aquellos con sentimiento positivo y aquellos con sentimiento negativo. El resultado ($t=-0.46$, $p=0.649$) mostró que, con un nivel de significancia $\alpha=0.05$, no se rechaza la hipótesis nula. Por lo tanto, no existe evidencia estadísticamente significativa para afirmar que el precio promedio de los libros difiere en función del sentimiento.

6.2 ¿El sentimiento del libro influye en el rating?

Se realizó una prueba t de Student para muestras independientes con el fin de comparar el rating promedio de los libros entre aquellos con sentimiento positivo y aquellos con sentimiento negativo. El resultado ($t=2.02$, $p=0.044$) mostró que, con un nivel de significancia $\alpha=0.05$, se rechaza la hipótesis nula. Por lo tanto, existe evidencia estadísticamente significativa para afirmar que el sentimiento del libro influye en el rating promedio.

6.3 Los libros con rating alto (mayores o iguales a la media) tienen un precio mayor a los de rating bajo?

Se realizó una prueba t de Student para muestras independientes con el fin de comparar el precio promedio de los libros entre aquellos con rating alto (mayor o igual al promedio) y aquellos con rating bajo (menor al promedio). El resultado ($t=0.74$, $p=0.457$) mostró que, con un nivel de significancia $\alpha=0.05$, no se rechaza la hipótesis nula. Por lo tanto, no existe evidencia estadísticamente significativa para afirmar que el rating influye en el precio promedio de los libros.

6.4 ¿Los libros con descripción larga tienen mayor disponibilidad que los libros con descripción corta?

Se realizó una prueba t de Student para muestras independientes con el fin de comparar la disponibilidad promedio de los libros entre aquellos con descripción larga (mayor o igual a la media) y aquellos con descripción corta (menor a la media). El resultado ($t=1.65$, $p=0.099$) mostró que, con un nivel de significancia $\alpha=0.05$, no se rechaza la hipótesis nula. Por lo tanto, no existe evidencia estadísticamente significativa para afirmar que la longitud de la descripción influye en la disponibilidad de los libros.

6.5 Los libros con mayor disponibilidad tienen menor rating que los libros con menor disponibilidad?

Se realizó una prueba t de Student para muestras independientes con el fin de comparar el rating promedio de los libros entre aquellos con alta disponibilidad (mayor o igual a la media) y aquellos con baja disponibilidad (menor a la media). El resultado ($t=0.23$, $p=0.815$) mostró que, con un nivel de significancia $\alpha=0.05$, no se rechaza la hipótesis nula. Por lo tanto, no existe evidencia estadísticamente significativa para afirmar que la disponibilidad influya en el rating de los libros.

6.6 El género influye en el sentimiento?

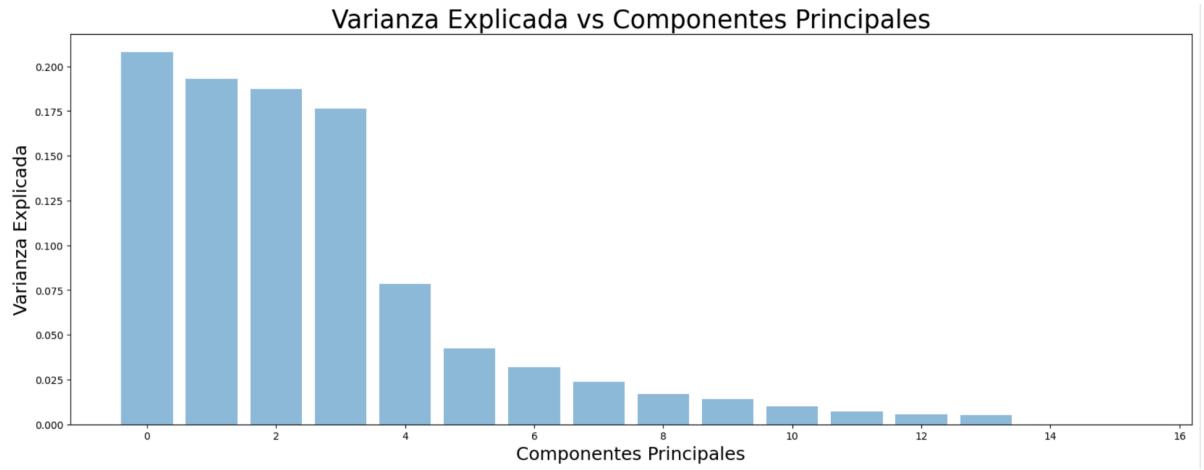
Se aplicó una prueba chi-cuadrada de independencia para evaluar la asociación entre el género del libro (categoría agrupada) y el sentimiento asignado. El resultado ($\chi^2=131.45$, $p<0.001$) mostró que, con un nivel de significancia $\alpha=0.05$, se rechaza la hipótesis nula. Por

lo tanto, existe evidencia estadísticamente significativa de que el género influye en el sentimiento de los libros.

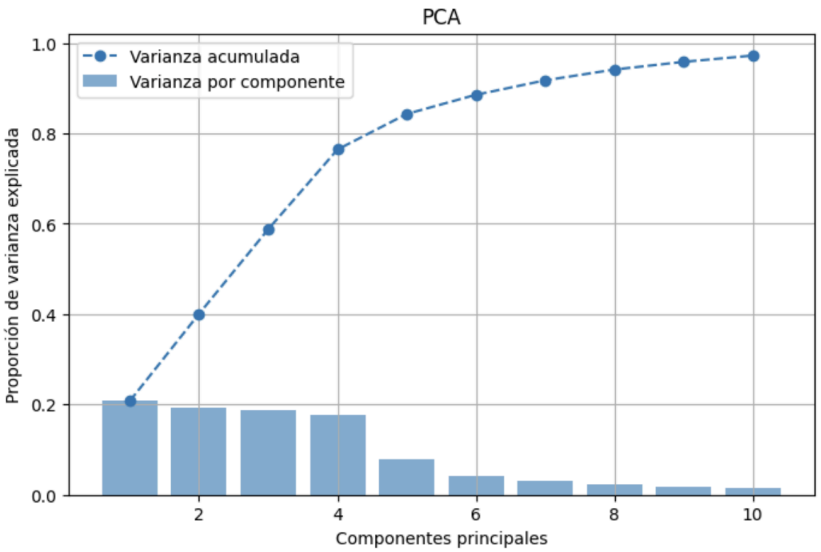
6.7 Los libros con menor disponibilidad tienen mayor precio que los libros con mayor disponibilidad?

Se realizó una prueba t de Student para muestras independientes con el fin de comparar el precio promedio de los libros entre aquellos con alta disponibilidad (mayor o igual a la media) y aquellos con baja disponibilidad (menor a la media). El resultado ($t=-0.28$, $p=0.777$) mostró que, con un nivel de significancia $\alpha=0.05$, no se rechaza la hipótesis nula. Por lo tanto, no existe evidencia estadísticamente significativa para afirmar que la disponibilidad influya en el precio de los libros.

8. PCA



El 100% de varianza de los datos puede explicarse mediante 16 componentes, aunque a partir del componente 14, la varianza explicada es cero.



El PCA muestra que con los primeros 7 componentes principales se explica aproximadamente el 90% de la varianza total del dataset.
No obstante, como las características son pocas, se decidió no implementar PCA en los modelos.

8. Construcción de los modelos

8.1 Seleccionar técnicas de modelado

Los objetivos del proyecto son ajustar los precios de los libros y clasificarlos bien, se propone usar modelos de regresión y clasificación (respectivamente). Las siguientes técnicas se utilizaron:

Predicción del precio (regresión)

- Modelos lineales: Se usa regresión lineal solo y con Lasso/Ridge para evaluar relaciones simples y poder reducir multicolinealidad.
- Modelos basados en árboles: Se usa Random Forest Regressor y XGBoost para capturar relaciones no lineales y suelen tener mejor desempeño con datos heterogéneos.

Clasificación de categorías (clasificación supervisada)

- RandomForest: es muy robusto frente a ruido y buenos con variables categóricas.
- Máquina de Soporte Vectorial: Es buena para datasets medianas y se evalúa adecuadamente cuando hay desbalance.
- Regresión lineal: Es fácil de interpretar.

Criterios de selección inicial

- Comparar modelos simples contra complejos.
- Evaluar interpretabilidad y precisión.
- Seleccionar los que den mejor rendimiento y resultados en pruebas de validación.

8.2 Diseñar las pruebas

Se diseña un procedimiento para evaluar la calidad del modelo. Este procedimiento sigue los siguientes pasos:

Separación de datos

Se separan los datos en 80% entrenamiento y 20% prueba de forma aleatoria pero consistente en cada ejecución.

Métricas de evaluación

Para el modelo de predicción de precios:

- RMSE (error cuadrática medio)

Para el modelo de clasificación de categorías:

- Se usa el reporte de clasificación que permite ver el desempeño del modelo por categoría, usa acurracy, precision, recall y f1-score.

Procedimiento de control:

Se pretende revisar e identificar sesgos dentro de los precios y también se planea checar el balance dentro de las clases para evitar que favorezca a las categorías mayoritarias. Este procedimiento de control se realiza a través de iteraciones de ajuste donde se verifican y cambian hiper parámetros para mejorar los modelos y sus resultados.

8.3 Construir los modelos

En esta etapa del proyecto se procede a entrenar los modelos de predicción y clasificación. Para lograrlo, se siguen los siguientes pasos:

1. Datos preparados: Debemos tener cargados todos los datos ya limpios, transformados y filtrados en subconjuntos aleatorios de entrenamiento y prueba (80% y 20% respectivamente), listos para ser pasado al modelo.
2. Todos los modelos se entrenan dentro de un pipeline, cuyo primer paso es un ColumnTransformer que aplica:
    a. StandardScaler a las variables numéricas, para estandarizarlas con media cero y desviación estándar uno. Útil para cuando los rangos entre variables difieren demasiado, ya que permite comparar distribuciones.

b. OneHotEncoder a las variables categóricas, para convertirlas en variables binarias.

3. Continúa el entrenamiento del modelo: Se entrenan diferentes modelos para la regresión y la clasificación:
    a. Para regresión:
        i. XGBoost con 100 árboles, profundidad 5 y una tasa de aprendizaje del 0.01.
        ii. Random Forest Regressor con 100 árboles y profundidad 5.
        iii. Regresión Lineal solo y con Lasso y Ridge (fuerza de regularización del 0.05).
    b. Para Clasificación sin undersampling:
        i. Máquina de Soporte Vectorial (SVM).
        ii. Random Forest Classifier con 100 árboles y profundidad 7.
        iii. Regresión Logística.
    c. Para clasificación con undersampling
        i. SVM.
        ii. Random Forest Classifier con 100 árboles y profundidad 7.
        iii. Regresión logística.
        iv. KNeighbors Classifier

9. Evaluación del modelo

Resultados de regresión

Modelo	RMSE	R^2
Regresión Lineal	14.014	-0.040
Lasso	13.780	-0.005
Ridge	14.009	-0.039
Random Forest Regressor	14.194	-0.067
XGBoost	14.044	-0.044

El mejor modelo es Lasso, a pesar de que no hay relación lineal entre variables. Es el modelo con menor RMSE (13.78) y la mejor R^2 (-0.0057). Cabe mencionar que

Lasso le da peso únicamente a description_length y es un valor que puede considerarse cero (0.04). En sí es el mejor modelo porque prácticamente solo considera el intercepto (35.28), el cual es un valor cercano al promedio (35.07). Esto nos dice que es mejor que demos como estimación el promedio.

Un detalle interesante es que el error es similar a la desviación estándar del precio (14 apróximadamente).

Resultados de clasificación

Previo al ajuste de cantidad de títulos por categoría (máximo 100 títulos por categoría)

Modelo	Accuracy
Regresión Logística	0.23
Random Forest Classifier	0.27
SVC (máquina de soporte vectorial)	0.29

Después del ajuste de cantidad de títulos por categoría

Modelo	Accuracy
Regresión Logística	0.28
Random Forest Classifier	0.27
SVC (máquina de soporte vectorial)	0.30
KNeighbors Classifier	0.26

De los modelos de clasificación, después de nivelar el número de registros por categoría, el mejor siguió siendo la máquina de soporte vectorial. Fue aquel con accuracy más alto (0.31) y con métricas de precision y recall balanceados.

Aún así el modelo no es bueno para la tarea.

10. Conclusiones y áreas de oportunidad

No se muestra ninguna relación entre las características de los libros y sus categorías o precios. Quiere decir que el tamaño de la descripción, cantidad disponible, calificación no influye en el precio o la categoría.

La cantidad de los datos es limitada por lo que perjudica los resultados de los modelos de aprendizaje de máquina, así como los resultados estadísticos.

El rango del precio es estrecho, lo cual hace que el promedio sea un buen estimador. Porcentualmente, podrá significar mucho, pero en términos de la cotidianidad del consumidor el precio no le afectará demasiado.

Debido a estas dos complicaciones, se recomienda especificar con mayor precisión las características de los libros. Datos que podrían ser de interés serían el tamaño del libro, número de páginas, autor, número de reseñas, pasta blanda/dura, densidad de página, tipo de libro, fecha de lanzamiento (antigüedad). Porque estos factores están directamente relacionados con el precio.

La adición de una sección de reseñas puede potenciar la predicción de categoría, debido a que la categoría y el precio no están relacionados fuertemente, necesitando de otras características para clasificar correctamente el texto.

El modelo de predicción del precio no muestra una diferencia significativa con el promedio del precio, por lo que no se recomienda desplegar el modelo.

Se recomienda un posterior análisis de los datos una vez añadidos los datos sugeridos en este informe.