Predicción de géneros de películas a partir del argumento o plot

Daniela Parra[a,c], Manuel Hernández [a,c], Paula Jaimes [a,c],
Lina Varón[a,c] , Jenny Ortiz [a,c]
Sergio Alberto Mora[b,c]

[a]*Estudiantes de Maestría en Analítica para la Inteligencia de Negocios*
[b]*Profesor, Departamento de Ingeniería Industrial*
[c]*Pontificia Universidad Javeriana, Bogotá, Colombia*

ENTENDIMIENTO DEL NEGOCIO

Background

Los sistemas de recomendación se han convertido en una herramienta fundamental en la industria del entretenimiento, especialmente en el sector cinematográfico y de streaming. Estos sistemas no solo mejoran la experiencia del usuario, sino que también tienen un impacto significativo en los resultados financieros de las empresas. Según un estudio publicado en el Journal of Marketing, los sistemas de recomendación pueden aumentar las ventas en hasta un 35% (Hosanagar et al., 2014). En el contexto de la industria cinematográfica, esto se traduce en un incremento sustancial del tiempo de visualización y la retención de suscriptores. Netflix, líder en el mercado de streaming, es un ejemplo paradigmático del poder de los sistemas de recomendación. La compañía estima que su algoritmo de recomendación ahorra más de $1 mil millones al año en retención de clientes (Gomez-Uribe & Hunt, 2016). Este sistema no solo sugiere contenido basado en el historial de visualización del usuario, sino que también personaliza las imágenes de portada de las películas y series para aumentar la probabilidad de que el usuario haga click.

La importancia de estos sistemas va más allá de la mera sugerencia de contenido. Influyen directamente en la toma de decisiones estratégicas de producción y adquisición de contenido. Por ejemplo, Netflix utiliza los datos de su sistema de recomendación para informar decisiones sobre qué tipo de contenido original producir. El éxito de series como "House of Cards" se atribuye en parte a la confianza que Netflix depositó en su algoritmo, que predijo una alta probabilidad de éxito basándose en los patrones de visualización de sus usuarios. La eficacia de estos sistemas también se refleja en métricas clave de negocio. Un estudio de McKinsey encontró que los servicios de streaming con sistemas de recomendación avanzados experimentan tasas de cancelación un 20% más bajas que aquellos con sistemas menos sofisticados (McKinsey & Company, 2021). En este contexto, contar con una herramienta que permita clasificar de manera precisa los géneros de las películas, basándose en sus descripciones, no solo enriquecería las recomendaciones ofrecidas, sino que también optimizaría la experiencia del usuario al personalizar el contenido de acuerdo con sus preferencias.

Business goal

- Identificar el género al que pertenece una película, basándose en la descripción de la trama de la misma.

Data mining goal

- Encontrar e implementar el algoritmo que permita predecir el género de la película basado en la descripción de la trama.

Data mining success criteria

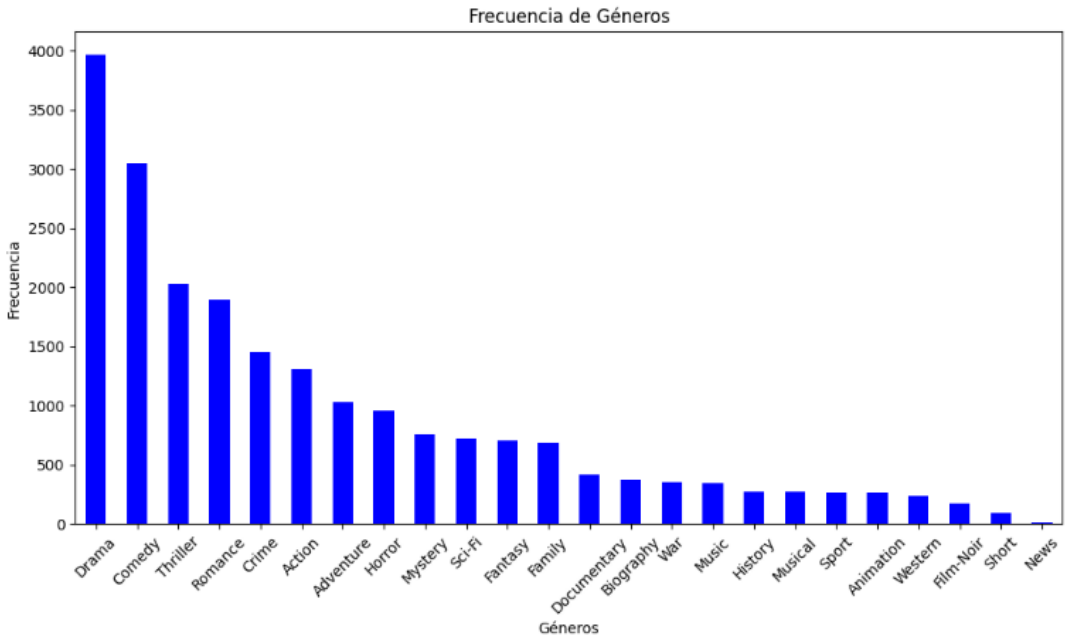
- KPI I: El AUC (Área Bajo la Curva) del modelo implementado debe superar el 0,89.

ENTENDIMIENTO DE LOS DATOS

La base de datos de entrenamiento proporcionada contiene información sobre un total de 7.895 películas y consta de 5 variables: Una de estas variables indica el al que pertenece cada película, lo que la establece como la variable objetivo del presente análisis; dos variables numéricas: el año de lanzamiento y la calificación o rating obtenido; y por último otras dos variables de tipo objeto: el título de la película y una descripción de la misma, siendo esta última la variable más esencial para realizar el entrenamiento del modelo.

Dicho esto, y teniendo en cuenta que hay películas que pueden tener más de una etiqueta de género, en el gráfico 1 se puede evidenciar el registro de 24 géneros distintos, dentro de los cuales el género de drama es el que mayor presencia tiene con una participación en 3.965 películas, seguido por comedia y thriller con 3.046 y 2.024 registros, respectivamente. Por otro lado, los géneros de Film-Noir, short y news presentan la menor representación, con solo 168, 92 y 7 registros, respectivamente.

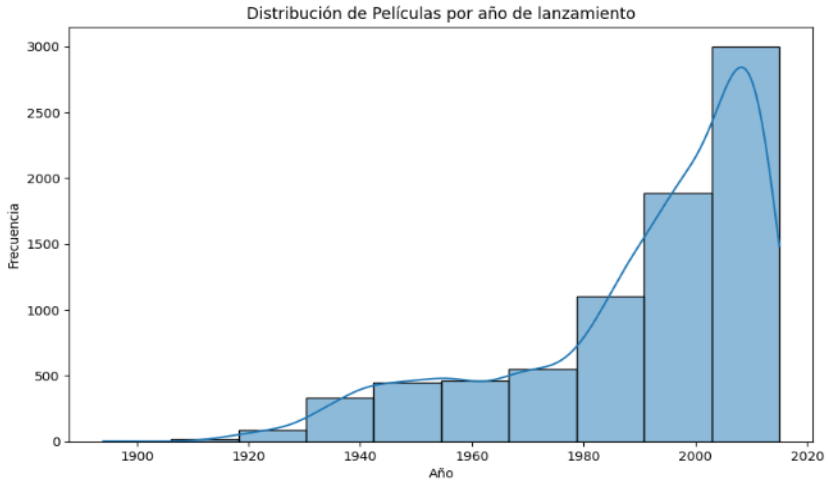
Gráfico 1. Distribución de películas por clasificación de géneros



Fuente: Elaboración propia

El análisis de la distribución de películas según su año de lanzamiento revela que el conjunto de datos abarca desde 1984 hasta 2015. El análisis de las medidas de posición indica que la mitad de las películas se estrenaron antes de 1997, y que el 75% de ellas se lanzaron antes de 2007. Esto sugiere que, hay una representación considerable de títulos de décadas anteriores. Este aspecto temporal es un factor relevante, ya que el lenguaje utilizado en las descripciones de películas puede haber evolucionado con el tiempo. En consecuencia, las descripciones más antiguas pueden emplear estilos de escritura y vocabulario diferentes a los de las películas más recientes.

Gráfico 2. Distribución de películas por año de lanzamiento

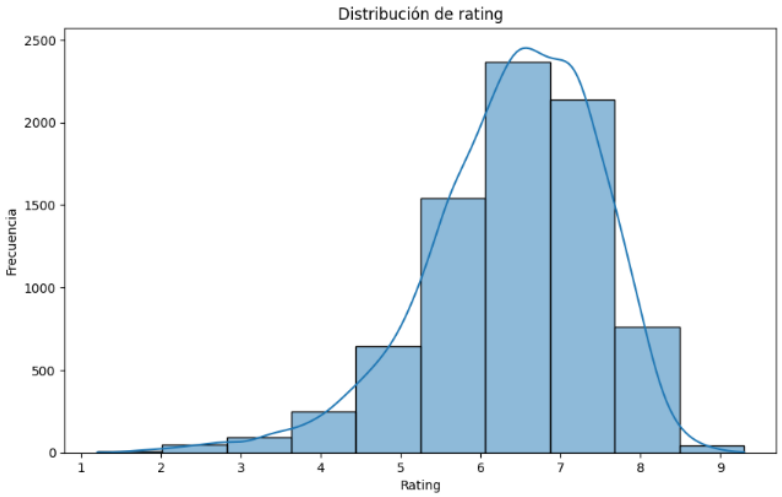


Count	Mean	Std	Min	25%	50%	75%	Max
7895	1989	22	1984	1980	1997	2007	2015

Fuente: Elaboración propia

Finalmente, y en cuanto al rating de las películas, se identificó un rango significativo de calificaciones, desde 1.2 hasta 9.3. No obstante, la media de 6.40 en ratings indica que, en general, las películas tienden a recibir calificaciones medianamente positivas. Además, la desviación estándar del rating (1.08) indica que las calificaciones están relativamente concentradas alrededor de la media, lo que significa que hay consistencia en las calificaciones.

Gráfico 3. Distribución de películas por rating



Count	Mean	Std	Min	25%	50%	75%	Max
7895	6,4	1,07	1,2	5,8	6,5	7,2	9,3

Fuente: Elaboración propia

PREPARACIÓN DE LOS DATOS

Pre-procesamiento del texto:

Inicialmente, sobre cada 'plot' se ejecuta una función de pre-procesamiento que consiste en aplicar lowercase (dejar todas las palabras en minúsculas), quitar las puntuaciones y realizar stemming (para reducir las palabras a su raíz a partir de las reglas de corte definidas en esta técnica).

Método de vectorización:

Una vez realizada la depuración del texto explicada previamente, se realizó la vectorización de cada plot, haciendo pruebas con CountVectorizer, TFIDF y embedings.

El método que generó el mejor resultado de 0.9008 de AUC fue TFIDF, el cual consiste en convertir cada documento (que en nuestro caso es el plot de la película) a una matriz de características que pondera la frecuencia de palabras en función de su importancia: La relevancia de un término en un documento combina su frecuencia en el documento (TF) con qué tan raro es en todo el corpus (IDF), dándole más importancia a términos frecuentes en un documento, pero poco comunes en el resto del conjunto.

Dentro de la función TFIDF se quitaron stopwords y se utilizaron n-gramas 1 y 2, es decir, se tuvieron en cuenta secuencias de una palabra (unigramas) y de dos palabras (bigramas). Adicionalmente, se aplicó una transformación sublineal a la frecuencia del término (sublinear_tf=True). Esto significa que, en lugar de usar directamente la frecuencia de aparición de las palabras en los documentos, se toma el logaritmo de la frecuencia, lo que permite mejorar los resultados dado que puede pasar que una palabra que aparece muy frecuentemente no necesariamente es más importante en el documento (por ejemplo, una palabra que aparece 100 veces no es necesariamente diez veces más relevante que una que aparece 10 veces). Con la transformación sublineal, se suaviza este impacto, permitiendo que palabras que aparezcan más veces sigan siendo importantes, pero con un peso más moderado.

También, se agregó el parámetro que establece el número mínimo de documentos en los que debe aparecer un término para ser incluido en la matriz TF-IDF. Así, min_df=2 ignora términos que aparecen en menos de dos documentos. De esta manera se puede filtrar ruido quitando palabras que no aporten al entrenamiento.

MODELAMIENTO

Para la construcción del modelo, se realizaron diferentes iteraciones en la forma de la preparación de los datos sobre SVM, árboles de decisión, XGBoost y redes neuronales para obtener un AUC mínimo de 0,89. Sin embargo, fue el modelo de regresión logística junto con la preparación explicada en el apartado anterior con la que se superó el AUC propuesto.

Regresión Logística

Para el entrenamiento del modelo, se realizó la división de la data, con un 80% de datos en entrenamiento y un 20% en test, posteriormente se utilizó la regresión logística para resolver un problema de clasificación multietiqueta (multi-label classification), donde una película puede ser clasificada en múltiples géneros.

La regresión se definió utilizando la clase OneVsRestClassifier, la cual divide el problema en múltiples problemas de clasificación binaria, uno por cada género. Cada uno de estos modelos, predice si una película pertenece o no a un género específico.

```
clf = OneVsRestClassifier(LogisticRegression(solver='sag', max_iter=15000, random_state=41))
```

Esta clase se utiliza en conjunto con LogisticRegression, el cual es un modelo lineal que predice la probabilidad de que un género pertenezca a una clase, ya que utiliza la función sigmoide para dar su resultado en forma de probabilidad. El solver=’sag’ es el optimizador que se utiliza para trabajar con grandes conjuntos de datos y características dispersas como es el caso de TF-IDF, el max_iter=15000 es un número alto de iteraciones que se define para asegurarse que el modelo converge.

Posterior a esto, se entra el modelo con clf.fit(X_train, y_train_genres) y se utiliza y_pred_genres = clf.predict_proba(X_test) para predecir las probabilidades de los géneros en el conjunto de test. Por último, se evalúa el AUC, el cual se encarga de medir que tan bien el modelo separa las clases. Con la línea de código roc_auc_score(y_test_genres, y_pred_genres, average='macro'), el average='macro' calcula el AUC de forma independiente para cada género y luego toma el promedio.

EVALUACIÓN

Se selecciona como mejor modelo la regresión logística, ya que al realizar la prueba sobre el AUC se obtiene un valor de 0.9008.

Esto nos permite concluir que, en los problemas de procesamiento de lenguaje natural, la rigurosidad de los métodos utilizados para el procesamiento y vectorización del texto pueden llegar a tener más peso sobre las métricas de rendimiento que la misma complejidad del modelo.

Modelo	Mejor AUC logrado
Regresión logística	0,9008
Random Forest	0.8550
XGBoost	0.8141
Red neuronal	0,7618

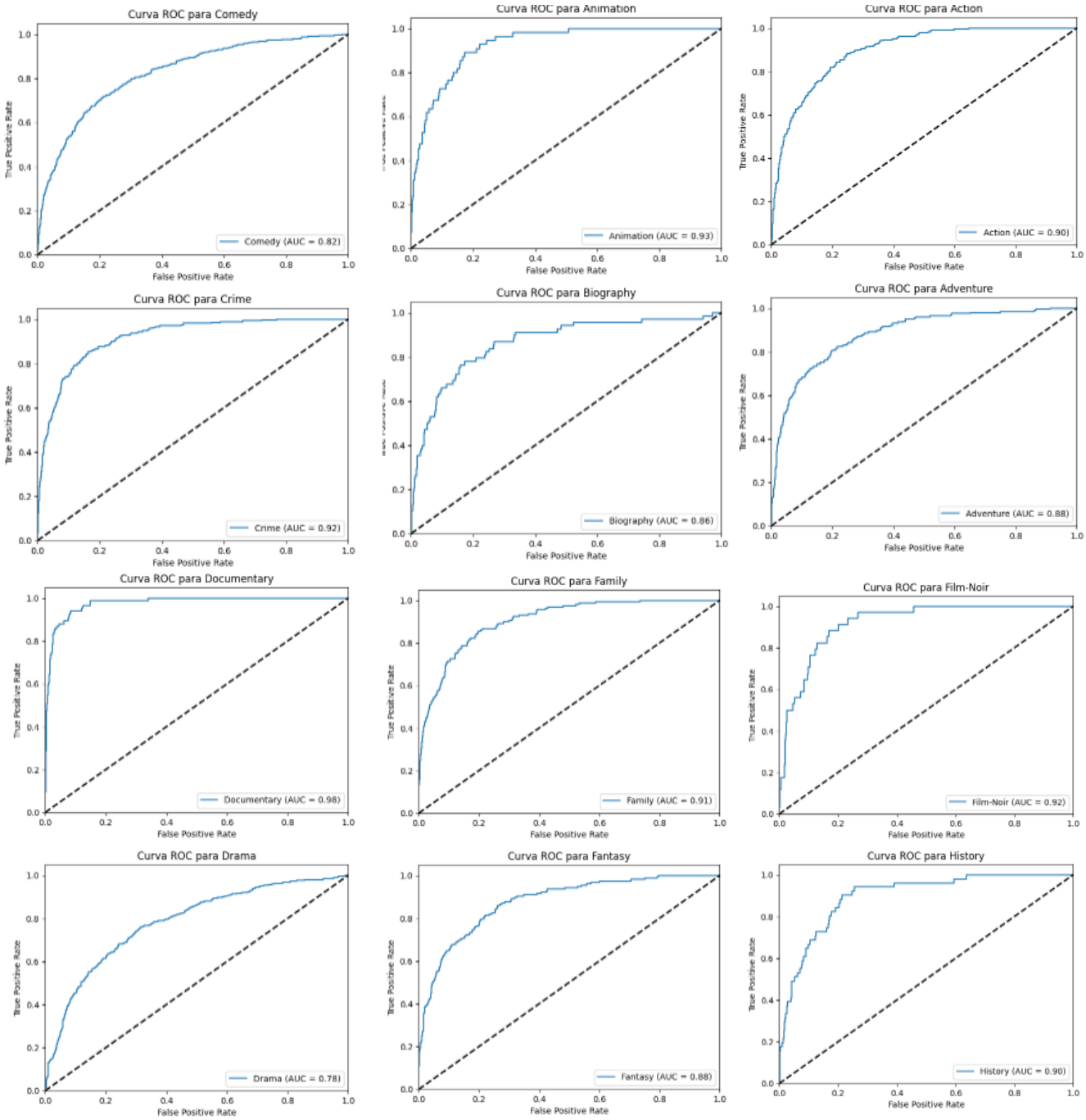
ANÁLISIS DE RESULTADOS POR DISTINCIÓN DE GÉNERO

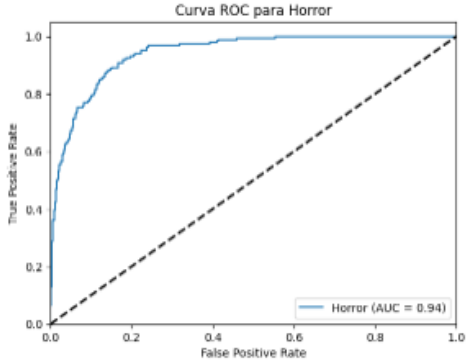
Además de la complejidad de las técnicas de procesamiento empleadas, se identificaron características intrínsecas de la base de datos que influyen en la variabilidad del rendimiento del modelo; esto se identificó mediante el análisis del AUC por género, el cual reveló que los géneros con mayor número de registros tienden a obtener un AUC más bajo. Tal y como se puede evidenciar en el gráfico 4, el género Drama, que es la clase mayoritaria en la data de entrenamiento con 3.965 registros, obtuvo el AUC más bajo (0.78). En contraste, los géneros con menor representación en la data, son los que mayor AUC individual obtienen, por ejemplo, el género de sport que tiene 261 registros obtuvo el AUC más alto de 0.99.

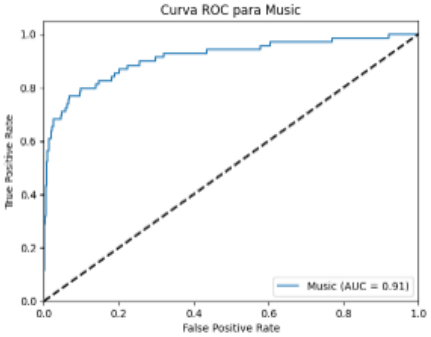
Esto sugiere que la heterogeneidad y el volumen de datos dentro de las clases más grandes pueden dificultar la capacidad del modelo para generalizar correctamente, mientras que las clases más pequeñas facilitan una mejor identificación de patrones; el modelo puede distinguir mejor las clases con menos datos ya que es más fácil capturar patrones distintivos en conjuntos de datos pequeños. En cambio, las clases más grandes tienden a presentar mayor diversidad en las descripciones, estilos narrativos y lenguaje, lo que incrementa la complejidad. Además, el hecho de tratarse de un problema

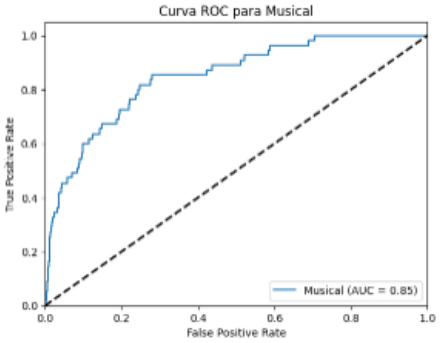
multicategoría incrementa la dificultad en la clasificación, ya que géneros como "Drama" pueden incluir desde películas románticas hasta thrillers emocionales, lo que da lugar a descripciones muy variadas y hace más difícil identificar patrones consistentes.

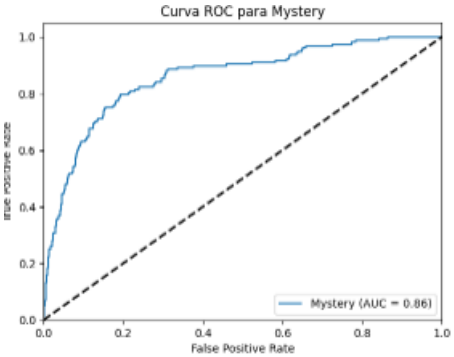
Gráfico 4. Análisis comparativo de AUC por distinción de género

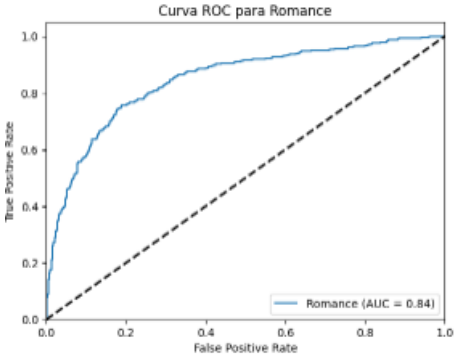


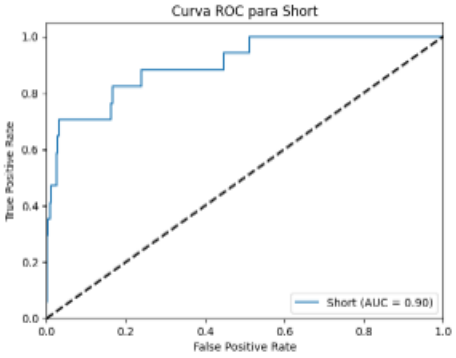

Curva ROC para Horror
True Positive Rate
False Positive Rate
Horror (AUC = 0.94)

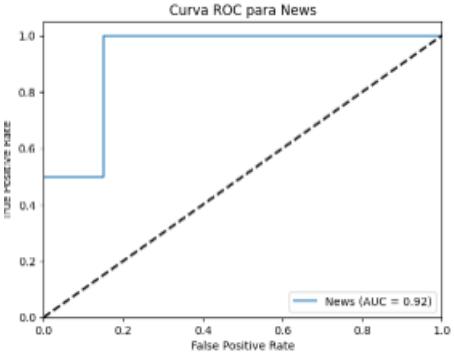

Curva ROC para Music
True Positive Rate
False Positive Rate
Music (AUC = 0.91)

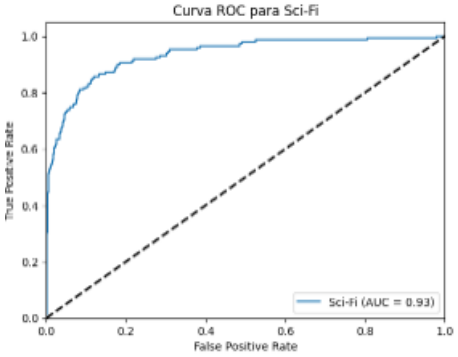

Curva ROC para Musical
True Positive Rate
False Positive Rate
Musical (AUC = 0.85)

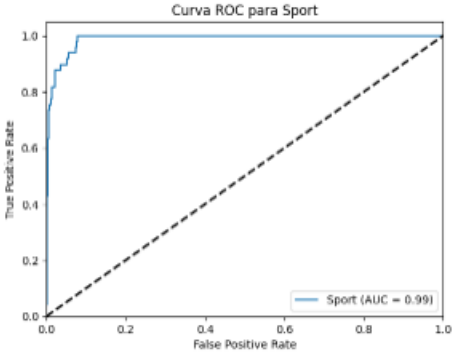

Curva ROC para Mystery
True Positive Rate
False Positive Rate
Mystery (AUC = 0.86)

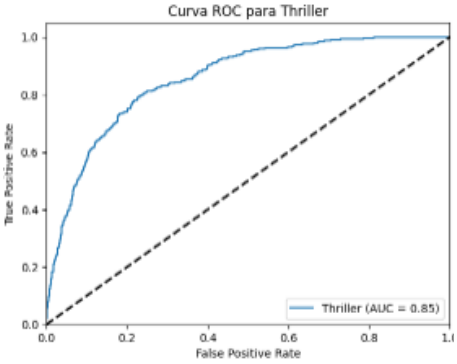

Curva ROC para Romance
True Positive Rate
False Positive Rate
Romance (AUC = 0.84)

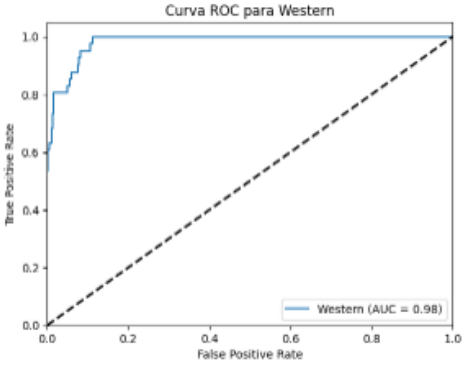

Curva ROC para Short
True Positive Rate
False Positive Rate
Short (AUC = 0.90)

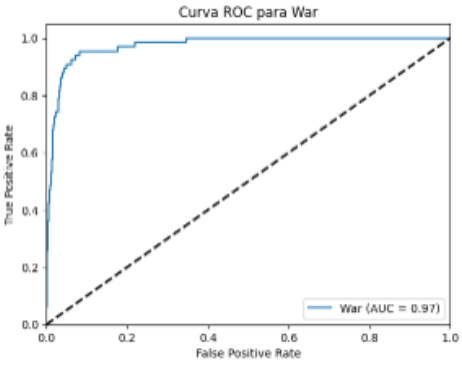

Curva ROC para News
True Positive Rate
False Positive Rate
News (AUC = 0.92)


Curva ROC para Sci-Fi
True Positive Rate
False Positive Rate
Sci-Fi (AUC = 0.93)


Curva ROC para Sport
True Positive Rate
False Positive Rate
Sport (AUC = 0.99)


Curva ROC para Thriller
True Positive Rate
False Positive Rate
Thriller (AUC = 0.85)


Curva ROC para Western
True Positive Rate
False Positive Rate
Western (AUC = 0.98)


Curva ROC para War
True Positive Rate
False Positive Rate
War (AUC = 0.97)

BIBLIOGRAFÍA

1. Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems, 6(4), 1-19.
2. Hosanagar, K., Fleder, D., Lee, D., & Buja, A. (2014). Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation. Management Science, 60(4), 805-823.
3. McKinsey & Company. (2021). The future of personalization—and how to get ready for it. McKinsey Digital.
4. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
5. Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.