**Proyecto II – Predicción de géneros de películas**

Daniel Felipe Rodríguez Cuellar a,c, Diana Katherin Ibáñez Contreras a,c, Erika Andrea Cárdenas Rodríguez a,c, Paola Andrea Castro Cardozo a,c y Oscar Eduardo Correcha Guzmán b,c

Sergio Alberto Mora Pardo b,c

aEstudiante de Maestría en Analítica para la Inteligencia de Negocios

bEstudiante de Maestría en Inteligencia Artificial

bProfesor, Facultad de Ingeniería

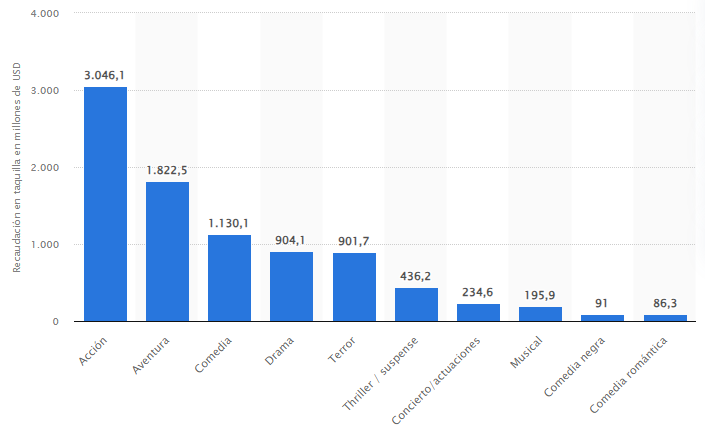
cPontificia Universidad Javeriana, Bogotá, Colombia

# Entendimiento del Negocio

Existen aproximadamente entre 15 a 20 géneros cinematográficos ampliamente reconocidos[[1]](#footnote-2), como comedia, drama, acción, ciencia ficción, romance, terror, aventura y documentales, entre otros. Cada género tiene características distintivas que influyen tanto en el desarrollo de la trama como en la reacción del público. Por ejemplo, las películas de acción suelen incluir tramas rápidas, escenas de persecución, conflictos y desafíos físicos, mientras que los dramas se centran en conflictos emocionales profundos, desarrollo de personajes y situaciones de la vida cotidiana. Las comedias buscan el entretenimiento a través del humor, mientras que las películas de ciencia ficción exploran futuros posibles, avances tecnológicos y dilemas filosóficos.

En el Diagrama 1 se puede observar que las películas que más recaudaron dinero en el 2023 en USA y Canadá fueron las del género de acción y aventura, obteniendo 3.046M de USD y 1.822M USD respectivamente.

Diagrama 1.Recaudación en taquilla de los diez principales géneros cinematográficos en Estados Unidos y Canadá en 2023



*Fuente.* (*Géneros cinematográficos según taquilla en Norteamérica en 2023*, s. f.)

El contexto del proyecto se enfoca en analizar y clasificar reseñas de diferentes géneros de películas. Esto es particularmente útil para organizar catálogos de películas en plataformas de streaming o bibliotecas, ayudándo a los usuarios a buscar según sus intereses, como insumo en algoritmos de recomendación para dar opciones personalizadas a los usuarios, para realizar labores de marketing y segmentación de usuarios por tipo de contenido, realizar análisis de tendencias en la industria del cine y realizar análisis de películas por categoría para los premios de cine.

Netflix fue una de las empresas que empezó a desafiar la industria en 1997, a través de las recomendaciones personalizadas, con un algoritmo llamado **Cinematch** que busca conocer las preferencias de los clientes a través de las calificaciones, interacciones y búsquedas. Netflix puso a prueba a los programadores y lanzó un concurso externo en el cual el equipo que mejorara la precisión del algoritmo se ganaría 1 millón de dólares, premio que ganó [BellKor’s Pragmatic Chaos,](https://gibsonbiddle.medium.com/a-brief-history-of-netflix-personalization-1f2debf010a1)un equipo de siete miembros que incluía a dos investigadores de AT&T, quienes lograron mejorar el modelo en un 10%. (Romo, 2023)

El análisis de las tramas permitirá identificar el género de las películas mediante la aplicación de técnicas de procesamiento de lenguaje natural (NLP). Estas técnicas facilitarán la extracción de características clave de las descripciones narrativas, lo que contribuirá a una clasificación precisa de las películas en sus géneros correspondientes.

## Objetivos de Negocio

1. Determinar la probabilidad de que una película pertenezca a un género específico, utilizando como base el análisis de su trama escrita, para optimizar el sistema de recomendación basado en la clasificación por géneros.

### Criterio de éxito del objetivo de negocio

Incrementar en un 10% las visualizaciones por parte de los usuarios, de recomendaciones de películas derivadas de la clasificación por géneros.

## Objetivos de Minería de Datos

1. Realizar el preprocesamiento del texto del plot, aplicando técnicas de procesamiento de lenguaje natural (NLP), para transformar las reseñas en datos estructurados que puedan ser utilizados en los modelos predictivos.
2. Desarrollar un modelo de clasificación para las películas según los géneros a los que corresponden, evaluando diferentes técnicas de machine learning, tales como Regresión Logística, Random Forest, XGBoost, LSTM.

### Criterios de éxito de los objetivos de minería de datos

Seleccionar el modelo que mejor rendimiento demuestre para distinguir los géneros de las películas del dataset de validación, según la métrica de AUC; por lo cual, se escogerá aquel en el que los AUCs tanto en entrenamiento como en prueba sean cercanos a uno y no distantes entre sí. El AUC mínimo objetivo es del 0,89%.

# Entendimiento de los Datos

El conjunto de datos incluye tramas de películas, es decir, las descripciones narrativas que resumen el argumento de cada película. Se dispone de dos conjuntos de datos en las siguientes urls de GitHub:

<https://github.com/sergiomora03/AdvancedTopicsAnalytics/raw/main/datasets/dataTraining.zip> (Datos de entrenamiento).

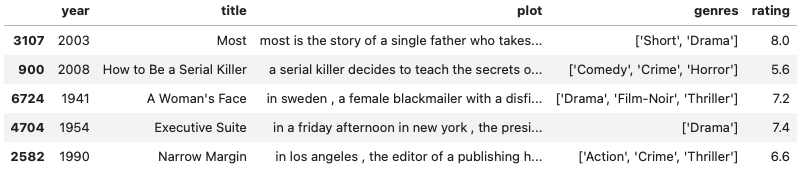
<https://github.com/sergiomora03/AdvancedTopicsAnalytics/raw/main/datasets/dataTesting.zip> (Datos de validación).

Los dataset contienen las siguientes variables:

* **year:** Año en que se lanzó la película.
* **title:** Título de la película.
* **plot:** Trama de la película.
* **genres:** Los géneros asociados con la película. NO está presente en ‘dataTesting’.
* **rating:** La calificación de la película. NO está presente en ‘dataTesting’.

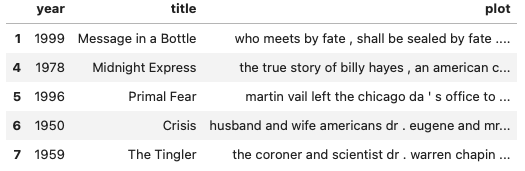
El archivo de ‘dataTraining’ (primeros registros, ver Tabla 1.) contiene 7.895 registros con 5 columnas (year, title, plot, genres, rating).

Tabla 1. Primeros registros dataTraining



El archivo de ‘dataTesting’ (primeros resgistros, ver Tabla 2.) contiene 3.383 registros con 3 columnas (year, title, plot).

Tabla 2. Primeros registros dataTesting



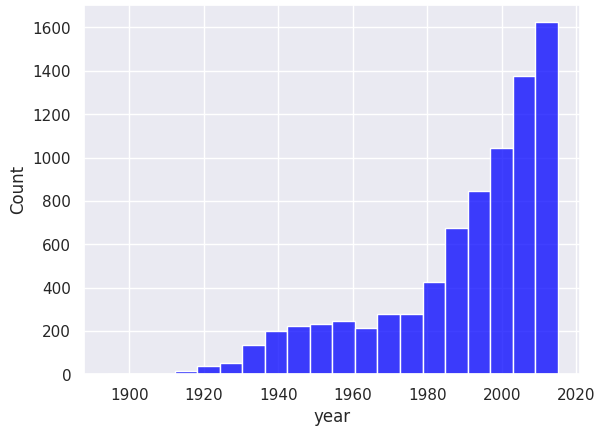
# Exploración de los Datos

A continuación, se presenta el análisis univariado y multivariado realizado en ‘dataTraining’:

## 3.1. Análisis univariado

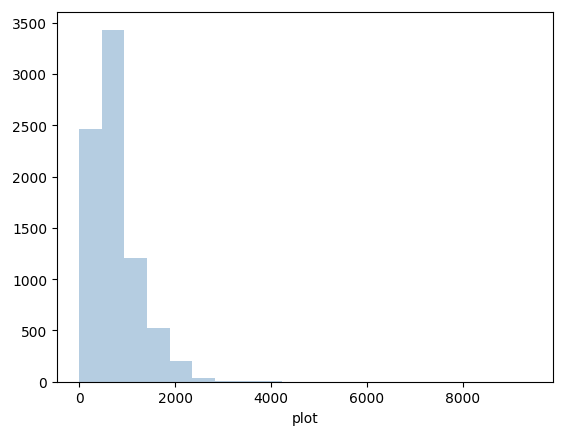
En el Diagrama 3, se evidencia para la variable ‘year’ una distribución asimétrica a la derecha o sesgo positivo, lo que significa que se filmaron más películas en los últimos 30 años.

Diagrama 2. Distribución películas por año.



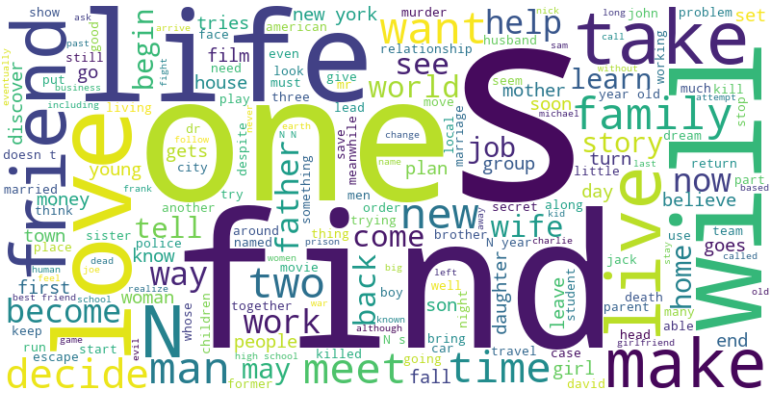
Para la variable ‘plot’, por medio del Diagrama 4 se ha representado la distribución de la cantidad de palabras. Se puede concluir que en su mayoría tienen menos de1000 palabras y que ninguno supera las 3000 palabras.

Diagrama 3. Histograma cantidad de palabras



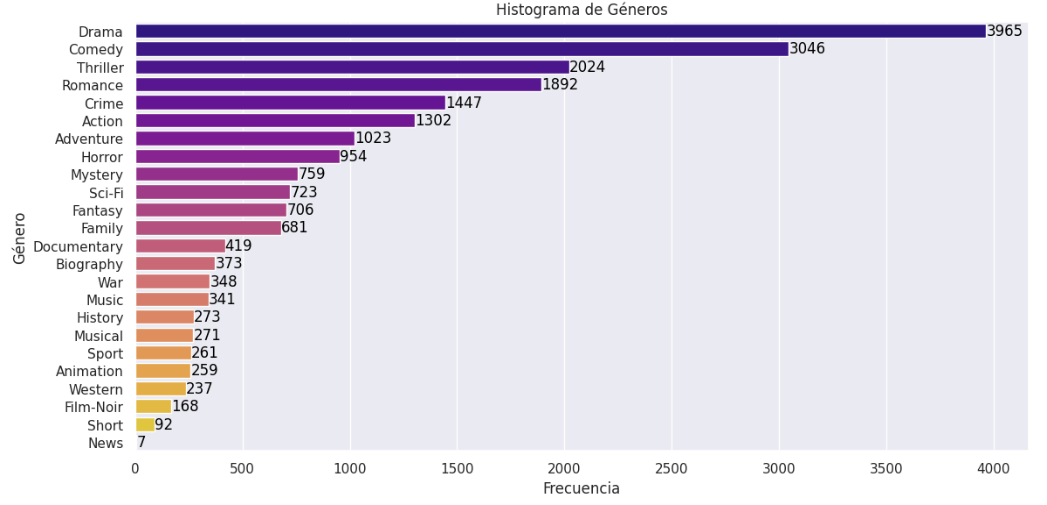
Por medio del word cloud del Diagrama 5, se identifican las palabras más comunes en los plot del dataset: ‘find’, ‘one’, ‘life’, ‘love’, ‘friend’, ‘will’, ‘take’, ‘make’ y ‘live’.

Diagrama 4. Nube de palabras



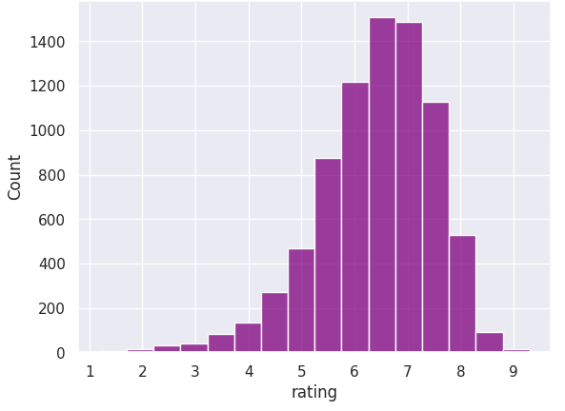
En el Diagrama 6, la variable ‘genres’ muestra que los géneros ‘Drama’ y ‘Comedy’, están significativamente más representados en comparación con otros géneros, lo que permite identificar que son las clases dominantes de nuestro dataset. Los menos representados son ‘Short’ y ‘News’. El número de combinaciones de géneros es de 1336.

Diagrama 5. Histograma de géneros



El Diagrama 7 para la variable ‘rating’, muestra que la mayoría de las películas ha recibido una evaluación entre 6 y 8.

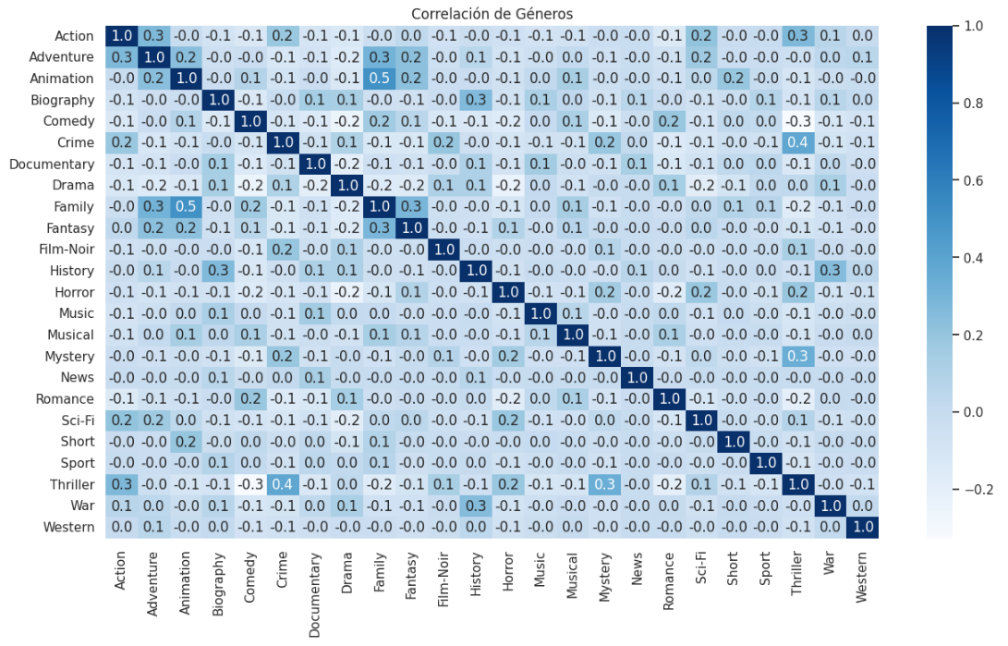
Diagrama 6. Histograma de rating



## 3.2. Análisis multivariado

Después de convertir la lista de etiquetas con el género de la película en una representación binaria, se generó la matriz de correlación de Pearson que se presenta en el Diagrama 8. Hay una mayor correlación positiva entre 'Family' y 'Animation' con 0.5. y una mayor correlación negativa entre 'Triller' y 'Comedy' con -0.3.

Diagrama 7. Correlación de géneros



# Preparación de los datos

Como parte de la preparación de los datos, se realizaron las actividades de limpieza de datos y pre-procesamento básico del texto, que se describen en seguida:

## 4.1. Limpieza de los datos

* Verificar si hay valores nulos o faltantes en los dataset. Sin embargo, no se identifica ninguno, ni en ‘dataTraining’ ni en ‘dataTesting’.
* Analizar si hay valores duplicados para evitar que información repetida sesgue los modelos a desarrollar. Como resultado se encuentra una película duplicada en ‘dataTraining’. En ‘dataTesting’ no se encuentran registros duplicados.

## 4.2. Pre-procesamiento básico de los datos

* Extraer la lista de etiquetas con el género de la película y transformarla en una representación binaria, en donde se evidencia que una película puede estar asociada a más de un género. De esta forma se convierte cada género en una columna y a cada película le asigna un 1 o un 0, dependiendo de si pertenece o no a ese género. Esto ayuda a que el modelo pueda predecir más de una clase para cada película.
* Convertir todo el texto del ‘plot’ a minúsculas, para evitar duplicidad por diferencias de mayúsculas y minúsculas.
* Eliminar la puntuación, los caracteres especiales y los posibles emojis en el texto, para reemplazarlos por un espacio. Solo se mantienen letras y números.
* Tokenizar, separando el texto en palabras individuales para así crear una lista de palabras. Con esto, se eliminan también los espacios.
* Remover las palabras que se consideran con poco significado (stop words).

# Pre-procesamiento de los datos por modelo, Modelado y Evaluación

Para el modelado y evaluación se realiza el entrenamiento con diferentes modelos de clasificación multiclase, ya que en nuestro caso tenemos múltiples géneros por película. Para la aplicación de los distintos modelos se realizaron tres tratamientos de datos distintos para evaluar su rendimiento. En seguida, se describen los intentos realizados:

## 5.1. Primer experimento: Lematización, TF-IDF Vectorizer y combinar ‘title’

Para este experimento, se incluyó al pre-procesamiento básico de los datos, lo siguiente:

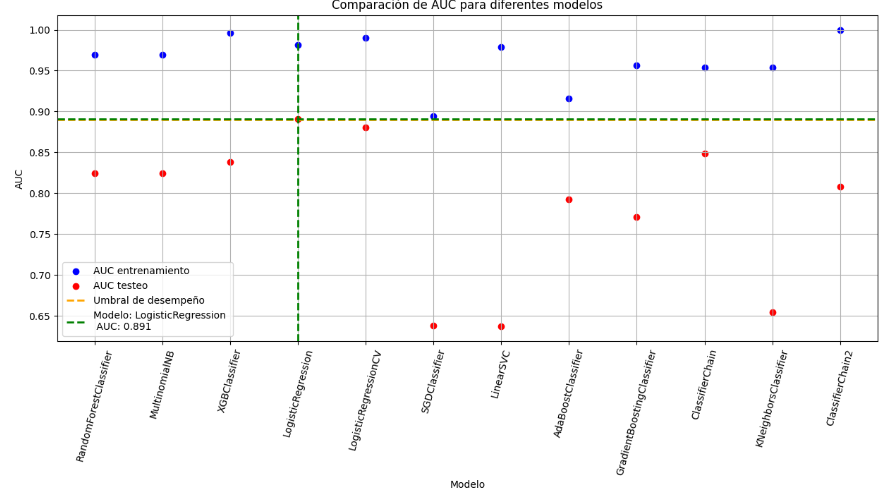
* Combinar el texto de las columnas ‘title’ y ‘plot’ en una columna para crear descripciones más completas.
* Eliminar las palabras que tienen menos de tres caracteres, ya que generalmente no aportan información significativa en el análisis del texto.
* Aplicar lematización con el lematizador de ‘WordNet', para reducir cada palabra a su forma base o raíz. Eso es útil para agrupar palabras con diferentes formas gramaticales como por ejemplo running se convierte en run.
* Realizar feature engineering, aplicando un vectorizador TF-IDF con n-gramas (1,3) para transformar el texto en una matriz de vectores numéricos, ponderando la importancia de las palabras en función de su frecuencia en el documento y en el corpus completo.

Luego, se dividió el dataset de entrenamiento en los conjuntos de entrenamiento y prueba, utilizando un 80% para entrenamiento y un 20% para pruebas.

Se desarrollaron los 12 modelos que se listan en la Tabla 3, con los resultados de la métrica por AUC tanto en entrenamiento como en prueba:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Modelo** | **AUC train** | **AUC test** |
| **0** | RandomForestClassifier | 0.969599 | 0.824126 |
| **1** | MultinomialNB | 0.969599 | 0.824126 |
| **2** | XGBClassifier | 0.996142 | 0.838152 |
| **3** | LogisticRegression | **0.981136** | **0.890810** |
| **4** | LogisticRegressionCV | 0.990007 | 0.880912 |
| **5** | SGDClassifier | 0.894456 | 0.638421 |
| **6** | LinearSVC | 0.978612 | 0.637556 |
| **7** | AdaBoostClassifier | 0.916098 | 0.792520 |
| **8** | GradientBoostingClassifier | 0.956628 | 0.770923 |
| **9** | ClassifierChain | 0.953491 | 0.848301 |
| **10** | KNeighborsClassifier | 0.954235 | 0.654497 |
| **11** | ClassifierChain2 | 1.000.000 | 0.807897 |

El modelo LogisticRegression obtuvo el mejor desempeño con un AUC de 0.981 en entrenamiento y 0.891 en prueba. Los modelos basados en cadenas de clasificadores (ClassifierChain) también mostraron resultados sólidos, especialmente ClassifierChain2 con un AUC perfecto en entrenamiento, pero un rendimiento ligeramente inferior en prueba (0.808). Sin embargo, los modelos como SGDClassifier y LinearSVC obtuvieron los peores resultados, mostrando un sobreajuste y un desempeño inferior en el conjunto de prueba. Se realiza una comparación de estos resultados mediante la gráfica del Diagrama 9:



## 5.2. Segundo experimento: Word Embedding y combinar ‘year’ y ‘rating’

Se incluyó al pre-procesamiento básico de los datos, lo siguiente:

* Realizar feature engineering, con el Universal Sentence Encoder (USE) versión 4 entrenado por Google, para convertir los textos en embeddings y así poder analizar cada palabra en su contexto.
* Combinar las variables adicionales como el año de la película y su rating.

Los datos se dividieron en conjuntos de entrenamiento y prueba, utilizando un 67% para entrenamiento y un 33% para pruebas.

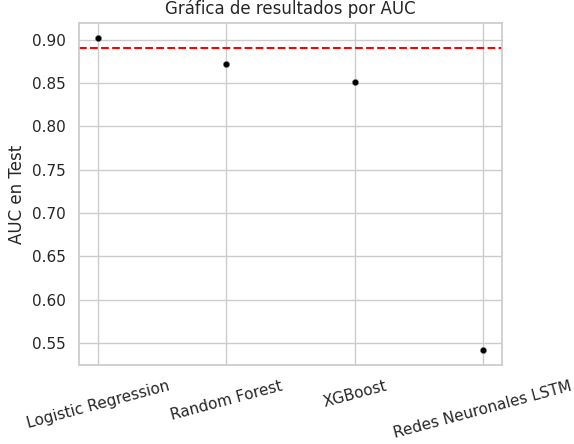
Se desarrollaron los 4 modelos que se listan en la Tabla 4, con los resultados de la métrica por AUC tanto en entrenamiento como en prueba:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Modelo** | **AUC\_Train** | **AUC\_Test** |
| **0** | Logistic Regression | **0.931710** | **0.901413** |
| **1** | Random Forest | 0.922219 | 0.872027 |
| **2** | XGBoost | 0.985749 | 0.851250 |
| **3** | Redes Neuronales LSTM | 0.546452 | 0.542434 |

El modelo de regresión logística presentó el mejor desempeño general con un AUC de 0.901 en el conjunto de prueba, seguido por el modelo Random Forest y XGBoost. Si bien el modelo de Random Forest mostró un buen ajuste en el conjunto de entrenamiento, evidenció cierto sobreajuste en los datos de prueba. Por otro lado, la red neuronal LSTM obtuvo el peor desempeño, lo que sugiere que el modelo no logró capturar las relaciones entre los datos de manera eficiente. La gráfica final muestra que el umbral de éxito fue superado únicamente por la regresión logística, destacándose como el modelo más robusto para este conjunto de datos.

En los modelos implementados, se utilizó *MultiOutputClassifier*, un clasificador para problemas de clasificación multietiqueta (multilabel), donde cada instancia (ejemplo) puede pertenecer a múltiples clases simultáneamente. Entrena un clasificador por cada etiqueta (o clase) presente en los datos. Para cada etiqueta, crea un modelo independiente y luego hace predicciones para cada uno de esos modelos. El resultado es que cada instancia puede ser clasificada en múltiples clases. Finalmente, se obtiene el AUC promedio de los AUC calculados por clase.

En el modelo Random Forest se incluyó el parámetro *class\_weight='balanced'* de scikit-learn, para manejar el desbalanceo de las clases a predecir, calculando automáticamente los pesos inversamente proporcionales a las frecuencias de las clases en los datos. Esto significa que las clases con menos instancias obtendrán un peso mayor, mientras que las clases con más instancias obtendrán un peso menor. Este ajuste ayuda a que el clasificador preste más atención a las instancias de la clase minoritaria durante el proceso de entrenamiento. Como resultado, puede mejorar la capacidad del modelo para predecir correctamente las clases menos representadas. Se realiza una comparación de estos resultados mediante la gráfica del Diagrama 10:



## 5.3. Tercer experimento: Word Embedding, solo con ‘plot’

En el último experimento, se optó por una preparación de datos solo considerando la variable de plot, sin añadir información adicional como en los anteriores intentos.

Se incluyó al pre-procesamiento básico de los datos, lo siguiente:

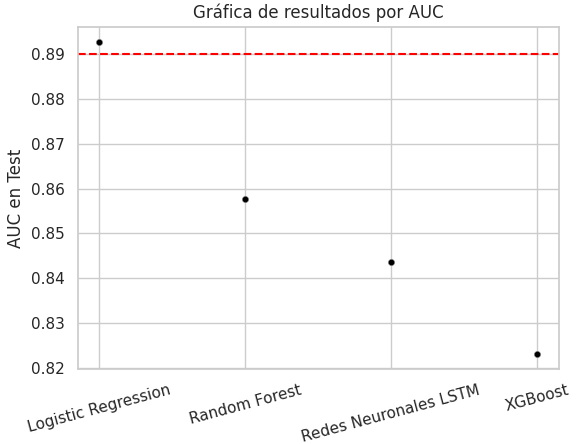
* Realizar feature engineering, con el Universal Sentence Encoder (USE) versión 4 entrenado por Google, para convertir los textos en embeddings y así poder analizar cada palabra en su contexto.

Los datos se dividieron en conjuntos de entrenamiento y prueba, utilizando un 67% para entrenamiento y un 33% para pruebas.

Se desarrollaron los 4 modelos que se listan en la Tabla 5, con los resultados de la métrica por AUC tanto en entrenamiento como en prueba:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Modelo** | **AUC\_Train** | **AUC\_Test** |
| **0** | Logistic Regression | **0.928439** | **0.892604** |
| **1** | Random Forest | 0.916614 | 0.857721 |
| **3** | Redes Neuronales LSTM | 0.878771 | 0.843580 |
| **2** | XGBoost | 0.985967 | 0.823202 |

El modelo de regresión logística resultó ser el más eficiente y confiable para este conjunto de datos, alcanzando el mejor desempeño en términos de AUC sin sobreajuste significativo. La evaluación de los resultados se muestra gráficamente con un umbral de éxito del AUC fijado en 0.89, donde solo la regresión logística superó este umbral, destacándose como la mejor opción para la clasificación multietiqueta en este problema. Se realiza una comparación de estos resultados mediante la gráfica del Diagrama 11:



# Conclusiones

**Experimento 1:** La lematización con el vectorizador TF-IDF de n-gramas (1,3), aplicadas al texto combinado de título y plot, permitió un buen rendimiento de la regresión logística (logrando el AUC mínimo propuesto de 0,89 en test) con balance entre simplicidad y efectividad.

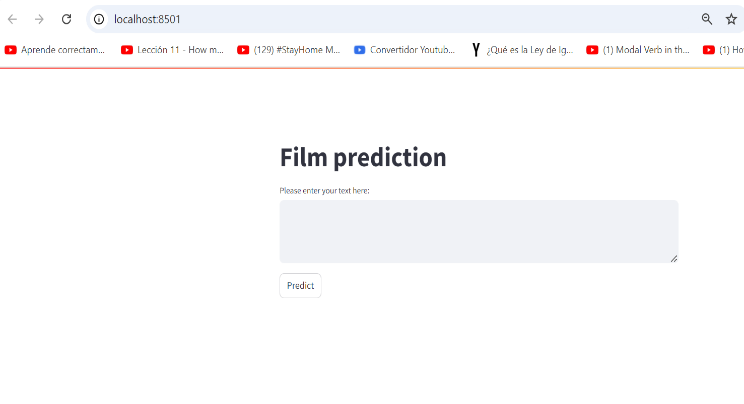
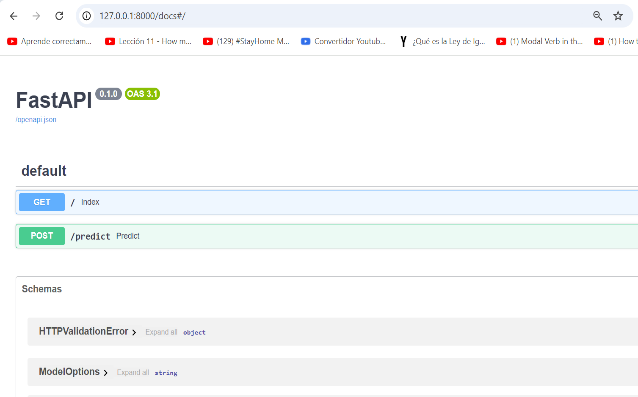
**Experimento 2:** Utilizando el USE v4 para embeddings en la variable ‘plot’ y combinando las variables ‘year’ y ‘rating’, se alcanzó el máximo AUC en test de los experimentos realizados, de 0,90 y sin sobreajuste significativo. Como en el experimento 1, el modelo más sencillo obtiene los mejores resultados para este problema.

**Experimento 3:** El embedding USE v4 aplicado a la variable ‘plot’ sin combinar otras variables, también logró el mínimo AUC propuesto de 0,89 en test con el modelo más sencillo de regresión logística.

# Anexos

Se anexan tres notebooks con los experimentos realizados para la predicción de géneros de las películas. También se anexa el archivo .csv con las predicciones del modelo de regresión logística bajo el Experimento 2 (el mejor modelo obtenido), para los datos en ‘dataTesting’.

Con el fin de mejorar la experiencia del usuario y optimizar la búsqueda de contenido, se ha incorporado una API. Esta API, construida con FastAPI y Streamlit, emplea algoritmos desarrollados en el proyecto para determinar el género de una película a partir de su sinopsis. La documentación de uso se encuentra disponible en el repositorio de GitHub.



# Bibliografía

*Géneros cinematográficos según taquilla en Norteamérica en 2023*. (s. f.). Statista. Recuperado 1 de octubre de 2024, de https://es.statista.com/estadisticas/637140/taquilla-de-las-peliculas-de-hollywood-por-genero-cinematografico-norteamerica/

Romo, unocero-J. C. P. (2023, octubre 2). *¿Cómo funcionan las recomendaciones de Netflix?* unocero. https://www.unocero.com/como-se-hace/como-funcionan-las-recomendaciones-de-netflix/

1. (*Géneros cinematográficos según taquilla en Norteamérica en 2023*, s. f.) [↑](#footnote-ref-2)