# Proyecto 1 Tópicos Avanzados de Analítica

## Predicción de géneros de películas

Diego Merlano Porto
David Oñate Acosta
Natalia Puello Acosta
Camilo Soto Zambrano
Adriana Zuica Restrepo

*Pontificia Universidad Javeriana, Bogotá, Colombia*

### 1. Adquisición de datos

Con el presente ejercicio, se pretende mediante la utilización de técnicas de procesamiento de lenguaje natural (NLP), el desarrollo de un modelo que permita predecir uno o varios géneros a los que pertenece una película, de acuerdo con el contenido de la trama. Para ello se utilizarán las bases de datos "dataTraining" y "dataTesting" disponibles en los siguientes enlaces:

https://github.com/sergiomora03/AdvancedTopicsAnalytics/raw/main/datasets/dataTesting.zip

https://github.com/sergiomora03/AdvancedTopicsAnalytics/raw/main/datasets/dataTraining.zip

Como paso previo, se hace un análisis exploratorio de la base de entrenamiento con el propósito de lograr un mayor entendimiento sobre las variables que componen el conjunto de datos y evaluar si se requiere hacer tratamientos especiales de valores nulos o incompletos.

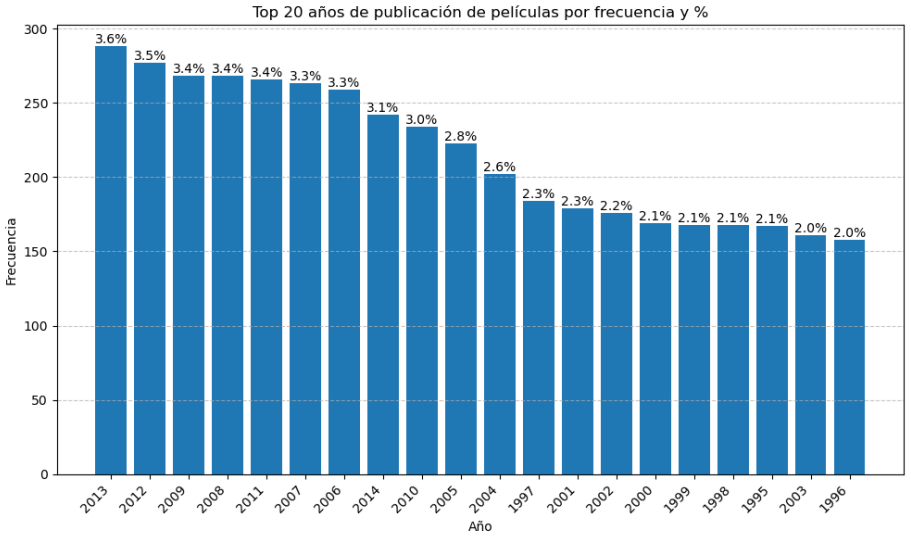
**Tabla 1. Diccionario de datos**

| Variable | Tipo | Rango | Descripción |
|---|---|---|---|
| Año | Numérica | 1894 - 2015 | Año de publicación de la película |
| Título | Texto | | Nombre de la película |
| Descripción | Texto | | Descripción de trama de la película |
| Género | Lista | | Géneros en los que clasifica |
| Puntuación | Numérica | 1.2 - 9.3 | Puntuación |

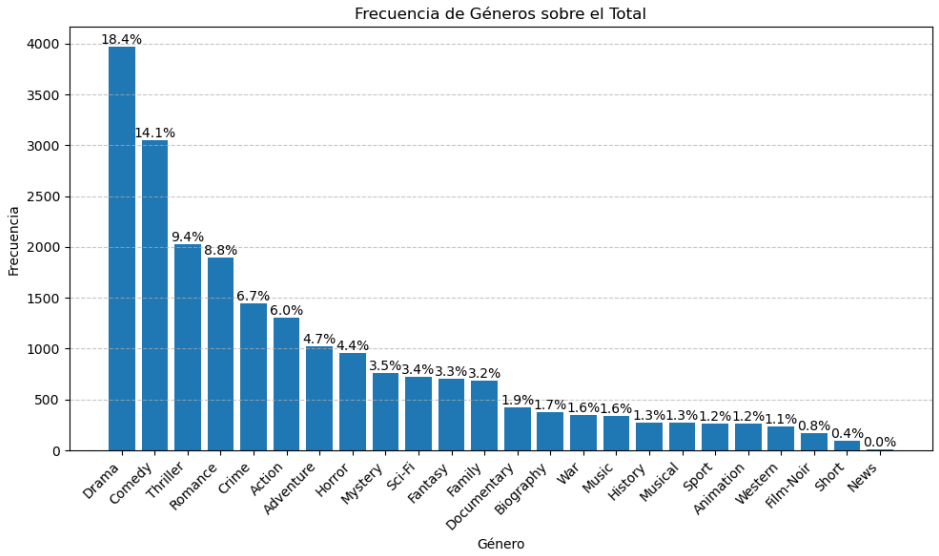
La revisión de la base de entrenamiento permite identificar que se cuenta con cinco variables (descritas en la Tabla 1) y 7859 datos sin registros nulos. Ahora bien, el conjunto de datos comprende un período de 121 años que abarcan desde 1894 hasta 2015, cuya distribución en cuanto a frecuencia y porcentaje de participación sobre el total de películas puede observarse en la gráfica siguiente. Asimismo, se representa en el gráfico posterior la distribución de los géneros de películas disponibles en la base evaluada.

**Gráfico 1. Histograma de variables descriptivas**

**Años con mayor cantidad lanzamientos**



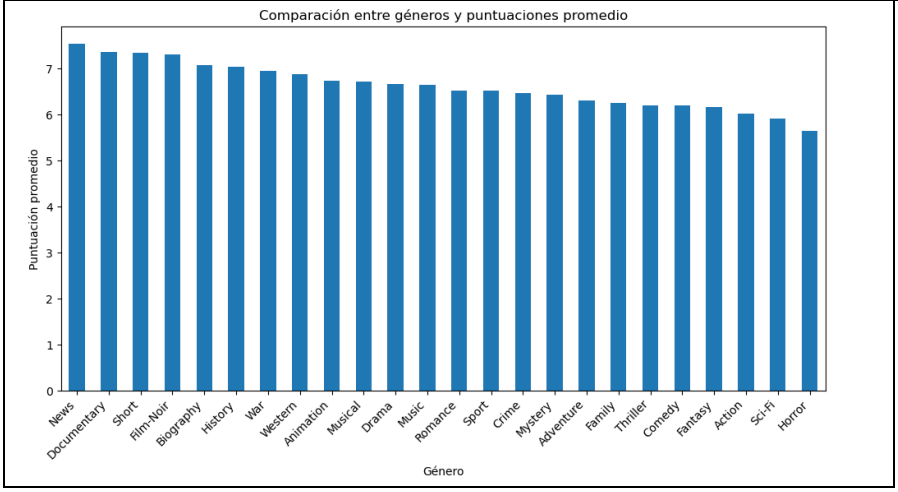

**Frecuencia de géneros de películas**




Del análisis gráfico anterior, puede observarse que no hay un año con una representación significativamente superior a otro, ya que siendo 2013 el año con mayor cantidad de películas lanzadas, su participación apenas alcanza un 3.6% del total disponible; más, sin embargo, hay una tendencia creciente en la cantidad de lanzamientos a partir del inicio del nuevo milenio, teniendo en cuenta que dentro del top 20 se encuentran todos los años desde el 2000 hasta el 2013. En lo que concierne al género de las películas de la base, se tienen 24 géneros diferentes, siendo *Drama* el género más común (repetido en 3965 películas), seguido por *Comedia* (con una frecuencia de

3046), mientras que el género con menor cantidad de repeticiones es *Noticias* apareciendo tan sólo en 7 registros.

**Gráfico 2. Puntuación promedio por género**




En el gráfico 2 se presenta una comparación de la puntuación promedio por cada género representado en la base de entrenamiento revisada, identificando a la categoría Noticias como la de mejor valoración del público, teniendo esta última la particularidad de ser la de menor frecuencia en el recuento realizado en el gráfico 1. En los siguientes lugares se encuentran los géneros de *documental* y *película corta*, mientras que las peor valoradas en promedio pertenecen a los grupos de *acción*, *ciencia ficción* y *horror*. Ahora bien, en términos generales, las puntuaciones de las películas se mueven entre 9.2 y 1.2 con una media de 6.4 puntos.

## 2. Limpieza de texto

En primera instancia, se examinó la existencia de valores nulos en todos los atributos, sin encontrar ningún caso donde hubiera que realizar algún tratamiento sobre este tipo de registros. Como paso siguiente, para mejorar el desempeño del modelo predictivo se aplicaron técnicas de limpieza de datos como la eliminación de los signos de puntuación u otros caracteres especiales, la conversión de todo el texto a minúsculas para facilitar la comparación, además de la eliminación de nombres propios y caracteres numéricos tanto ordinales como cardinales. Lo anterior permitirá hacer más eficiente la dimensionalidad a partir del vocabulario existente, hacer más fácil la interpretabilidad al tratar únicamente el texto relevante y obtener una mayor eficiencia computacional al reducir considerablemente el número de palabras a procesar en el desarrollo del modelo predictivo.

Una muestra de la importancia de aplicar la limpieza de texto puede observarse en la tabla 2, en la cual se representan los caracteres especiales con mayores repeticiones en la base analizada.

| Tabla 2. Frecuencia caracteres especiales | |
|---|---|
| **Caracter** | **Frecuencia** |
| , | 57327 |
| . | 46614 |
| ‘ | 15571 |
| - | 10466 |
| (: | 2072 |
| ): | 2059 |

### 3. Preprocesamiento de texto

Para preparar los datos antes de usarlos en modelos, se realizaron las siguientes modificaciones:

- Conversión a minúsculas
- Eliminación StopWords
- Eliminación de caracteres no alfabéticos
- Eliminación de espacios en blanco
- Eliminación de lista adicional de palabras adicionales
- Lematización (Se optó por utilizar la lematización en lugar del stemming. Esta decisión se basó en que, tras analizar diversos escenarios, se observó que la lematización de las palabras contribuía a una mejora significativa en la precisión de las predicciones.)

Transformación de Texto a Representaciones Vectoriales

Para este proyecto, se implementaron dos métodos principales de vectorización: vectorización básica y TF-IDF.

La vectorización básica transforma una colección de documentos textuales en una matriz de frecuencias de términos. Esto implica contabilizar la ocurrencia de cada palabra en los documentos y, a partir de estos datos, generar una matriz numérica que refleje dichas frecuencias.

Por otra parte, el método TF-IDF (Term Frequency-Inverse Document Frequency) asigna valores numéricos a las palabras utilizando dos componentes clave:

TF (Term Frequency): Evalúa la frecuencia con la que aparece una palabra en un documento dado. Una palabra que se repite a menudo en un documento obtendrá un valor TF más alto en ese documento.

IDF (Inverse Document Frequency): Determina la relevancia de una palabra en el conjunto total de documentos. Las palabras comunes en muchos documentos reciben un valor IDF

bajo, mientras que términos raros o únicos en documentos específicos tienen un IDF más alto.

Estas técnicas permiten una eficaz representación numérica del texto, facilitando su posterior análisis y procesamiento en tareas de modelado predictivo.

## 4. Modelo

En la fase de modelado de nuestro proyecto, evaluamos 3 modelos de aprendizaje automático para identificar aquel con el mejor desempeño en términos del Área bajo la Curva ROC (AUC). Los modelos incluyen:

- XGBoost: Parte de la familia de métodos de Gradient Boosting, este algoritmo es altamente eficiente para abordar tanto clasificación como regresión. Destaca por su rapidez y precisión, incluso con datos de gran complejidad.
- Regresión Logística: Es un algoritmo de clasificación que se emplea tanto en clasificaciones binarias como multiclase. Establece una relación entre variables independientes y una variable dependiente categórica, utilizando la función logística para modelar la probabilidad de los diferentes resultados posibles.
- Random forest con grid search para encontrar los mejores parámetros, es un algoritmo de aprendizaje automático que construye múltiples árboles de decisión durante el entrenamiento y combina sus predicciones para mejorar la precisión y evitar el sobreajuste.

Métrica de Evaluación

Área Bajo la Curva ROC (AUC): Esta métrica es fundamental en la evaluación de modelos de clasificación. Representa la relación entre la Sensibilidad (tasa de verdaderos positivos) y 1 - Especificidad (tasa de falsos positivos) a través de diferentes puntos de corte o umbrales de decisión. Un AUC cercano a 1 señala un alto grado de precisión en la clasificación, indicando que el modelo puede diferenciar eficazmente entre las clases. Por el contrario, un AUC alrededor de 0.5 implica un rendimiento que no es mejor que el de una clasificación al azar. Por tanto, el AUC proporciona una medida integral de la capacidad del modelo para separar correctamente las clases positivas de las negativas.

## 5. Evaluación

Para determinar la eficacia de los modelos descritos, adoptamos una metodología convencional de partición de datos. Comenzamos asignando el 75% de nuestro conjunto de datos al entrenamiento de los modelos, reservando el 25% restante para su validación.

A continuación, se expone las métricas de evaluación resultantes de aplicar los modelos al conjunto de entrenamiento. Estas métricas son indicativas del comportamiento de los

modelos al ser aplicados sobre los datos con los que fueron ajustados, ofreciendo una visión preliminar de su rendimiento.

| Tabla 3. AUC modelo Train | |
|---|---|
| **Modelo** | **AUC** |
| XGBoost | 99% |
| Regresión logística | 96% |
| Random Forest | 96% |

Los resultados de los modelos evaluados en el conjunto de datos de prueba, que incluye información no utilizada durante el entrenamiento. Las métricas obtenidas son fundamentales para determinar la capacidad de los modelos de generalizar y funcionar eficazmente con nuevos datos.

| Tabla 4. AUC modelo Test | |
|---|---|
| **Modelo** | **AUC** |
| XGBoost | 86% |
| Regresión logística | 89% |
| Random forest | 81% |

La información derivada de estas evaluaciones es crucial para comprender la efectividad y la fiabilidad de los modelos que hemos propuesto. Utilizando estos insights, procederemos a tomar decisiones informadas y a implementar los ajustes necesarios con el fin de alcanzar las metas establecidas en nuestro estudio analítico. Aunque todos los modelos han mostrado una capacidad por encima del punto de corte del AUC, la Regresión Logística se distingue por su excepcional desempeño.

## 6. Conclusiones

Durante nuestra investigación en el campo del procesamiento de lenguaje natural, hemos aplicado técnicas avanzadas de aprendizaje automático utilizando Python: Regresión Logística, XGBoost y Random Forest. A través de un riguroso proceso de pruebas y análisis, hemos descubierto que la Regresión Logística sobresale notablemente, demostrando ser el modelo más efectivo para la clasificación de géneros cinematográficos. Este modelo alcanzó un destacado valor de AUC del 89%, evidenciando su superioridad en el desempeño.