PREDICCIÓN DE GENEROS PARA PELÍCULAS

ASOCIACIÓN
BORCELLE

Adquisición de datos y Análisis descriptivo

Adquisición de datos:

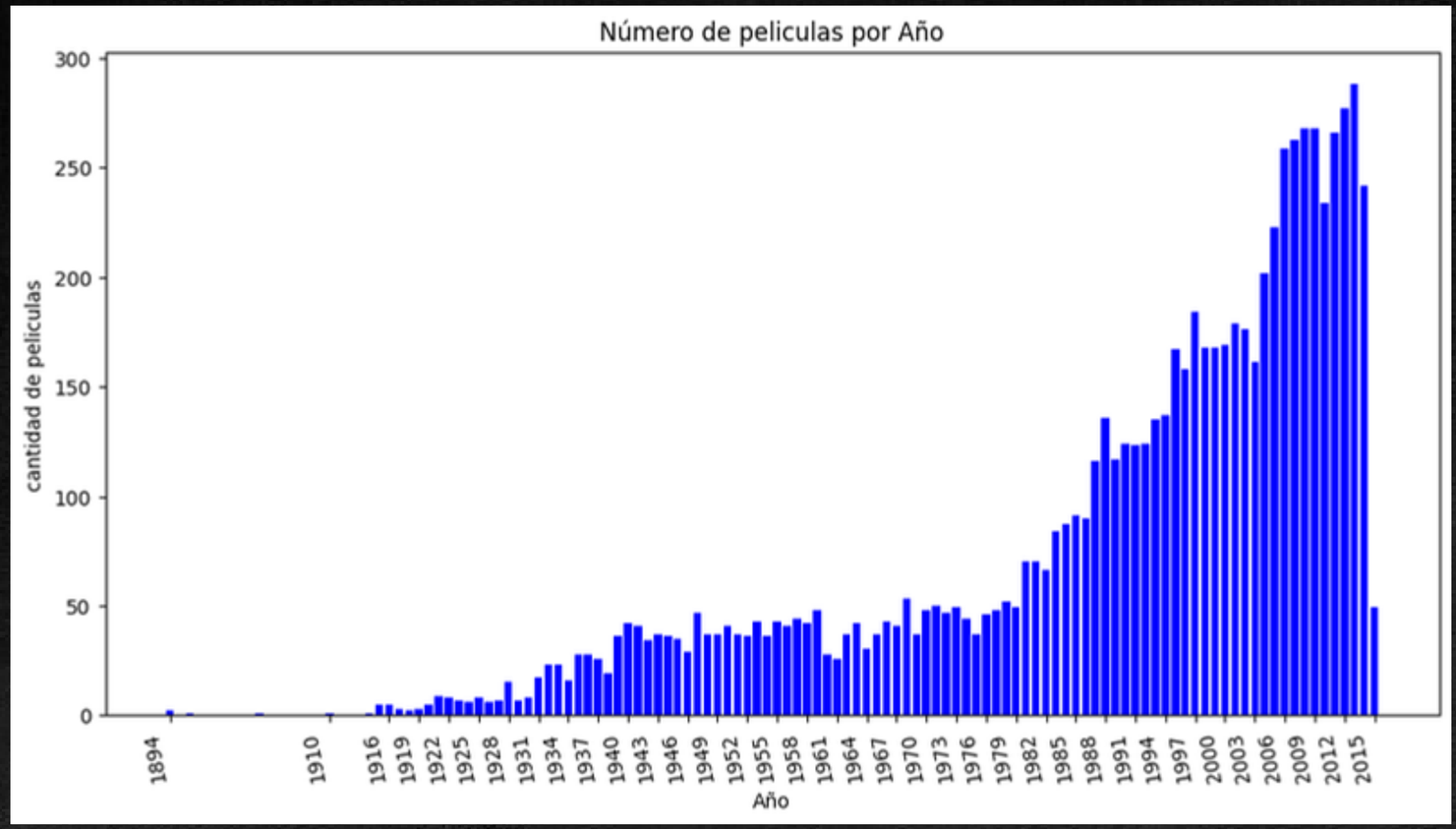
- Repositorio de GitHub
- Biblioteca de Pandas

Base de Datos

- Campos : Year, Title, Plot, Genres & Rating
- Periodo de estudio: 1894 al 2015.
- Palabras más repetidas por genero: Find, One, Life.

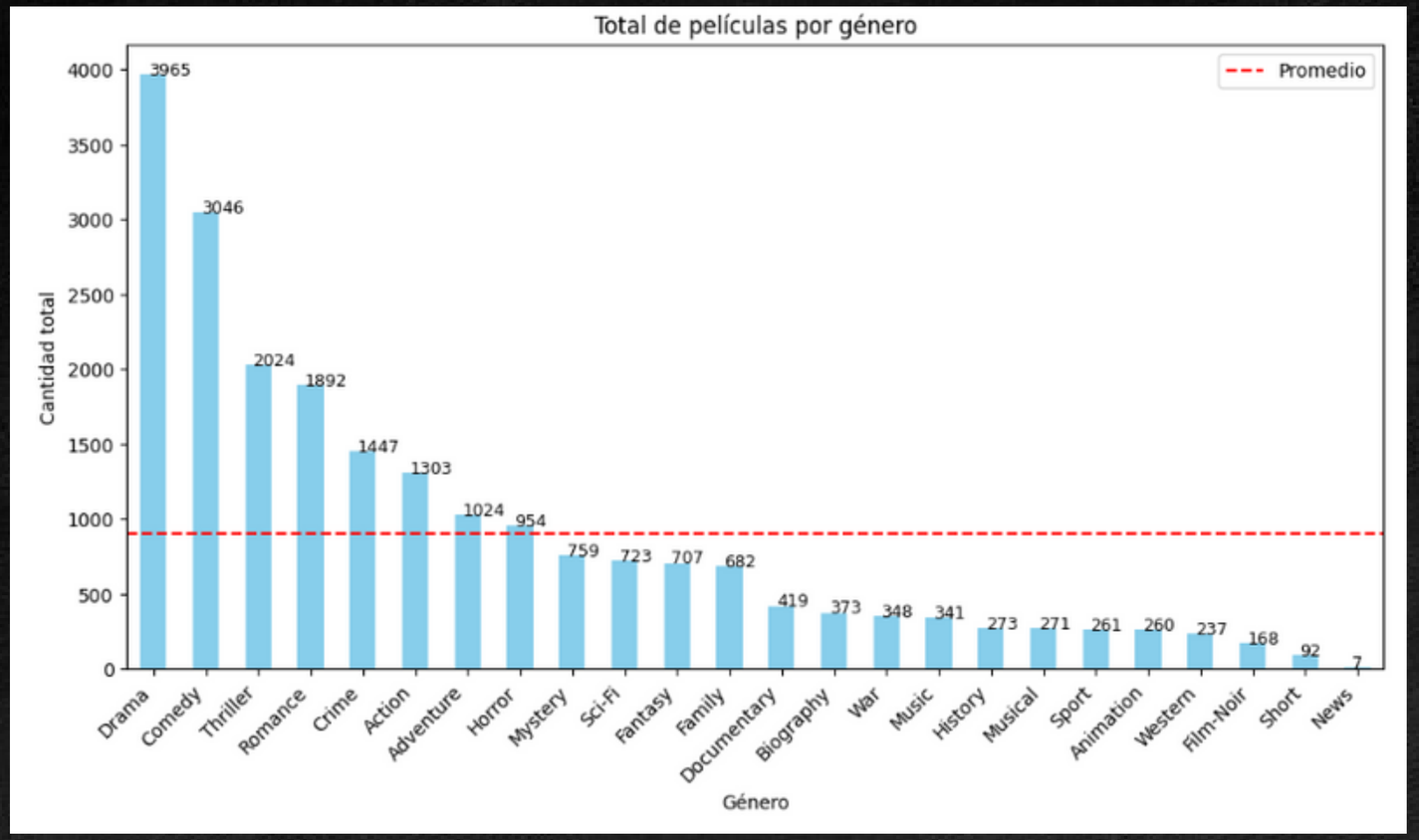
Adquisición de datos y Análisis descriptivo

Se observa un crecimiento de películas estrenadas por año llegando a su máximo pico en el 2013 cercano a las 280 películas



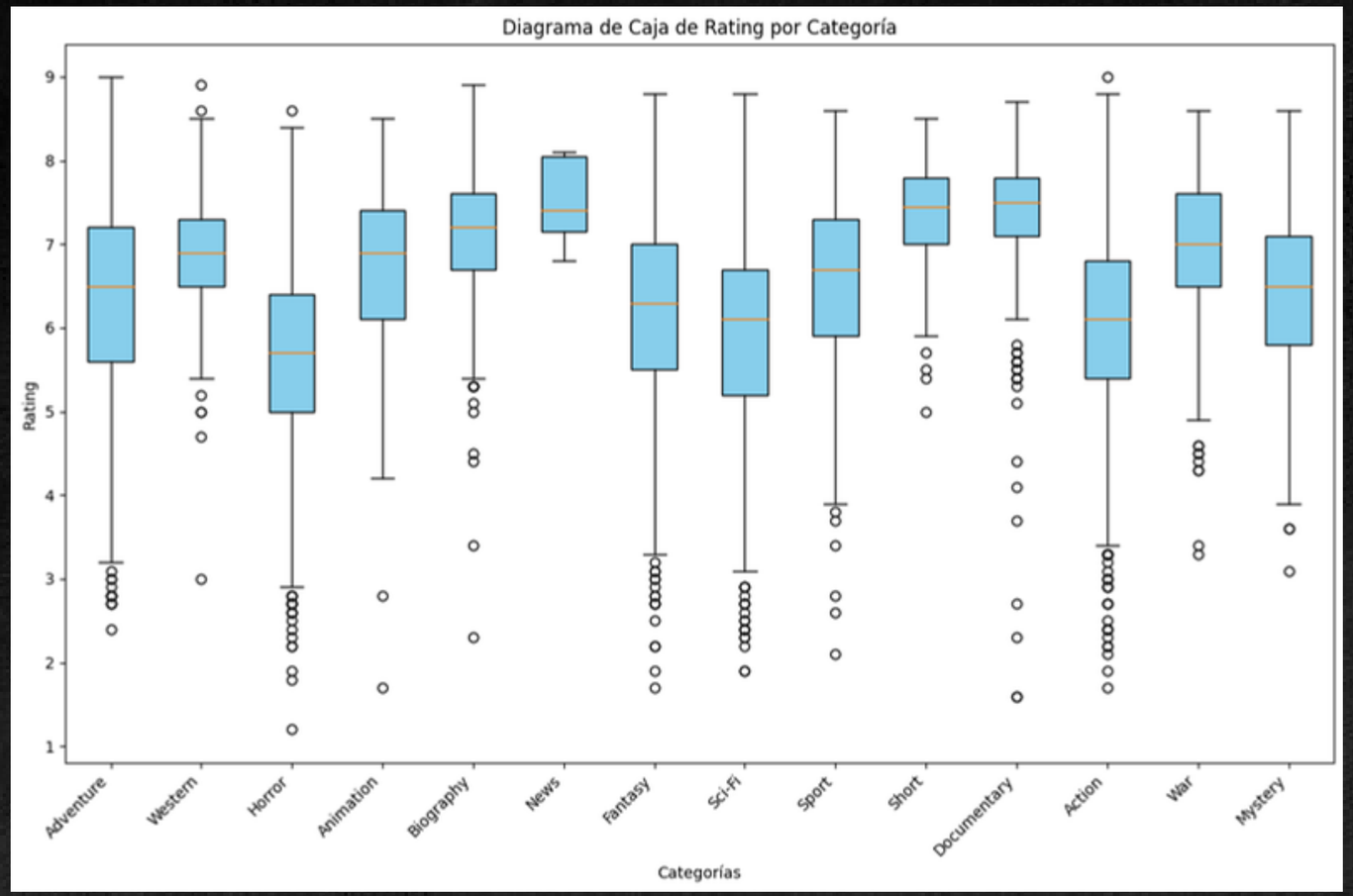
Adquisición de datos y Análisis descriptivo

Se presentan 24 géneros distintos de películas, en donde pueden combinar entre 2 a 4 géneros y se evidencia que el drama, la comedia, thriller, romance, crimen, acción, aventura y horror son las categorías más lanzadas (por encima del promedio).



Adquisición de datos y Análisis descriptivo

Las películas que exhiben el rating promedio más alto (7,8), aunque se identifican 13 documentales como valores atípicos con ratings inferiores a 6. En contraste, las películas de horror muestran los ratings promedio más bajos (5,8), con 10 películas que se encuentran fuera del rango intercuartílico, llegando incluso a un rating mínimo de 1,2.



Limpieza de texto de la trama

1. Transformación de texto a minuscula
2. Eliminación de stopwords
3. Eliminación de signos de puntuación
4. Eliminación de Palabras repetidas en multiples generos
5. Aplicación de lematización o estematización

Criterios de Modelación

Vectorizador

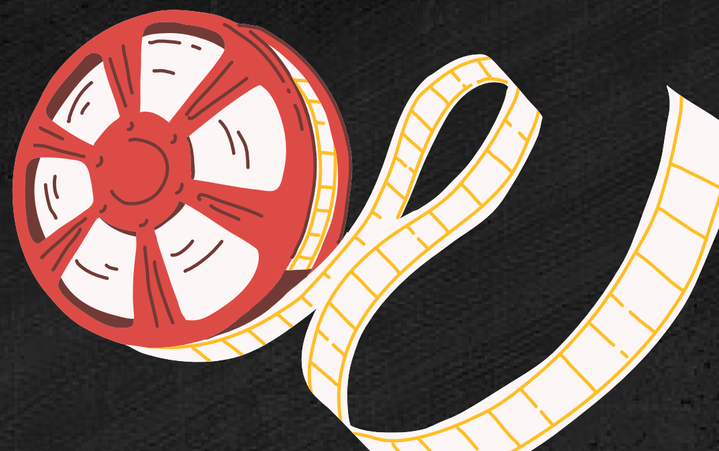
- Word2Vect
- TdIfVect
- CountVectorizer
- Modelo pre-entrenado wiki

Hiperparámetros

- Rango del n-grama
- Max Features
- Aplicación de grid search

Clasificadores

- SVM
- Regresión logística multicategoría
- Gradient Boosting Classifier
- Bayes
- Random forest
- XGBoost
- KNN



Grid Search

TDIF Vectorizer
Max Features: (8000, 10000, 15000, 20000, 30000)
Rango del n-grama: (1, 1), (1, 2), (1, 3), (2, 1), (2,2)

Regresión Logística
C: 0.1, 1, 9, 10, 11
solver: liblinear, lbfgs, newton-cg, sag, saga

Mejores Hiperparámetros
TDIF Vectorizer
Max Features: 8.000
Rango del. n-grama: (1, 3)
Regresión Logística
C: 11
solver: lbfgs
AUC Promedio: 0.89

Resultados Modelos

Clasificadores	Vectorizador	Max Futures	N - Grams	AUC
Reg Log	TdIfVect	8000	(1, 3)	0.89
NB	CountVect	2000	(1, 2)	0.82
XG BOOST	CountVect	8000	(1, 8)	0.85
SVM	TdIfVect	10000	(1, 2)	0.85
Gradient Boosting Classifier	TdIfVect	8000	(1, 3)	0.79
KNN	TdIfVect	8000	(1, 3)	0,68
RANDOM FOREST	CountVect	10000	1	0,82