UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Instituto de Cálculo

Herramientas de Modelado Estadístico para Ciencia de Datos 2025

TP2: Modelos mixtos, splines penalizados y causalidad

Aaron Bernal Huanca
Casiel Joshua Estragó
Ramiro Lipszyc

Profesor: Gonzalo Chebi

Buenos Aires, 17 de julio de 2025

Introducción

Dadas 4000 películas o series de una plataforma de streaming, se busca predecir la calificacion de IMDB a partir de otras covariables como por ejemplo, género, actores y directores, descripción, el mismo titulo, país de origen, rate de edad y duración.

Ejercicio 1

En esta sección realizamos un análisis exploratorio de los datos.

Ítem a

Veamos cómo se distribuyen los puntajes de películas en función de su género y determinar cuáles son las categorías que suelen recibir mejores evaluaciones por parte del público.

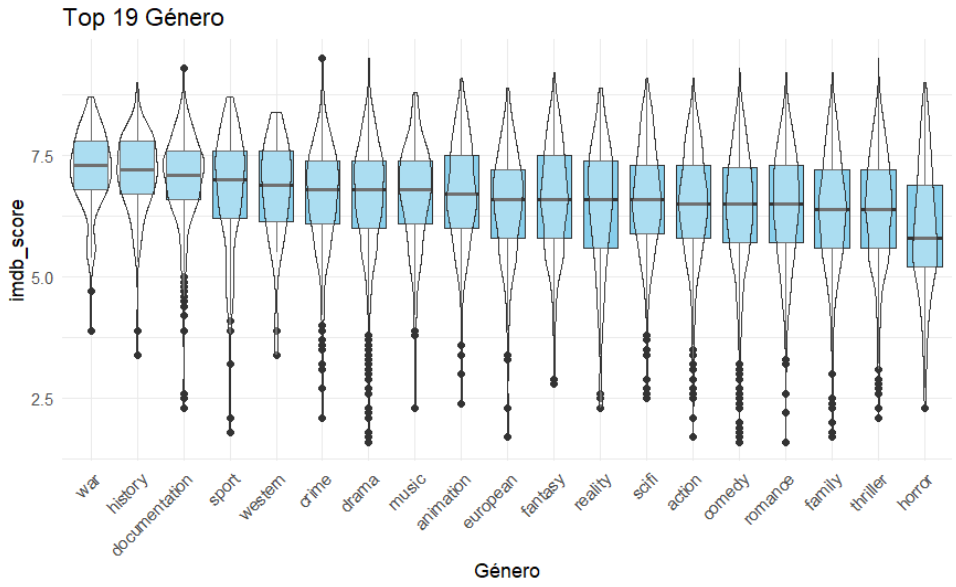

Figura 1: Top 19 mejores géneros de mediana

En el gráfico vemos que claramente hay géneros asociados a mejores puntajes que otros. “Guerra” es el mejor género en el sentido de que si elegimos dos títulos al azar y uno es de guerra, lo más probable es que el de guerra tenga mejor puntaje.

Ítem b

A continuación, observamos cómo se comportan los puntajes de IMDb en relación con actores y directores. La idea es identificar si existen personas asociadas a mejores o peores puntajes. Para ello, filtramos los actores y directores con mayor a 5 apariciones y analizamos sus distribuciones de puntaje.

Figura 2: Top 20 mejores vs peores actores ordenados por la mediana

Las cajas muestran la dispersión de los puntajes, y el gráfico de violín nos da una idea más detallada sobre la densidad de estos puntajes.

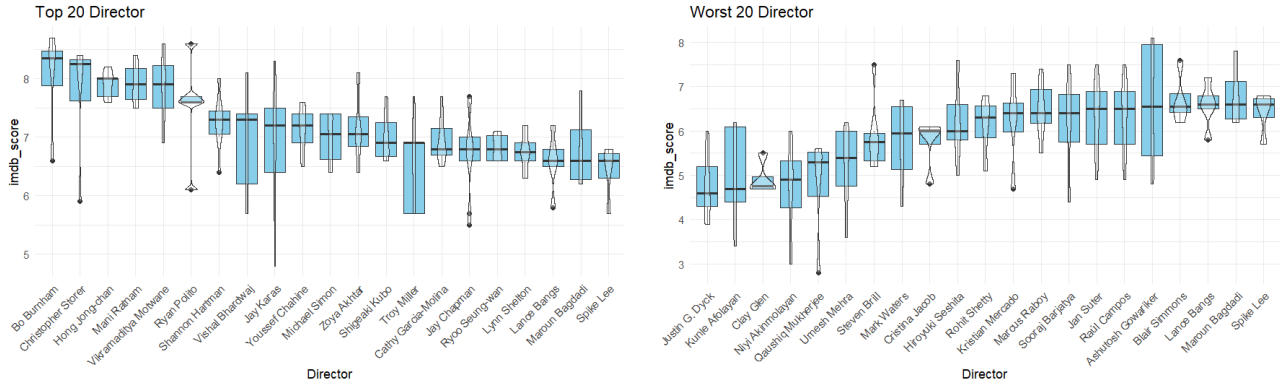

Figura 3: Top 20 mejores vs peores directores (los gráficos no comparten escalas)

Aquí es más difícil afirmar si hay actores o directores asociados a mejores o peores puntajes, porque a medida que filtramos por cantidad de apariciones, empiezan a quedar muy pocos actores y directores. En la mayoría de los casos (especialmente en actores), los puntajes no parecen estar centrados alrededor de nada, y el gráfico de violín no tiene una masa central cerca de la mediana, como sí pasaba con los géneros. Este análisis es informativo solamente si la distribución de puntajes da algo no uniforme teniendo muchas muestras, porque uniformidad es lo mismo que no tener información. Este no es el caso con actores y directores.

Ítem c

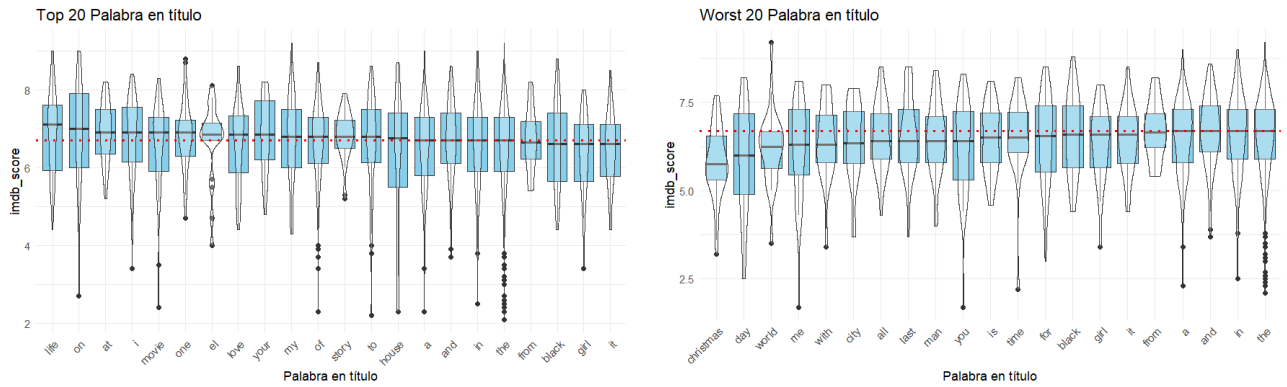

Figura 4: Top 20 mejores vs peores palabras en los títulos (los gráficos no comparten escalas). En rojo, la mediana global

Vemos que hay palabras con medianas grandes (> 7), pero las más altas no están muy por encima de la mediana global de 6.6, vease Figura 4. Las más bajas (como "navidad" tanto en títulos como descripciones) sí están considerablemente por debajo en proporción. Aprovechamos para aclarar que la palabra "s" viene de que en inglés el posesivo se escribe 's y el regex que separa palabras cuenta el apóstrofe como algo que interrumpe la palabra.

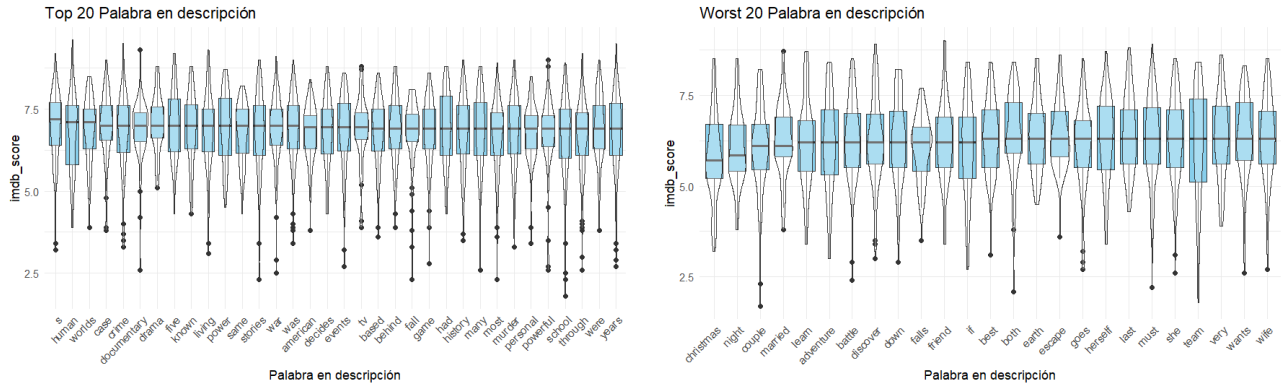

Figura 5: Top 20 mejores vs peores palabras en descripcion (los gráficos no comparten escalas)

Extra

Proponemos hacer el mismo análisis con las demás categorías: países, restricción de edad y si el título es película o serie.

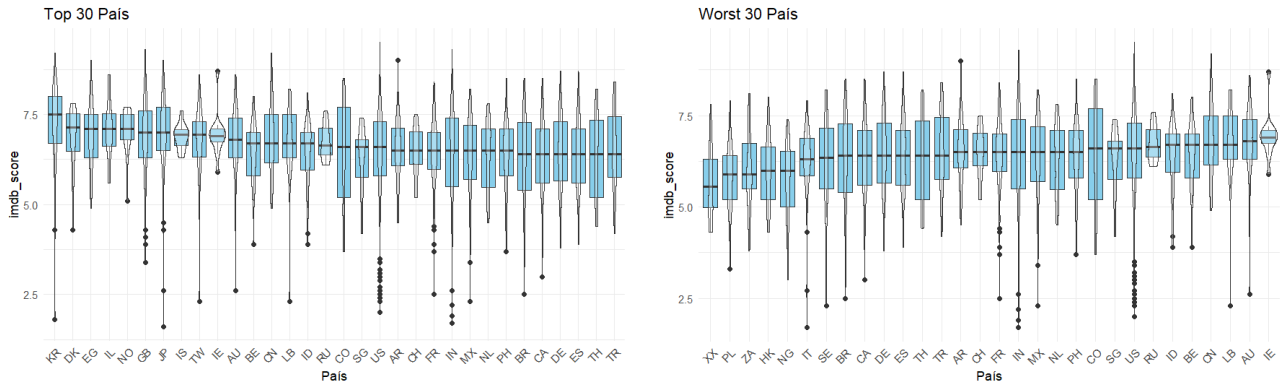

Figura 6: Top 30 mejores vs peores paises (los gráficos no comparten escalas)

Consideramos que podría haber algo informativo en estos datos, y eso va a ser tenido en cuenta cuando hagamos el ejercicio 5 a la hora de proponer modelos.

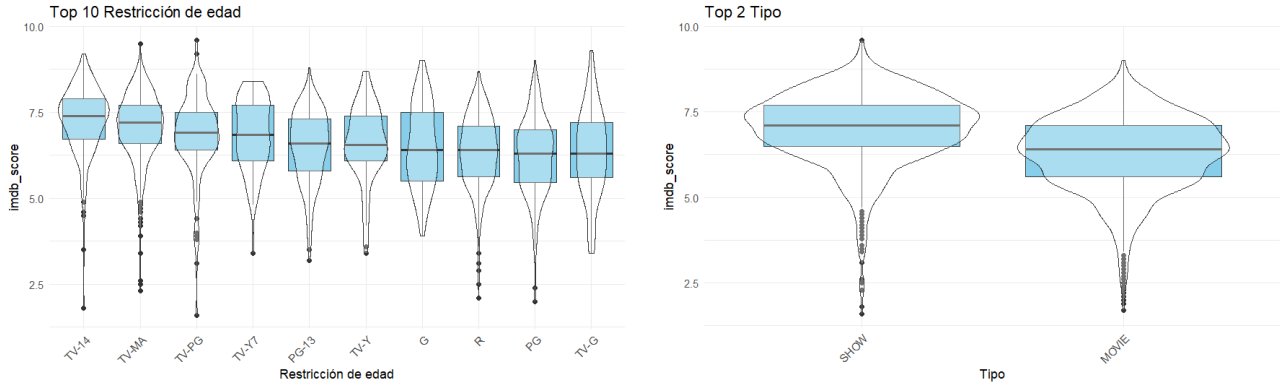

Figura 7: Restricción de edad y tipo de titulo vs imdb score (los gráficos no comparten escalas)

En Figura 7, observamos individualmente una simetria entre los rates y el tipo de titulo, con una cola ligeramente más pasada a puntajes bajos. Grupalmente la variabilidad no es significativa.

Ejercicio 2

Como hay algunas producciones que fueron realizadas en más de un país, vamos a duplicar cada una por cada país que tenga registrado. Esto en parte implica asumir que cada país contribuyó "por igual", cosa que no sabemos. Para los fines de este TP lo dejamos así. Más adelante, en los modelos 3 y 4 este approach se cambia.

Ítem a

Planteamos un modelo de efectos fijos donde cada país tiene un coeficiente específico, y con eso obtuvimos un error cuadrático medio de 1,4016

Ítem b

Usando efectos aleatorios obtuvimos 1,3786, ligeramente mejor que anterior.

Ítem c

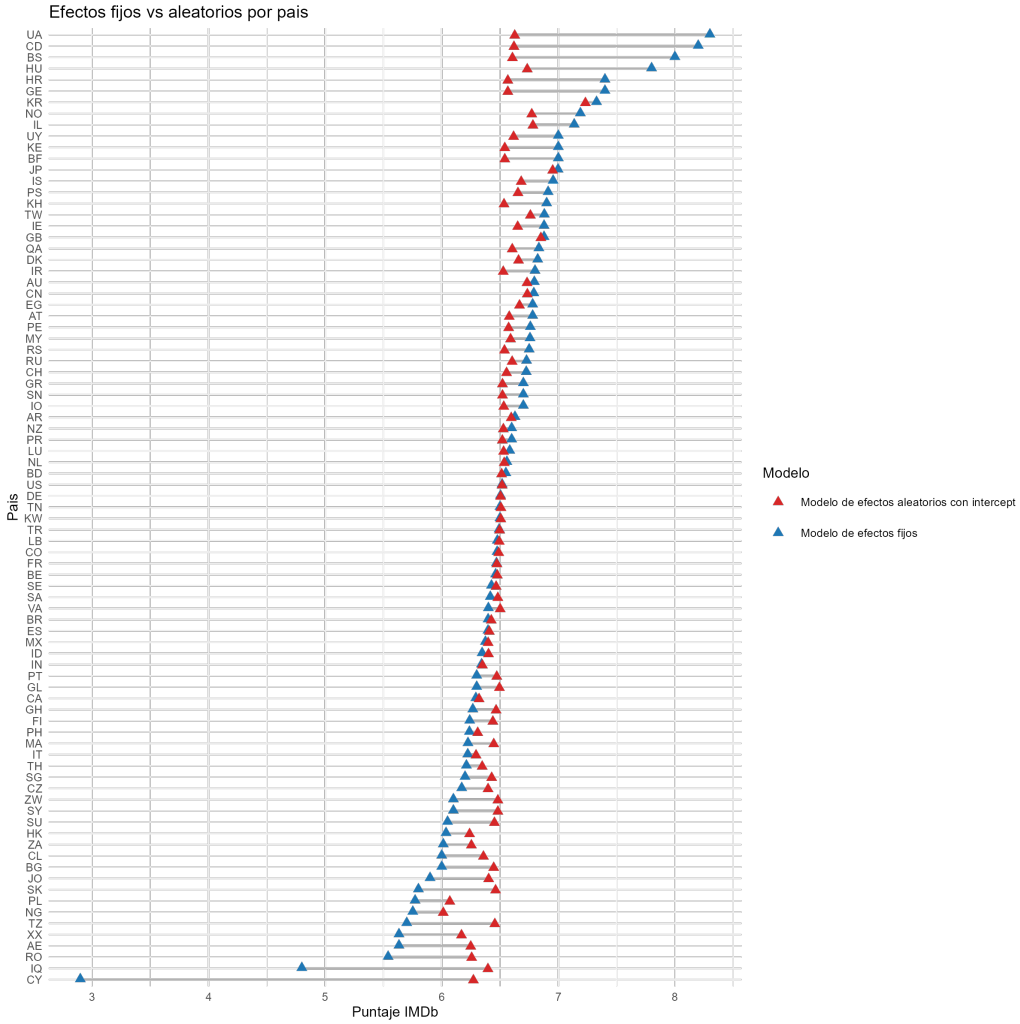

Figura 8: Efectos fijos vs aleatorios por país

Para notar la incidencia que tienen ambos efectos sobre el score IMDB en cada país, visualizamos la relación entre la cantidad de producciones y la diferencia entre ambos modelos:

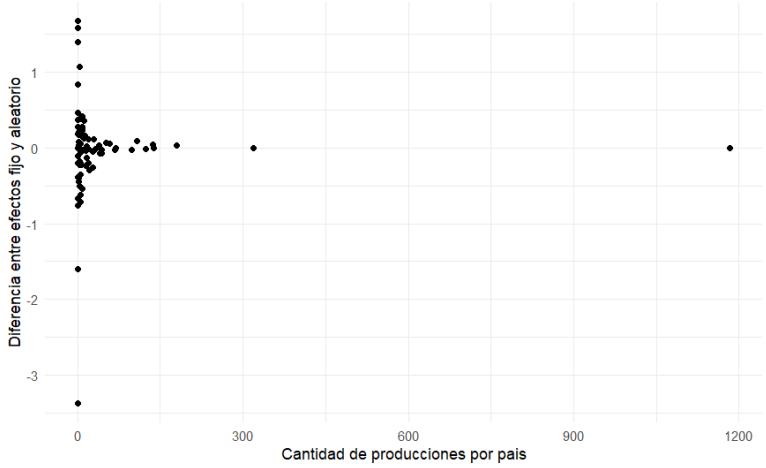

Figura 9: Cantidad de producciones vs la diferencia entre los efectos

Podemos ver que el modelo de efectos aleatorios tiene a contraer las estimaciones de aquellos países que tienen pocos datos hacia una "media global" (en este caso, pocas producciones). Este ultimo gráfico lo refleja bien, pues a menor cantidad de producciones, mayor es la corrección de los efectos aleatorios en contraste con los fijos. En cambio, para aquellos países que cuentan con mas producciones, el modelo de efectos aleatorios "confía" más, por así decir, en las medias muestrales.

Ejercicio 3

En este gráfico analizamos cómo varía la relación entre el año de lanzamiento y el puntaje con splines cúbicos no penalizados para distintos $k = 1, 2, 3, 5, 10, 20, 50$. Se observa que con pocos nodos las curvas son más estables, mientras que con más nodos las curvas se vuelven más irregulares y capturan fluctuaciones locales, aunque llegando a sobreajustar en algunas zonas con pocos datos.

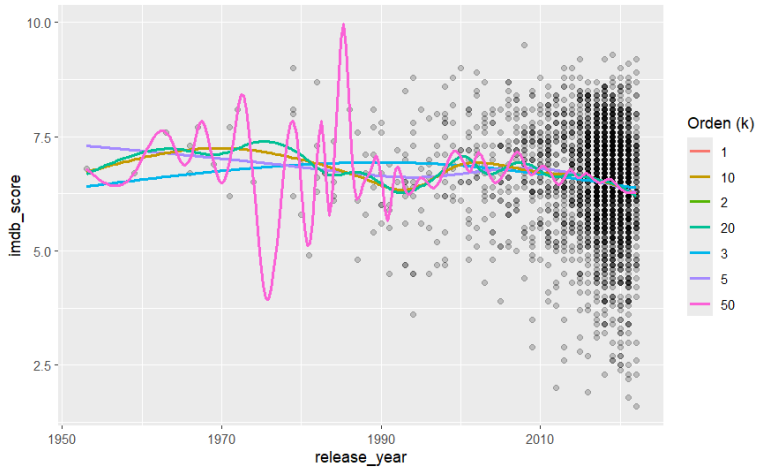

Figura 10: Comparación de la regresión con splines de orden k

Ejercicio 4

Dado el siguiente DAG y sea $Z \subset \{Año, Duracion, Pais\}$, buscamos los posibles Z tal que si condicionamos se puede calcular el efecto causal promedio de la variable *Comedia* sobre el *Score*.

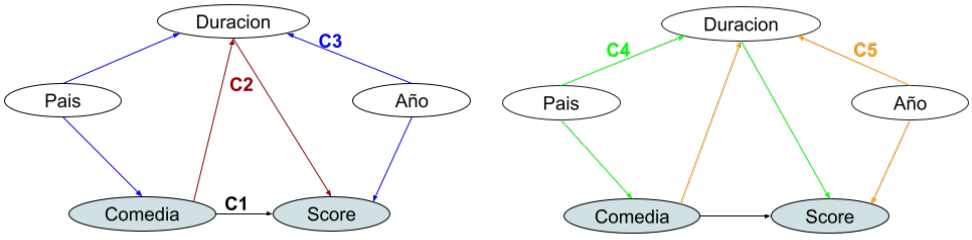

Figura 11: DAG y los posibles caminos entre *Comedia* y *Score*

Sean $X \equiv Comedia$ y $Y \equiv Score$, basta ver que que X, Y están d-separados por el conjunto Z, es decir, que todos los caminos entre X, Y están bloqueados por Z.

Definition 0.1 (Camino bloqueado)**.** *Un camino C entre X, Y esta bloqueado por un conj de v.a. Z si se cumple alguna de las siguientes condiciones:*

- $\exists W \in Z \cap C$ *tal que W W es nodo central de una chain o de un confunder. (1)*

- $\exists W \in Z - C$ *tal que* W W *es nodo central de un collider y descendientes*$(W) \cap Z = \emptyset$ *(2)*

Luego:

Por camino C_2 en (citar figura del DAG), *Duracion* tiene que estar en Z porque es nodo central de una chain y por condición (1).

Por camino C_3, tanto *Pais* y *Anio* son nodo central de un confunder, por lo tanto, alguno de los dos deber estar en Z.

Por camino C_4, *Duracion* ya bloquea porque es un nodo central de una chain, entonces *Pais* y *Anio* pueden estar o no en Z.

Sin embargo, por camino C_5, *Anio* tiene que estar en Z porque es un nodo central de un confunder y *Duracion* (collider) esta en Z.

Posibles Z:

- $\{Año, Duracion, Pais\}$
- $\{Año, Pais\}$

Ejercicio 5

Presentamos algunos modelos y los testeamos la subsección Resultados

Modelos 1 y 2

Para estos modelos, la idea es convertir las variables categóricas en algo numérico, intentando aprovechar lo visto en el ejercicio 1.

- Definimos $R, D, P, G, W_T, W_{,D}\, A, T$ los conjuntos de actores, directores, países, géneros, palabras en títulos, palabras en descripciones, restricciones de edad y tipo respectivamente.
- Definimos P_k (lo llamamos peso) con $k \in K$ así: $P_k = \frac{1}{n+1}(\mu_K + \sum_{i=1}^{n} a_i^k)$ con n la cantidad de apariciones de k en las películas o series, a^k el vector de scores asociados a cada aparición un k (por ejemplo, la lista de puntajes de todos los títulos en los que aparece un director), y μ_K es el promedio de los scores asociados a todos los elementos de K.
- Si $k \subseteq K$ tomamos $P_k = \frac{1}{\#k} \sum_{j \in k} P_j$

De esta forma, la idea es tener una predicción estándar para el valor que aportan los actores, directores, palabras, etc. desconocidos e ir desplazándola a medida que haya más observaciones, y que toda esta información no numérica se pueda traducir en algo que sí es numérico.

Modelos 3 y 4

Transformamos las covariables genero a one-hot (una por clase) y definimos como país de origen al primer país en su lista de países de producción. Por ultimo, borramos las columnas de texto.

En el modelo 3 usamos una función de link logística para mapeando los scores del $[0, 10] \to [0, 1]$, como desventaja los puntajes cerca borde son inalcanzables (improbables al mismo tiempo).

El modelo 4 es una regresión lineal con la desventaja que acepta valores fuera del rango $[0, 10]$ (nuevamente, improbable para datos de testeo razonables) Además, muestra que variables con estadísticamente significativas y si favorece o empeora el score. Esto respaldado los hallazgos del ejercicio 1. Por ejemplo, Corea tiene un coeficiente de 0.7114 que se puede interpretar informalmente como que "ser de Corea sube el puntaje 0.7".

Resultados

Modelo	MSE
Modelo 1	1.2781
Modelo 2	1.2781
Modelo 3	1.0191
Modelo 4	1.0321

Cuadro 1: Comparación de error cuadrático medio (MSE) entre los modelos

Finalmente, seleccionamos el modelo 3 para generar las *predicciones.csv* de los datos nuevos.

Conclusiones

Concluimos que nuestro modelo obtiene una performance decente respecto al MSE y debería ser capaz de distinguir entre buenas y malas películas. Sugerimos seguir explorando nuevos modelos e incorporar modelos de lenguaje para procesar la descripción y el titulo de manera contextual mediante la ayuda de representaciones como BERT.