**Propuesta de investigación**

**“**Predicción sesgo de género en películas extendiendo el test Bechdel”

**Alumnas**:

Angela Navajas McCormick

Legajo: 33542

Tiziana Sol Paciente

Legajo: 33377

**Materia:**

Ciencia de Datos

Lic. en Ciencias del Comportamiento

**Docentes:**

Maria Noelia Romero

Ignacio Spiousas

**Fecha de entrega:**

7 de Diciembre, 2024

## Introducción

La propuesta de investigación consiste en un análisis de las características que poseen diferentes películas, con el objetivo de predecir si estas producciones serian clasificadas por el test de Bechdel cómo “tienen o no tienen sesgo de género” en base a las variables que la prueba toma como importantes y extendiendo el análisis de predicción al sumar otros predictores que pueden llegar a ser relevantes sobre el resultado predicho.

El test de Bechdel propone observar la presencia de mujeres en películas y medir el sesgo de género en la producción cinematográfica en base a 3 criterios: si tiene al menos 2 mujeres que tienen un nombre en pantalla, si hay interacción entre ellas y por último hace foco en el tipo de interacción que tienen: si la conversación se trata de algo por fuera de un hombre. El test propone que si la película cumple con las 3 variables, entonces la producción no tendría sesgo de género y de esta manera determina el rol que poseen las mujeres dentro de la trama, el nivel de desarrollo que tienen los personajes femeninos.

La pregunta de investigación consiste en que conocer cuales son las características que determinan si una película pasa o no el test de Bechdel y por ende, si las mujeres tienen un rol más allá de secundario y tienen una profundidad en el personaje. Por otro lado, se busca observar cómo varía la representación de personajes femeninos en las películas según el género y el año de producción.

La motivación del trabajo es explorar los patrones relacionados con la representación de las mujeres dentro de la industria cinematográfica, observar si hay determinados patrones que determinan el rol que estas toman dentro de la trama, utilizando el Test de Bechdel como base para este análisis. El aporte de este análisis es expandir los predictores que pueden estar relacionados con el tipo de representación que tienen las mujeres en la trama.

## Literatura previa

Dentro de los antecedentes, Lakhotia, *et al*. (2019) en base a las críticas que posee el test de ser muy un análisis muy simplificado, las autoras proponen tomar en cuenta algunos parámetros que se relacionan con la representación femenina en películas, como por ejemplo la cantidad de diálogos entre mujeres en la película, el género de diálogo y las categorías gramaticales que se utilizan en esta conversación. Para esto utilizaron el método de PCA, con el que analizaron el guión de 342 películas para testear si los parámetros agregados al criterio de Bechdel son significativos para calcular el puntaje de representación femenina.

Por otro lado, Lindner y Schulting (2017) estudian un mecanismo potencial que puede contribuir a la subrepresentación de mujeres en películas al considerar si las críticas que recibe penalizan o premian la presencia de personajes femeninos independientes. Para esto, testean si las películas que pasan el test Bechdel tienen mayor o menor puntaje en Metacritic (una medida de calidad basada en reseñas de críticos) controladas por variables que pueden influir en el puntaje, como el género, el presupuesto de producción, la participación de un actor reconocido y el hecho de ser una secuela.

Sumado a esto, Johann Valentowitsch (2022) expone que el test de Bechdel tiene cada vez más presencia y uso dentro de la academia como indicador de la representación de mujeres en películas. Expone que en estudios previos se observó como pasar o no el test impacta sobre la taquilla en películas en EEUU, por lo que el autor extiende el análisis de este efecto de películas de Hollywood sobre la taquilla internacional, demostrando que pasar el test implicaba una mejora significativa en el rendimiento en taquilla de la película.

**Base de datos**

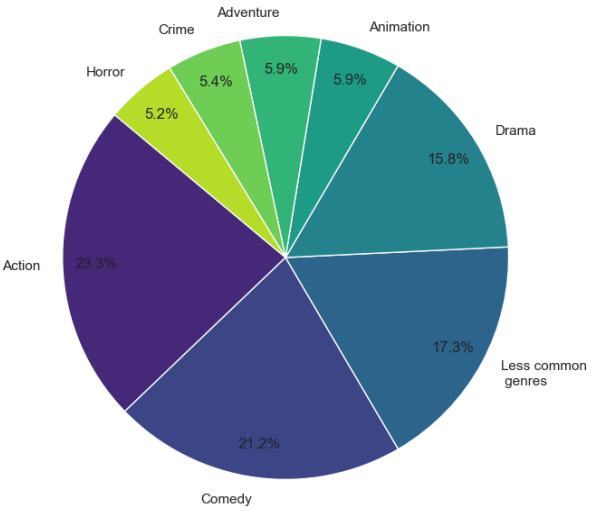
La base de datos a utilizar es llamada movies.csv, fue obtenida de https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-03-09 que forma parte de una propuesta de tidytuesday, de libre acceso.

Esta base de datos tiene 34 variables sobre características de películas que pueden tener incidencia sobre el resultado de pasar o no el test Bechdel, como por ejemplo, el año en el que produjo la película, el título, el presupuesto en moneda nacional e internacional con el que se hizo la película, el presupuesto normalizado al año 2013, la década, el tipo de plot de la película, la calificación de la película, el idioma, en que país se produjo, el director de la producción,el escritor, los actores involucrados, el género, el tipo de película, entre otros.

Para profundizar acerca de cómo está conformada la base se realizaron algunas estadísticas descriptivas. En principio podemos ver en la Figura 1 que más del 50% de la base está compuesta por películas de comedia, acción y drama. Dada la amplia cantidad de géneros, decidimos centrarnos en los géneros más frecuentes por lo que se generó una categoría que unifica los géneros menos frecuentes tales como: misterio, documentales, musicales. familiares, entre otros. Esta unificación forma el 17.3% de la base. El resto de la base se divide en películas de aventura, crimen, horror.

Se realizaron gráficos analizando el porcentaje de películas que pasaron o no el test dividiendo por género. En la figura 2 podemos ver que el género acción tiene el mayor porcentaje de películas que no pasan el test, junto con las películas animadas. Los géneros que tienen un porcentaje mayor de películas que pasan el test son las películas de terror, de comedia y de drama.

Figura 1: Porcentaje de películas por género



Después se realizó un gráfico para analizar dentro de las películas que no pasan, en qué categoría caen para no pasarlo (Figura 3). Se ve claramente que el mayor porcentaje de películas que no pasan, no lo pasan porque los personajes mujeres no hablan entre si, salvo las películas de terror.

Figura 2: Porcentaje de películas que pasaron o no el test Bechdel por género

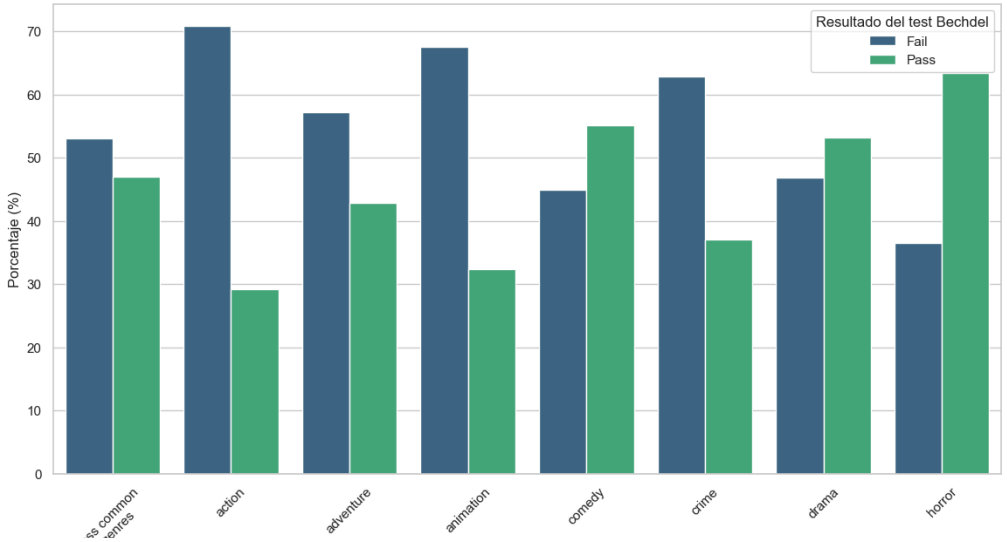
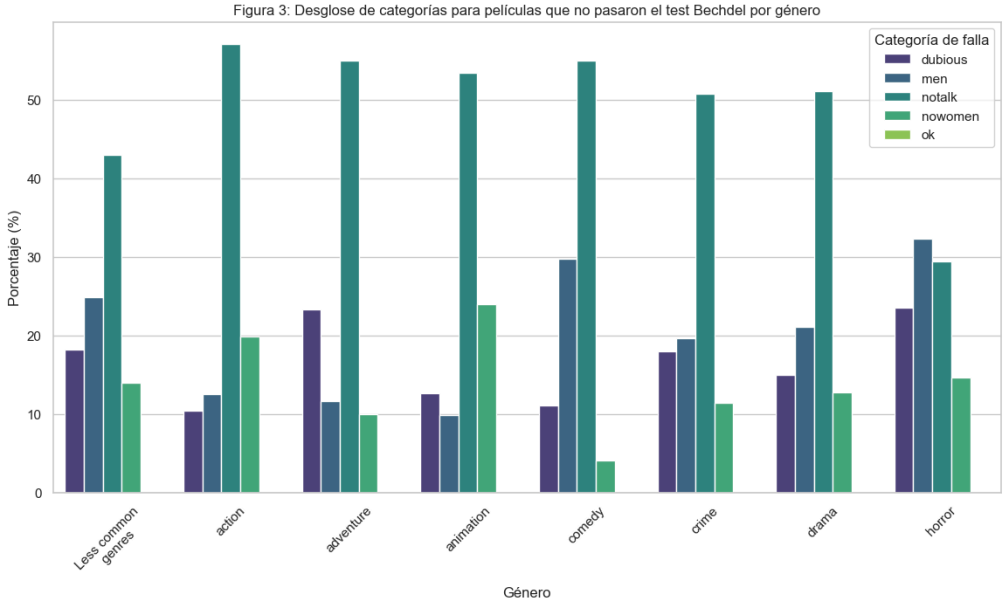


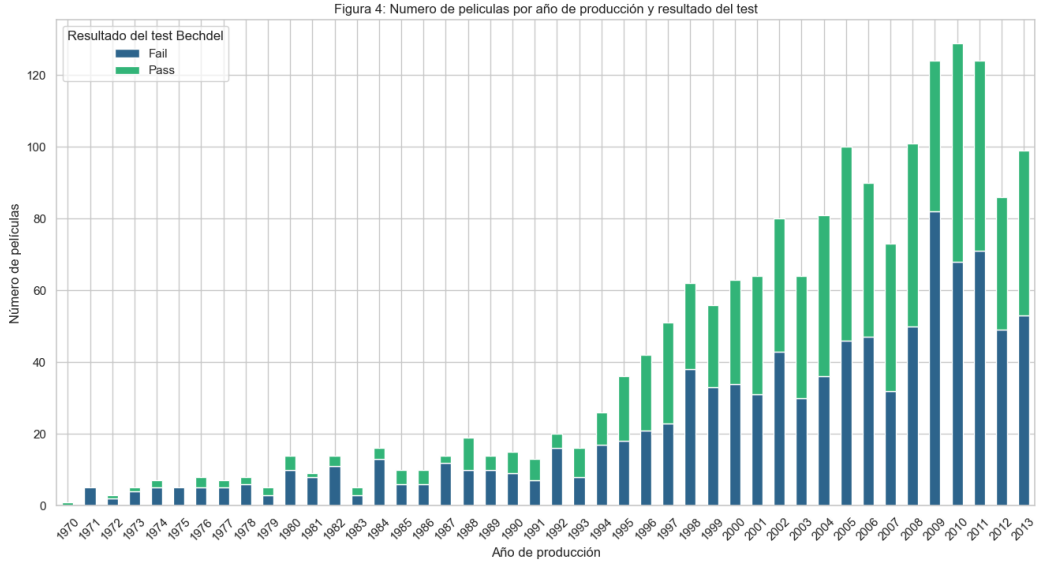

Figura 3: Desglose de categorías para películas que no pasaron el test Bechdel por género




Después realizamos un análisis de las películas con respecto al año de producción. Primero se realizó un histograma para ver la distribución de películas por año. Para profundizar el análisis se agregó en el histograma el resultado del test binario (Fail, Pass), es decir, el número de películas por año, que pasó y no pasó el test. En este gráfico podemos ver

en principio que la cantidad de películas que tiene la base de datos por año aumentó dramáticamente en los últimos veinte años del análisis. En los primeros treinta años, si bien la cantidad de películas es reducida, podemos ver que el desempeño del test es bastante malo. A partir de 1995 podemos ver que el resultado empieza a mejorar pero que casi que en ningún año parece haber más del 50% de películas que sí lo pasen.

Figura 4: Numero de peliculas por año de producción y resultado del test




**Metodología**

Para comenzar con la aplicación de los modelos se realizará una limpieza de los datos, Como primer paso para poder utilizar y analizar los datos con el objetivo de asegurarnos el mejor funcionamiento de los modelos en las próximas etapas. Se eliminan los valores nulos, los datos sin sentido y se tratarán los valores faltantes y los outliers. Por último se realizaron los cambios necesarios de las variables categóricas a dummies utilizando One-Hot Encoding, para poder utilizarlas en los métodos propuestos.

Después de realizar la limpieza, se profundizará la etapa exploratoria (vista en las figuras 1 a 4), para seguir analizando las características de las películas con respecto del test. Además, una vez hechas las variables dummies se realizarán matrices de correlación,

teniendo en cuenta la colinealidad causada por la transformación, para comprender la relación entre las variables e identificar los patrones necesarios a tener en cuenta.

Para empezar con el análisis propuesto, se realizarán clusters de k-media para agrupar las películas en grupos homogéneos basados en características importantes de los datos. Las variables que se utilizarán para este análisis serán en principio, género y presupuesto. A través de esta segmentación, se pretende captar las posibles diferencias que se pueden encontrar entre películas con mayor y menor presupuesto. En general se pretende encontrar patrones posibles asociados con el desempeño en el test y de esta forma proporcionar nueva información relevante para los modelos predictivos. Se eligió k-medias dado que este modelo nos permitirá obtener resultados claros y rápidos.

En principio, se empleará un método de clasificación para predecir la variable de interés, un modelo de regresión logística. Este método es ampliamente utilizado en problemas de clasificación binaria por lo que lo usaremos para predecir si una película cumple o no el test de Bechdel. El modelo no requiere de una distribución normal de los datos por lo que lo hace adecuado a nuestra base. Sospechamos que las variables pueden llegar a sufrir de la presencia de ciertas correlaciones entre variables por lo que un método no tan sensible a estas es esencial. También vamos a tomar ventaja de la posibilidad de interpretación del modelo para evaluar el impacto de cada variable en la probabilidad de pasar el test.

Por otro lado, se empleará un método de ensamble llamado CART (classification and Regression Trees) con el objetivo de identificar las variables más relevantes para predecir si una película pasa o no el test. Este método de clasificación utiliza particiones recursivas binarias para crear regiones en el espacio de predictores, a cada región se le asigna una predicción basada en el promedio de las clases observadas que perteneces a ella. Las particiones se van a efectivizar utilizando el índice de gini para cada una. Se eligió este

método debido a su claridad a la hora de visualizar las formas en que las variables influyen en la probabilidad final de que una película pase el test. Por otro lado, el modelo prioriza los predictores más relevantes por lo que nos resulta clave para identificar qué variables son las que tienen mayor influencia en la probabilidad. Para evitar un posible sobreajuste que resulta del método se implementará un Tree Pruning, proceso que nos permitirá reducir la profundidad del árbol de CART, eliminando ramas que no tengan una contribución significativa a la predicción. Aparte de reducir el Trade off de sesgo y varianza este proceso simplificará la estructura del árbol, focalizándose en las variables cruciales del análisis.

Para ambos modelos se dividirá la base de datos en conjuntos de entrenamiento y prueba para evaluar el desempeño del modelo. La partición se hará siguiendo el estándar de 70% para entrenamiento y 30% para prueba. Con respecto al desempeño del modelo se pretende evaluar en principio la precisión global (Accuracy), que mide el porcentaje de observaciones clasificadas correctamente en ambas clases, tanto para si pasan el test como si no. Este resultado se pretende complementar con una matriz de confusión para así evaluar de manera detallada la capacidad predictiva de los modelos. Por último, se graficara la curva ROC, para ver de forma clara la capacidad del modelo en discriminar entre pasar o no pasar el test. Para asegurar la estabilidad del modelo se implementará un procedimiento de validación cruzada utilizando un k=10 para así obtener las métricas promedio en las diferentes particiones de la base y asegurarnos de reducir lo más posible un sobreajuste. Finalmente los resultados de estas evaluaciones permitirán identificar cuál modelo presenta un mejor desempeño predictivo y poder definir cuales son las variables más relevantes sobre la probabilidad de pasar el test.

**Conclusiones y limitaciones**

A modo de conclusión, luego de aplicar los métodos anteriormente mencionados, esperamos encontrar qué variables son las más influyentes sobre la representación de personajes femeninos en películas, sumado a identificar si el agregar los nuevos predictores al test inicial de Bechdel modifica la clasificación de las películas si tienen una representación de mujeres más allá de un papel secundario.

En tanto a las limitaciones que podríamos enfrentar, una de las principales se encuentran se relaciona con el test propuesto es la simplicidad de las 3 variables que toma, el estudiar otras variables y la interacción que tienen entre estas generaría un análisis más profundo y preciso sobre la representación de las mujeres. Sumado a esto, en la base de datos sobre la que trabajamos no hay variables que podrían ser relevantes para el estudio, como por ejemplo la cantidad de interacciones entre mujeres, cuánto duró, el tipo de palabras que fueron utilizadas. Creemos que sería interesante para aumentar la comparación entre películas y representaciones, poder tener variables sobre las interacciones de los personajes hombres dentro de la película, como por ejemplo el porcentaje de interacciones entre hombres a comparación con el porcentaje de interacciones entre mujeres, el tipo de interacciones que se tiene entre hombre y mujeres, entre otras variables. Por otro lado, la base de datos no se encuentra actualizada con películas de la última década, por lo que agregarlas sería de interés para ver cómo fue evolucionando el ratio de películas que pasan o no el test durante el último tiempo.

Por último y relacionado a la metodología elegida, tanto Logit como CART son poco robustos ante la colinealidad entre los predictores, por lo que puede verse afectada la clasificación. En el caso que se detecte correlación entre estas variables, se podría evaluar el uso de regularización en en la regresión logística usando Ridge para manejar la

multicolinealidad y en en el caso de CART, se podría complementar con Random Forest para ver distintos árboles con distintas relaciones entre la variable dependiente y las variables independientes.

**Bibliografía**

Lakhotia, R., Nagesh, C. K., & Madgula, K. (2019). Identifying Missing Component in the Bechdel Test Using Principal Component Analysis Method. arXiv preprint arXiv:1907.03702.

Lindner, A. M., & Schulting, Z. (2017). How Movies with a Female Presence Fare with Critics. Socius, 3. https://doi.org/10.1177/2378023117727636

Valentowitsch, J. (2023). Hollywood caught in two worlds? the impact of the bechdel test on the international box office performance of cinematic films. Marketing Letters, 34(2), 293-308. doi:https://doi.org/10.1007/s11002-022-09652-5