Informe Final De Proyecto

Proyecto de Predicción de Recaudación de Películas

Integrantes:

Steven Alipio Berrio - CC. 1036661504

Sergio Andrés zapata Ruiz – CC. 1037648161

Profesor:

Raúl Ramos Pollan

Semestre 2023-2

1. Introducción

El enfoque inicial del proyecto se mantuvo consistente desde la primera entrega, pero se amplió para incluir un análisis más detallado de diversas variables que podrían influir en la predicción de la recaudación de películas al igual que el manejo de los modelos para poder analizar el dataset en conformidad a lo que se estaba pidiendo. El objetivo principal sigue siendo desarrollar un modelo de inteligencia artificial capaz de predecir con precisión los ingresos de una película basándose en características clave con una mejor precisión y ampliando los datos a más columnas si es posible.

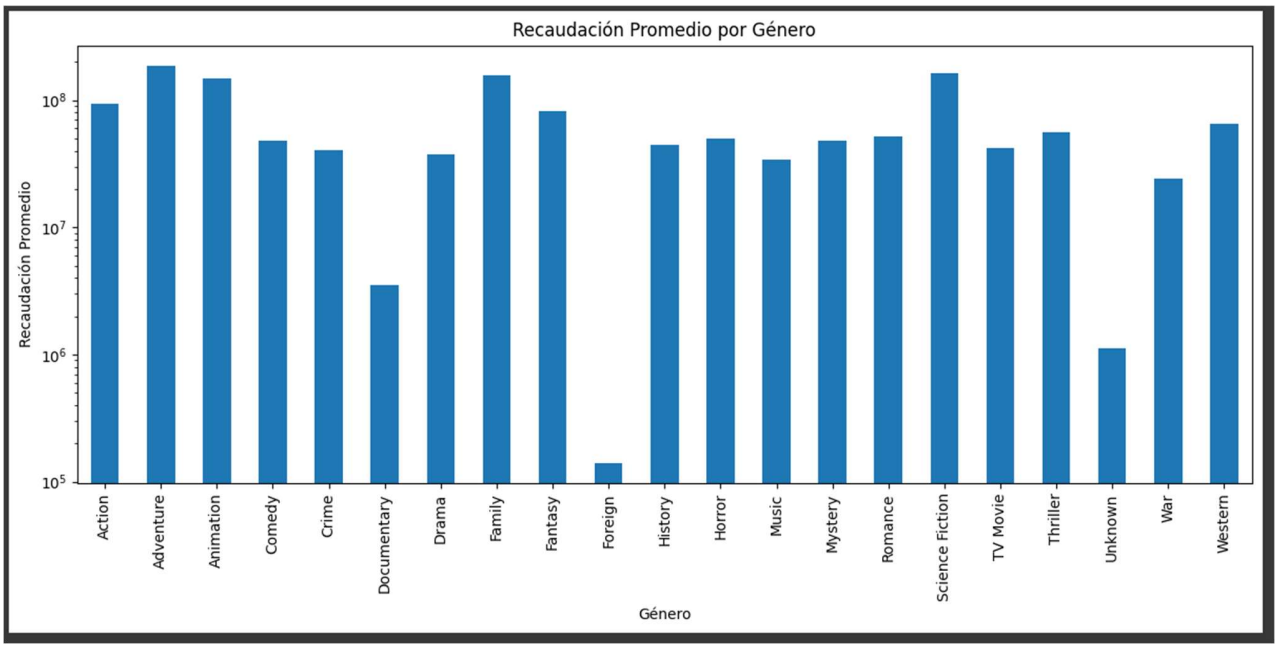
2. Exploración Descriptiva del Dataset

La exploración descriptiva reveló que variables como popularidad, presupuesto y duración son las más influyentes en las predicciones de la mayoría de este tipo de contenido visual como lo son las películas. Se observó que factores como el elenco, la compañía productora y el género no variaban significativamente en relación con los ingresos generados como lo hacían las anteriores variables ya mencionadas.

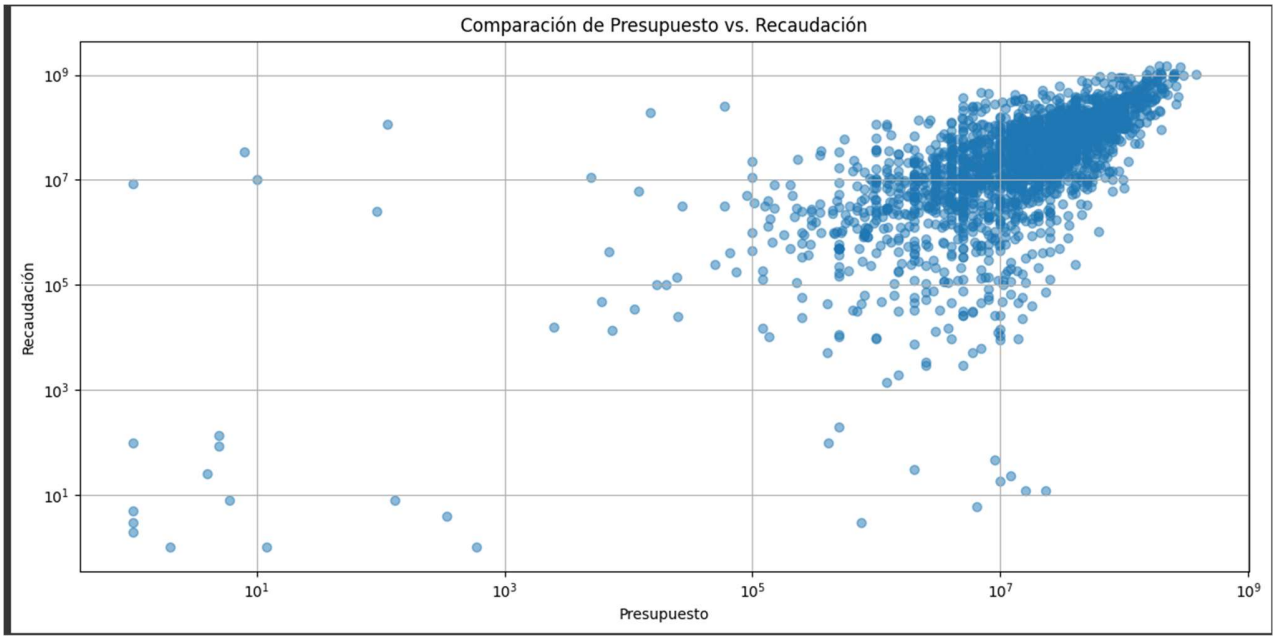
En nuestro datase realizamos la exploración de los datos presentes listando las columnas más importantes y sus características para la toma de los datos y así poder realizar nuestras predicciones mas acertadamente y tener unos resultados mas óptimos para el proyecto.

#	Column	Non-Null Count	Dtype
0	id	3000 non-null	int64
1	belongs_to_collection	604 non-null	object
2	budget	3000 non-null	int64
3	genres	2993 non-null	object
4	homepage	946 non-null	object
5	imdb_id	3000 non-null	object
6	original_language	3000 non-null	object
7	original_title	3000 non-null	object
8	overview	2992 non-null	object
9	popularity	3000 non-null	object
10	poster_path	2999 non-null	object
11	production_companies	2844 non-null	object
12	production_countries	2945 non-null	object
13	release_date	3000 non-null	object
14	runtime	2998 non-null	float64
15	spoken_languages	2980 non-null	object
16	status	3000 non-null	object
17	tagline	2403 non-null	object
18	title	3000 non-null	object
19	Keywords	2724 non-null	object
20	cast	2987 non-null	object
21	crew	2984 non-null	object
22	revenue	3000 non-null	int64

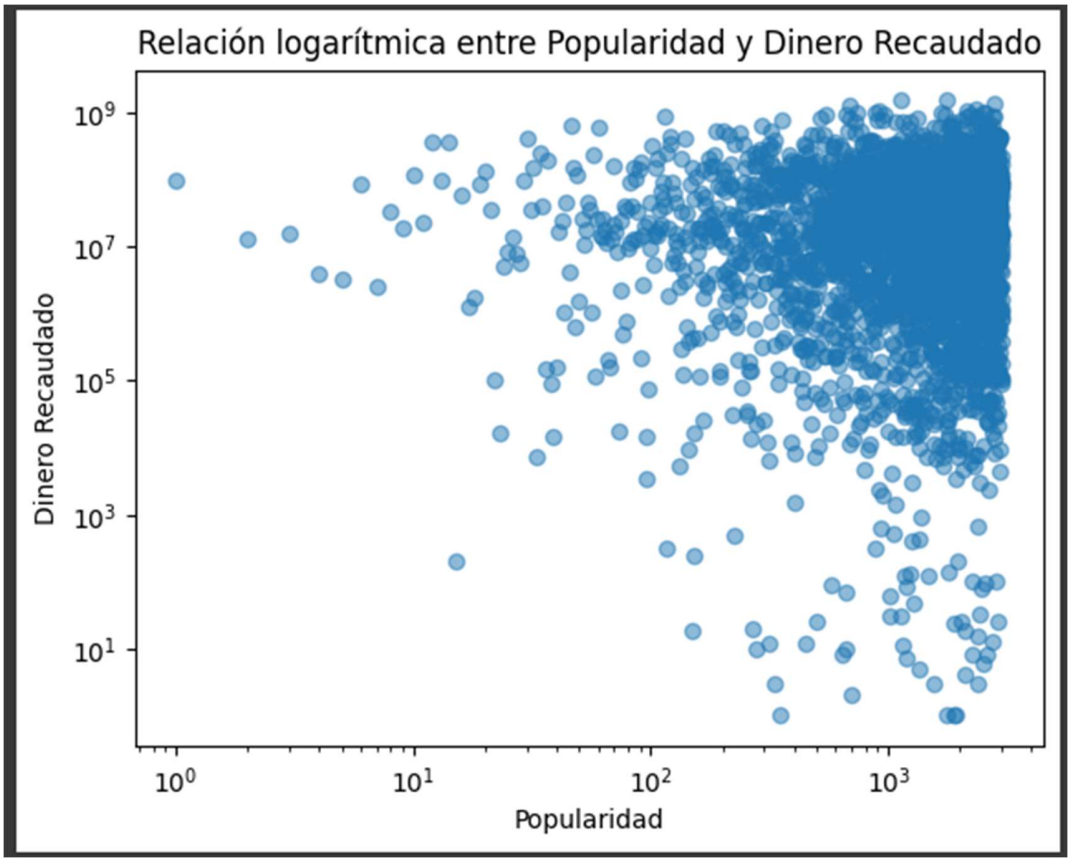
Se realizaron algunas comparaciones entre algunas de las columnas presentes para tener mejor información del comportamiento de los datos registrados en el dataset.



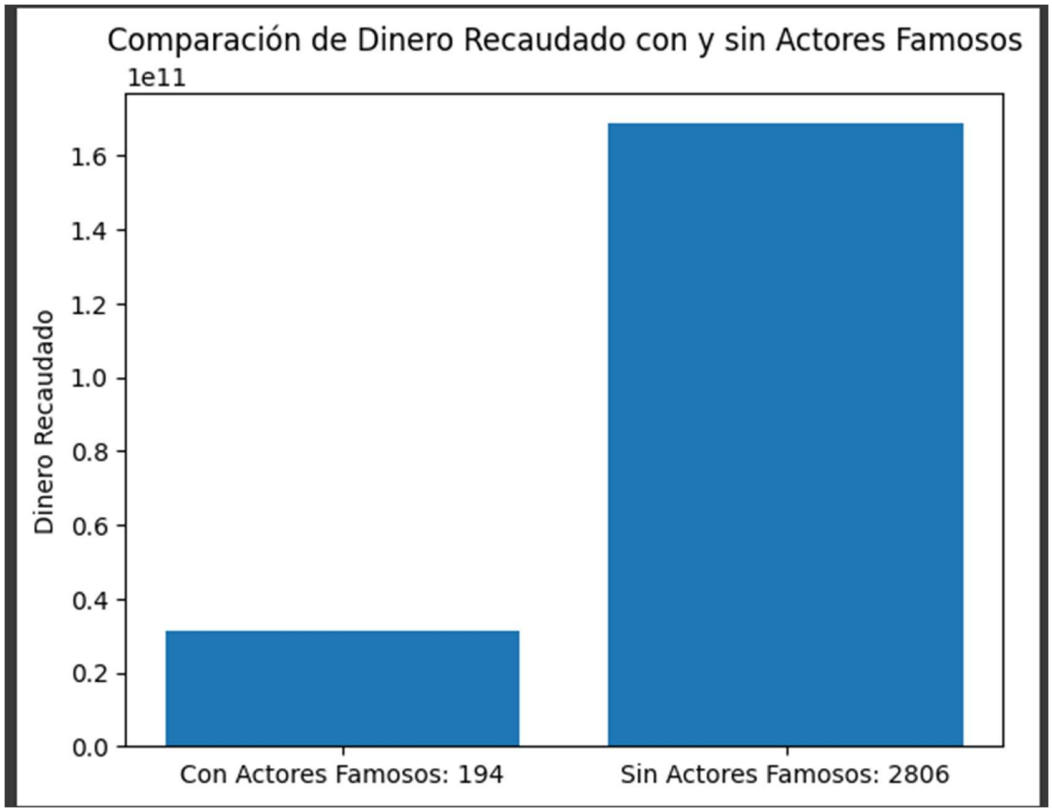
También se pudo realizar comparaciones entre dos variables que son el tema de importancia en nuestro proyecto que es hallar la recaudación según el presupuesto dedicado



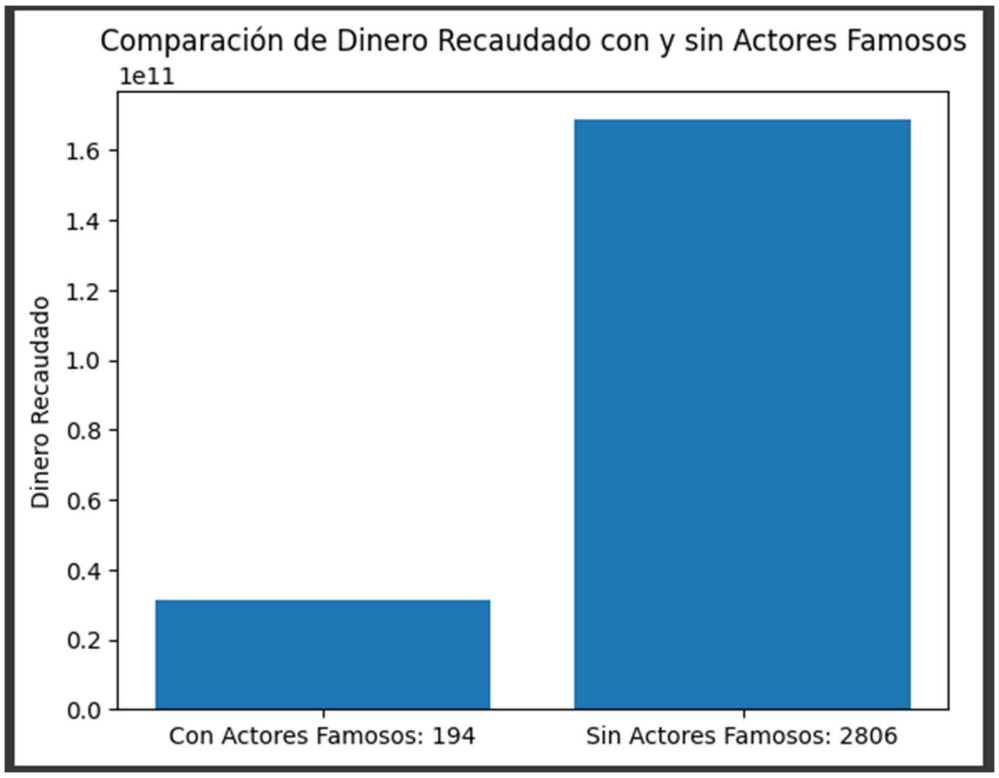
Y también la popularidad que alcanzo la cinta en un tiempo definido para realizar el promedio comparándolo con el dinero recaudado



En base a esto analizamos el dinero recaudado con y sin actores famosos

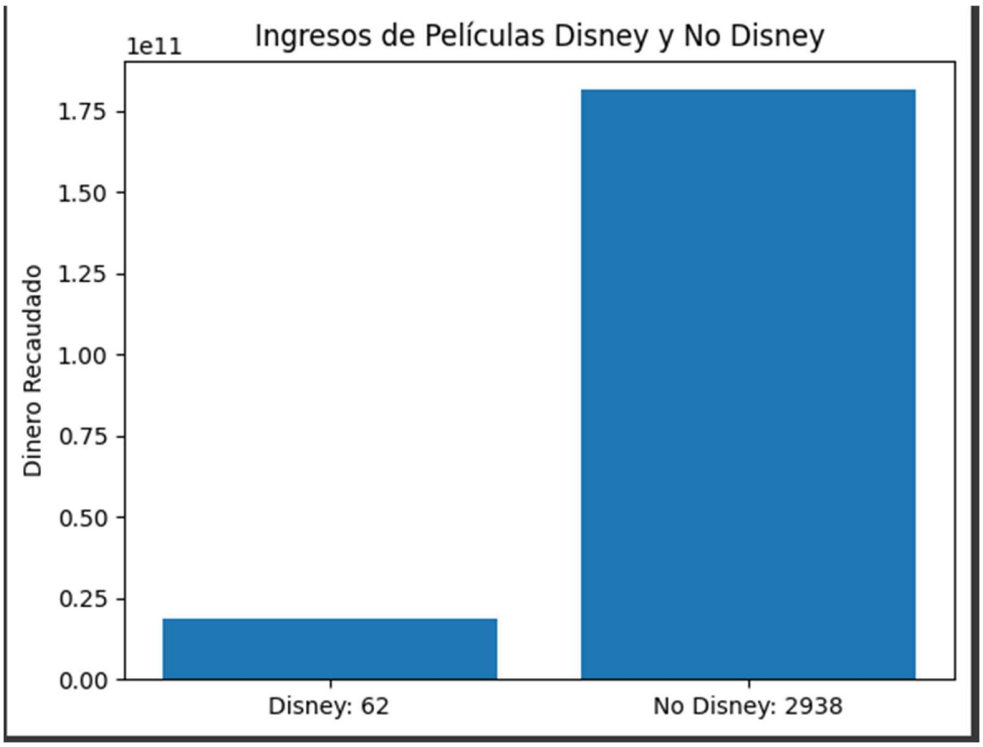


Después validamos los ingresos totales de las películas en base a su casting de actores

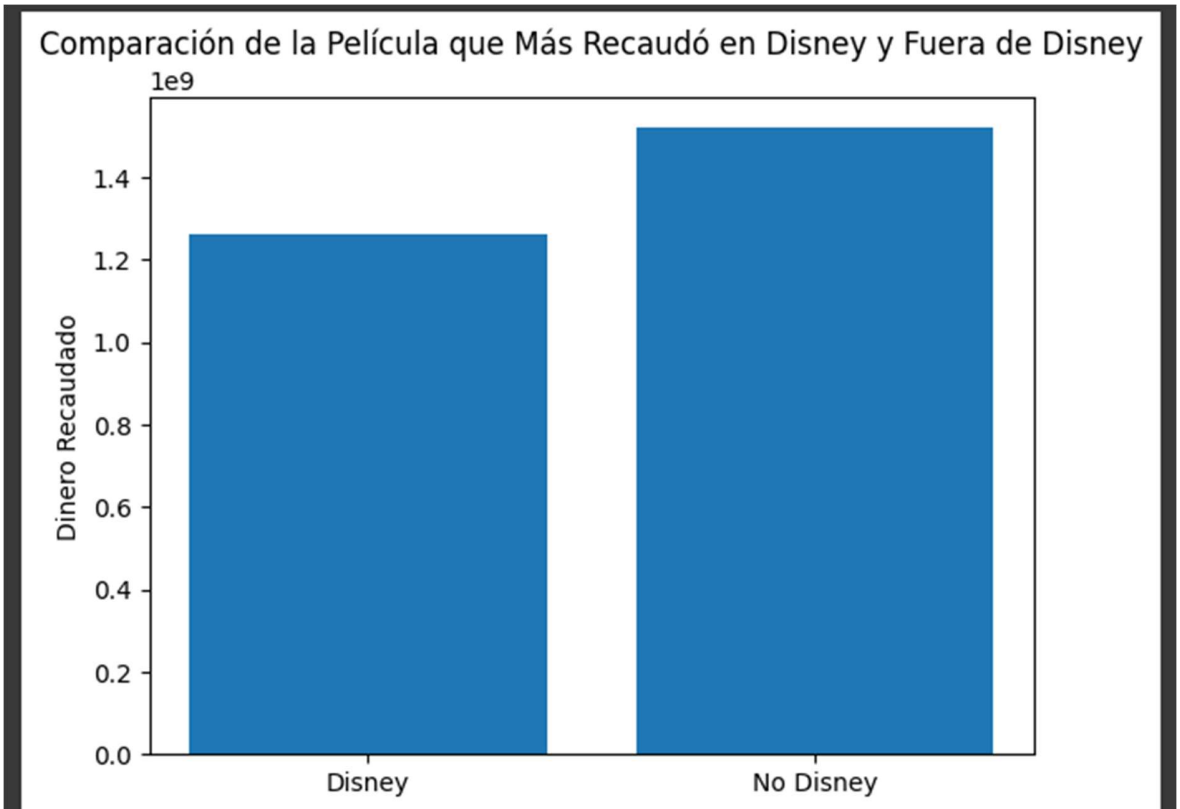


otros datos que se tuvieron en cuenta para la exploración de los datos fue la empresa que trabajo en el desarrollo de la cinta

teniendo en cuenta que Disney es de las famosas tomamos en cuenta esta como referencia para realizar la comparación



Al igual que una comparación entre las películas que se comparación en base a esta productora y notamos la diferencia entre algunas de las producciones mas famosas en contraste con las de Disney

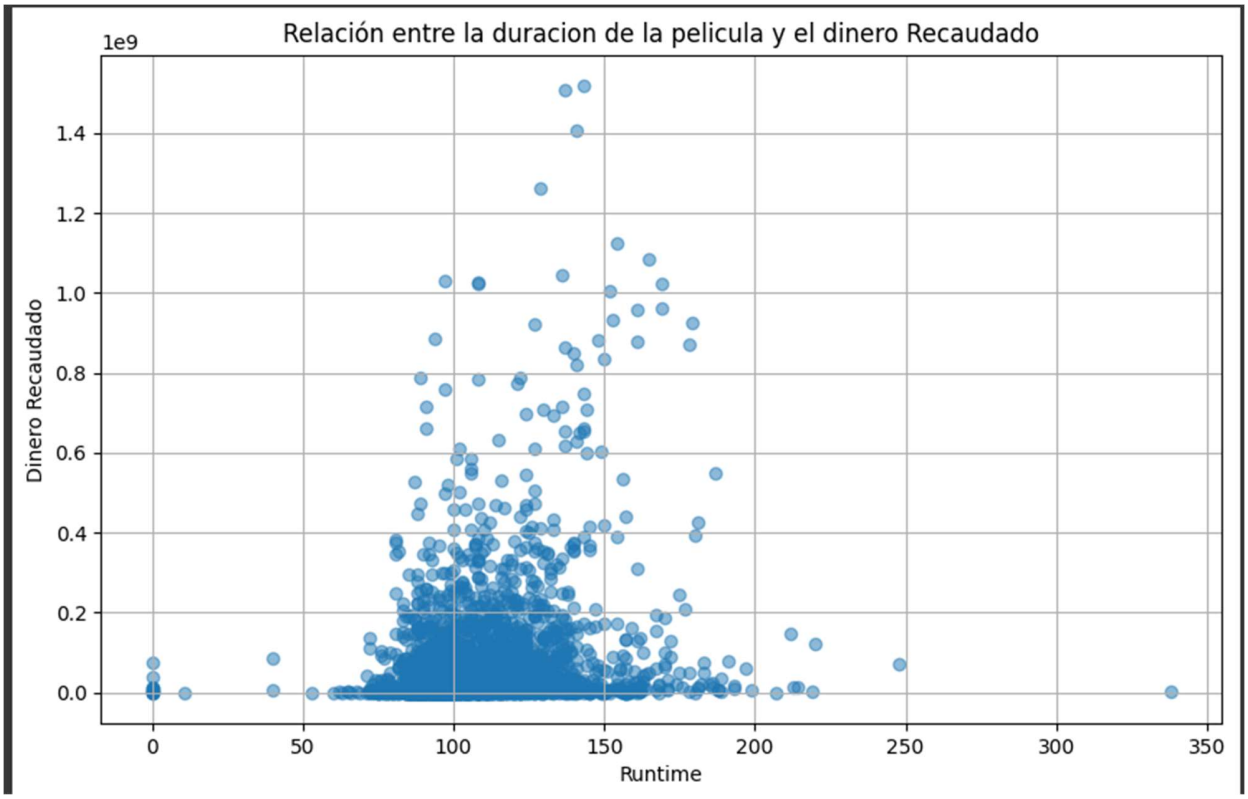



```
Película más recaudada de Disney:
title      Beauty and the Beast
revenue              1262886337
Name: 684, dtype: object

Película más recaudada fuera de Disney:
title      The Avengers
revenue      1519557910
Name: 1126, dtype: object
```

En ultima instancia se evaluó la duración de las películas para ver cuál era la preferencia de el publico en general comparado a las demás

A lo que observamos que se prefiere una duración media de 1 hora y media en la mayoría de los casos como se muestra en la grafica



3. Iteraciones de Desarrollo

Se implementaron dos modelos de inteligencia artificial: Random Forest y Support Vector Machine (SVM). Aunque ambos modelos proporcionaron resultados, ninguno alcanzó un nivel de precisión satisfactorio. Se notó que las variables consideradas menos influyentes, como el elenco y la compañía productora, no generaban un impacto significativo en la precisión del modelo. Los porcentajes de error rondaron el 61%, indicando que hay margen para mejorar el rendimiento de los modelos trabajados en el dataset que elegimos.

Modelo random forest :

Para el modelo random forest se analizará primordialmente las variables de ('budget', 'popularity', 'runtime') puesto se les considera los factores más influyentes. A partir de este entendimiento, se integrarán más columnas al modelo, esperando que mejore la calidad de la predicción

Para este modelo no aroojaron resultados obtenidos en las siguientes variables :

Error Cuadrático Medio en el conjunto de evaluación: 1229573966511585.0

Error Porcentual Absoluto Medio en el conjunto de evaluación: 83.00649221608263%

Es evidente que el error es demasiado elevado, con lo cual se concluye que a pesar de que parece que estas variables son las más influyentes a primera vista, no son suficientemente descriptivas en un ámbito general.

A partir de este entendimiento, se integrarán más columnas al modelo, esperando que mejore la calidad de la predicción

Se puede apreciar un menor porcentaje de error, sin embargo, la diferencia es mínima y este sigue siendo relativamente alto, a pesar de la inclusión de nuevas variables que parecieran ser más significativas e influyentes en el valor del dinero recaudado por estas películas

Error Cuadrático Medio en el conjunto de evaluación: 6616851962288012.0

Error Porcentual Absoluto Medio en el conjunto de evaluación: 61.45531854162958%

Modelo con SVM:

Para el modelo con SVM

Se realizará una prueba con el modelo SVM con la intención de comparar resultados.

Con este modelo nos da un resultado de:

Error Cuadrático Medio en el conjunto de evaluación: 1.9866483239291616e+16

Error Porcentual Absoluto Medio en el conjunto de evaluación: 93.01%

Lo cual es significativamente alto en comparación al anterior modelo

4. Retos y Consideraciones de Despliegue

Uno de los mayores desafíos fue el manejo de datos que se nos entregó en el datasaset, especialmente la preparación para el manejo de datos tales como el entrenamiento del modelo debido a la presencia de datos nulos. Este aspecto requirió un enfoque cuidadoso para asegurar la calidad de los datos de entrada y que nos diera un resultado optimo a la hora de trabajar en nuestro proyecto.

También a la hora de manejar este proyecto se nos dificulto un poco debido al poco conocimiento que teníamos sobre este tipo de procedimientos para el manejo de data sets como su implementación y desarrollo dentro de una herramienta como lo es colab

5. Conclusiones

La evaluación reveló que la popularidad y el presupuesto son factores más influyentes en el éxito de una película, mientras que el elenco, el género y la productora tienen un peso secundario. El modelo SVM mostró un rendimiento inferior al modelo Random Forest, lo que sugiere que la sensibilidad a la escala y la ajuste de hiperparámetros podrían ser áreas clave para mejorar.

De la presente evaluación es posible concluir que el modelo de Support Vector Machine (SVM) con el conjunto de datos proporcionado parece tener un rendimiento significativamente inferior en comparación al modelo de Random Forest. Existe una multiplicidad de factores que pueden contribuir a esta discrepancia, uno de estos radica en la sensibilidad a la escala de las características en SVM, del mismo modo, la necesidad de ajustar adecuadamente los hiperparámetros y el manejo de valores nulos. Podríamos afirmar entonces que en comparación con Random Forest, que puede ser más robusto y menos sensible a la escala, es recomendable centrarse en ajustar y mejorar el modelo de Random Forest, considerando entonces la exploración de diferentes configuraciones de hiperparámetros y asegurándose de que los datos se preparen adecuadamente previo a entrenar el modelo.

En conclusión, se recomienda centrarse en el desarrollo continuo del modelo Random Forest, explorando ajustes de hiperparámetros y asegurando una preparación adecuada de los datos. La mejora en la precisión del modelo podría lograrse mediante la optimización de factores clave identificados durante el proceso.