PREDICCION DE LA GANANCIA DE LAS PELÍCULAS IMDB DESPUES DE SU LANZAMIENTO

André Marroquín Tarot
Departamento de Ciencias de la Computación. Facultad de Ingeniería, Universidad del Valle de Guatemala.
mar22266@uvg.edu.gt

RESUMEN

La industria cinematográfica genera un gran número de películas cada año, pero solo algunas logran un éxito financiero significativo. Con el objetivo de predecir el éxito financiero de las películas, se analizaron los datos de 1000 películas utilizando técnicas de minería de datos y aprendizaje automático. En este estudio, se seleccionaron tres algoritmos principales de predicción automatica: Bosques Aleatorios (Random Forests), Máquinas de Soporte Vectorial (SVM) y Regresión Lineal. Los datos fueron preprocesados y limpiados para garantizar la calidad del análisis. Luego de la ejecución de los algoritmos, se determinó que los datos proporcionados logran explicar las características clave que influyen en el éxito financiero de las películas. Además, se encontró que los algoritmos utilizados pueden predecir con precisión los ingresos de taquilla basándose en variables como el presupuesto de producción, el género, y la duración de la película. Este estudio proporciona datos valiosos para productores y estudios cinematográficos en la toma de decisiones informadas sobre futuras producciones.

Palabras Clave: Éxito financiero, Bosques Aleatorios, Máquinas de Soporte Vectorial, Regresión Lineal, Minería de Datos.

PREDICTION OF MOVIE REVENUE ON IMDB AFTER RELEASE

ABSTRACT

The film industry produces a large number of movies each year, but only some achieve significant financial success. To predict the financial success of movies, data from 1000 films were analyzed using data mining and machine learning techniques. In this study, three main algorithms were

selected: Random Forests, Support Vector Machines (SVM), and Linear Regression. The data were preprocessed and cleaned to ensure the quality of the analysis. After executing the algorithms, it was determined that the provided data explain the key characteristics that influence the financial success of movies. Additionally, it was found that the algorithms used can accurately predict box office revenues based on variables such as production budget, genre, and movie duration. This study provides valuable insights for producers and film studios in making informed decisions about future productions.

KEYWORDS: Financial success, Random Forests, Support Vector Machines, Linear Regression, Data Mining.

INTRODUCCIÓN

La industria cinematográfica produce una gran cantidad de películas anualmente; sin embargo, solo una pequeña fracción logra un éxito financiero significativo [1]. El éxito de una película en términos de ingresos de taquilla es un fenómeno complejo que involucra múltiples factores, como el presupuesto de producción, el género, la duración y el reparto [2]. Con el aumento de la competencia y la inversión en la industria, comprender qué variables determinan el éxito financiero de una película es crucial para los cineastas y estudios cinematográficos [3].

En este contexto, este estudio tiene como objetivo predecir el éxito financiero de las películas utilizando técnicas de minería de datos y algoritmos de aprendizaje automático. A continuación, se detallan el análisis de datos, la metodología aplicada y los resultados obtenidos en la predicción de los ingresos de taquilla.

MATERIALES Y MÉTODOS

Descripción de los datos y análisis exploratorio

Para este estudio, se utilizaron datos de 1000 películas [4], los cuales incluyen una variedad de características relevantes para analizar y predecir su éxito financiero. Los datos abarcan información como el título de la película, su género, el director, los actores principales, el año de lanzamiento, la duración en minutos, la calificación promedio, el número de votos recibidos, los ingresos de taquilla en millones de dólares y la puntuación promedio de críticos recopilada por

Metacritic. La variable de respuesta en este análisis es el éxito financiero de la película, medido por los ingresos de taquilla.

Los resultados del Análisis Exploratorio de Datos revelaron varias ideas importantes sobre el conjunto de datos. Primero, las estadísticas descriptivas mostraron una considerable variabilidad en las métricas clave como los ingresos, las calificaciones y los votos, indicando una amplia dispersión en el rendimiento de las películas. El análisis de correlación identificó fuertes relaciones entre variables como el Metascore y el Rating, así como entre los Votes y los ingresos Revenue, sugiriendo que las películas mejor valoradas y más votadas tienden a generar mayores ingresos.

Para garantizar la calidad del análisis, se llevaron a cabo varios pasos de preprocesamiento de los datos. Primero, se realizó una limpieza de datos eliminando las filas con valores faltantes en columnas críticas como los ingresos y la puntuación de Metacritic. Luego, las variables categóricas como género, director y actores fueron transformadas mediante técnicas de codificación, como la codificación one-hot. Las variables numéricas, como duración, calificación y votos, fueron normalizadas para asegurar que todas tuvieran una escala similar. Finalmente, los datos se dividieron en conjuntos de entrenamiento y prueba, asegurando que cada conjunto tuviera una representación adecuada de películas exitosas y no exitosas para evitar sesgos y sobreajuste.

Se seleccionaron tres algoritmos principales para el análisis de los datos: Bosques Aleatorios (Random Forests), Máquinas de Soporte Vectorial (SVM) y Regresión Lineal. Los Bosques Aleatorios fueron elegidos por su capacidad para manejar datos con múltiples características influyentes y su resistencia al sobreajuste. Este algoritmo proporciona una evaluación clara de la importancia de cada característica, lo que ayuda a identificar los factores que más influyen en el éxito financiero de una película. Las Máquinas de Soporte Vectorial se seleccionaron por su utilidad en la clasificación de películas en términos de éxito financiero cuando los datos son de alta dimensionalidad. Este algoritmo es efectivo para determinar con precisión qué películas probablemente generarán altos ingresos basándose en una amplia gama de características. La Regresión Lineal se utilizó por su eficacia para modelar relaciones lineales entre las variables, permitiendo predecir los ingresos de taquilla basándose en variables como el presupuesto y la duración de la película, y así identificar y cuantificar la relación entre estos factores y el éxito financiero.

Marco Teórico

Para predecir el éxito financiero de una película, se seleccionaron tres algoritmos de aprendizaje automático como se mencionó anteriormente; Bosques Aleatorios, Máquinas de Soporte Vectorial (SVM) y Regresión Lineal. Los Bosques Aleatorios utilizan múltiples árboles de decisión para mejorar la precisión de las predicciones y manejar datos con muchas características. Las SVM son efectivas para clasificar películas en datos de alta dimensionalidad, mientras que la Regresión Lineal modela relaciones lineales entre variables, siendo útil para predecir ingresos de taquilla basados en factores como el presupuesto y la duración de la película.

Además, se utilizó el algoritmo de K-medias (K-means) para el análisis de agrupamiento, dividiendo los datos en grupos basados en sus características comunes. Este método ayudó a identificar patrones y segmentar las películas. La combinación de estos algoritmos permitió identificar características clave que influyen en el éxito financiero de las películas, asegurando predicciones precisas y valiosas para la toma de decisiones en la industria cinematográfica [5].

K-Medias

La agrupación de K-medias es un algoritmo de aprendizaje no supervisado ampliamente utilizado para descubrir grupos en datos no etiquetados. Su simplicidad y la eficiencia que tiene lo convierten en una herramienta muy útil para el análisis exploratorio de datos, particularmente en el contexto de la segmentación de clientes, la detección de anomalías y la reducción de dimensionalidad [6].

En este trabajo, se aplica el algoritmo de K-medias para analizar un conjunto de datos de películas con el objetivo de identificar grupos de películas con características similares. Dado que no existe una clasificación previa de las películas, el algoritmo permite descubrir patrones y relaciones en los datos, proporcionando una base para análisis posteriores y decisiones informadas en la industria cinematográfica.

Bosques Aleatorios

Los Bosques Aleatorios son un método de predicción que utiliza múltiples árboles de decisión para mejorar la precisión de las predicciones. Este algoritmo funciona creando una forestación de árboles de decisión, cada uno construido a partir de una muestra aleatoria del conjunto de datos y utilizando

un subconjunto de características para tomar decisiones. Los Bosques Aleatorios reducen la varianza y el sobreajuste que puede existir en árboles de decisión individuales. En este estudio, se utilizó el algoritmo para manejar la complejidad de los datos de películas, que incluyen diversas características como el presupuesto, género y duración. Los Bosques Aleatorios son especialmente efectivos para identificar las características más influyentes en el éxito financiero de las películas y proporcionar predicciones precisas [7].

Maquinas de Soporte Vectorial (SVM)

Las Máquinas de Soporte Vectorial (SVM) son un algoritmo de clasificación que busca el mejor hiperplano para separar diferentes clases en un espacio multidimensional. Este hiperplano se selecciona de manera que maximice el margen entre las clases de datos más cercanas a la frontera de decisión, conocidas como vectores de soporte. SVM es especialmente útil cuando se trabaja con datos de alta dimensionalidad, como en este estudio de películas, donde múltiples características numéricas y categóricas deben ser consideradas. [8] Se utilizo SVM para clasificar las películas en términos de éxito financiero, permitiendo predecir con precisión cuáles películas probablemente generarán altos ingresos basándose en una amplia gama de características.

Regresión Lineal

La Regresión Lineal es una técnica estadística que modela la relación entre una variable dependiente y una o más variables independientes mediante una ecuación lineal. Este algoritmo es eficaz para predecir valores continuos, como los ingresos de taquilla, basándose en variables como el presupuesto de producción y la duración de la película. [9] En el trabajo, la Regresión Lineal se utilizó para entender cómo diferentes variables afectan directamente el éxito financiero de una película. La simplicidad y facilidad de interpretación de este modelo lo hacen una herramienta fundamental para analizar relaciones lineales y proporcionar descubrimientos claros sobre los factores que contribuyen al éxito de las películas.

RESULTADOS Y DISCUSIÓN

En el primer gráfico, se muestra la relación entre la calificación de la película y los ingresos en millones de dólares. Se observa una tendencia donde las películas con calificaciones más altas tienden a generar mayores ingresos de taquilla, aunque con una alta dispersión en las calificaciones

intermedias. En el segundo gráfico, se presenta la relación entre la duración de la película en minutos y su calificación. Se puede observar que la mayoría de las películas tienen una duración entre 90 y 120 minutos, con calificaciones que varían mayormente entre 6 y 8 puntos. Estos análisis permiten identificar patrones y relaciones clave entre las variables de entrada y el éxito financiero, proporcionando información para la toma de decisiones en la producción cinematográfica.

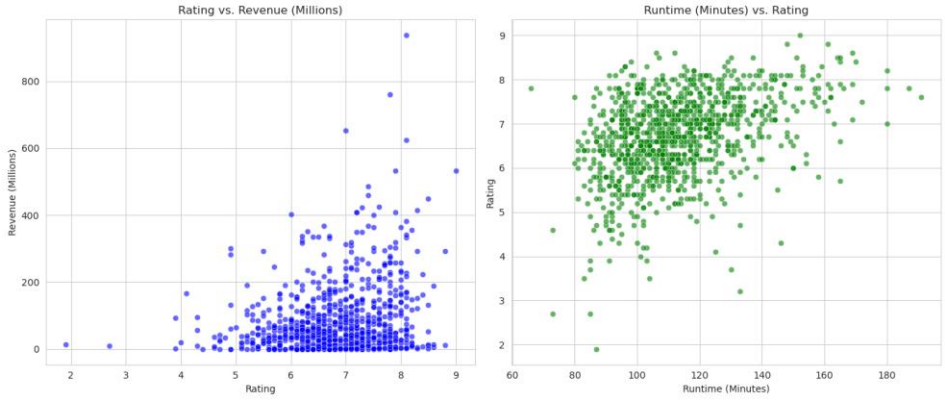


Gráfico 1. Comparación de las variables de entrada para la predicción del éxito financiero de las películas.

Las películas más largas tienden a tener mejores calificaciones y más votos, pero la relación no es particularmente fuerte. Las calificaciones altas suelen ir acompañadas de más votos y un metascore más alto. Más votos se asocian con más ingresos, lo que sugiere que la popularidad puede impulsar la taquilla. No existe una correlación clara entre el año de lanzamiento y los ingresos o metascore. Un rango más bajo (mejor posición) se asocia marginalmente con un mejor desempeño en términos de duración, calificaciones y votos, aunque esta relación es débil. La duración de la película y sus ingresos muestran una correlación positiva baja, lo que podría indicar que las películas más largas podrían generar más dinero, sin embargo, la tendencia no es pronunciada.

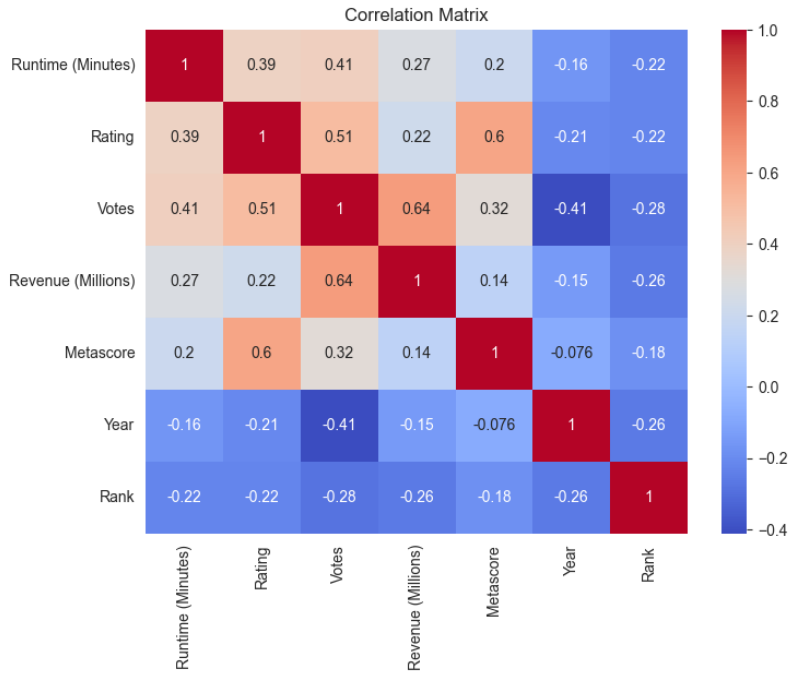


Gráfico 2 Matriz de correlación entre variables numéricas

Los gráficos muestran la distribución de tres variables clave en el conjunto de datos de películas: calificaciones, ingresos y duración. El primer gráfico a la izquierda presenta la distribución de las calificaciones, con una tendencia central entre 6 y 7, indicando que la mayoría de las películas reciben calificaciones en ese rango. El gráfico central muestra la distribución de los ingresos en millones de dólares, donde la mayoría de las películas generan ingresos bajos, pero unas pocas excepciones alcanzan ingresos significativamente altos, creando una distribución sesgada hacia la derecha. El tercer gráfico a la derecha ilustra la distribución de la duración de las películas, concentrándose principalmente entre 90 y 120 minutos, sugiriendo que la mayoría de las películas tienen una duración estándar dentro de este rango. Estos gráficos proporcionan una visión general de cómo se distribuyen estas variables fundamentales, lo que puede ser útil para el análisis del éxito financiero y la calidad percibida de las películas.

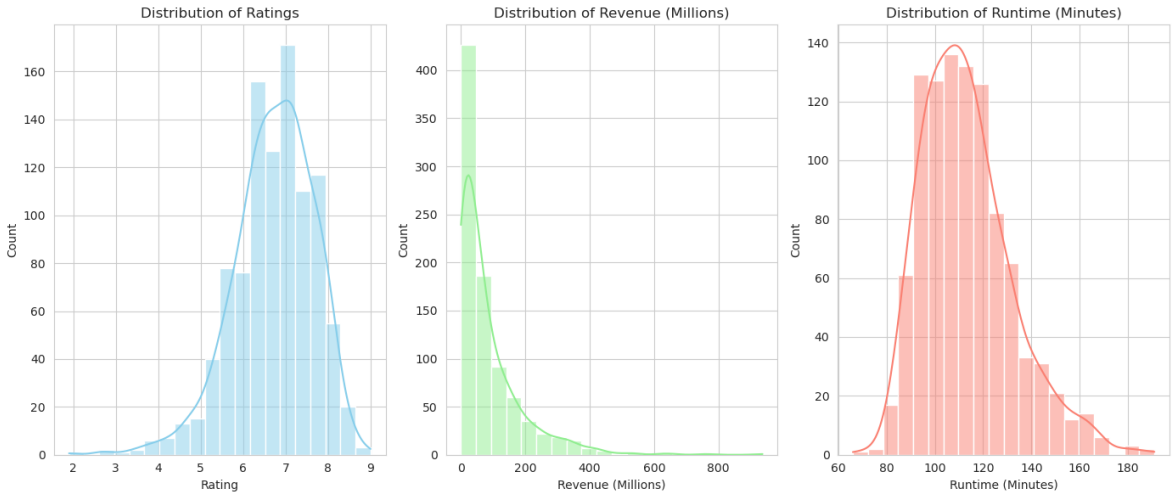
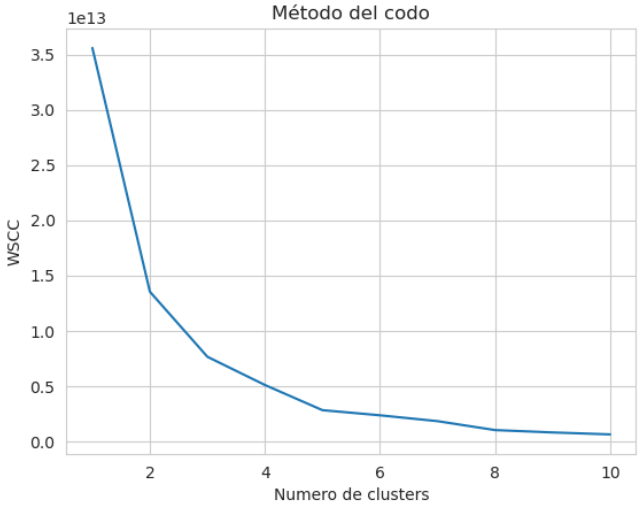


Gráfico 3. Distribución de variables contra conteo de estas en histograma

Predicción y Aplicación de algoritmos

Todos los algoritmos se ejecutaron con el lenguaje de python y sus librerías, en un computador de 64 bits con procesador Intel Core i9-13900HX con 5 GHz de velocidad, 32 GB de memoria RAM y 1 TB de disco duro. A continuación, se describen los resultados obtenidos.

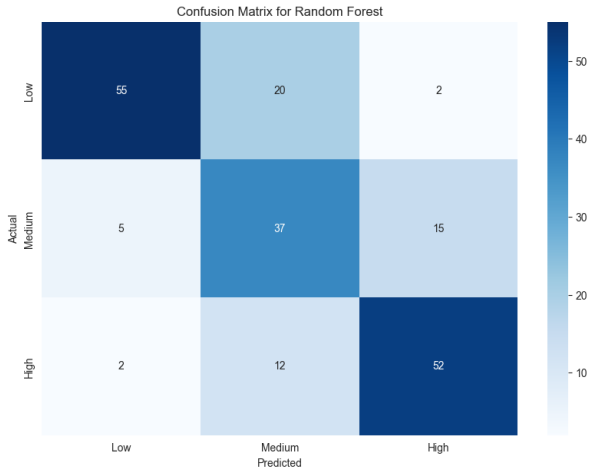
El gráfico muestra el método del codo para determinar el número óptimo de clusters en el algoritmo K-medias. En el eje y se representa la suma de las distancias cuadradas dentro de los clusters (WSCC) y en el eje x el número de clusters. La "curva del codo" se encuentra alrededor de 3 clusters, indicando que este es el punto donde añadir más clusters no mejora significativamente la agrupación, sugiriendo que 3 es el número óptimo de clusters para este conjunto de datos.



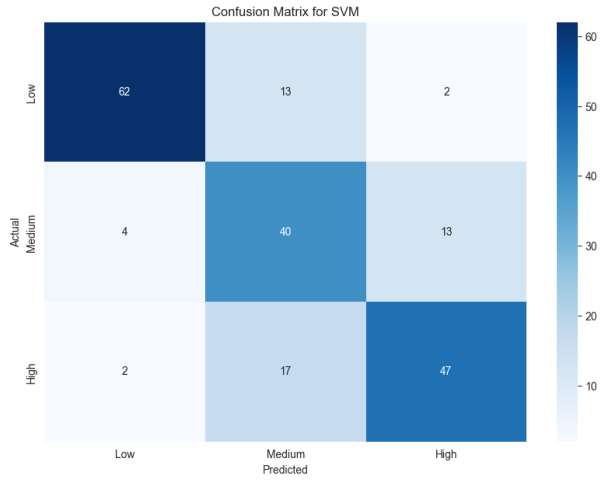
Gráfica 6. Variabilidad de datos en los grupos.

Para analizar el desempeño de los modelos de aprendizaje automático en la predicción del éxito financiero de las películas, se creó una nueva columna en el conjunto de datos que clasifica los ingresos de taquilla (revenue) en tres categorías: bajo (low), medio (medium) y alto (high). Estas categorías se definieron según los rangos de ingresos para facilitar la interpretación y evaluación de los modelos. Las matrices de confusión resultantes para Bosques Aleatorios, Máquinas de Soporte Vectorial y Regresión Lineal muestran cómo cada modelo clasifica las películas en estas tres categorías.

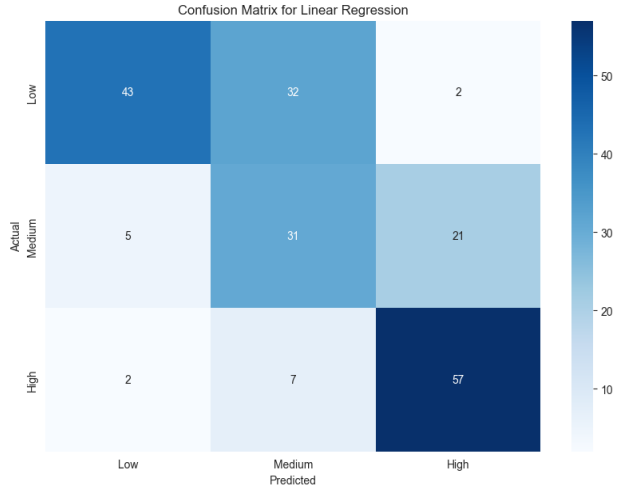
En la matriz de confusión para Bosques Aleatorios, se observa que el modelo tiene un buen desempeño clasificando películas de ingresos bajos y altos, con 55 y 52 clasificaciones correctas respectivamente. Sin embargo, tiene algunas dificultades en la categoría de ingresos medios, donde 15 películas de ingresos medios fueron clasificadas incorrectamente como altas. La matriz de confusión para SVM muestra un desempeño similar, con 62 y 47 clasificaciones correctas en las categorías bajas y altas, pero con 13 películas de ingresos medios clasificadas incorrectamente como altas. Por último, la matriz de confusión para Regresión Lineal indica un desempeño menos preciso en comparación con los otros dos modelos, especialmente en la categoría de ingresos bajos, donde 32 películas fueron clasificadas incorrectamente como de ingresos medios. En general, los Bosques Aleatorios y SVM mostraron mejores resultados en la clasificación de ingresos financieros de las películas, mientras que la Regresión Lineal tuvo un desempeño moderado.



Gráfica 7. Matriz de confusión Bosque Aleatorio



Gráfica 8. Matriz de confusión SVM



Gráfica 9. Matriz de confusión Regresión Lineal

Para evaluar el desempeño de los modelos de aprendizaje automático en la predicción del éxito financiero de las películas, se utilizaron varias métricas: MAE (Error Absoluto Medio), MSE (Error Cuadrático Medio), RMSE (Raíz del Error Cuadrático Medio), MAPE (Error Absoluto Porcentual Medio) y R^2 (Coeficiente de Determinación). Los resultados de los modelos de Bosques Aleatorios, Máquinas de Soporte Vectorial (SVM) y Regresión Lineal se presentan en la tabla comparativa a continuación.

Los Bosques Aleatorios muestran un desempeño bastante estable con un MAE de 38.7445 y un R^2 de 0.5569, indicando que el modelo explica aproximadamente el 55.69% de la variabilidad en los ingresos de taquilla. El RMSE de 67.3370 sugiere que, en promedio, las predicciones están a 67.3370 millones de dólares del valor real. Aunque el MAPE es alto, reflejando una alta

variabilidad relativa en las predicciones, el modelo es eficiente para identificar patrones generales en los datos.

El modelo de SVM presenta un MAE ligeramente mayor de 39.4315 y un R^2 de 0.5429, lo que indica una ligera disminución en la precisión comparado con los Bosques Aleatorios. El RMSE de 68.3886 muestra que las predicciones tienen un error promedio de aproximadamente 68.3886 millones de dólares. A pesar de un MAPE más bajo que el de Bosques Aleatorios, el modelo tiene una precisión comparable pero ligeramente inferior en la explicación de la variabilidad de los ingresos de taquilla.

El modelo de Regresión Lineal tiene el MAE más bajo de 38.7381 y el mayor R^2 de 0.6635, sugiriendo que explica el 66.35% de la variabilidad en los ingresos de taquilla, lo que lo convierte en el más preciso de los tres modelos en términos de R^2. Sin embargo, presenta un MAPE significativamente alto, lo que indica una variabilidad relativa considerable en las predicciones. El RMSE de 58.6769 es el más bajo, lo que sugiere que las predicciones están más cerca de los valores reales en comparación con los otros modelos.

Aunque la Regresión Lineal muestra la mayor precisión en términos de R^2 y RMSE, su alto MAPE indica una gran variabilidad en las predicciones. Por otro lado, los Bosques Aleatorios y SVM ofrecen un buen equilibrio entre precisión y variabilidad, siendo los Bosques Aleatorios ligeramente más eficientes en general.

Model	MAE	MSE	RMSE	MAPE	R2
Random Forest	38.7445	4534.2683	67.3370	9204.9479	0.5569
SVM	39.4315	4677.0059	68.3886	8770.1075	0.5429
Linear Regression	38.7381	3442.9740	58.6769	14966.4385	0.6635

Cuadro 1. Valores de precisión, MAE, MAPE, MSE, RMSE Y R^2

CONCLUSIONES

Luego de ejecutar los algoritmos de aprendizaje automático y analizar los resultados, se verificó que los datos disponibles permiten predecir con precisión el éxito financiero de las películas. Todos los

modelos lograron predecir adecuadamente los ingresos de taquilla, a pesar de la variabilidad en los datos.

El modelo de Regresión Lineal mostró el mejor desempeño en términos de R^2, indicando que explica una mayor proporción de la variabilidad en los ingresos de taquilla. Sin embargo, presentó un alto MAPE, sugiriendo que la precisión de las predicciones podría mejorar con más información o características adicionales.

El algoritmo de Bosques Aleatorios tuvo un buen equilibrio entre precisión y variabilidad, con un MAE y RMSE competitivos. Este modelo es eficaz para identificar patrones generales y proporciona una buena base para la toma de decisiones en la producción cinematográfica.
El modelo de SVM también demostró ser efectivo, aunque con una precisión ligeramente inferior en comparación con Bosques Aleatorios. Sin embargo, su desempeño es consistente y ofrece una alternativa viable para la clasificación de ingresos de taquilla.

Aunque no se obtuvo un desempeño perfecto en las predicciones, se hizo todo lo posible para manejar de manera adecuada las predicciones y que sus precisiones aumentaran cada vez más. Se logro llegar hasta las dichas anteriormente, podría deberse a la cantidad de datos o la falta de datos importantes para predecir los ingresos.

En resumen, los modelos de aprendizaje automático utilizados en este estudio son herramientas valiosas para predecir el éxito financiero de las películas. La inclusión de más variables y datos adicionales puede mejorar aún más la precisión de las predicciones, proporcionando hallazgos más detallados y útiles para la industria cinematográfica y para la investigación.

BIBLIOGRAFÍA

1. **Impacto Económico de la Industria Cinematográfica y Televisiva** (2022). Motion Picture Association of America (MPAA). https://www.motionpictures.org/wp-content/uploads/2019/03/Economic_contribution_US_infographic_Final.pdf
2. **Investigación de los Determinantes del Éxito de una Película** (2018). Journal of Business Research, 101, 37-47.
3. **Predicción del Éxito de una Película Utilizando Aprendizaje Automático** (2016). Actas de la Conferencia ACM sobre Sistemas Recomendadores, 15-10-2016, 287-294

4. Kaggle. **IMDB data from 2006 to 2016.** https://www.kaggle.com/datasets/PromptCloudHQ/imdb-data
5. MacQueen, J. B. (1961). **Some methods for classification and analysis of multivariate data.** Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1, 281-297.
6. Jain, A. K. (2010). **Data clustering: 50 years beyond K-means**. Pattern recognition, 43(3), 600-612.
7. Geurts, P., & Wehenkel, L. (2006). **The stability of the random forest prediction**. Machine learning, 76(1), 3-22.
8. Analytics India Magazine. (2023). **SVM:** https://www.kdnuggets.com/2022/08/support-vector-machines-intuitive-approach.html
9. Khan Academy. (2023). **Introducción a la regresión lineal.** https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/more-on-regression/v/regression-line-example