UNIVERSIDAD TEGNOLÓGICA DE PANAMÁ

LICENCIATURA EN INGENIERIA EN SISTEMA Y COMPUTACIÓN

PROBABILIDAD APLICADA A TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIÓN

CINEMETRICS: MODELADO PREDICTIVO DE CALIFICACIÓNES CINEMATOGRÁFICAS CON MACHINE LEARNING

Estudiantes:
Aldahir Aguilar 8-1029-1115
Andrés Flores 8-1025-1254
Diego García 8-1034-95
Luis Torné 8-1032-1644
Rashell Vidal 8-1028-643

Profesor:
Juan Marcos Castillo, PhD

Salón: 1IL124

30 de julio del 2025

ÍNDICE

INTRODUCCIÓN

El presente proyecto tuvo como propósito desarrollar un modelo predictivo capaz de estimar el posible éxito de una película próxima a estrenarse, utilizando variables clave como el género, la participación de directores y actores principales, sus calificaciones previas y los premios obtenidos. Esta propuesta surgió del interés por aplicar métodos estadísticos y de análisis comparativo a un área cercana y motivadora como el cine.

Para alcanzar este objetivo, se construyó una base de datos personalizada a partir de información extraída de The Movie Database (TMDb), la cual permitió recopilar más de 40,000 registros con características relevantes de cada producción. Esta base fue limpiada, organizada y convertida a un formato adecuado para su análisis.

Durante el desarrollo del proyecto se clasificaron las variables en cualitativas y cuantitativas, se aplicaron estadísticas descriptivas y se generaron visualizaciones gráficas que ayudaron a identificar tendencias y relaciones significativas, como la influencia del género o del prestigio del elenco en el desempeño de una película. Asimismo, se realizaron pruebas con modelos predictivos básicos, con el fin de explorar la viabilidad del enfoque propuesto.

En conjunto, este trabajo integró técnicas de análisis de datos, probabilidad aplicada y razonamiento lógico, permitiendo no solo acercarse a una posible predicción del éxito cinematográfico, sino también fortalecer habilidades esenciales para la práctica profesional.

JUSTIFICACIÓN

1. Identificación del Problema o la Oportunidad

El cine es una de las industrias más dinámicas del entretenimiento, pero también una de las más impredecibles. Muchas producciones cuentan con grandes inversiones y no logran el impacto esperado, mientras que otras con menor promoción logran gran éxito. Esta situación presenta una oportunidad de análisis: aplicar herramientas comparativas para predecir el posible rendimiento de una película próxima a estrenarse, basándonos en patrones de producciones anteriores. Además, como grupo, identificamos una oportunidad de trabajar con un tema que nos interesa y motiva: el cine.

2. Objetivos y Resultados Esperados

Objetivo general:

- Predecir el posible éxito de una película por estrenarse, utilizando un análisis comparativo basado en variables clave.

Objetivos específicos:

- Identificar factores comunes en películas exitosas.
- Comparar dichos factores con los de la película seleccionada.
- Estimar la probabilidad de éxito de la película en base a la opinión del público.

Resultados esperados:

- Un modelo comparativo que relacione variables como género, elenco, directores, clasificación, entre otras.
- Un informe que presente conclusiones claras sobre el potencial éxito de la película.

3. Evaluación de Beneficios y Costos

Beneficios:

- Aplicación práctica de análisis comparativo y razonamiento lógico.
- Desarrollo de habilidades de interpretación de datos reales.
- Mayor motivación al trabajar con un tema de interés común.
- Aporte al entendimiento de cómo ciertos factores influyen en el éxito cinematográfico.

Costos:

- Tiempo invertido en la recolección y análisis de datos.
- Recursos tecnológicos como software para manipulación de datos (ya disponibles).

  Dado que los costos son bajos y los beneficios educativos y analíticos son altos, el proyecto resulta favorable.

4. Análisis de Viabilidad

- Técnica: Contamos con las herramientas necesarias (Excel, internet, bibliografía) y los conocimientos para realizar análisis comparativos.
- Económica: No se requiere inversión monetaria significativa.
- Operativa: El grupo trabaja de forma colaborativa y se han distribuido las tareas para cumplir con los plazos establecidos.

  Por lo tanto, el proyecto es viable en sus dimensiones meramente analíticas y educativas.

5. Exploración de Alternativas

Se consideraron otras temáticas como predicciones deportivas o análisis de tendencias musicales, pero se decidió por el cine porque representa un área de interés compartido por todos los integrantes y permite trabajar con datos accesibles y variados. Esta elección favorece el compromiso y la comprensión del análisis.

6. Alineamiento Estratégico

Este proyecto se alinea con los objetivos formativos de nuestra carrera, ya que integra análisis de datos, razonamiento lógico y trabajo en equipo. También promueve la aplicación práctica de conceptos aprendidos en clases, fortaleciendo nuestras habilidades de investigación y presentación.

7. Análisis de Impacto

El impacto esperado es positivo tanto para el aprendizaje del grupo como para futuras aplicaciones del análisis comparativo en otros contextos. Además, el proyecto tiene el potencial de despertar interés en el uso de datos en áreas no tradicionales como el cine. Los riesgos son mínimos, y se gestionan a través de una buena planificación y comunicación en el grupo.

ANTECEDENTES

Durante el desarrollo del presente proyecto, se identificó una limitación importante: la inexistencia de una base de datos pública que se ajustara específicamente a los objetivos del análisis probabilístico y estadístico planteado. A raíz de esta necesidad, se tomó la decisión de construir una base de datos propia y personalizada, orientada al estudio del cine y sus distintas variables relevantes.

Para ello, se realizó una breve investigación exploratoria que permitió identificar a The Movie Database (TMDb) como una fuente confiable y robusta de información cinematográfica. Gracias a su sistema de acceso mediante API, fue posible diseñar un script en Python encargado de automatizar la recolección de datos. Este script fue configurado para extraer únicamente las variables pertinentes al proyecto, tales como título, año, género, calificaciones, estudio, actores principales, directores, y detalles sobre nominaciones y premios. El proceso de recolección fue extenso y requirió cerca de ocho horas de ejecución para compilar una base de datos compuesta por más de 40,000 registros, posteriormente almacenados en un archivo SQL estructurado.

Una vez generada la base de datos, se procedió a su análisis y limpieza. Esto implicó la verificación de formatos, la eliminación de registros incompletos o inconsistentes y la normalización de valores, con el fin de garantizar que la información fuera clara, coherente y lista para su análisis estadístico. Durante este proceso también se exploraron diversas herramientas de visualización, lo cual permitió identificar patrones y relaciones entre las variables.

Finalmente, en respuesta a las necesidades del equipo y como un reto adicional, se desarrolló un segundo script para convertir la base de datos SQL a formato Excel, facilitando así su manipulación, análisis gráfico y la elaboración de estudios probabilísticos por parte de todos los integrantes del proyecto.

DEFINICIÓN DEL PROBLEMA

La industria cinematográfica contemporánea, marcada por la revolución digital y el auge de las plataformas de streaming, enfrenta un desafío fundamental: la impredecibilidad del éxito de sus producciones. En un mercado globalizado donde cada lanzamiento representa inversiones millonarias y compite por la atención de audiencias cada vez más fragmentadas, la capacidad de anticipar el desempeño de una película se convierte en un factor estratégico clave. Este proyecto al que nombramos como “CineMetrix” se enfoca en desarrollar un modelo predictivo que permita estimar el éxito cinematográfico, utilizando como variable principal el rating de audiencia en IMDb, considerado un indicador confiable de aceptación popular.

Para construir este modelo, analizaremos variables fundamentales como el reconocimiento de premios del director y actores principales, características esenciales de la producción como género cinematográfico y clasificación por edad, incluyendo también presupuesto y año de lanzamiento. La selección de estas variables se basa en estudios previos que identificaron su correlación con el rendimiento de las películas, así como en su disponibilidad para análisis cuantitativo.

Este problema adquiere una doble relevancia, tanto académica como práctica. Por un lado, permite explorar la aplicación de técnicas estadísticas y de machine learning en un dominio cultural tradicionalmente dominado por criterios subjetivos, aportando un enfoque cuantitativo innovador. Por otro, los resultados ofrecen herramientas valiosas para productores, distribuidores y plataformas de streaming, quienes podrían utilizar el modelo predictivo para optimizar sus decisiones de inversión, marketing y distribución.

La investigación busca responder preguntas clave mediante el uso de algoritmos de machine learning como: ¿Qué combinación de factores artísticos como el prestigio del director o el reparto y comerciales como el género y clasificación por edad influyen más en la recepción del público según IMDb? ¿Existe un patrón discernible entre el reconocimiento de premios y el éxito popular? ¿Pueden modelos supervisados, como regresión lineal avanzada o random forests, predecir con precisión aceptable el rendimiento de una película? La implementación de estas técnicas no solo mejorará la precisión del modelo, sino que también permitirá identificar relaciones no lineales y patrones ocultos en los datos, ofreciendo percepciones más robustas en el mundo del entretenimiento.

ANÁLISIS CON DIFERENTES MODELOS ESTOCÁSTICOS

a. Determinación de la base de datos

Se utiliza una base de datos con información de películas, incluyendo características como:

- actor_rating, actor_nominado, actor_ganador
- año y mes de lanzamiento
- clasificación por edad, género
- actor principal (reducido a los 50 más frecuentes)
- imdb_pelicula (variable objetivo, continua entre 1 y 10)

Fuente: archivo CSV previamente cargado (Base_de_datos_peliculas.csv)

b. Preprocesamiento y limpieza

- Se eliminaron filas con valores nulos en las variables relevantes.
- Se filtraron los actores principales a los 50 más frecuentes para reducir dimensionalidad.
- Se codificaron variables categóricas con One Hot Encoding (clasificación, género, actor).
- Se dividió en entrenamiento (80%) y prueba (20%).
- Se aplicó estandarización (StandardScaler) para modelos sensibles a escala (SVR, redes neuronales).

c. Análisis descriptivo

- Se generaron gráficas KDE y mapas de correlación para las variables numéricas.
- imdb_pelicula presenta una distribución sesgada hacia valores entre 6 y 8.
- Se observaron correlaciones moderadas entre actor_rating, actor_nominado y imdb_pelicula.
- La mayoría de películas pertenecen a géneros populares y clasificación PG-13 o R.

Se pueden incluir estos gráficos si deseas que los genere nuevamente.

d. Selección de variables

Variables predictoras seleccionadas:

- Numéricas: actor_rating, actor_nominado, actor_ganador, año, mes
- Categóricas codificadas: clasificación_edad, género, actor_principal (top 50)

Variable objetivo:

- imdb_pelicula (regresión continua)

e. Selección de modelos estocásticos

Se compararon varios modelos, entre ellos algunos estocásticos o que incluyen componentes de aleatoriedad:

Modelo	Tipo	Estocástico	Observaciones clave
Linear Regression	Determinístico	No	Bajo riesgo de sobreajuste
Ridge / Lasso	Regularizado	No	Mejor generalización
Support Vector (SVR)	Estocástico	Parcial	Necesita escalado, costoso computacional
Random Forest	Estocástico	Sí	Preciso pero propenso a sobreajuste
Decision Tree	Estocástico	Sí	Fácil de interpretar, sensible a ruido
MLP Regressor (NN)	Estocástico	Sí	Alto riesgo de sobreajuste

Se midieron:

- R^2 en entrenamiento y prueba
- Error absoluto medio (MAE)

Ya determinamos que Random Forest y MLP muestran indicios de sobreajuste.

CONCLUSIONES

1. En lo personal fue un trabajo de mucho estrés y presión, pero el trabajo en equipo, la coordinación y el apoyo de los integrantes puede dar un buen trabajo si no los proponemos, fue una experiencia enriquecedora y sabemos que este pequeño desafío nos ayudara a ser mejores profesionales y estar preparados para el tedioso "**Campo Laboral**".
2. Random Forest fue el modelo más efectivo. Se concluyó que el modelo Random Forest superó a todos los demás en términos de desempeño predictivo, alcanzando los mejores valores de R^2 (0.724) y MAE (0.462). Esto demuestra que es una herramienta sólida para predecir calificaciones de películas en IMDb, especialmente cuando se trata de calificaciones altas.
3. Existe un componente de error aleatorio inevitable. Se reconoció que, aun con buenos modelos y variables, siempre hay incertidumbre en la predicción debido a factores no observados (como el contexto cultural o la fama momentánea de un actor). Por eso, se asumió la existencia de un componente estocástico, típico en todo modelo de regresión.
4. El análisis por subgrupos reveló patrones importantes. La comparación de predicciones por género, actor y director permitió identificar sesgos sistemáticos, como la subestimación del género terror o diferencias puntuales en actores con alta variabilidad. Este análisis granular ayudó a mejorar la interpretación del modelo y su comportamiento en casos específicos.
5. Las métricas estadísticas facilitaron la evaluación objetiva. Se utilizaron métricas como R^2, MAE y RMSE para evaluar la calidad del modelo desde un enfoque probabilístico. Estas medidas ofrecieron una visión cuantitativa y comparativa del rendimiento, permitiendo validar empíricamente la robustez y estabilidad del modelo ante distintos conjuntos de datos.

RECOMENDACIONES

1. Gestionar el tiempo de manera equitativa y favorable.
2. La comunicación del equipo es importante en todos los aspectos de este proyecto.
3. Definir sus variables y tener claro que es lo que desean lograr con el modelo predictivo y el proyecto en sí.
4. Documentar cada uno de los avances, archivos, bibliografías y todo lo que se logre y se utilice a lo largo de la realización del proyecto.
5. En lo que conlleva a la visualización, está claro que una base de datos en tabla es mucho menos atractiva y algo complicada de entender a simple vista, así que la realización de gráficos o alguna otra manera más simple de visualizar tal cantidad de datos es una gran opción y ayuda para la comprensión del equipo y todo aquel que desee evaluar o entender el proyecto realizado.

BIBLIOGRAFÍA

https://api.themoviedb.org/3

(además de las aclaraciones proporcionadas por una IA, tuvimos una ayuda externa de una persona, que nos guio para conseguir los mejores resultados en el modelo predictivo, mejorar nuestros script y hacer lo posible para lograr un gran trabajo)

ANEXOS

https://github.com/MichiStar25/Cinemetrix-Proyecto-de-Probabilidad.git