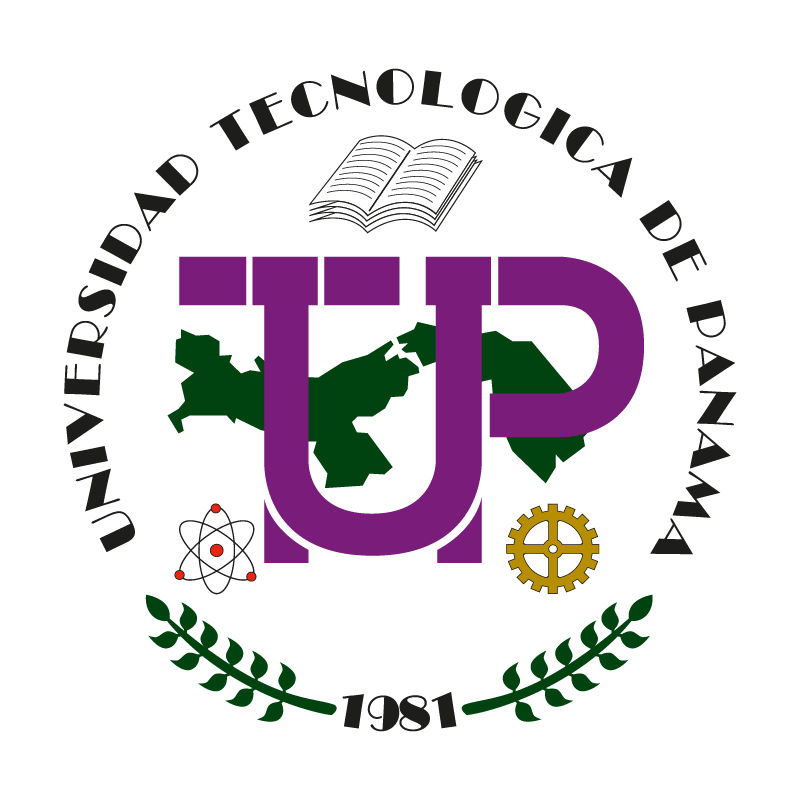
UNIVERSIDAD TEGNOLÓGICA DE PANAMÁ 

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

La base de datos de este proyecto contiene información detallada sobre distintas películas, con el objetivo de analizar los factores que influyen en su éxito, medido a través del puntaje otorgado por usuarios en TMDb. Cada fila representa una película diferente, y se incluyen variables que describen aspectos artísticos y comerciales.

Entre las variables más importantes se encuentran:

* genero, que indica el tipo de película (como comedia, drama o acción).
* clasificacion\_edad, que señala si es apta para todo público o solo para adultos.
* actor\_nominado y director\_ganador, que muestran si los protagonistas han sido reconocidos con premios.
* anio y mes, que indican cuándo fue lanzada la película.
* y la variable principal imdb\_pelicula, que muestra su calificación promedio en IMDb.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Tratamiento en el código** | **Tipo estadístico** | **Uso en probabilidad** |
| id, titulo, actor\_principal, director, estudio | Eliminadas del modelo | Categóricas sin valor cuantificable | No se consideran; no aportan a inferencias de éxito |
| anio y mes | Incluidas directamente | Numéricas (discretas) | Permiten estimar si hay una probabilidad mayor de éxito en ciertos años o meses (por ejemplo, películas estrenadas en febrero o en 2020) |
| clasificacion\_edad | Transformada a binaria con edad\_adulto (1 si es 18+) | Categórica binaria | Permite evaluar si hay más éxito cuando una película es para adultos o para todo público. P(éxito) |
| genero | Codificado con *MultiLabelBinarizer* a múltiples columnas como genre\_Comedia, genre\_Drama, etc. | Variables dummy | Permite estimar la probabilidad condicional de éxito por género, como P(imdb > 7) |
| actor\_rating y director\_rating | Eliminadas del modelo final | Numéricas (float) | Aunque fueron eliminadas, son variables altamente correlacionadas con el éxito. Pueden estimarse distribuciones como:  P (imdb > 8 actor\_rating > 7) |
| actor\_nominado, actor\_ganador, director\_nominado, director\_ganador | Incluidas directamente como booleanas (0 o 1) | Categóricas binarias | Muy útiles en probabilidad condicional:  P(éxito actor\_ganador = 1)  P (imdb alto director\_nominado = True), pero los datos booleanos no es exactamente el mejor dato para un modelo predictivo y mas como el que buscamos conseguir. |
| Estos eventos se pueden modelar como eventos condicionados y comprobar su influencia real. |  |  |  |

Esta base de datos permite construir modelos estadísticos para predecir el éxito de futuras películas y descubrir patrones entre el contenido y su aceptación por parte del público. El modelo utiliza un conjunto de técnicas estadísticas y de machine learning (específicamente Random Forest) para predecir el rating IMDb de una película. A continuación, explico cómo se trabaja con cada variable desde el punto de vista de probabilidad y estadística, también se explica el cómo funcionaron en los scripts para su evaluación.

Luego de la descripción de las columnas de la base de datos, podemos decir y asegurar que esta nos permitirá construir modelos estadísticos para predecir el éxito de futuras películas y descubrir patrones entre el contenido y su aceptación por parte del público. Gracias a que fue una base de datos que extrajimos para nuestra conveniencia y con las variables deseadas, es grato clasificar y explicar de que se trata cada columna de esta extensa base de datos de 40k.