Universidad de Guadalajara

Centro Universitario de Ciencias Exactas e Ingenierías

Proyecto Final

Paramount+

215768177 — Sarabia Anaya Andrea Iyari

220288515 — Barba Cardenas Carlos Alberto

219750027 — Gutierrez Macias Juan Ricardo

Ingeniería en Informática

D01 — Minería de Datos

Israel Román Godinez

2023A

Entendimiento del negocio:

a. Objetivos del negocio

i. Antecedentes

Este proyecto utiliza un conjunto de datos que lista todas las películas y series disponibles en Paramount+, con la información recopilada hasta marzo de 2023. Toda la información recopilada es de Estados Unidos.

Al haber utilizado información de un repositorio, no se tiene información exacta acerca de la empresa, pero en general, Paramount+ es un servicio de streaming propiedad y operado por Paramount Streaming, una filial de Paramount Global.

La problemática principal describe que se ha tenido una serie de películas y series que no han tenido una meta de visualización esperada por los usuarios debido a que muchas de los contenidos que se tienen, no encajan con lo que la mayoría de personas prefiere ver.

ii. Objetivo comercial o de investigación del proyecto

El objetivo principal es saber qué tan exitosa puede llegar a ser una película o serie teniendo en cuenta las películas y series que han estado dentro de los contenidos con mayor tasa de visualizaciones. Esto ayudaría a la empresa a no perder suscriptores y generar una mayor cantidad de contenido de calidad y con los gustos que la mayoría de las personas tienen.

iii. Criterios de rendimiento

El criterio de éxito que se planea tener es tener una cantidad mayor de suscriptores, elevar las suscripciones un 10% y una creación de contenidos nuevos y de calidad en un 15%.
Para clasificar una regresión como exitosa deberá evaluarse la popularidad real del proyecto evaluado y compararla con el resultado real contemplando un porcentaje de error de más o menos 8%.

b. Valoración de la situación

i. Inventario de recursos

contamos con dos conjuntos de datos extraidos de la base de datos Paramount+ Movies and TV Shows publicada en kaggle en la direccion https://www.kaggle.com/datasets/dgoenrique/paramount-movies-and-tv-shows?resource=download recuperada el 18 de abril del 2023.

- El primero llamado titles.csv con 3182 instancias de contenidos multimedia alojados en el servicio paramount+, estas instancias poseen los atributos de id, titulo, tipo, descripción, año de lanzamiento, clasificación por edad, duración, género, país de origen, temporadas, id de imdb, calificación en imdb, cantidad de votos en imdb, popularidad en tmdb, puntuación en tmdb.

- El segundo dataset llamado credits.csv que contiene 51195 instancias que describen los roles de distintos actores o directores en distintos proyectos alojados en paramount+, sus atributos son el id de la persona, id según tmdb, nombre, personaje y su rol en la producción.

Para la realización del proyecto se utilizarán tres equipos portátiles con las siguientes características:

Equipo	Procesador	Tarjeta de video	RAM	Sistema operativo
1	intel Core i5-8250U	Gráficos integrados intel UHD Graphics 920	12 GB	Windows 10
2	Ryzen 5-3500U	Gráficos integrados radeon vega mobile Gfx	8 GB	Windows 11
3	AMD A5	Gráficos integrados Radeon R5	12 GB	Windows 10

Como recursos de software utilizaremos Microsoft excel para la visualización de los documentos .csv y Orange Data Mining para su análisis .

ii. Requisitos, supuestos y restricciones

Requisitos:

- Generar un documento con las conclusiones de nuestro análisis de datos sobre la rentabilidad del nuevo proyecto que se presentará a Paramount+.
- Entregar el documento antes del 29 de Mayo.

Supuestos:

Es posible predecir la popularidad de un nuevo proyecto audiovisual basándose en las estadísticas de visualización de los contenidos anteriores que se encuentran en el servicio de streaming.

Restricciones:

- No se dispone de persona de contacto para posibles consultas sobre los datos.
- No se usará el dataset completo.

iii. Riesgos y contingencias

Riesgos:

- El proyecto, al tratar de predecir valores subjetivos como la popularidad de un producto a lanzarse, puede no tener una correlación matemática que nos permita estimar con precisión los datos esperados.
- Los días feriados influyen en la planificación realizada.
- No cumplir con el plazo de entrega.
- Alguno de los archivos necesarios se ve corrompido.
- No es posible recuperar un backup o no existe tal backup.

iv. Terminología

KNN: El algoritmo de k vecinos más cercanos (también conocido como KNN o k-NN, por sus siglas en inglés K-Nearest Neighbours) es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual.

Regresión Lineal: Técnica estadística para determinar la relación entre variables y permite predecir a partir de un muestreo de datos aleatorio. Se adapta a una amplia variedad de situaciones. La regresión ajustada con el error cuadrático medio más bajo se elige como el modelo final.

Dataset: Un conjunto de datos es una colección de datos habitualmente tabulada. En el caso de datos tabulados, un conjunto de datos contiene los valores para cada una de las variables organizadas como columnas que corresponden a cada miembro del conjunto de datos, que están organizados en filas.

v. Análisis de costo/beneficio

El costo de desarrollo del proyecto contempla los siguientes conceptos:

Concepto	Unidades	Costo
Equipo de cómputo con licencia de windows incluida	3	$40,000.00 MXN
Licencia de office 365	3	$2,700.00 MXN
Fracción de costo de servicio de luz eléctrica por un mes	1	$250.00 MXN
Servicio de internet con proveedor totalplay por un mes	1	$550.00 MXN
Salario de científico de datos jr	3	$75,000.00 MXN

Total: $118,500 MXN

Como beneficio para la empresa, al descartar producciones infructuosas puede evitar pérdidas calculadas en millones de dólares para la empresa, citando como ejemplo la reciente producción del estudio paramount, Babylon cuyo costo de producción (sin contar gastos por publicidad) fue de más de 80 millones de dólares de los cuales recuperó solo 63,4 millones representando una pérdida aproximada del 20%.

Tras conversiones de divisas podemos concluir que de evitar una producción que no sea del agrado del público la empresa tendría un beneficio porcentual de 99.06% haciéndolo completamente rentable.

c. Objetivos de la minería de datos

i. Metas de la minería de datos

El problema de minería de datos con el que se va a trabajar es la regresión lineal pues para poder generar contenido de las películas y series que pueden tener éxito y que sean de calidad, es necesario tener conocimiento de los contenidos generados anteriormente que han tenido una mayor tasa de visualizaciones. Entonces, teniendo en cuenta que se utilizará una regresión lineal, pues es preciso tener un aproximado de tiempo en el que se aplica la regresión, para esto, se decidió tomar la información de 3 años atrás pues desde su lanzamiento se han ido generado y subiendo contenidos diferentes, además, creemos que en ese tiempo podemos recopilar una buena cantidad de datos que nos pueden ayudar a cumplir con el objetivo principal.

ii. Criterio de rendimiento

Al tener un problema de regresión lineal, el algoritmo utilizado sería K-Nearest Neighbor (KNN) y para esto, es necesario calcular el error cuadrático medio o MSE. Para la evaluación se necesitará un 8% de los datos.
En el caso de las mediciones subjetivas podemos definir que:
El modelo es fácil de entender y de aplicar para el cliente, con esto, se tiene una ventaja en cuanto a la comprensión de lo que se quiere predecir. El modelo también proporciona información relevante y útil sobre el problema que se está resolviendo, además se puede confiar en los resultados del modelo para tomar decisiones importantes. El modelo es adecuado para el problema que se está resolviendo y se ajusta a los datos de forma precisa y adecuada.

d. Plan de proyecto

i. Plan de proyecto

1. Entender el dominio del problema y los datos.
2. Aprender a manejar la herramienta Orange.
3. Utilizar la metodología “KNN”.
4. Plantear el problema.
5. Seleccionar el subconjunto de datos a analizar.
6. Limpieza, preparación, transformación y carga de los datos.
7. Aplicar la(s) técnica(s) de data mining.
8. Evaluar los resultados obtenidos y refinar el proceso incluyendo posibles iteraciones de los pasos 5 a 8.
9. Generar el documento final.
10. Presentar al cliente.

Se cuenta con un mes y medio para la realización del proyecto y se contempla dividirlo de la siguiente manera:

- **17 a 21 de Abril:** Planteamiento de objetivos y obtención de datos.
- **24 a 28 de Abril:** Limpieza, transformación y carga de los datos.
- **2 a 12 de Mayo:** Aplicación de metodologías para el análisis de datos.
- **15 a 19 de Mayo:** Rectificación de los resultados y corrección de errores.
- **22 a 26 de Mayo:** Creación del documento con los resultados.
- **29 de Mayo:** Presentación de los resultados.

ii. Valoración de herramientas técnicas

Las herramientas propuestas para la realización del proyecto se tomaron en cuenta por su accesibilidad y conocimiento de uso del equipo que realizará la investigación, lo cual las hace elegibles tanto por su costo como por la familiaridad del equipo, permitiendo hacer un análisis de utilidad de la información.

Entendimiento de los datos

a. Recolección de datos iniciales

i. Requerimiento de los datos

El dataset *Paramount+ Movies and TV Shows* fue obtenido de *Kaggle* el 18 de abril de 2023 desde el siguiente link. Este dataset, acorde a su propia descripción, lista todas las series y películas disponibles en *Paramount+* Estados Unidos según *JustWatch* en Marzo de 2023.

La disponibilidad de los datos está sujeta a la disposición del archivo en Kaggle, aunque solo se planea utilizar la versión que fue descargada para este proyecto y se crearon copias de seguridad para prevenir cualquier fallo del archivo.

ii. Criterios de selección

para considerar una instancia como seleccionable debe cumplir con los siguientes criterios:

- Tener un año de lanzamiento posterior al 2020.
- Pertenecer al documento titles.csv.
- Tener los atributos de titulo, tipo, año de lanzamiento, clasificación por edad, tiempo, genero, pais de produccion, puntaje de imdb, votos en imdb, popularidad en tmdb y puntaje de tmdb (se excluyen los id, descripción, temporadas, e id de imdb).

iii. Inserción de datos

Los datos fueron obtenidos de un solo origen de datos (mencionado en secciones anteriores). En la página se listan dos archivos planos, *titles.csv* y *credits.csv*; sin embargo, para propósitos de este proyecto solo se usará el archivo de *titles.csv*.

Se descargaron tres copias del archivo en caso de ser necesario un respaldo de emergencia.

b. Descripción de los datos

i. Análisis volumétrico

1. Número de instancias

A pesar de que el dataset contiene 3,182 instancias, estas datan desde 1912 hasta el presente año; pero debido a los cambios generados por la pandemia, se considera que solo las instancias generadas a partir de 2020 son relevantes para este proyecto, por lo que solo se usarán 358 instancias.

2. Número de atributos

Para este modelo se seleccionaron 8 atributos de los 15 disponibles. Los atributos seleccionados fueron incluidos al proporcionar información de identificación de la instancia o contener datos computables para la clasificación o procesamiento de los datos de las instancias.

Los atributos discriminados se dejaron de lado por ser datos nominales o de identificación única que no aportan al modelo por su volumen e irrelevancia.

3. Número de instancias por clase

Dados los criterios de selección contamos con 358 instancias restantes las cuales cumplen con tener todos los atributos llenos.

ii. Definición del dominio del atributo

Diccionario de datos

Relación	concepto	tipo concepto	tipo de dato	tamaño	dominio	descripción	ejemplo
N/A	titles	entidad	N/A	N/A	N/A	conjunto de datos que contiene información sobre películas y series	N/A
titles	title	campo	text	N/A	alfanumérico	Título del producto	Run & Gun
titles	type	campo	text	5	alfanumerico + ´[,]	clasificación del producto como “MOVIE” o “SHOW”	SHOW
titles	release_year	campo	int	4	0…9	año de lanzamiento del producto	2020
titles	age_certification	campo	text	5	alfanumérico, -	clasificación de edad para el producto	TV-14
titles	runtime	campo	int	3	0…9	tiempo en pantalla del producto	90
titles	genres	campo	text	N/A	alfanumérico	generos de la obra, ofrecen una clasificación del contenido	['comedy', 'action', 'family', 'fantasy', 'scifi', 'animation', 'crime']
titles	production_countries	campo	text	N/A	alfanumérico	países de producción del producto	['GB', 'US', 'CA']
titles	imdb_score	campo	float	N/A	0…9, .	calificación del producto según imdb	6.5

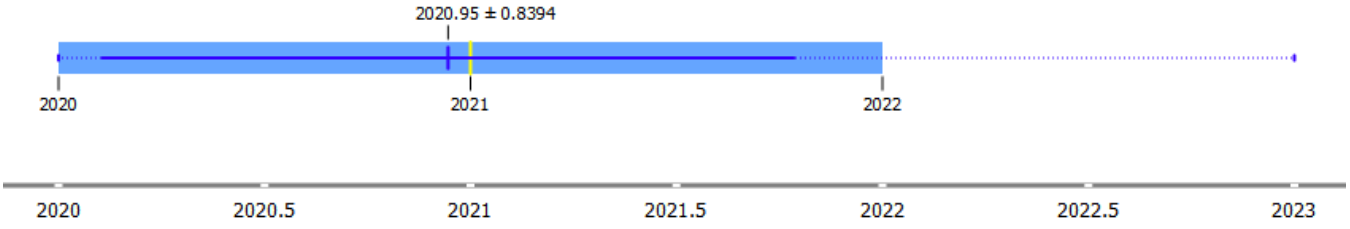
c. Exploración de los datos

i. Análisis univariable (por atributo)

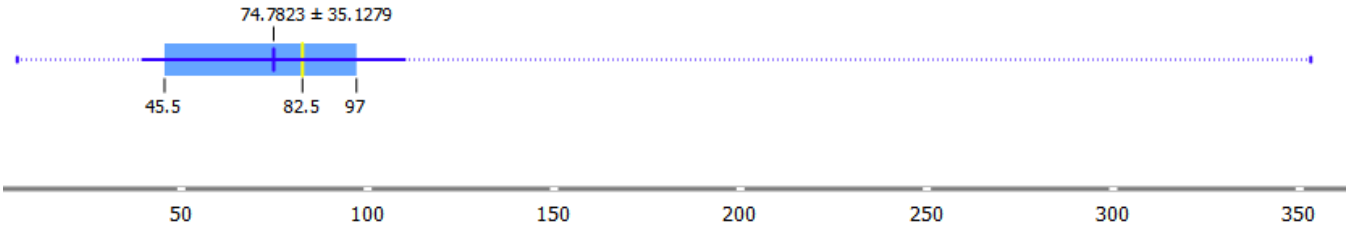
1. Resumen estadístico

Para los valores numéricos se crearon 3 diagramas de caja y bigote o box plot

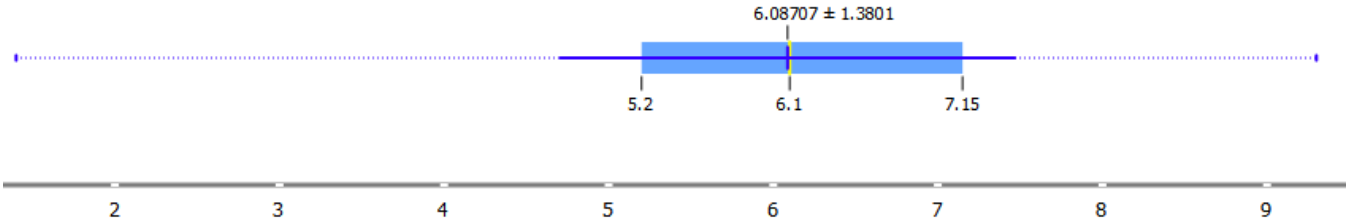
Para el atributo “release year”



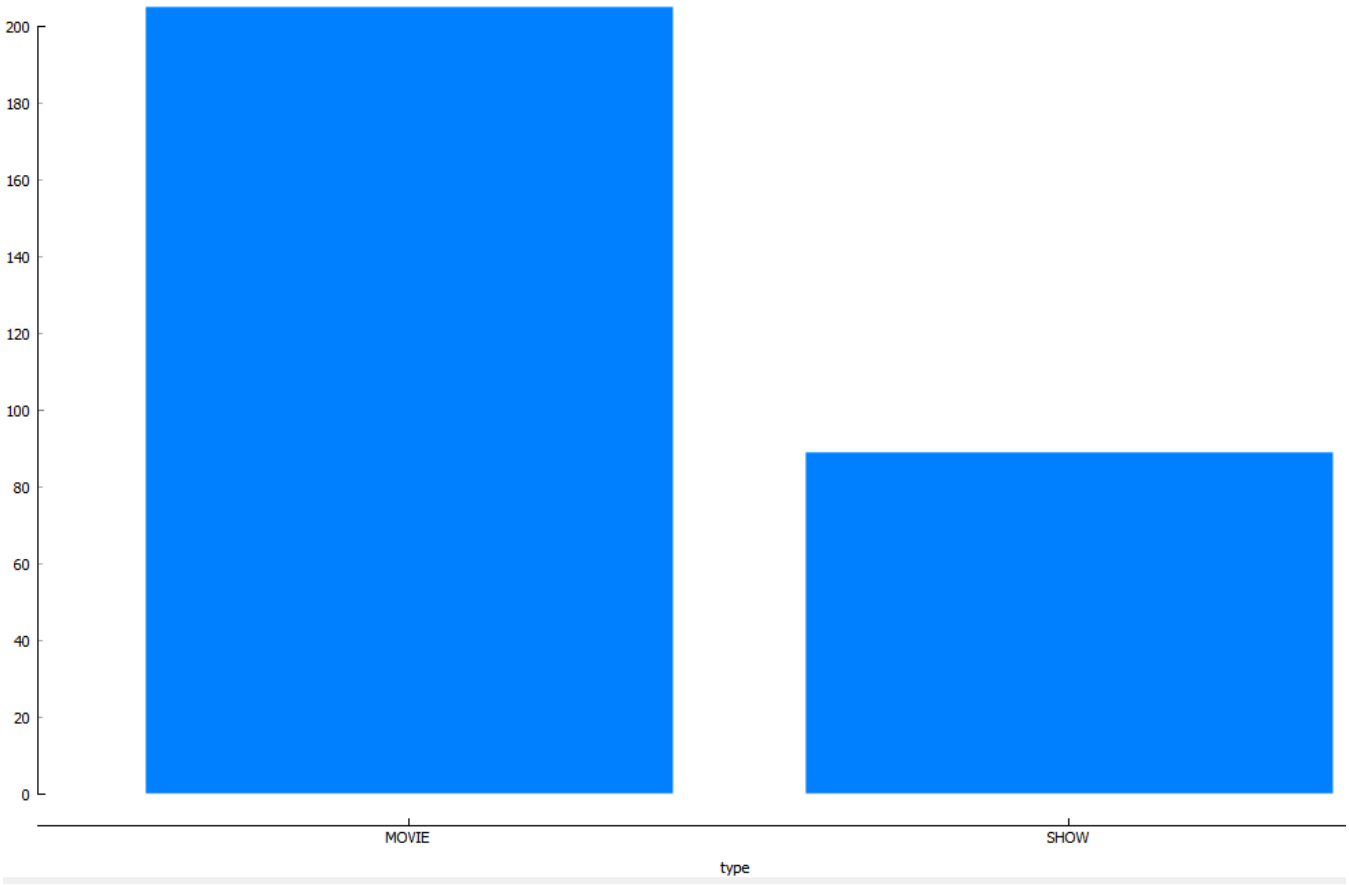
Para el atributo “runtime”



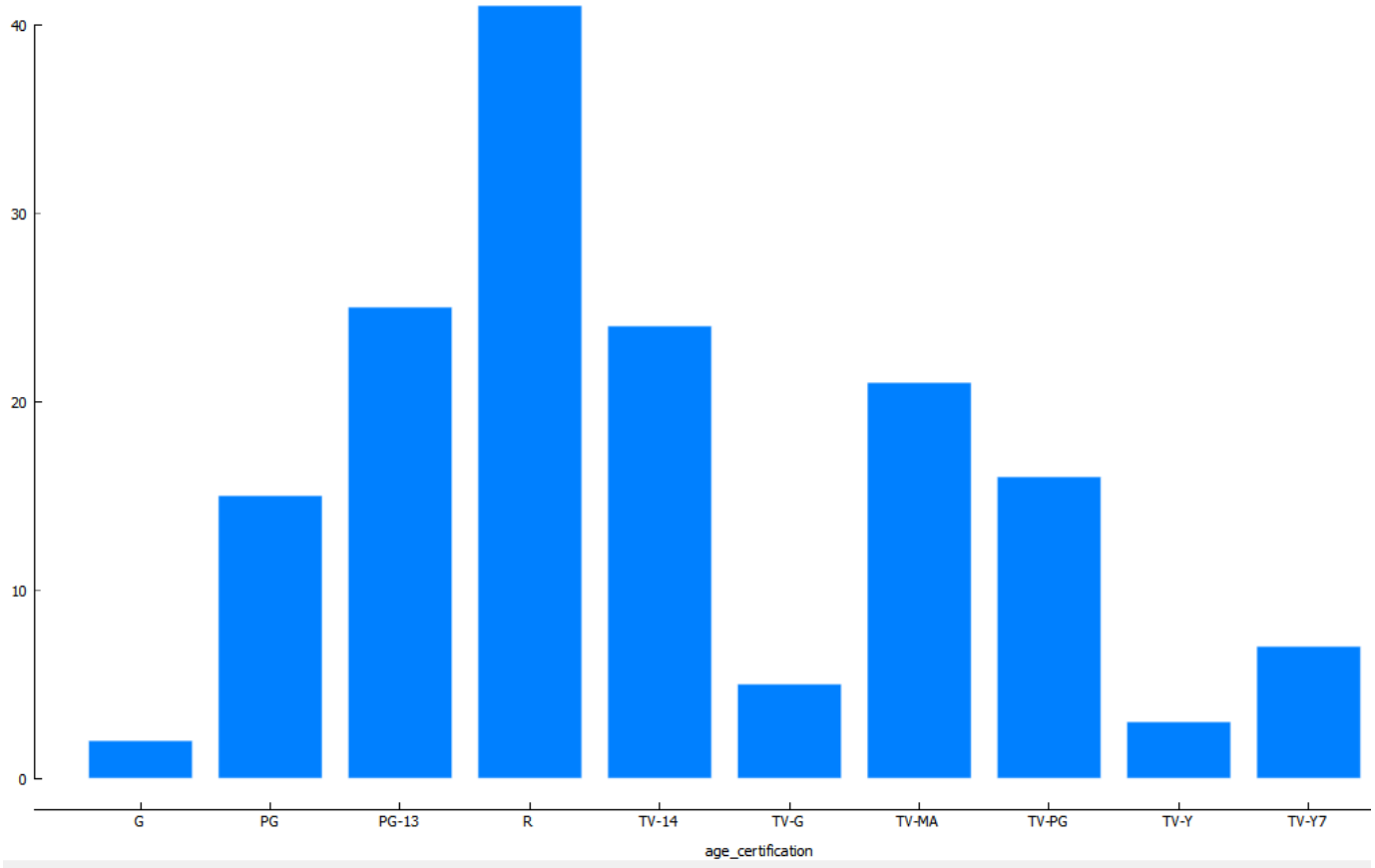
Para el atributo “imdb score”



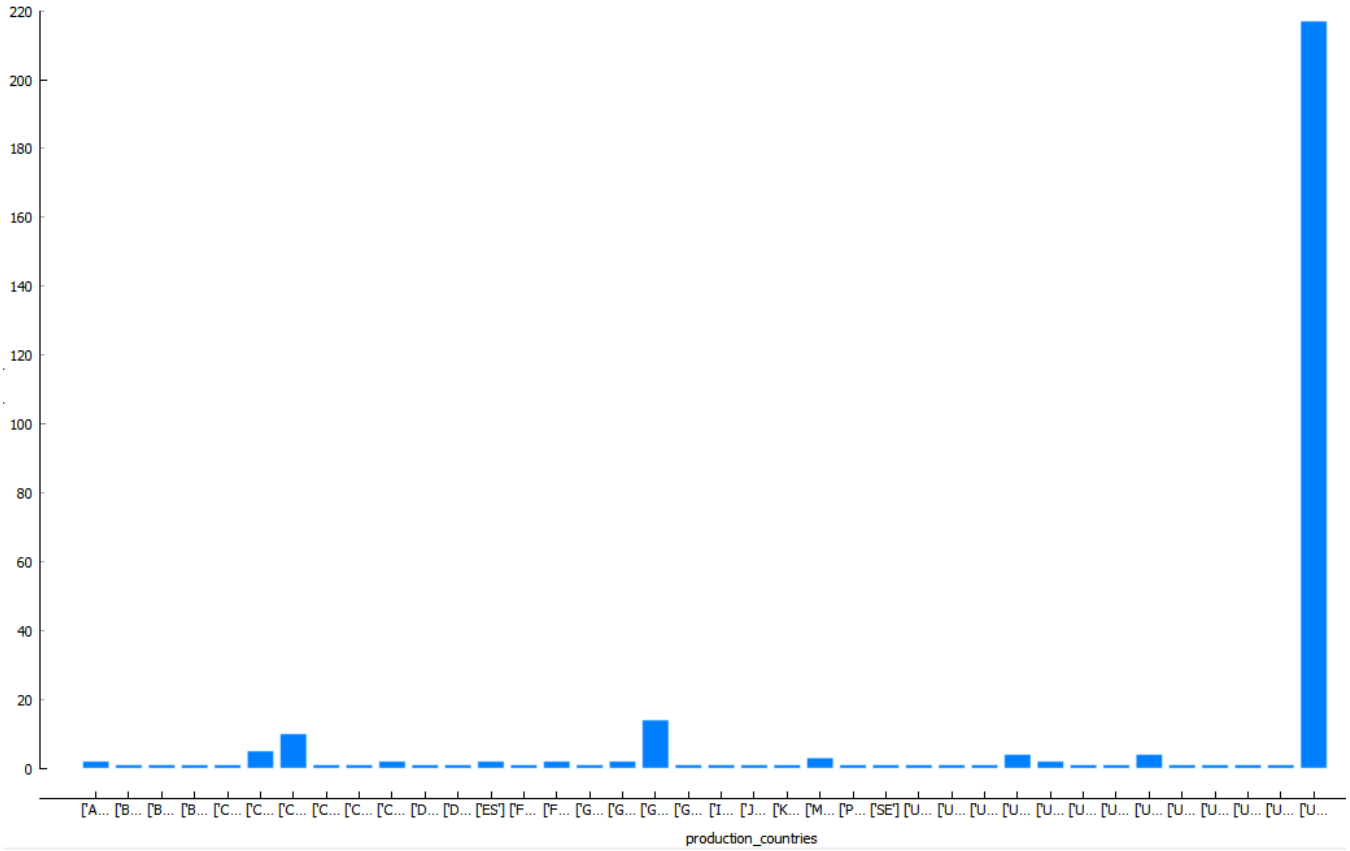
Para los valores categóricos se crearon 4 histogramas.
Para el atributo “type”:



Para el atributo “age of certification”:



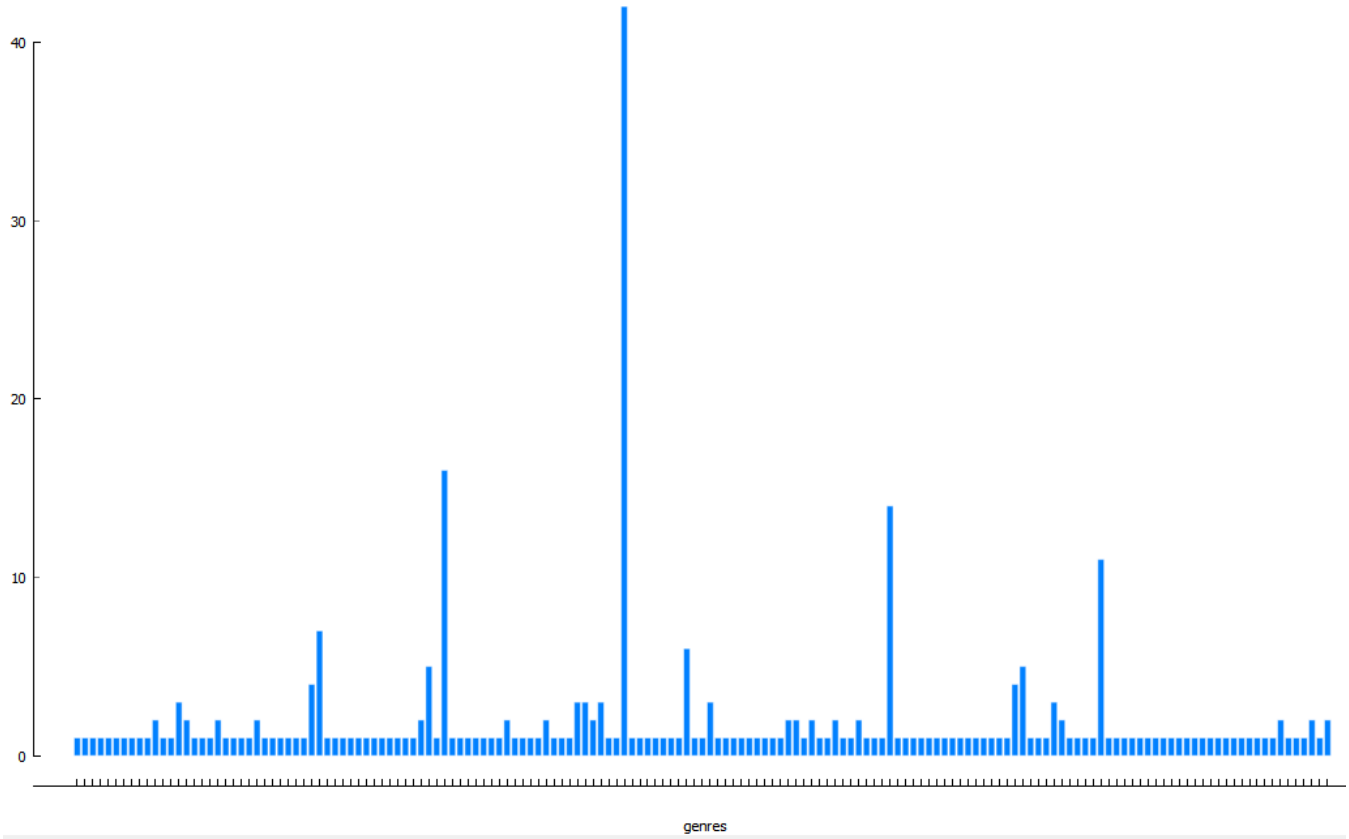
Para el atributo “Production countries”:




```
['BE', 'EE', 'RS', 'SE', 'GB']
['CA', 'US']
['CA']
['CN', 'US']
['GB', 'US', 'CA']
['GB', 'US']
['GB']
['JP', 'US']
['MX']
['US', 'AU']
['US', 'CA', 'GB']
['US', 'CA', 'JP']
['US', 'CA']
['US', 'CN', 'GB']
['US', 'GB', 'CA']
['US', 'GB']
['US', 'JP']
['US', 'KR']
['US', 'MX']
['US']
[]
```

En orden se enlistan los países de producción

Y para el atributo “genres”:



2. Determinación de sesgo

En cada uno de los valores existe un sesgo que se determina por la media y la mediana, dependiendo de donde se sitúan, se confirma si está sesgado negativamente o positivamente.

- Para el atributo “release year” se aprecia un sesgo negativo
- Para el atributo “runtime” existe un sesgo negativo
- Para el atributo “imdb score”existe un sesgo negativo, sin embargo es muy cercano a ser un conjunto simétrico.

ii. Análisis bivariable

En el análisis bivariable de Numéricos vs Numéricos usando el coeficiente de correlación de Pearson, tenemos lo siguiente:

1	+0.173	imdb_score	release_year
2	-0.110	imdb_score	runtime
3	+0.105	release_year	runtime

En este caso podemos observar que la correlación que existe entre estos tres atributos tiende a ser una correlación no lineal pues están más cercanos al 0

- Para el caso de “type” y “production countries”

```
Valor de chi-cuadrado: 36.16117610657803
Coeficiente de contingencia de Tschuprow: 0.33094686546007557
```

- Para el caso de “type” y “age certification”

```
Valor de chi-cuadrado: 158.99999999999997
Coeficiente de contingencia de Tschuprow: 0.5924469406482269
```

- Para el caso de “type” y “genres”

```
Valor de chi-cuadrado: 213.36351055083583
Coeficiente de contingencia de Tschuprow: 0.6484857830331447
```

- Para el caso de “age certification” y “production countries”

```
Valor de chi-cuadrado: 199.55061316782812
Coeficiente de contingencia de Tschuprow: 0.6358587989685568
```

- Para el caso de “age certification” y “genres”

```
Valor de chi-cuadrado: 1168.464110876058
Coeficiente de contingencia de Tschuprow: 0.8938508968335397
```

- Para el caso de “production countries” y “genres”

```
Valor de chi-cuadrado: 6631.367045454545
Coeficiente de contingencia de Tschuprow: 0.9785434974840062
```

El código utilizado para encontrar los valores de chi-cuadrado y el coeficiente de contingencia de Tschuprow es el siguiente:

```
import pandas as pd

from scipy.stats import chi2_contingency


# Cargar el archivo CSV

data = pd.read_csv('C:/Users/RICARDO/Desktop/8vo Semestre/Mineria de datos/nuevo_dataset.csv')


# Selección de las columnas que se utilizarán para el análisis

column1 = 'COLUMNA 1'

column2 = 'COLUMNA 2'


# Crear una tabla de contingencia utilizando las dos columnas

contingency_table = pd.crosstab(data[column1], data[column2])


# Calcular el valor de chi-cuadrado

chi2, p, _, _ = chi2_contingency(contingency_table)


# Calcular el coeficiente de contingencia de Tschuprow

tschuprow = (chi2 / (chi2 + len(data))) ** 0.5


# Imprimir los resultados

print(f 'Valor de chi-cuadrado: {chi2}')

print(f 'Coeficiente de contingencia de Tschuprow: {tschuprow}')
```

Procesamiento de datos

i. Integración de los datos

1. Fusión de datos

En el caso de nuestro proyecto no aplicaría esta fase

2. Adición de datos

En el caso de nuestro proyecto no aplicaría esta fase

ii. Selección de los datos

1. Creación de conjuntos específicos para cada tarea de descubrimiento de datos

Para este proyecto se seleccionaron 8 atributos de los 15 atributos del dataset original, por lo que se eliminaron atributos que no eran indispensables para el proyecto que tenemos planeado. Por consiguiente se creó un dataset específico con los atributos que sí nos ayudarán para el proyecto

Objetivo del conjunto de datos	Descripción del objetivo	
Restricciones de las instancias	Dominio	Justificación de selección
title	Alfanumérico	El título nos sirve para saber cuáles son las películas o series que más éxito tienen
type	Alfanumérico	Nos sirve para saber si es una película o una serie
release_year	0...9	Nos ayuda a tener en cuenta el año en que se realizó y partir de ahí
age_certification	Alfanumérico, -	Es la clasificación de la película o serie para entender cuánto alcance tiene
runtime	0...9	Es la duración de la película para tener en cuenta el tiempo de los

		más exitosos
genres	Alfanumérico	Este atributo nos ayuda a identificar cuál género es el más visto
production_countries	Alfanumérico	Nos ayuda a identificar el país en el que más se produce contenido
imdb_score	0...9	Nos sirve para saber la calificación que tiene en IMDb

2. Selección de subconjuntos de datos (Muestreo)

Utilizando un muestreo aleatorio sin reemplazo, de los 3182 registros que se tienen en el conjunto de datos original se seleccionó una cantidad menor de los datos para poderlos procesar de una manera más eficiente, por lo que se seleccionaron 358 instancias y se separaron en un documento aparte.

iii. Limpieza de datos

1. Eliminación de atributos con poca variabilidad

Se eliminaron datos de identificación como lo es “id” (general) y “imdb_id”, además de la eliminación de datos que no son indispensables como “description” o “seasons”. Además de eliminar también “imdb_votes”, “tmdb_score” y “imdb_popularity” ya que no son necesarios para el modelo.

2. Identificación de valores erróneos (outliers y typos)

En este caso, del muestreo que se obtuvo, no se encontraron valores con errores en escritura.

3. Detección de valores faltantes

Se encontraron algunos valores faltantes por lo que se procedió a rellenar el valor faltante de los valores numéricos con la mediana y los valores categóricos con la moda. Si el valor se conoce o es verificable por una fuente externa, se sustituye por el valor real.

4. Eliminación de falsos predictores

Al no tener una clase en específico, no se tiene una correlación con la que nos puedan dar falsos predictores, sin embargo, no existe una correlación entre los atributos numéricos tan alto como para considerarlo un falso predictor.

1	+0.173	imdb_score	release_year
2	-0.110	imdb_score	runtime
3	+0.105	release_year	runtime

5. Errores de dominio, tipo o formato.

Tampoco se encontraron errores en los dominios, en los tipos o en el formato de los datos. Pero si esto ocurriese, la medida que se tomaría es eliminar la instancia o corregir los datos si es que se tiene información recuperada de fuentes externas.

Modelado

a. Selección de técnicas de modelado.

I. **Lista de las técnicas que se utilizarán para la evaluación del modelo**

Se está utilizando un modelo KNN para realizar una regresión lineal, por lo que optamos por utilizar una métrica como el error cuadrático medio (MSE) para evaluar el rendimiento del modelo.

II. **Descripción de las razones por las que fueron seleccionados dichos modelos.**

La métrica utilizada nos permite evaluar el rendimiento del modelo al proporcionar información acertada sobre la diferencia entre nuestras predicciones y los valores reales que se tienen, permitiéndonos medir la variabilidad de los datos y ajustar nuestro modelo acorde a esto. Por esto, creemos que el modelo utilizado es el más adecuado para nuestro proyecto, además de ser un modelo fácil de entender y muy eficiente

b. Generación de plan de prueba

I. **Lista de las diferentes pruebas que se realizan para cada uno de los conjuntos de datos.**

Se realizó una prueba con el conjunto de datos que se definió al principio, con 140 instancias y 11 atributos, sin embargo se optó por desechar este dataset ya que al tener inconsistencias, no podría funcionar.

Para una segunda prueba, se modificó el dataset para dejar el conjunto actual de 358 instancias y 8 atributos con el que se hizo una prueba con el modelo KNN descrito anteriormente.

II. **Indicar los números de iteraciones y técnicas de validación a ser utilizadas (por cada conjunto de datos).**

Para el primer conjunto de datos se realizaron 10 iteraciones con un 66% de los datos para su entrenamiento y utilizando como técnica de validación Leave One Out.

Para el segundo conjunto y conjunto final, se realizaron también 10 iteraciones con un 66% de datos para entrenamiento y utilizando Leave One Out.

c. Construcción de modelos

I. **Selección de parámetros: Indicar los parámetros seleccionados para cada uno de los modelos.**

- Como primer parámetro, en cada uno de los conjuntos se utilizaron 2 “K” o 2 “vecinos”
- Como métrica se definió la distancia Euclidiana en cada conjunto

II. **Ejecución de cada una de las pruebas definidas en el punto anterior.**

En la prueba con el conjunto inicial los resultados fueron los siguientes:

Model	MSE	RMSE	MAE	R2
kNN	2.471	1.572	1.220	-0.452

En la prueba con el conjunto de datos final los resultados fueron:

Model	MSE	RMSE	MAE	R2
kNN	2.192	1.481	1.151	-0.151

III. **Interpretación de los resultados, desde el punto de vista de minería de datos, no del negocio.**

En la primera prueba se observa que el error cuadrático medio puede ser de 2.471 mayor o menor que el resultado de predicción que establecimos.

En la segunda prueba se obtiene un error cuadrático medio de 2.192 mayor o menor que el resultado inicial establecido. Con esto, podemos darnos cuenta que con el conjunto con más instancias y menos atributos existe un error menor que el conjunto inicial.

IV. Presentar conclusiones sobre los patrones encontrados.

En las dos pruebas con el modelo podemos ver resultados muy similares teniendo en cuenta que no tienen los mismos datos, pero el resultado podría ser muy cercano uno del otro. Utilizando las mismas métricas incluso, cambia un poco el resultado del modelo, sin embargo, el modelo no es tan preciso, aunque tampoco está muy alejado de un resultado bastante cercano a 0, por lo que podemos concluir que se puede mejorar más el modelo y hacerlo más preciso.

d. Evaluación de modelos

I. Presentar los resultados anteriores en una tabla comparativa para ver el desempeño de todos.

Al tener un solo modelo, este punto no aplicaría así que el único resultado que se tiene es el siguiente:

Model	MSE	RMSE	MAE	R2
kNN	2.192	1.481	1.151	-0.151

II. Identificar el o los mejores modelos encontrados en la etapa anterior y explicarlos en términos del negocio.

Creemos que el mejor modelo es KNN por el problema que se presenta, ya que es más fácil saber si alguna película o serie puede llegar a ser exitosa mejorando el modelo y con un rango menor de éxito y al ser valores numéricos, es más fácil comparar la tasa de éxito en cada película.

III. Comentarios sobre el desempeño de cada uno de los modelos e indicar sus puntos de vista sobre por qué tiene el desempeño que tienen (bueno o malo).

El modelo utilizado tiene un buen desempeño pues no tiene un error muy alto, sin embargo se podría mejorar con un procesamiento de los datos mucho más estricto.