# Proyecto 1. Entrega 2.
# Algoritmos de Aprendizaje no Supervisado

## INSTRUCCIONES:

El estudio "**CineVision Studios**" está complacido con lo que descubrió en el análisis exploratorio que su equipo le entregó. Sin embargo, le han surgido nuevas interrogantes. En la conformación de su equipo de Data Science, uno de los candidatos, en las entrevistas, le ha comentado que existen algoritmos que permiten detectar patrones no evidentes en los datos que le proporcionó y "juntar" variables para que sea más pequeño. Le ha pedido que indague un poco más en los datos y extraiga información interesante.

**Descripción de la consultoría:**

- Aplicación de algoritmos de Clustering
  - Aplicar varios algoritmos de agrupamiento y seleccionar el que tenga mejor calidad para interpretar los grupos
- Reglas de Asociación
  - Aplicar un algoritmo de reglas de asociación y sacar las reglas más interesantes.
- Análisis de Componentes Principales
  - Teniendo en cuenta la cantidad de variables que se tiene, podría resultar trabajoso usar modelos de predicción más adelante, por lo que se quiere una alternativa de las variables que tienen la mayor variabilidad.

Utilice el data set al que le hizo el análisis exploratorio en la hoja de trabajo anterior.

**Resultados Esperados**

Al aplicar este análisis, **CineVision Studios** puede comprender mejor el entorno obteniendo "insights" que le permitan explotar mercado que aún no ha sido descubierto por la competencia.

**Presentación de resultados**

La compañía espera **un informe** con todos los hallazgos que arrojó la aplicación de los algoritmos, estos deben estar bien explicados y se deben apoyar de gráficos que sostengan estas explicaciones. También le han pedido que entregue **el código** (R o python) utilizado pues su equipo de data science, una vez contratado lo utilizará para futuras producciones. Le están pidiendo que muestre todos los aportes del equipo en el repositorio de github.

## DESCRIPCIÓN DEL DATASET

El dataset contiene datos de 10000 películas obtenidos de la plataforma "The movie DB".

**Variables:**

- Id: Id de la película
- popularity: Índice de popularidad de la película calculado semanalmente
- budget: El presupuesto para la película.
- revenue: El ingreso de la película.
- original_title: El título original de la película, en su idioma original.
- originalLanguage: Idioma original en que se encuentra la película
- title: El título de la película traducido al inglés
- homePage: La página de inicio de la película
- video: Si tiene videos promocionales o no

- director: Director de la película
- runtime: La duración de la película.
- genres: El género de la película.
- genresAmount: Cantidad de géneros que representan la película
- productionCompany: Las compañías productoras de la película.
- productionCoAmount: Cantidad de compañías productoras que participaron en la película
- productionCompanyCountry: Países de las compañías productoras de la película
- productionCountry: Países en los que se llevó a cabo la producción de la película
- productionCountriesAmount: Cantidad de países en los que se rodó la película
- releaseDate: Fecha de lanzamiento de la película
- voteCount: El número de votos en la plataforma para la película.
- voteAvg: El promedio de los votos en la plataforma para la película
- actors: Actores que participan en la película (Elenco)
- actorsPopularity: Índice de popularidad del elenco de la película.
- actorsCharacter: Personaje que interpreta cada actor en la película
- actorsAmount: Cantidad de personas que actúan en la película
- castWomenAmount: Cantidad de actrices en el elenco de la película
- castMenAmount: Cantidad de actores en el elenco de la película.

## ACTIVIDADES

**1. Clustering**

1.1. Haga el preprocesamiento del dataset, explique qué variables no aportan información a la generación de grupos y por qué. Describa con qué variables calculará los grupos.

1.2. Analice la tendencia al agrupamiento usando el estadístico de Hopkings y la VAT (Visual Assessment of cluster Tendency). Esta última hágala si es posible, teniendo en cuenta las dimensiones del conjunto de datos. Discuta sus resultados e impresiones.

1.3. Determine cuál es el número de grupos a formar más adecuado para los datos que está trabajando. Haga una gráfica de codo y explique la razón de la elección de la cantidad de clústeres con la que trabajará.

1.4. Utilice los algoritmos k-medias y clustering jerárquico para agrupar. Compare los resultados generados por cada uno.

1.5. Determine la calidad del agrupamiento hecho por cada algoritmo con el método de la silueta. Discuta los resultados.

1.6. Interprete los grupos basado en el conocimiento que tiene de los datos. Recuerde investigar las medidas de tendencia central de las variables continuas y las tablas de frecuencia de las variables categóricas pertenecientes a cada grupo. Identifique hallazgos interesantes debido a las agrupaciones y describa para qué le podría servir.

**2. Reglas de Asociación**

2.1. Obtenga reglas de asociación interesantes del conjunto de datos usando el algoritmo “A priori”. Recuerde discretizar las variables numéricas. Genere reglas con diferentes niveles de confianza y soporte. Discuta los resultados. Si considera que debe eliminar variables

porque son muy frecuentes y con eso puede recibir más "insights" de la generación de reglas. Hágalo y discútalo.

**3. Análisis de Componentes Principales**

3.1. Estudie si es posible hacer transformaciones en las variables categóricas para incluirlas en el PCA, ¿valdrá la pena?

3.2. Estudie si es conveniente hacer un Análisis de Componentes Principales. Recuerde que puede usar el índice KMO y el test de esfericidad de Bartlett.

3.3. Haga un análisis de componentes principales con las variables numéricas, discuta los resultados e interprete los componentes.

## EVALUACIÓN

**Nota**: Tiene que poderse comprobar su aporte al trabajo grupal a través de "commits". Si no existen al menos 3 "commits" con su aporte significativo no va a tener nota en esta entrega. Utilice una herramienta que permita registrar los aportes de cada uno.

- **(40 puntos) Clustering**
  - **(10 puntos)** Determinación de la cantidad de grupos: Utiliza un procedimiento adecuado para determinar la cantidad de grupos que deberían formarse de acuerdo con el conjunto de datos. Explica en que se basa para seleccionar el número de grupos (clusters), interpretando los resultados del método usado. Se basa en gráficas para apoyar su decisión.
  - **(10 puntos)** Agrupamiento: Utiliza los algoritmos de agrupamiento sugeridos. Muestra el resultado generado por cada uno y los compara. Determina la calidad de los grupos arrojados por cada algoritmo. Discute los resultados y determina cuál va a usar y por qué, para explorar e interpretar los grupos.
  - **(20 puntos)** Interpretación de los grupos: Hace un análisis de los grupos generados. Explica los hallazgos interesantes que arrojaron. Muestra los elementos que utilizó para describir los grupos generados, medidas de tendencia central, tablas de frecuencia, etc. Explica como estos elementos ayudan a explicar los grupos.
- **(20 puntos) Análisis de componentes Principales**
  - Estudia la matriz de correlación, la agrega y explica lo que observa en ella
  - Determina si es posible usar la técnica de análisis factorial para hallar las componentes principales
  - Determina si vale la pena aplicar las componentes principales interpretando la prueba de esfericidad de Bartlett
  - Obtiene los componentes principales y explica cuántos seleccionará para explicar la mayor variabilidad posible.
  - Interpreta los coeficientes principales.
- **(15 puntos)** Reglas de asociación
  - Construye reglas de asociación usando el algoritmo a priori.
  - Prueba con varios valores de confianza y soporte, y decide si quitar o no características para obtener mejores hallazgos.
  - Discute sobre las reglas de asociación más interesantes teniendo en cuenta sus niveles de confianza y soporte.
- **(25 puntos)** Hallazgos y conclusiones.

- Hace un resumen de los hallazgos que arrojó el agrupamiento.
- Llega a conclusiones sobre el análisis de componentes principales
- Determina las reglas de asociación más interesantes.
- Propone sugerencias a **CineVision Studios** para nuevos desarrollos y mejora de áreas teniendo en cuenta los descubrimientos que hizo.

## MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con la información requerida por el estudio cinematográfico.
- Script de R o de Python que utilizó debidamente organizado y comentado (Si utilizó jupyter notebooks o rmd debe añadir el html que se genera)
- Enlace de controlador de versiones utilizado.

## FECHAS DE ENTREGA

- Viernes, 14 de febrero de 2025 23:59:
  - (60 puntos) Clustering y PCA.
- Domingo 16 de febrero de 2025 23:59
  - (40 puntos) Documento completo incluyendo Reglas de Asociación, hallazgos y conclusiones

**NOTA**: Para poder tener nota completa debe entregar las asignaciones en el tiempo adecuado. No se calificará el avance de la entrega si no fue subido en tiempo, aunque esté en el repositorio.

**Sugerencia**: El segundo día de clase de la semana tendrá un tiempo de aclaración de dudas con el profesor, se le sugiere que avance en la resolución del laboratorio en los pasos del contenido teórico visto en la clase presencial para que aclare todas sus dudas al respecto en dicho espacio