Hoja de Trabajo 2.
Clustering

INSTRUCCIONES:

Utilice el data set al que le hizo el análisis exploratorio en la hoja de trabajo anterior. Debe comparar los resultados generados por cada algoritmo de clustering. Genere un informe con el análisis del funcionamiento de los algoritmos. Determine, según las métricas de calidad que algoritmo hizo el mejor agrupamiento e interprete los clusters generados por él. Añada un apartado donde describa qué le pareció interesante de la información generada con el agrupamiento y de qué forma indagaría más en esa línea. Guarde el código que ha utilizado para hacer esta hoja de trabajo. Los lenguajes que tiene permitido usar son R o Python. **La calificación de cada ejercicio tomará en cuenta tanto lo escrito en el informe como el código.**

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. 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.
2. Analice la tendencia al agrupamiento usando el estadístico de Hopkings y la VAT (Visual Assessment of cluster Tendency). Discuta sus resultados e impresiones.
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.
4. Utilice los algoritmos k-medias y clustering jerárquico para agrupar. Compare los resultados generados por cada uno.
5. Determine la calidad del agrupamiento hecho por cada algoritmo con el método de la silueta. Discuta los resultados.
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.

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 de la hoja de trabajo. Utilice una herramienta que permita registrar los aportes de cada uno.

- **(10 puntos)** Preprocesamiento: Elaboró el preprocesamiento de los datos necesario para utilizar los algoritmos de clustering.
- **(5 puntos)** Explicación del preprocesamiento: Explica por qué hace transformaciones en el conjunto de datos.
- **(15 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 clústeres, interpretando los resultados del método usado. Se basa en gráficas para apoyar su decisión.
- **(15 puntos)** Clustering: Utiliza los algoritmos de agrupamiento sugeridos. Muestra el resultado generado por cada uno y los compara.
- **(15 puntos)** Calidad del agrupamiento: 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.
- **(30 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.
- **(10 puntos)** Trabajo que sigue: Describe el trabajo que desarrollará a partir de la generación de grupos, las tendencias que investigará partiendo de lo que descubrió.

MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con la descripción del proceso de agrupamiento:
  - Descripción del preprocesamiento
  - Explicación de la selección del número adecuado de grupos
  - Comparación de los algoritmos de clustering, incluyendo la calidad del agrupamiento que hizo cada uno.
  - Interpretación de los grupos
  - Descripción del trabajo futuro.
- 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)
- Link de controlador de versiones utilizado.