Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
Mineria de datos
Ciclo 1 de 2,025

Proyecto 1

Gonzalo Enrique Santizo Vega – 21504

Link repositorio github:

https://github.com/GonzaloSantizo/Proyecto1MineriaData

1.

Las siguiente variables no son parte de nuestro análisis, ya que estas son variables cualitativas y no son muy relevantes a la hora de analizar nuestra muestra de datos:

1. Id: Id de la película
2. original_title: El título original de la película, en su idioma original.
3. originalLanguage: Idioma original en que se encuentra la película
4. title: El título de la película traducido al inglés
5. homePage: La página de inicio de la película
6. video: Si tiene videos promocionales o no
7. director: Director de la película
8. genres: El género de la película.
9. productionCompany: Las compañías productoras de la película.
10. productionCompanyCountry: Países de las compañías productoras de la película
11. productionCountry: Países en los que se llevó a cabo la producción de la película
12. releaseDate: Fecha de lanzamiento de la película
13. actors: Actores que participan en la película (Elenco)
14. actorsCharacter: Personaje que interpreta cada actor en la película

Las variables que vamos a utilizar serían las siguientes, ya que estás variables son cuantitativas y nos ayuda mejor a calcular las distancias entre puntos para la formación de grupos:

1. Presupuesto (budget): Indica la cantidad de dinero invertida en la producción de la película.
2. Ingresos (revenue): Indica la cantidad de dinero recaudada por la película en taquilla.

3. Duración (runtime): Indica la duración de la película en minutos.

4. Popularidad de los actores (actorsPopularity): Indica la popularidad de los actores de

la película.

5. Popularidad (popularity): Indica la popularidad general de la película.

6. Promedio de votos (voteAvg): Indica la calificación promedio de la película por parte

de los usuarios.

7. Cantidad de votos (voteCount): Indica la cantidad de votos que ha recibido la película.

8. Cantidad de géneros (genresAmount): Indica la cantidad de géneros que tiene la

película.

9. Cantidad de compañías de producción (productionCoAmount): Indica la cantidad de

compañías de producción que participaron en la película.

10. Cantidad de países de producción (productionCountriesAmount): Indica la cantidad

de países que participaron en la producción de la película.

11. Cantidad de actores (actorsAmount): Indica la cantidad de actores que aparecen en la

película.

12. Cantidad de mujeres en el reparto (castWomenAmount): Indica la cantidad de mujeres

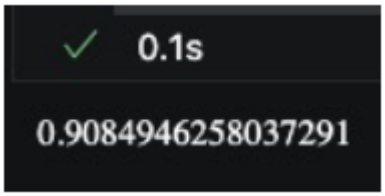
que aparecen en el reparto de la película.

13. Cantidad de hombres en el reparto (castMenAmount): Indica la cantidad de hombres

que aparecen en el reparto de la película.

2. Analice la tendencia al agrupamiento usando el estadístico de Hopkings y la VAT

(Visual Assessment of cluster Tendency). Discuta sus resultados e impresiones.



Al analizar la tendencia con el estadístico de Hopkins, pudimos observar que el conjunto de

datos con mayor tendencia al agrupamiento eran las primeras 11 variables de las que

elegimos en el inciso anterior. Dándonos un valor de 0.90, el cual, en la escala de hopkins,

al estar más cerca del 1, nos indica que puede existir una alta tendencia al agrupamiento

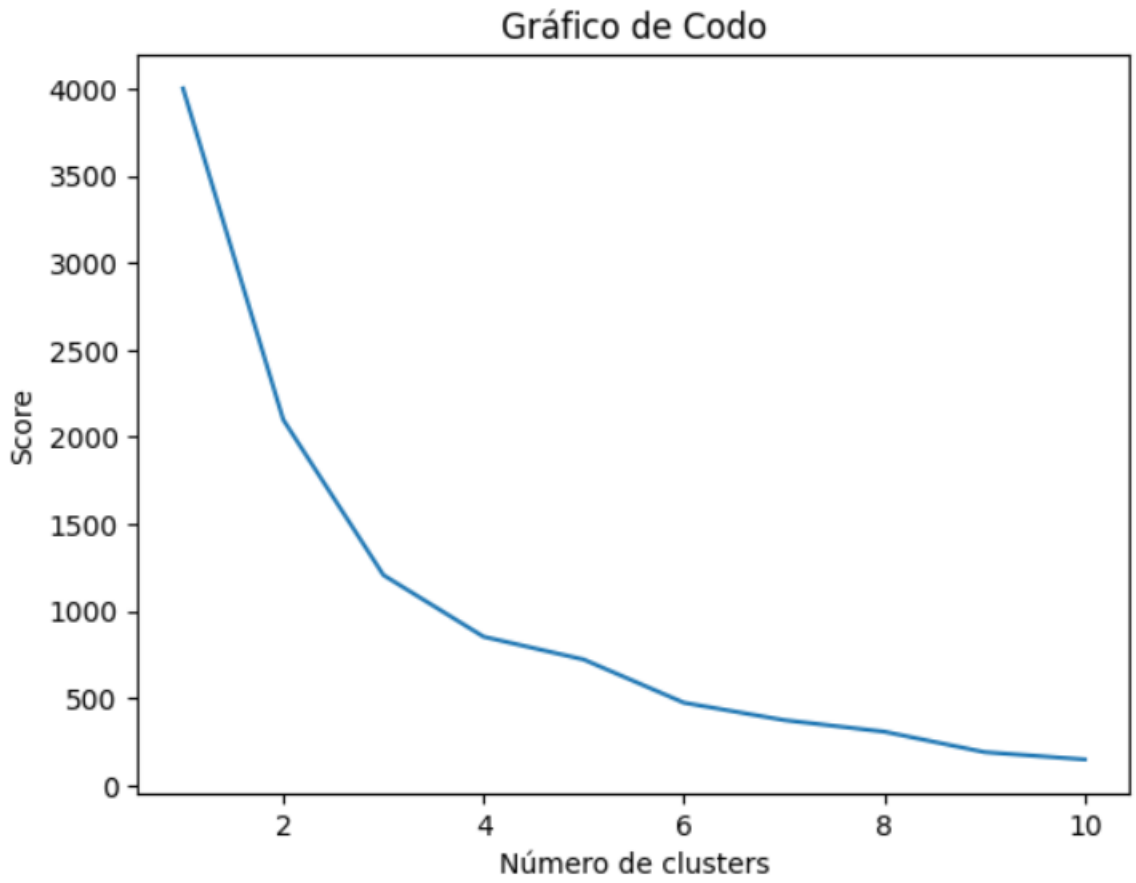
entre estos campos. Quitando los campos de popularity, avgCount, ya que estos no se

agrupaban bien con los demás.

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.

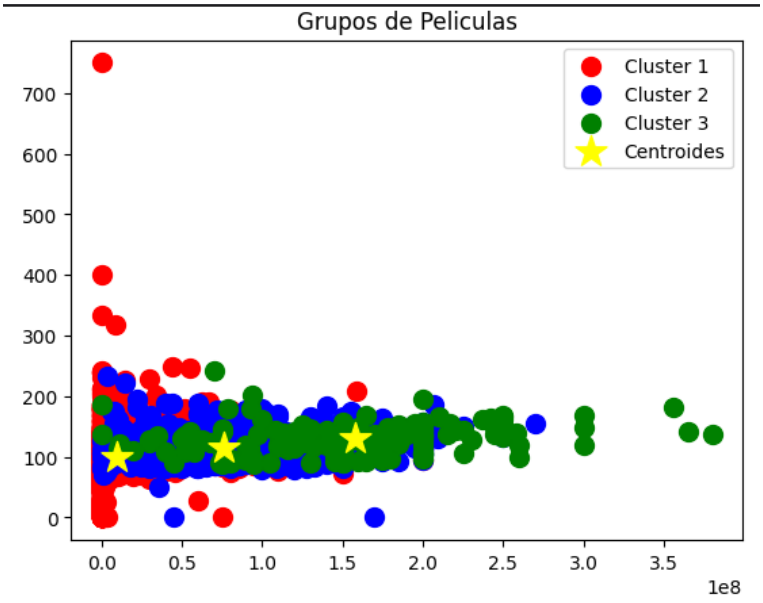


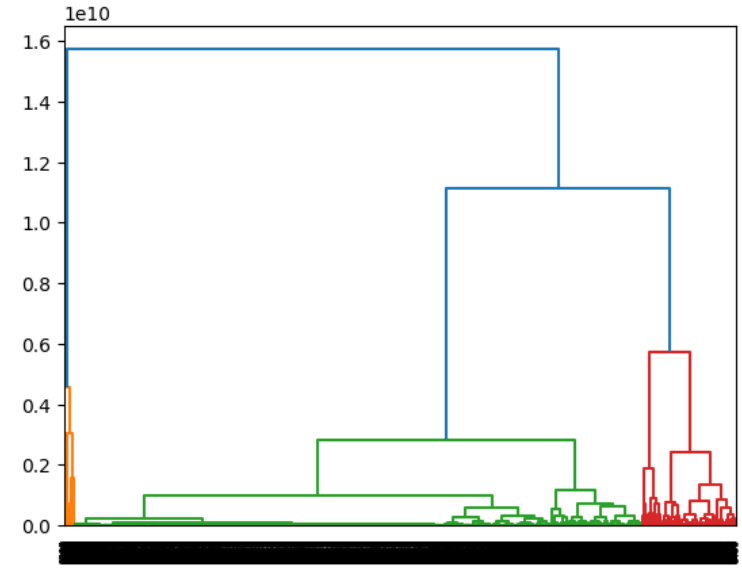
Al ver la gráfica podemos observar que el codo de la gráfica se encuentra en 3, por lo que

elegiremos 3 grupos para realizar el agrupamiento de nuestros datos.

4. Utilice los algoritmos k-medias y clustering jerárquico para agrupar. Compare los

resultados generados por cada uno.





En las gráficas se puede ver que la repartición de los grupos está algo desbalanceada, ya

que la distancia intra grupos de algunos puntos es más pequeña con otros grupos que con

el grupo al que pertenecen; como se puede ver en la gráfica Kmeans. Se puede ver un gran

margen entre los puntos del grupo verde, comparado con los del grupo rojo en el gráfico de

clustering jerárquico, haciendo alusión a la gran extensión que presentó un grupo y a la falta

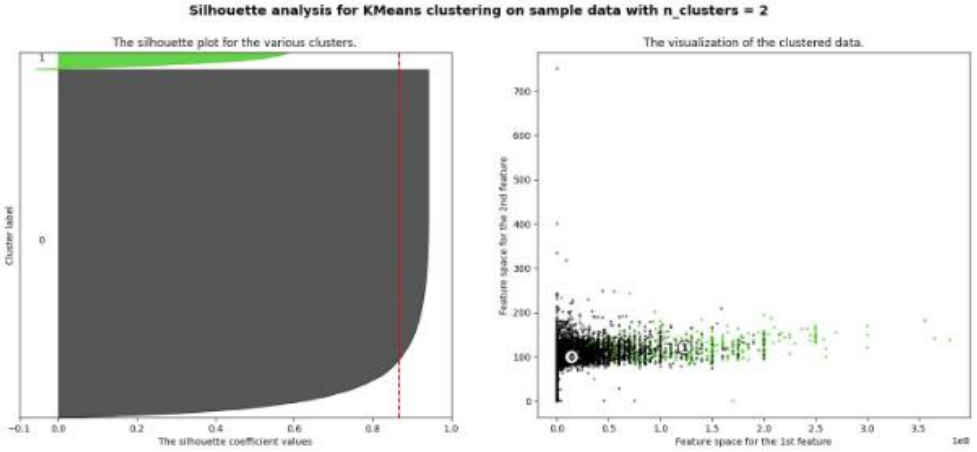
de ordenamiento que poseen comparados con la distancia de los mismo hacia el nodo

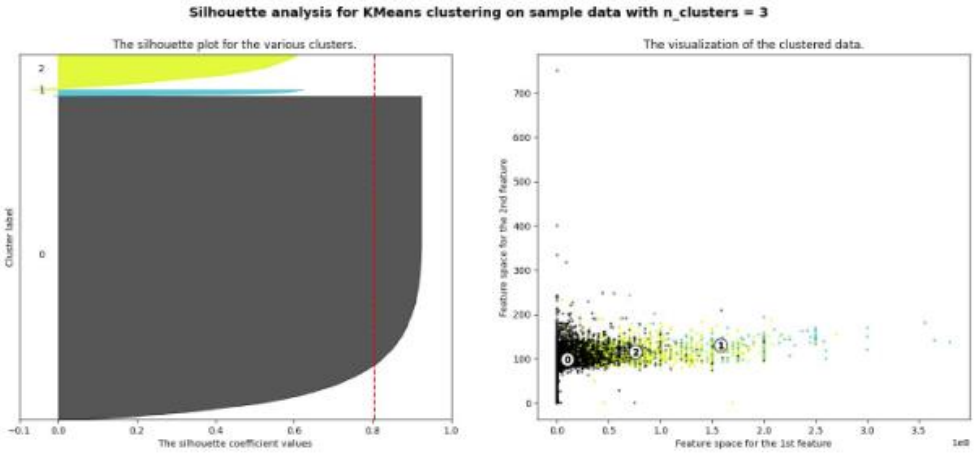
central.

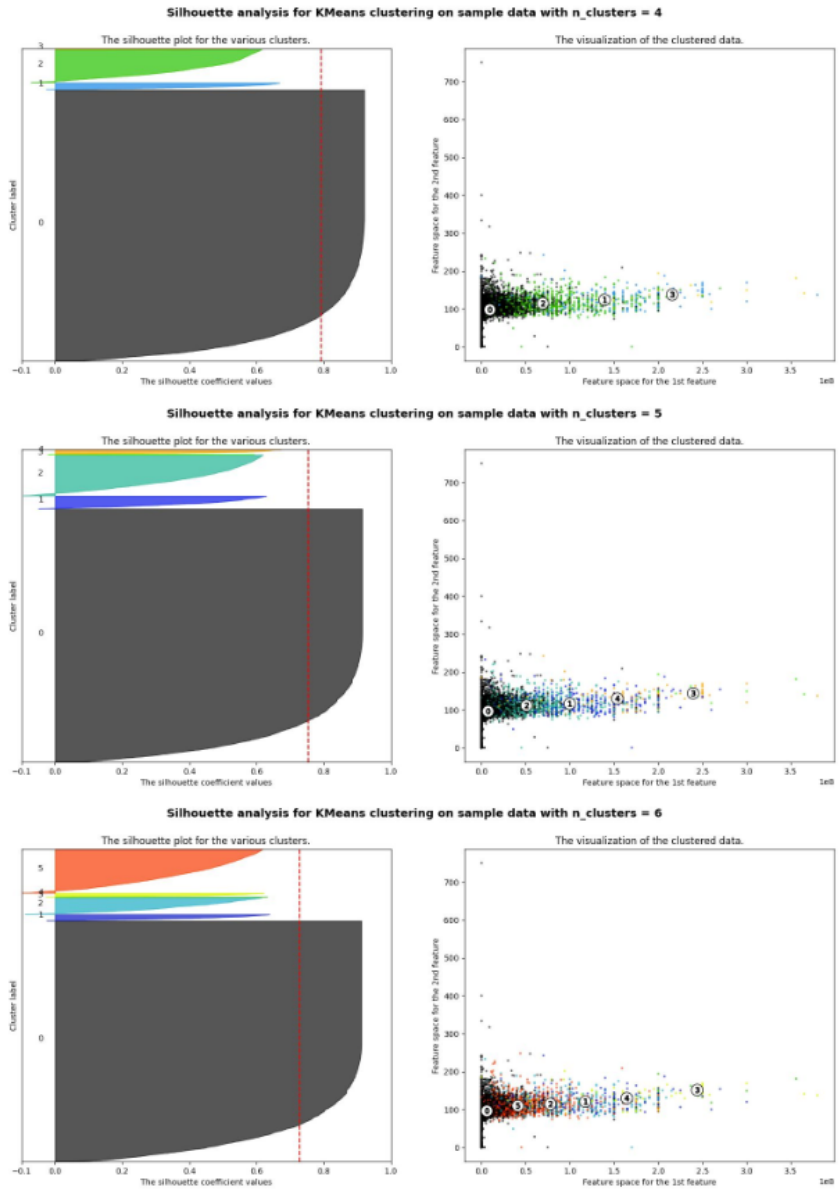
5. Determine la calidad del agrupamiento hecho por cada algoritmo con el método de la

silueta. Discuta los resultados.

```
For n_clusters = 2 The average silhouette_score is : 0.8676886012029503
For n_clusters = 3 The average silhouette_score is : 0.8042355525411609
For n_clusters = 4 The average silhouette_score is : 0.7918381097126703
For n_clusters = 5 The average silhouette_score is : 0.753959296020232
For n_clusters = 6 The average silhouette_score is : 0.7279634308847808
```







Después de ejecutar el código proporcionado, obtendremos un gráfico que consiste en dos

partes principales: la trama de la silueta y la visualización de los datos agrupados.

Trama de la silueta: Esta parte del gráfico muestra la medida de la silueta para cada

muestra individual y cómo se distribuyen en los diferentes clústeres. Cada barra en la trama

de la silueta representa una muestra, donde su longitud indica qué tan similar es esa

muestra a su propio clúster en comparación con los clústeres vecinos. El coeficiente de

silueta promedio para todos los puntos se muestra como una línea punteada vertical roja.

Se busca que la mayoría de las barras estén por encima de esta línea, lo que indica una buena separación y cohesión de los clústeres.

Visualización de los datos agrupados: Esta parte del gráfico muestra cómo se agrupan los datos en el espacio de características original. Cada punto en el gráfico representa una muestra de datos, y los colores indican a qué clúster pertenecen según el algoritmo de KMeans. También se muestran los centroides de cada clúster como círculos blancos, con etiquetas numéricas que indican el índice del clúster.

Para concluir podemos observar que la mejor agrupación de los datos estan en los clusters 2 y 3 ya que se observa un valor aproximado de 0.8 lo cual nos indica que es bueno para poder agrupar los datos. Un valor alto y cercano a 1 indica una buena separación entre los clústeres, mientras que un valor cercano a 0 indica que los clústeres se superponen. Un valor negativo sugiere que los puntos pueden haber sido asignados incorrectamente a los clústeres