Proyecto Final “Queveo py”

Pablo Gutiérrez, Yohany Quintero

Inteligencia Artificial

Pontificia Universidad Javeriana, Bogotá, Colombia

1. Introducción

El conjunto de datos de películas es una herramienta valiosa para realizar análisis en la industria cinematográfica. Este conjunto de datos incluye una amplia gama de información como el título de la película, el distribuidor, el año de lanzamiento, el género, la recaudación en taquilla, la duración y la clasificación. El conjunto de datos incluye 918 muestras y 11 características, lo que lo convierte en un conjunto de datos relativamente grande. Los datos fueron recolectados a través de varias fuentes, incluyendo IMDb y Rotten Tomatoes. Se desea obtener el conjunto de datos óptimo para su manejo a través de diferentes estrategias de limpieza y preprocesamiento con el fin de comparar dos tipos de técnicas para el agrupamiento según las características similares que tenga las películas.

2. Desarrollo

Para el proyecto, las dos técnicas que se usaron para el agrupamiento de las películas fueron KMeans y Aglomerative Clustering. Primeramente, al KMeans se le asignaron diferentes cantidades de clústeres y diferentes semillas de inicialización para poder evaluar con el coeficiente de silueta cual tenía el mejor desempeño. Se tomaron las columnas con las que se deseaba aplicar el método para entrenar el modelo con los datos y obtener las etiquetas de cada muestra. Posteriormente con el Aglomerative Clustering se asignan los parámetros e igualmente se obtienen las etiquetas para cada muestra. Antes y después de la implementación de los métodos se realizaron diferentes tareas, como el preprocesamiento y la limpieza de los datos; y la normalización y el PCA respectivamente.

3. Resultados

En las siguientes imágenes se pueden apreciar los diferentes resultados de coeficiente de silueta para cierto número de clústeres, además de la visualización de la agrupaciones que se realizan con 2 o 3 clústeres:

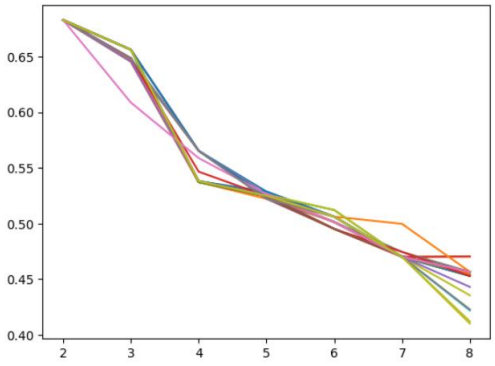
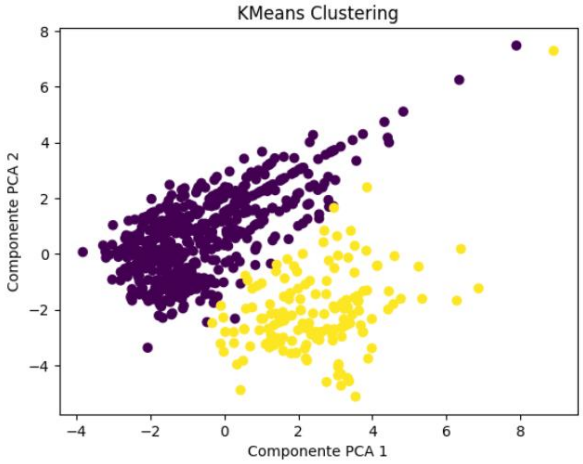


Ilustración 1. Coeficiente de silueta según clúster KMeans.

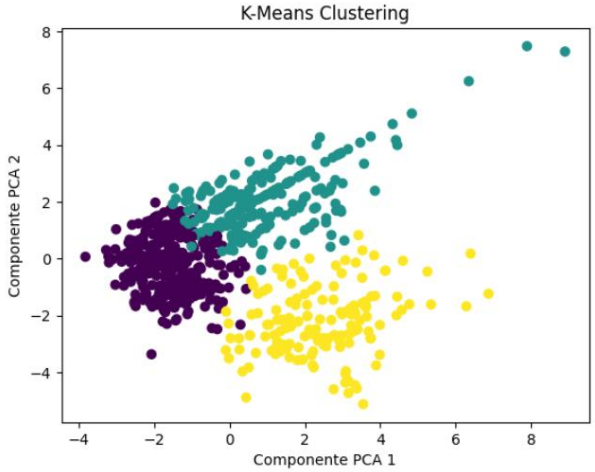
```
Coeficiente de Silueta: 0.6870275748396851
```

Ilustración 2. Agglomerative Clustering 2 clusters.



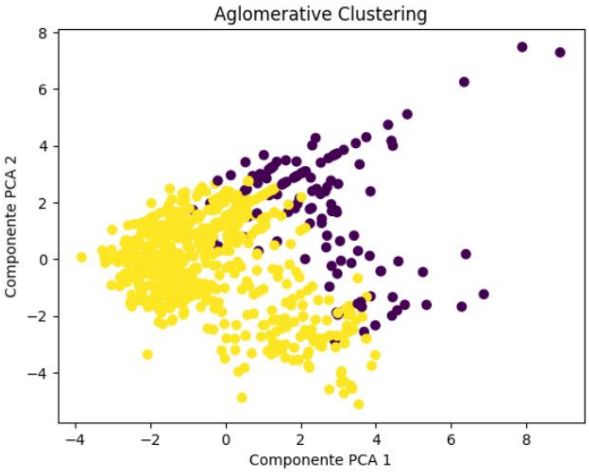

```
movies.Cluster1.value_counts()
```

```
0    583
1    161
Name: Cluster1, dtype: int64
```



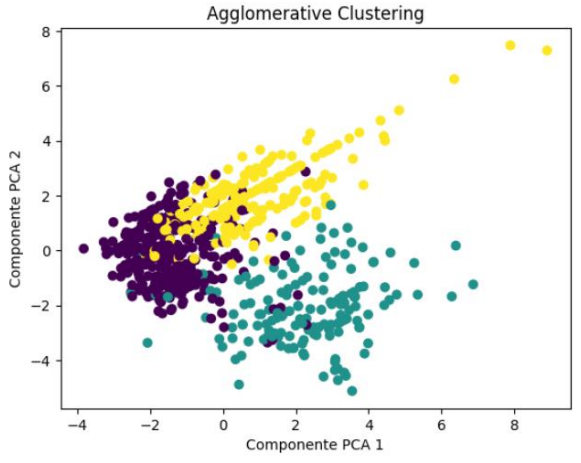

```
movies.Cluster1.value_counts()
```

```
0    369
1    220
2    155
Name: Cluster1, dtype: int64
```




```
movies.Cluster2.value_counts()

0    579
1    165
Name: Cluster2, dtype: int64
```




```
movies.Cluster2.value_counts()

0    368
2    211
1    165
Name: Cluster2, dtype: int64
```

4. Conclusiones

- Es necesario limpiar de manera correcta los datos, ya que la presencia de diferentes valores NaN o valores no válidos puede generar un mal desempeño del programa.
- Resulta mejor separar los arreglos con datos en su interior, como en nuestro caso los géneros, ya que si no se realiza podrían existir muchas combinaciones que no se manejen de manera adecuada.
- Es necesario normalizar los datos previamente a realizar la técnica de PCA.

- El PCA a pesar de que permite reducir la dimensionalidad de los datos, disminuye el coeficiente de silueta ya que, dependiendo de la varianza deseada, no toma en su totalidad todas las características.
- A pesar de que los resultados en el número de películas perteneciente a los clústeres resultan similar en ambas técnicas, gráficamente es posible notar que en el KMeans la separación entre los grupos resulta más notoria, mientras que en Agglomerative, principalmente para los 3 clústeres, no se puede diferenciar mucho y se ve datos encima de otros.