Precios de suscripciones de Netflix en diferentes países (2021)

Jimena Isaura Medina Padilla

2022-06-05

Capitulo I: Introducción

El Clustering Jerárquico (agrupamiento jerárquico o Hierarchical Clustering en inglés), es un método de data mining para agrupar datos El algoritmo de clúster jerárquico agrupa los datos basándose en la distancia entre cada uno y buscando que los datos que están dentro de un clúster sean los más similares entre sí. En una representación gráfica los elementos quedan anidados en jerarquías con forma de árbol. La mejor forma para explicarlo de manera gráfica es mediante un dendograma.

El objetivo de este proyecto es analizar mediante el cluster jerárquico una matriz de datos referente a los precios de subscripciones mensuales de la plataforma de streaming Netflix en diversos paises, con la finalidad de establecer si existen similitud entre el costo mensual en sus diferentes paquetes de servicio.

Para la elaboracion del diagrama jerárquico se utilizaron diversas paqueterías, entre ellas *dendextend* y *circlize* con la función "hclust" cuya función es realizar un análisis de conglomerados jerárquicos utilizando un conjunto de diferencias.

Capitulo II: Tratamiento de la matriz

Descripción de la matriz de datos

Los datos recabados para este proyecto fueron tomados del repositorio de bases de datos pertenecientes a Kaggle, con derechos de autoría a el corporativo Prasert Kanawattanachai, puesto que esta empresa recabó los datos para la matriz empleada. La matriz de datos trata de los precios en subscripciones a la plataforma de streaming Netflix de diferentes países en sus 3 paquetes; básico, estándar y premium.

Exploración de la matriz

1.-Se carga la paquetería y la ruta de la matriz de datos

```
library(readxl)
ruta_base <- "/Users/jimenamedina/Desktop/Netflix_data.xlsx"
BDN <- read_excel(ruta_base)
```

2.-Se obtiene su dimensión

```
dim(BDN)
```

```
## [1] 44  7
```

La matriz tiene una dimensión de 44 datos y 7 variables.

3.- Nombre de las variables

```
colnames(BDN)
```

```
## [1] "Country"                    "Total Library Size"
## [3] "No. of TV Shows"            "No. of Movies"
## [5] "Cost Per Month - Basic"     "Cost Per Month - Standard"
## [7] "Cost Per Month - Premium"
```

Los nombres de las variables son los siguientes: [1] “País” [2] “Tamaño total de la biblioteca” [3] “Número de programas de televisión” [4] “Número de películas” [5] “Costo por mes - Básico” [6] “Coste por mes - Estándar” [7] “Coste por mes - Premium”

4-.Se verifica que no existan datos perdidos

```
anyNA(BDN)
```

```
## [1] FALSE
```

No presenta datos perdidos

5.-Tipo de variables

```
str(BDN)
```

```
## tibble [44 x 7] (S3: tbl_df/tbl/data.frame)
##  $ Country                  : chr [1:44] "Argentina" "Australia" "Bolivia" "Brazil" ...
##  $ Total Library Size       : num [1:44] 4760 6114 4991 4972 6797 ...
##  $ No. of TV Shows          : num [1:44] 3154 4050 3155 3162 4819 ...
##  $ No. of Movies            : num [1:44] 1606 2064 1836 1810 1978 ...
##  $ Cost Per Month - Basic   : num [1:44] 3.74 7.84 7.99 4.61 9.03 7.91 7.07 4.31 8.99 9.03 ...
##  $ Cost Per Month - Standard: num [1:44] 6.3 12.12 10.99 7.11 11.29 ...
##  $ Cost Per Month - Premium : num [1:44] 9.26 16.39 13.99 9.96 13.54 ...
```

La matriz de datos presenta una variable cualitativa nominal (“países”), tres del tipo cuantitativa discreta (“Tamaño total de la biblioteca”, “Número de programas de televisión” y “Número de películas”) y tres del tipo cualitativa continua (“Costo por mes - Básico”, “Coste por mes - Estándar” y “Coste por mes - Premium”).

6.- Previsualización de los primeros 6 datos de la matriz

```
head(BDN)
```

```
## # A tibble: 6 x 7
##   Country    `Total Library S~ `No. of TV Show~ `No. of Movies ` `Cost Per Month ~
##   <chr>                  <dbl>            <dbl>           <dbl>              <dbl>
## 1 Argentina               4760             3154            1606               3.74
## 2 Australia               6114             4050            2064               7.84
## 3 Bolivia                 4991             3155            1836               7.99
## 4 Brazil                  4972             3162            1810               4.61
## 5 Bulgaria                6797             4819            1978               9.03
## 6 Canada                  6239             4311            1928               7.91
## # ... with 2 more variables: Cost Per Month - Standard <dbl>,
## #   Cost Per Month - Premium <dbl>
```

7.-Cálculo de la matriz de distancia de Mahalonobis (primeras 15 observaciones)

```
dist.BDN<-dist(BDN[,2:6])
head(dist.BDN, 15)
```

```
##  [1] 1686.9934  326.0400  294.3220 2657.0719 1905.2093  329.5575  325.2793
##  [8]  323.9758 3062.9624  327.4539 2187.0503  903.9704  852.6847 1149.6374
## [15]  260.7555
```

```
round(as.matrix(dist.BDN)[1:6, 1:6],3)
```

8.-Convertir los resultados del cálculo de la distancia a una matriz de datos y que indique 3 dígitos.

```
##          1        2        3        4        5        6
## 1    0.000 1686.993  326.040  294.322 2657.072 1905.209
## 2 1686.993    0.000 1454.008 1468.761 1032.109  319.753
## 3  326.040 1454.008    0.000   33.354 2459.816 1703.615
## 4  294.322 1468.761   33.354    0.000 2470.736 1714.482
## 5 2657.072 1032.109 2459.816 2470.736    0.000  756.260
## 6 1905.209  319.753 1703.615 1714.482  756.260    0.000
```

Capitulo III: Metodología de análisis

Elaboración del dendogrma

Para realizar el dendodogrma de este proyecto se utilizó la función **hclust** que se encarga de realizar un análisis de conglomerados jerárquicos utilizando un conjunto de diferencias para los n objetos que se agrupan. Inicialmente, cada objeto se asigna a su propio grupo y luego el algoritmo procede de manera iterativa, en cada etapa uniendo los dos grupos más similares, continuando hasta que haya un solo grupo. En cada etapa, las distancias entre los conglomerados se vuelven a calcular mediante la fórmula de actualización de disimilitud de Lance-Williams según el método de conglomerado particular que se utilice.

1.-Cálculo del dendrograma

```
dend.BDN<-as.dendrogram(hclust(dist.BDN))
dend.BDN
```

```
## 'dendrogram' with 2 branches and 44 members total, at height 5999.89
```

2.- Se realiza un dendograma exploratorio con la función hclust

```
DN <- BDN[2:6]
N<- dist(DN)
Nf <- hclust(N)
Nf
```

```
##
## Call:
## hclust(d = N)
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 44
```

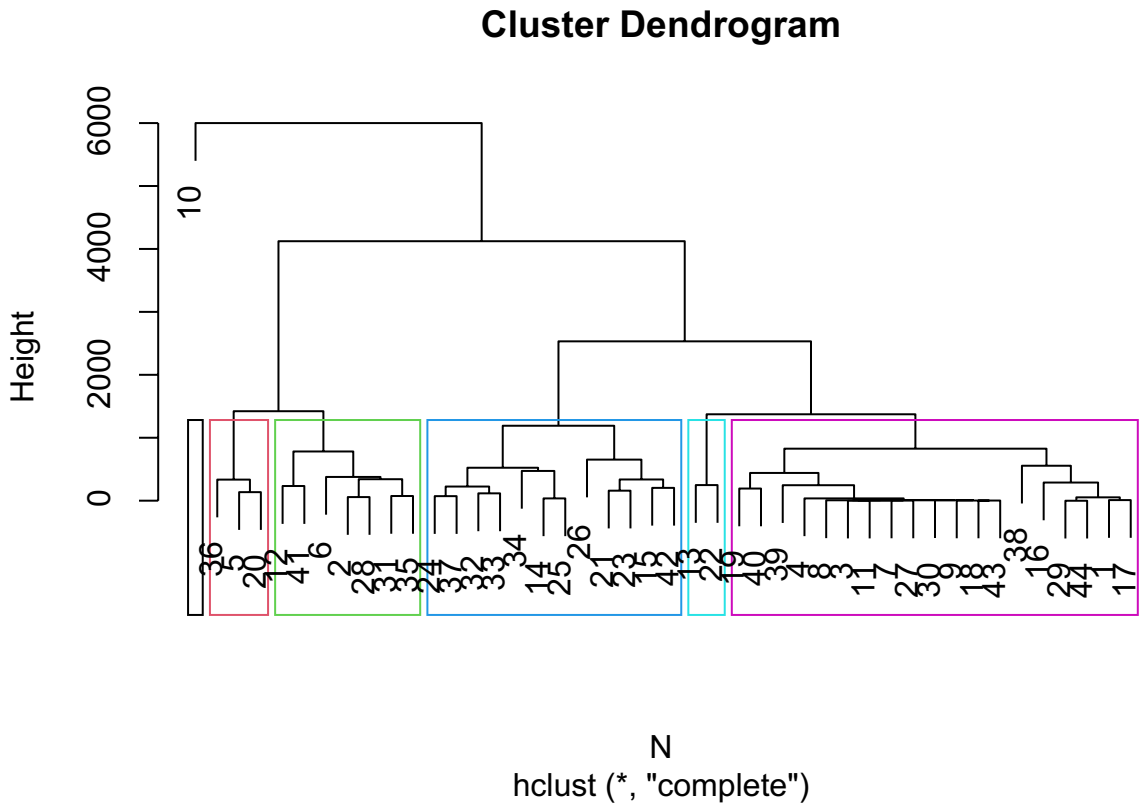
Capitulo IV: Resultados

1.- Cargar la librería para generar el dendograma

```
library("dendextend")
```

2.-Gráfico exploratorio del dendograma

```
plot(Nf)
rect.hclust(Nf, k = 6,
            border = 1:6)
```



En este dendograma se observa una partición final de 6 conglomerados, que ocurre a un nivel de similitud de aproximadamente 1500. El primer conglomerado (rectángulo negro) se compone de una observación. El segundo conglomerado (rectángulo rojo) está compuesto de tres observaciones. El tercer grupo (rectángulo verde) contiene siete observaciones. El cuarto conglomerado (rectángulo azul marino) consta de doce observaciones. El quinto (rectángulo azul cielo) presenta dos observaciones. Y el último conglomerado (rectángulo rosa) se compone de diecinueve observaciones.

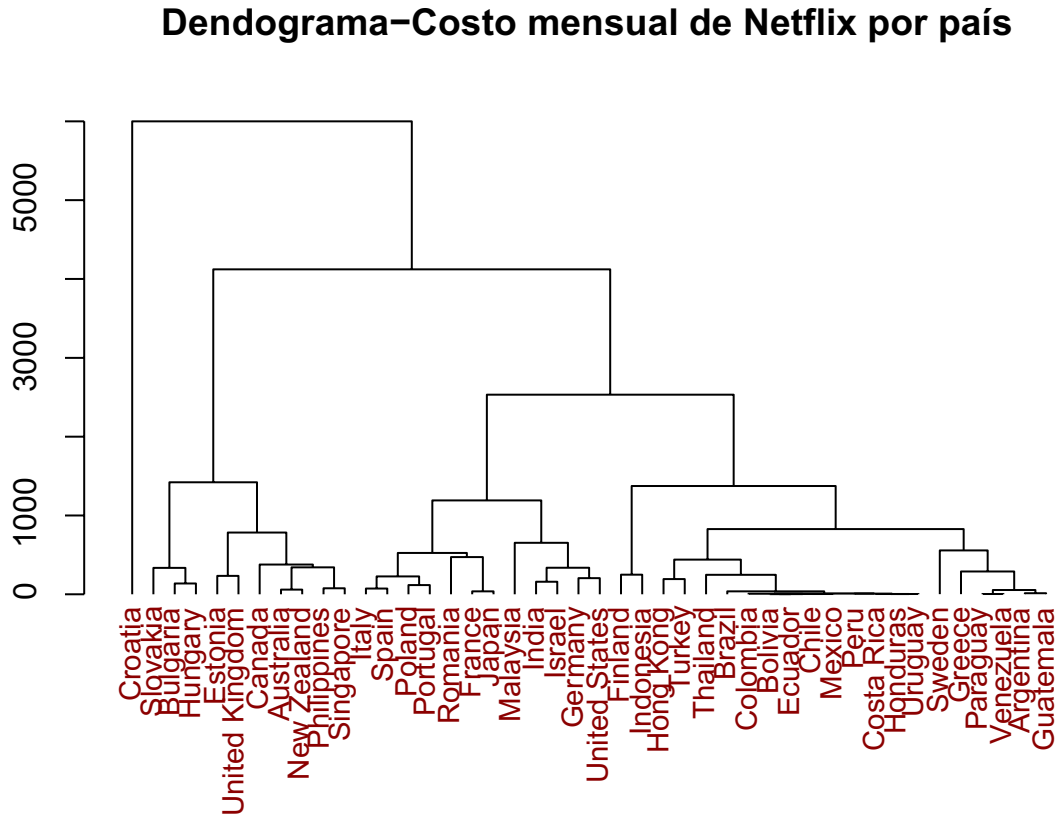
3.- Generación del dendrograma con etiquetas

3.1.-Guardar las etiquetas en un objeto “P”

```
P=labels(dend.BDN)
labels(dend.BDN)=BDN$Country[P]
```

3.2.-Color de las etiquetas

```
dend.BDN %>%
  set(what="labels_col","red4") %>%
  # Tamaño de las etiquetas
  set(what="labels_cex",0.9) %>%
  plot(main= "Dendograma-Costo mensual de Netflix por país")
```



Con la generación de etiquetas se puede apreciar el dendograma con los nombres de los países. En el primer conglomerado se observa que solo contiene a "Croacia". En el segundo conglomerado se observa que se encuentra "Eslovaquia", "Bulgaria" y "Hungría". En el tercer conglomerado se delimita desde "Estonia" hasta "Singapur". En el conglomerado cuatro se observa que se delimita desde "Italia" hasta "Estados Unidos". En el quinto grupo se encuentra "Hong Kong (China)" y "Turquía". Mientras que en el último conglomerado se observa delimitado desde "Tailandia" hasta "Guatemala".

4.-Dendograma circular

```
library("circlize")
```

```
circlize_dendrogram(dend.BDN, labels_track_height=NA,
                    dend_track_height=0.3)
```

Este dendogrma presenta una forma distinta para visualizar los conglomerados y sus respectivas observaciones, donde el círculo interior pertenece a las ramificaciones de los grupos y el círculo exterior con etiquetas son las observaciones.

Capitulo V: Conclusiones

Con base a los resultados anteriormente expuestos se puede llegar a la conclusión de que los países que presentan una mayor similitud en cuanto a los costos de suscripción mensual en sus diversos paquetes de la plataforma Netflix en el 2021 son los países pertenecientes al sexto conglomerado que son: “Tailandia”, “Brasil”, “Colombia”, “Bolivia”, “Ecuador”, “Chile”, “México”, “Perú”, “Costa Rica”, “Honduras”, “Uruguay”, “Suecia”, “Grecia”, “Paraguay”, “Venezuela”, “Argentina” y “Guatemala”, donde la mayoría pertenecen a América Latina. Y el conglomerado que presenta mayor diferencia en los costos de la plataforma es el primero, donde solo se encuentra “Croacia”.

Referencias

Anónimo. (2019). Algoritmos de Data Mining para agrupar datos – Clustering Jerárquico. Obtenido de Estrategias de Trading: https://estrategiastrading.com/clustering-jerarquico/

Prasert Kanawattanachai . (Enero de 2021). Netflix subscription fee in different countries. Obtenido de Kaggle: https://www.kaggle.com/datasets/prasertk/netflix-subscription-price-in-different-countries

Villardón, J. L. (2007). Introducción al análisis de clúster. Departamento de Estadística, Universidad de Salamanca.: 22p.