Proyecto de investigacion

Horacio Alberto

2022-06-06

Capitulo 1

Introduccion.

Mi base de datos trata sobre series de Netflix que contiene 5 variables las cuales son: “nombre de los directores, en el año que estuvo la serie, tiempo de cada capítulo, punto de popularidad de cada serie, puntos ganados” para la cual decidí trabajar con la variable de nombre de directores.

El objectivo es crear un dendograma donde se pueda reflejar claramente mi base de datos sobre los directores de series de netflix en el cual se puedan ver la similitudes de cada uno de ellos.

Capitulo 2

Tratamiento de la matriz

```
install.packages("cluster.datasets")
library("cluster.datasets")
```

```
library(readxl)
```

Bajamos la matriz de datos

```
series <- read_excel("series.xlsx")
```

Cambiamos el nombre de la matriz

```
SET=series
```

EXPLORACION DE KA MATRIS

DIMENSION

```
dim(SET)
```

```
## [1] 39  5
```

NOMBRE DE LAS VARIABLES

```
names(SET)
```

```
## [1] "name"            "release_year"    "runtime"         "tmdb_popularity"
## [5] "tmdb_score"
```

TIPO DE VARIABLES

```
str(SET)
```

```
## tibble [39 x 5] (S3: tbl_df/tbl/data.frame)
##  $ name           : chr [1:39] "Robert De Niro" "Jodie Foster" "Albert Brooks" "Harvey Keitel" ...
##  $ release_year   : num [1:39] 1945 1976 1975 1979 1973 ...
##  $ runtime        : num [1:39] 48 113 91 94 133 30 102 170 104 110 ...
##  $ tmdb_popularity: num [1:39] 0.6 27.6 18.2 17.5 95.3 ...
##  $ tmdb_score     : num [1:39] 8.4 8.2 7.8 7.8 7.7 8.3 7.5 7.6 6.2 7.5 ...
```

verificar si hay datos perdidos o no hay

```
anyNA(SET)
```

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

Capitulo 3

Metodologia

La técnica de análisis cluster o análisis de conglomerados consiste en clasificar a los individuos en estudio formando grupos o conglomerados (cluster) de elementos, tales que los individuos dentro de cada conglomerado presenten cierto grado de homogeneidad en base a los valores adoptados sobre un conjunto de variables. En el análisis cluster, a diferencia del análisis discriminante (donde los grupos están establecidos a priori y la función discriminante permite reasignar los elementos a los grupos), los conglomerados son desconocidos y el proceso consiste en su formación de modo óptimo, aglutinando unidades homogéneas.

Calculo de la matriz de distancia de Mahalonobis

```
dist.SET<-dist(SET[,2:5])
```

Convertir los resultados del Calculo de la distancia a una matriz de datos y me indique 3 digitos.

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

```
##         1      2      3      4       5       6
## 1   0.000 76.914 55.314 59.650 130.325  32.431
## 2  76.914  0.000 23.947 21.733  70.682  84.581
## 3  55.314 23.947  0.000  5.050  87.839  61.525
## 4  59.650 21.733  5.050  0.000  87.263  64.941
## 5 130.325 70.682 87.839 87.263   0.000 131.978
## 6  32.431 84.581 61.525 64.941 131.978   0.000
```

Calculo del dendrograma

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

Se Instala y Carga la Paqueteria para el Dendrograma

```
install.packages("dendextend")
library(dendextend)
```

Guardar las etiquetas en un objeto “L”

```
L=labels(dend.SET)
labels(dend.SET)=SET$name
```

Capitulo 4

Resultados

Dendorgrama:

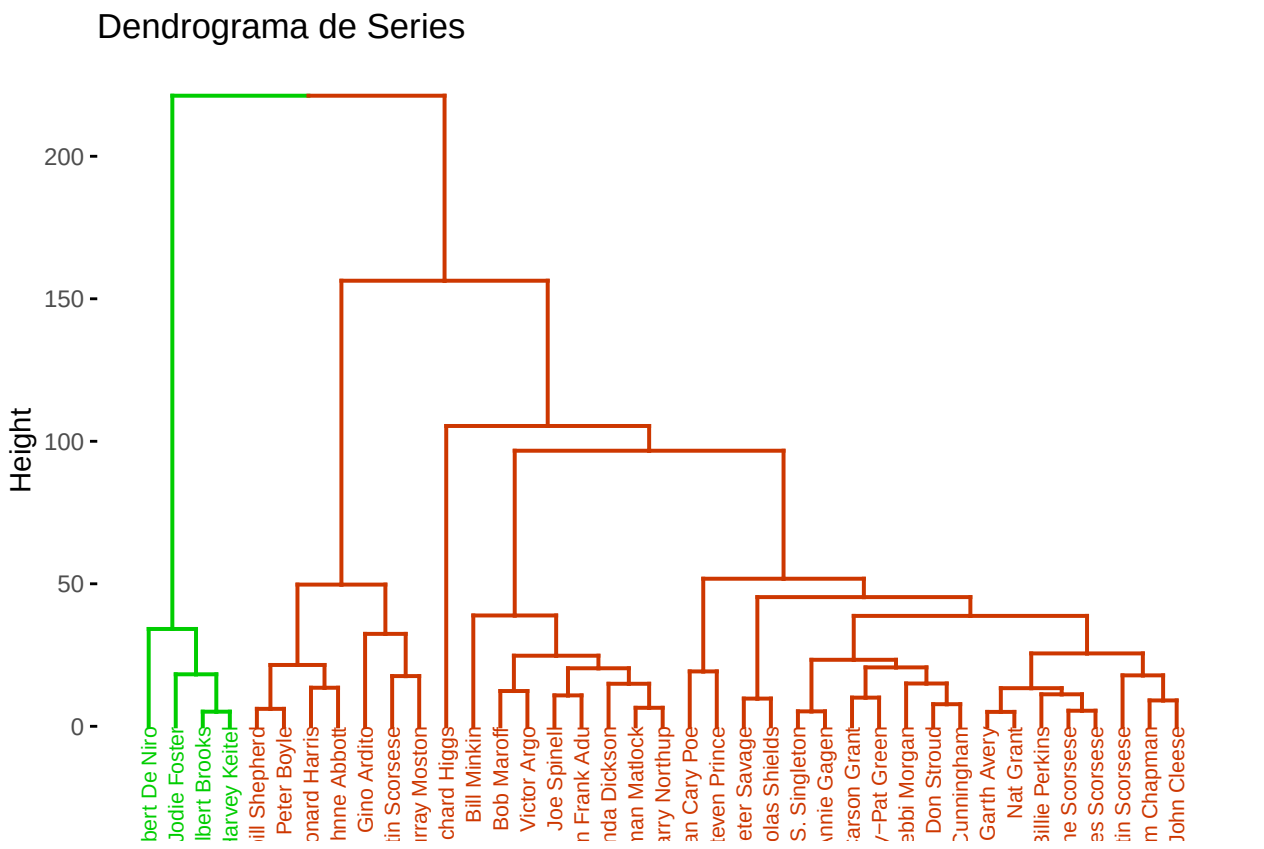
Un dendrograma es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado (asemejándose a las ramas de un árbol que se van dividiendo en otras sucesivamente). Este tipo de representación permite apreciar claramente las relaciones de agrupación entre los datos e incluso entre grupos de ellos aunque no las relaciones de similitud o cercanía entre categorías. Observando las sucesivas subdivisiones podemos hacernos una idea sobre los criterios de agrupación de los mismos, la distancia entre los datos según las relaciones establecidas, etc. También podríamos referirnos al dendrograma como la ilustración de las agrupaciones derivadas de la aplicación de un algoritmo de clustering jerárquico.

Creacion del Dendograma

```
install.packages("factoextra")
install.packages("ggplot2")
```

```
library(factoextra)
library(ggplot2)
```

```
fviz_dend(dend.SET,
          k = 2,cex=0.55,border=2:10,k_colors = c("#00CD00","#CD3700")) +
  labs(title = "Dendrograma de Series")
```



Capitulo 5

Conclusion

En el primer grupo comparten la similitud de que los 4 directores han hecho solo 1 sola serie, los cuales son: Jodie Foster, Albert Brooks y Harvey Keitel tienen el tiempo de popularidad por episodio y los dos últimos nombrados comparten el puntuaje de popularidad por temporada.

En el segundo grupo comparten que han hecho de 2 – 3 series cada director al igual que tienen las similitudes en la parte de los años que iniciaron con las filmaciones como también el tiempo de cada episodio que es alrededor de 70-150 minutos por cada uno pero por igual se nota que tiene los minutos de popularidad por episodio y al final se reflejan los puntos de cada serie generados de cada episodio.

Por lo tanto, los resultados no dan a entender cuáles fueron las similitudes de cada director de series de Netflix en los cuales se dividieron en 2 grupos, los cuales tenían su variables " año de la serie, minutos de cada episodio, cuantos minutos de popularidad tuvieron y cuantos puntos de reheating tuvo la serie".

Referencias

VÍCTOR SOEIRO (2022) https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

https://es.wikipedia.org/wiki/Dendrograma

https://www.ugr.es/~mvargas/2.RESUMENANLISISCLUSTER.pdf