https://datascience-project.eu/

# Plantilla de ficha de formación

| | |
|---|---|
| **Título** | Análisis Cluster (o de conglomerados) |
| **Palabras clave (metaetiquetas)** | Unidades estadísticas, conglomerados, intra-conglomerados, inter-conglomerados, índice de disimilitud, distancia de fusión, dendograma . |
| **Idioma** | Español |
| **Objetivos / Metas / Resultados de aprendizaje** | **El objetivo de este módulo es introducir y explicar la técnica del Análisis de Conglomerados.**<br><br>**Al final de este módulo podrás:**<br><br>- **Conocer la lógica del Análisis de Conglomerados**<br>- **Conocer los requisitos**<br>- **Realizar un análisis de conglomerados** |

| **Curso de formación:** | |
|---|---|
| **Alfabetización en ciencia de datos** | |
| **Módulo de visualización de datos y análisis visual** | X |
| **Introducción a la ciencia de datos para las ciencias humanas y sociales** | |
| **Ciencia de datos para siempre** | |
| **Periodismo de datos y storytelling** | |

| | |
|---|---|
| **Descripción** | En este módulo de formación se le presentará la técnica de análisis multidimensional denominada Cluster Analysis, también denominada análisis de grupo de conglomerados.<br>Los análisis de conglomerados se utilizan para agrupar unidades estadísticas que tienen características en común y asignarlas a categorías no definidas a priori. Los grupos que se formen deben ser lo más homogéneos posible en el interior (intra-cluster) y heterogéneos en el exterior (inter-cluster).<br>La aplicación de este tipo de análisis es múltiple: informática, medicina, biología, marketing.<br><br>La última parte del módulo estará dedicada a la aplicación del análisis de conglomerados con el software R. |

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

| Contenidos dispuestos en 3 niveles | **1. INTRODUCCIÓN**<br>Los análisis de conglomerados se utilizan para agrupar unidades estadísticas que tienen características en común y asignarlas a categorías no definidas a priori. Los grupos que se formen deben ser lo más homogéneos posible internamente (intra-cluster) y heterogéneos externamente (inter-cluster).<br>Los análisis de conglomerados son procedimientos que constan esencialmente de cuatro fases:<br><br>- Elección de variables<br>- Recopilación de datos<br>- Procesamiento de datos<br>- Verificar y usar los resultados<br><br>**2. REQUISITOS DEL ANÁLISIS CLÚSTER**<br><br>Se pueden utilizar varios tipos de variables en el análisis de conglomerados:<br><br>- Variables descriptivas (ejemplo: demográficas, socioeconómicas, geográficas)<br>- Variables de comportamiento (es decir, aquellas variables que responden preguntas: qué, cuándo, dónde, cómo y por qué)<br>Así que hablamos tanto de variables cualitativas como cuantitativas. La muestra disponible para el análisis de conglomerados deberá ser suficientemente numerosa, identificable, suficientemente estable, fácilmente accesible y suficientemente rentable.<br><br>**3. Cómo realizar un análisis de conglomerados**<br><br>**3.1 Matriz de disimilitud (o distancia), D**<br><br>Partimos de nuestra **matriz de datos X** , con dimensión nxp, y la transformamos en una **matriz de disimilitud D** , con dimensión nxn. Esto es útil para saber cuántas unidades estadísticas son diferentes entre sí y, por lo tanto, útil para elegir qué variables se deben considerar en el análisis. |
|---|---|

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & & x_{1,p} \\ & x_{i,k} & \\ x_{n,1} & & x_{n,p} \end{pmatrix} \Longrightarrow \mathbf{D} = \begin{pmatrix} d_{1,1} & & d_{1,n} \\ & d_{i,j} & \\ d_{n,1} & & d_{n,n} \end{pmatrix}$$

Como podemos ver la matriz **D** es una matriz simétrica que has 0 a lo largo de la diagonal principal, ya que la distancia de un punto consigo mismo es cero.

Para calcular las distancias entre los puntos, se usa el índice $d_{ij}$, que mide el grado de similitud entre i y j.

Hay diferentes índices you can use toque calculan estas distancias. Dependiendo del tipo de variable que estés usando, se debe usar uno en particular

**3.2 Distancias**

- Al utilizar **variables cuantitativas** nos referimos al **grado de disimilitud**. Hay varias formas de calcularlo:
  **Distancia Euclidea** :
  S se refiere a la teoría de Pitágoras , que es sensible a valores atípicos. Se calcula como:

$$d_{i,j} = \left[\sum_k (x_{i,k} - x_{j,k})^2\right]^{\frac{1}{2}}$$

  **Distancia Manhattan**:
  También llamada City Block, resulta ser más robusta que la distancia euclidiana y por lo tanto en se prefiere utilizar esta si es posible. Se calcula:

$$d_{i,j} = \sum_k |x_{i,k} - x_{j,k}|$$

En el cálculo de distancias, siempre se tienen en cuenta las unidades de medida de las variables. El efecto de la medición se puede eliminar

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

mediante la estandarización de la **matriz X** en la **matriz Z** , que vendrá dada por:

$$Z_k = \frac{(X_k - M_k)}{S_k}$$

Una vez que la matriz esté estandarizada, la usaremos para calcular el índice de disimilitud. La distancia Manhattan será:

$$d_{i,j} = \sum_k \frac{1}{S_k} |z_{i,k} - z_{j,k})|$$

donde $\frac{1}{S_k}$ indica una ponderación.

La estandarización se realiza si queremos dar el mismo peso a todas las variables. Si por el contrario se considera adecuado que una variable tenga un peso mayor que las demás, entonces no se realizará la estandarización.

- Al utilizar **variables Binarias** , es decir **,** variables que tienen solo dos categorías (**variables cualitativas** ). A las variables binarias se les asigna el estado 0 y 1. Con este tipo de variables calculamos **el grado de similitud,** es decir, la similitud entre i y j.
  Las variables binarias se dividen en:
  **Variables Binarias Simétricas**, **BS:** los dos estados (0 y 1) tienen la misma importancia.
  **Variables Binarias Asimétricas, BA**: se da más importancia al estado 1 que al estado 0.

  **Índice de Zubin** :
  Se utiliza para **Variables Binarias Simétricas** y se calcula sumando las frecuencias de concordancia y las frecuencias de discordancia , luego se divide por el total.

  $$s = \frac{(a+d)}{p}$$

  **Índice Jaccard** :

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Se utiliza para **variables binarias asimétricas** y se calcula dividiendo la frecuencia de concordancia por la diferencia entre el total y la frecuencia de discordancia.

$$s = \frac{a}{(p-d)}$$

**3.3 Tipos de clústeres**

Existen diferentes tipos de clústeres según el enfoque que se desee utilizar para crear grupos.

Los algoritmos jerárquicos realizan fusiones o divisiones sucesivas de datos. Una vez que un objeto se ha unido a un grupo, su asignación es irrevocable.

- **Clústeres aglomerativos o agregativos (bottom-up**):
  El objetivo es agrupar muchos clústeres y obtener un único clúster que contenga todos los presentes desde el principio.

- **Clústeres divididos o divisores (de arriba hacia abajo):**
  En este caso partimos de un solo clúster y el objetivo final es dividirlo en muchos clústeres.

**3.4) Tipos de Enlaces entre Unidades Estadísticas**

Los clústeres se pueden formar a través de diferentes tipos de enlaces:

- Enlace **simple o simple**
- Enlace **completo**
- Vinculación **media**

El **enlace simple** utiliza la técnica "del vecino más cercano". El grado de proximidad entre dos grupos se establece teniendo en cuenta la distancia mínima entre los puntos. En otras palabras, se tienen en cuenta las unidades que están más próximas entre sí. Este vínculo, sin embargo, a pesar de ser el más rápido de lograr a nivel computacional, crea grupos demasiado homogéneos entre ellos.

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

El **enlace completo** utiliza, en cambio, la técnica del "vecino más lejano". Considera las similitudes/distancias entre los grupos más distantes (por lo tanto, los menos parecidos entre sí). En la práctica, implica que solamente la distancia máxima entre los puntos se tiene en cuenta. Este enlace, a pesar de ser el más lento desde el punto de vista computacional, crea grupos muy heterogéneos entre sí y homogéneos internamente.

El **enlace o vínculo medio** para la creación de clústeres utiliza la distancia media mínima. En la práctica, primero se calcula la distancia media entre todas las observaciones y luego se tiene en cuenta la más pequeña. Este enlace también es lento desde el punto de vista computacional, pero es robusto y menos sensible a los valores atípicos.

El enlace **Ward** se puede utilizar con datos cuantitativos. Esta técnica minimiza la varianza dentro de los grupos al homogeneizarlos. En la práctica, este método maximiza la homogeneidad interna (o minimiza la heterogeneidad interna) y maximiza la heterogeneidad entre clusters.

**3.5 Dendograma y distancia de fusión**

Una vez elegido el enlace que mejor representa los datos estudiados, se procederá a calcular el **dendograma.** Podemos visualizar a través de un **gráfico de árbol** cómo se han distribuido las unidades estadísticas. En cada paso, la distancia entre los grupos tiende a aumentar y, por lo tanto, es necesario elegir una **regla de parada.** Esta regla nos permite elegir el número de grupos que queremos obtener. Puede utilizarse la técnica de tala de árboles a través de la gráfica de **distancias de fusión** (o alturas), que indica dónde se crean los conglomerados. Gráficamente observamos el punto en el que registramos un mayor salto. Esta parte se retomará luego en la parte del módulo dedicada al software R.

**4. Ejemplo con software R**

El análisis de conglomerados tiene como objetivo identificar la mejor distribución posible, en términos de número y composición, de un conjunto de elementos en grupos para que estos sean: lo más homogéneos posible dentro de ellos y lo más diferentes posible entre sí. Estas construcciones pueden llevarse a cabo tanto en función de la

Co-funded by the Erasmus+ Programme of the European Union

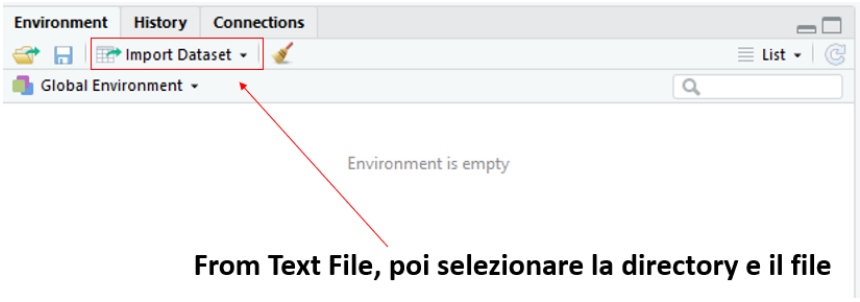
Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

elección de las estrategias de agrupación, como en relación con el criterio elegido para la medida de similitud/desigualdad.

Conjunto de datos:

| Nazioni | Cereali | Riso | Patate | Zucchero | Verdure | Vino | Carne | Latte | Burro | Uova |
|---|---|---|---|---|---|---|---|---|---|---|
| Belgio | 72,2 | 4,2 | 98,8 | 40,4 | 103,2 | 20,9 | 102 | 80 | 7,7 | 14,2 |
| Danimarca | 70,5 | 2,2 | 57 | 39,5 | 50 | 22 | 105,8 | 145,2 | 4,1 | 14,3 |
| Germania | 71,3 | 2,3 | 74,1 | 37,1 | 83,1 | 22,8 | 97,2 | 90,7 | 6,9 | 14,8 |
| Grecia | 109,8 | 5,4 | 90 | 30 | 229,5 | 25,3 | 77,1 | 63,1 | 0,9 | 11,3 |
| Spagna | 71,4 | 5,8 | 107,8 | 26,8 | 191,7 | 43 | 102,1 | 98,4 | 0,6 | 15,3 |
| Francia | 73 | 4,3 | 78,2 | 34,1 | 95 | 64,5 | 110,5 | 98,9 | 8,9 | 15 |
| Irlanda | 93,4 | 3,2 | 151,5 | 34,8 | 55 | 3,9 | 105 | 185,9 | 3,4 | 11,4 |
| Italia | 110,2 | 4,8 | 38,6 | 27,9 | 181,9 | 61,6 | 88 | 65 | 2,4 | 11,1 |
| Olanda | 54,6 | 5 | 86,7 | 39,7 | 99 | 14 | 89,4 | 136,2 | 5,4 | 10,7 |
| Portogallo | 86 | 5,7 | 106,6 | 29,4 | 100 | 57 | 75,5 | 96 | 1,5 | 7,7 |
| RegnoUnito | 74,3 | 4,5 | 94,1 | 39,8 | 60 | 10,4 | 74,4 | 129,3 | 3,2 | 10,8 |
| Austria | 68,7 | 4,2 | 62,6 | 37,1 | 81,9 | 34,3 | 93,4 | 121,3 | 4,3 | 13,4 |
| Finlandia | 70,1 | 5,4 | 61,6 | 35,7 | 52,6 | 10,2 | 65 | 208,4 | 5,8 | 10,9 |
| Islanda | 79,7 | 1,9 | 50,2 | 54,9 | 50 | 6,2 | 71,7 | 205,6 | 4,6 | 11,3 |
| Norvegia | 76,9 | 3,5 | 73,2 | 37,3 | 48,3 | 6,6 | 54,9 | 176,5 | 2,1 | 11,3 |
| Svezia | 69,3 | 4,3 | 70 | 37,5 | 48,5 | 12,3 | 60,5 | 154,1 | 5,7 | 12,9 |

Importamos el conjunto de datos:




En los ***nombres de las filas,*** seleccione la redacción: " ***usar la primera columna*** " para tener las etiquetas de los individuos y las variables en los gráficos.

En ***decimal*** seleccionamos: "***coma***".

Con el comando:

**X<-as.matrix(dataset_name)**

Creamos un objeto **X** : el conjunto de datos utilizado en el análisis.

Estandarizamos la matriz **X**:

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

**Z<-scale(X)**

A continuación, calculamos la distancia entre los elementos, podemos usar la distancia euclidiana o la distancia de Manhattan.

Respectivamente los comandos son:

**d<-distance(Z)**

**D<-round(D,2)**

**d_m <- dist (Z, method=" manhattan ")**

**d_m<-round(d_m, 2)**

Nota: el comando redondear nos permite redondear hacia arriba a la cifra significativa que prefiramos, en este caso a la segunda.

Luego pasamos a la elección del vínculo entre los elementos.

Comencemos con el **enlace único (o simple)** :

**hc_s <- hclust ( d,methosd ="single")**

Podemos mostrar un **resumen de los resultados del** enlace simple con el comando:

**summary( hc_s )**

Podemos visualizar el **dendograma** con la función plot:

**plot(hc_s)**

Para decidir dónde cortar el gráfico de árbol, utilice el comando **cutree.** La elección de cuántos grupos obtener al mostrar el punto de fusión a través del **gráfico de sedimentación** del enlace de la distancia de fusión. Los comandos son:

**n<-nrow(X)**

**n_clus<-seq(n-1,1)**

**hc_s$merge**

**hc_s$height**

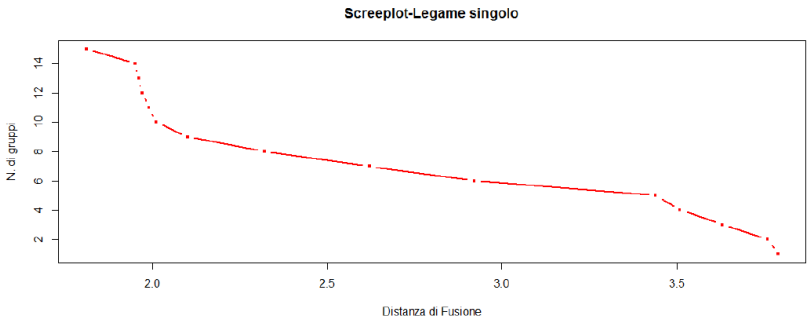
**d_fus_s<-hc_s$height**

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

**plot(d_fus_s,n_clus,"b", main="Screeplot Single bond", xlab="Melting Distance", ylab="Number of groups",cex=0.6, col="red",lwd=2.5)**

Gráficamente:




Los puntos de fusión (hc_s$merge) y las alturas (hc_s$height) pueden visualizarse juntos usando el comando **cbind** . El comando $merge muestra, para cada paso, del algoritmo de agrupación y el par de elementos fusionados según el enlace elegido. Los valores precedidos por "-" indican un elemento único, mientras que los valores positivos representan grupos formados en pasos anteriores.
Así, en el primer paso, el primer clúster estará formado por el par (13, 16), correspondiente a los modelos de Finlandia y Suecia, mientras que el tercer clúster (paso 10) estará formado por los elementos del clúster 2 (Grecia, Italia) más el elemento 1 (Francia). El campo $height muestra la distancia considerada para la fusión entre elementos/grupos.

**cbind(hc_s$merge,hc_s$height)**

```
> cbind(hc_s$merge,hc_s$height)
      [,1] [,2] [,3]
 [1,]  -13  -16 1.81
 [2,]   -2   -3 1.95
 [3,]   -1    2 1.96
 [4,]  -15    1 1.97
 [5,]  -11    4 1.99
 [6,]   -9    5 2.01
 [7,]  -12    3 2.10
 [8,]    6    7 2.32
 [9,]   -6    8 2.62
[10,]   -4   -8 2.92
[11,]  -14    9 3.44
[12,]   -7   11 3.51
[13,]  -10   12 3.63
[14,]   10   13 3.76
[15,]   -5   14 3.79
```

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

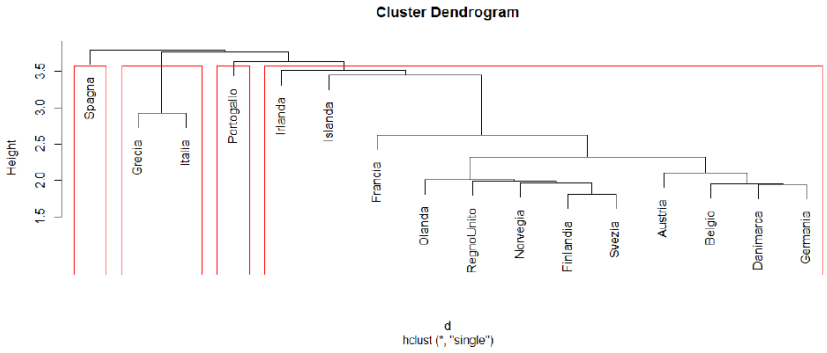
Para cortar el árbol usamos el comando cutree, en k ponemos el punto donde la distancia de fusión toma una tendencia horizontal:

**groups <- cutree(hc_s, k=4)**

**plot(hc_s)**

**rect.hclust(hc_s, k=4, border="red")**

El dendograma será:




Podemos decir que este tipo de enlace no es bueno, porque hay clusters que contienen elementos únicos y un clusters que es demasiado homogéneo dentro de él.

Procedemos con los demás enlaces de la misma forma.

Enlace completo:

**hc_c<-hclust(d,method="compl")**

**summary(hc_c)**

**plot(hc_c)**

**n<-nrow(X)**

**n_clus<-seq(n-1,1)**

**hc_c$merge**

**hc_c$height**

**d_fus_c<-hc_c$height**

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Screeplot de distancias de fusión para el enlace completo:

**plot(d_fus_c,n_clus,"b", main="Screeplot Full Bond", xlab="Melting Distance", ylab="N. of groups",cex=0.6, col="red",lwd=2.5)**

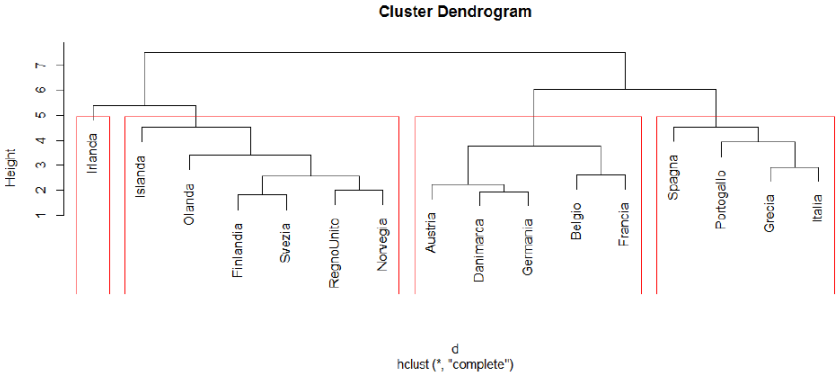
**cbind(hc_c$merge,hc_c$height)**

Recortando el gráfico de árbol para el enlace Completo, a k le atribuiremos la cifra según el screeplot de distancias de fusión:

**groups <- cutree(hc_c, k=4)**

**plot(hc_c)**

**rect.hclust(hc_c, k=4, border="red")**




Vinculación media:

**hc_a<-hclust(d,method="average")**

**summary(hc_a)**

**plot(hc_a)**

**n<-nrow(X)**

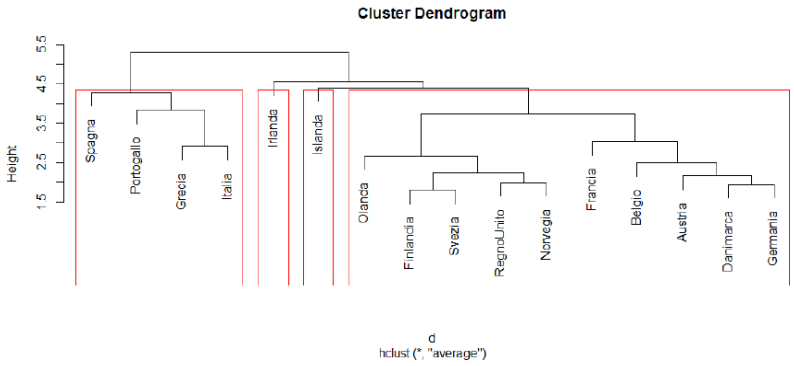
**n_clus<-seq(n-1,1)**

**hc_a$merge**

**hc_a$height**

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

| | |
|---|---|
| | **d_fus_a<-hc_a$height**<br><br>Screeplot de distancias de fusión para el enlace medio:<br><br>**plot(d_fus_a,n_clus,"b", main="Screeplot Mean bond", xlab="Melting Distance", ylab="N. of groups",cex=0.6, col="red",lwd=2.5)**<br><br>Cortando el eje para el varillaje medio, a k le asignaremos la cifra según el screeplot de distancias de fusión:<br><br>**groups <- cutree(hc_a, k=4)**<br><br>**plot(hc_a)**<br><br>**rect.hclust(hc_a, k=4, border="red")**<br><br>  |
| **Autoevaluación (preguntas y respuestas de opción múltiple)** | 1. La matriz de distancias:<br>a) tiene una diagonal principal compuesta de 0s<br>B) tiene una diagonal principal compuesta de 1s<br>c) tiene una diagonal principal con las distancias entre i y j<br><br>2. ¿Cual de estas distancias es más robusto, o insensible a valores extremos?<br>A) Índice Jaccard<br>B) City block |

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

| | |
|---|---|
| | C) Distancia Euclídea<br><br>3. La estandarización debería hacer posible:<br>A) Eliminar las frecuencias más altas<br>B) Eliminar el efecto de la unidad de medición<br>C) Dar diferente peso a las variables |
| **Recursos (videos, enlaces a referencias)** | |
| **material relacionado** | |
| **PPT relacionado** | |
| **Bibliografía** | Johnson, S. C. (1967). Hierarchical clustering schemes, Psychometrika, 32, 241-254.<br><br>Pollice, A. (2013). Statistica multivariata, http://www.uniba.it/ricerca/dipartimenti/dse/dipartimento/personale/personale-docente/pollice/stat_mult/disp10.pdf<br><br>Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function, Journal of American Statistical Association, 58, 236-244. |
| **Proporcionado por** | [ Unisalento ] |

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.