https://datascience-project.eu/

Fișă de învățare

Titlu	Analiza cluster
Cuvinte cheie (meta tags)	Unități statistice, Cluster, intra-cluser, inter-cluster, indice de disimilaritate, distanță de agregare, dendogramă.
Limbă	Română
Obiective / Scop **Rezultatele învățării**	**Scopul acestui modul este de a introduce și de a explica tehnica Analizei Cluster.** **La finalul acestui modul, vei fi capabil să:** - **Cunoști logica Analizei Cluster** - **Cunoști cerințele** - **Realizezi o Analiză Cluster**

Curs:	
Data Science Literacy	
Vizualizarea Datelor și Modulul de Visual Analytics	X
Introducere în Data science pentru Științe sociale	
Data Science for good	
Data Journalism și Storytelling	

Descriere	În acest modul de învățare va fi prezentată tehnica multidimensională a Analizei Cluster, cunoscută și sub numele de Analiză automată a grupurilor. Analizele cluster sunt utilizate pentru a grupa unitățile statistice care au caracteristici comune și pentru a le aloca pe categorii care nu sunt definite a priori. Grupurile formate trebuie să fie cât mai omogene în interior (intra-cluster) și cât mai eterogene între ele (inter-cluster). Aplicațiile acestui tip de analiză se regăsesc în mai multe domenii: informatică, medicină, biologie, marketing. Ultima parte a modulului este dedicată aplicațiilor analizei cluster cu ajutorul software-ului R.
Conținutul este organizat pe 3 niveluri	**1. INTRODUCERE** Analiza cluster este utilizată pentru a grupa unități statistice care au caracteristici comune și pentru a le aloca pe categorii care nu sunt

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

definite a priori. Grupurile formate trebuie să fie cât mai omogene în interior (intra-cluster) și cât mai eterogene între ele (inter-cluster). Analizele cluster sunt proceduri care constau în patru etape:

- Alegerea variabilelor
- Colectarea datelor
- Procesarea datelor
- Verificarea și utilizarea rezultatelor

2. CERINȚELE (IPOTEZELE) ANALIZEI CLUSTER

Mai multe tipuri de variabile poti fi utilizate în analiza cluster :

- Variabile descriptive (exemplu: demografice, socio-economice, geografice)
- Variabile comportamentale (acele variabile care răspund la întrebările: ce, când, unde, cum și de ce)

Deci vom discuta mai departe atât despre variabile calitative, cât și cantitative.

Eșantionul disponibil pentru analiza cluster trebuie să fie suficient de mare, identificabil, destul de stabil, ușor accesibil și suficient de util.

3. Cum se efectuează Analiza Cluster

3.1 Matricea de proximitate (sau Matricea de Distanță), D

Se pornește de la **matricea de date X**, cu dimensiunile nxp și o transformăm într-o **matrice de proximitate, D,** cu dimensiunile nxn. Aceasta din urmă este utilizată pentru a afla câte unități statistice sunt diferite între ele și deci este utilă în alegerea variabilelor care ar trebui luate în calcul în analiză.

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

După cum se poate observa, matricea **D** este o matrice simetrică, care pe diagonala principală are elemente zero, întrucât distanța de la un punct la el însuși este zero.

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Pentru a calcula distanțele dintre puncte se utilizează indicele $d_{i,j}$, reprezentând o măsură a gradului de similitudine dintre i și j.

În funcție de tipul variabilelor utilizate, există mai multe feluri de indici care se pot utiliza pentru a calcula aceste distanțe.

3.2 Tipuri de distanțe

- Atunci când utilizăm **variabile cantitative,** ne referim la **gradul de disimilaritate.** Există mai multe moduri pentru a-l calcula:
  Distanță Euclidiană:
  Are ca punct de referință teorema lui Pitagora și se dovedește a fi sensibilă la valori extreme (outlieri). Modul de calcul este:

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

  Distanța Manhattan:
  Distanță de tip City Block, este mai robustă decât distanța Euclidiană, de aceea se preferă utilizarea acesteia atunci când este posibil. Modul de calcul este:

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

În calculul distanțelor, unitățile de măsură ale variabilelor sunt luate în considerare întotdeauna. Efectul unităților de măsură poate fi eliminat prin standardizarea **matricii X** în **matricea Z**, obținută astfel :

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

Odată ce matricea este standardizată, va fi utilizată pentru calculul indicatorilor de disimilaritate. Distanța Manhattan devine:

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

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

unde $\frac{1}{S_k}$ este ponderea.

Standardizarea este realizată dacă se dorește acordarea aceleiași ponderi tuturor variabilelor; dacă, pe de altă parte, se consideră oportun ca o variabilă să aibă o pondere mai mare decât altele, atunci standardizarea nu va avea loc.

- Atunci când se utilizează **variabile binare**, deci variabile care au doar două stări posibile (variabile care fac parte din categoria **variabilelor calitative**). Celor două stări posibile le sunt atribuite valorile 0 și 1. Pentru acest tip de variabile, se calculează gradul de similaritate, adică similaritatea dintre i și j. Variabilele binare se împart în:
  Variabile binare simetrice, BS: cele două stări (0 și 1) au aceeași importanță.
  Variabile binare asimetrice, BA: mai multă importanță este acordată stării 1 decât stării zero.

  Indicele Zubin:
  Este utilizat pentru **variabilele binare simetrice**, este calculat prin adunarea frecvențelor de concordanță și discordanță, împărțit la total.

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

  Indicele Jaccard:
  Se utilizează pentru **variabilele binare asimetrice**, calculat prin împărțirea frecvențelor de concordanță la diferența dintre total și frecvențele de discrodanță.

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

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3.3 Tipuri de clustere

Există mai multe tipuri de clustere în funcție de abordarea utilizată în crearea grupurilor.

Algoritmii ierarhici efectuează agregări sau divizări succesive ale datelor; odată ce un obiect a fost alocat unui cluster, această asignare este irevocabilă.

- **Clustere obținute prin amalgamare sau agregare (bottom-up**): Scopul este gruparea clusterelor și obținerea unui singur cluster care să le conțină pe toate.

- **Clustere obținute prin dezagregare sau divizare (top-down):** În acest caz se pornește de la un singur cluster și scopul este împărțirea acestuia în mai multe clustere.

3.4 Tipuri de agregare între unități statistice

Clusterele pot fi formate utilizând diferite metode de agregare:

- Agregare **simplă** sau singulară
- Agregare **completă**
- Agregare **medie**

Agregarea simplă utilizează metoda ”celor mai apropiați vecini”. Gradul de proximitate dintre două clustere se stabilește luând în calcul distanța minimă dintre puncte. Cu alte cuvinte, se iau în considerare unitățile care sunt cele mai apropiate unele de altele. Totuși, această metodă de agregare, deși este cea mai rapidă la nivel computațional, creează grupuri care sunt prea omogene între ele.

Agregarea completă utilizează, în schimb, metoda ”celor mai îndepărtați vecini”. Consideră similaritatea/distanța dintre cele mai îndepărtate grupuri (deci cele care sunt cel mai puțin asemănătoare între ele). În practică, distanța minimă maximă dintre puncte este luată în considerare. Această agregare, desi este cea mai lentă din punct de vedere computațional, creează grupuri foarte eterogene între ele și omogene în interior.

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Agregarea medie în crearea clusterelor utilizează metoda distanțelor medii dintre perechi. În practică, mai întâi se calculează distanța medie dintre toate observațiile și apoi cea mai mică dintre acestea se ia în considerare. Acest tip de agregare este de asemenea lent din punct de vedere computațional, dar este unul robust, mai puțin sensibil la valorile extreme.

Metoda lui Ward poate fi utilizată pentru variabile cantitative. Această metodă minimizează variabilitatea din interiorul claselor, omogenizându-le. În practică, acestă metodă maximizează omogenitatea internă (sau minimizează eterogenitatea internă) și maximizează eterogenitatea externă.

3.5 Dendograma și distanța de agregare

Odată aleasă metoda optimă de agregare, se va obține **dendograma**. Se poate vizualiza printr-un **grafic de tip arbore** modul în care au fost distribuite unitățile statistice. La fiecare pas distanța dintre clustere are tendința de a crește și de aceea este necesară definirea unei **reguli de stop** a algoritmului. Această regulă permite alegerea numărului de clase pe care le obținem. Se poate utiliza tehnica de secționare a arborelui, pe baza **distanței sau a pasului de agregare** (distanțele care indică unde sunt create clusterele). Grafic, se observă punctul în care distanța de agregare are valoarea cea mai mare. Această parte a analizei va fi dezvoltată în secțiunea modulului dedicată software-ului R.

4. Exemplu în software R

Analiza cluster își propune să identifice cea mai bună distribuție posibilă, în termeni de număr și componență, a unui set de elemente în clase, astfel încât acestea să fie : cât mai omogene în interior și cât mai diferite între ele. Aceste construcții pot fi realizate ținând cont atât de alegerea strategiilor de grupare a elementelor, cât și în legătură cu criteriile alese pentru a măsura similaritatea sau disimilaritatea.

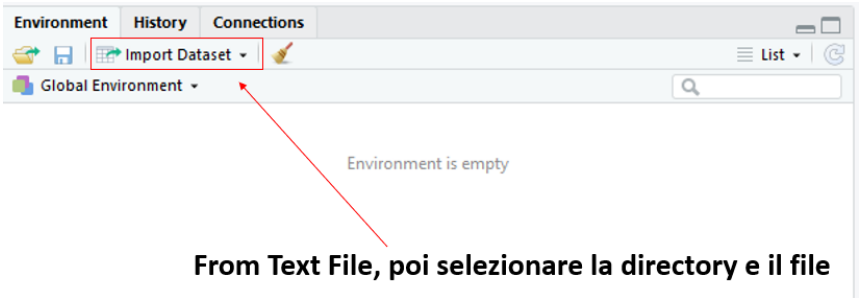
Setul de date:

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

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

Se importă setul de date:



From Text File, poi selezionare la directory e il file

Pentru ***row names*** se selectează: "***use first column***" pentru a avea etichetele atât pentru observații, cât și pentru variabile în grafice.

Pentru specificarea ***zecimalelor*** se selectează: "***comma***".

Cu ajutorul comenzii:

X<-as.matrix(nume_set_date)

Se atribuie matricea **X**, ca obiect, setului de date utilizat în analiză

Se standardizează matricea **X**:

Z<-scale(X)

În continuare, se calculează distanța dintre obiecte, pentru care putem folosi fie distanță Euclidiană, fie distanță Manhattan.

Comenzile pentru fiecare distanță în parte sunt:

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

d<-dist(Z)

D<-round(D,2)

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

d_m<-round(d_m, 2)

NB: comanda ”round” permite rotunjirea prin adaos la cifra preferată, în acest caz la a doua.

Apoi alegem metoda de agregare dintre elemente.

Vom începe cu **agregarea simplă:**

hc_s<-hclust(d,method="single")

Se poate afișa un **sumar al rezultatelor** metodei agregate cu ajutorul comenzii:

summary(hc_s)

Vizualizarea **dendogramei** se poate face cu funcția plot:

plot(hc_s)

Pentru a decide unde se va secționa arborele, utilizăm comanda **cutree**. Alegerea numărului de clase se poate face în urma afișării punctului de agregare în cadrul scree-plot aferent distanței de agregare. Secvența de comenzi este următoarea:

n<-nrow(X)

n_clus<-seq(n-1,1)

hc_s$merge

hc_s$height

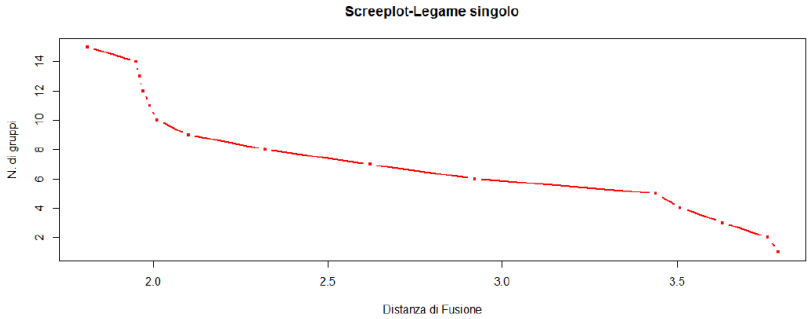
d_fus_s<-hc_s$height

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)

Grafic, se obține:

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



În continuare se pot vizualiza punctele de agregare (hc_s$merge) și distanțele (hc_s$height), iar pentru a le vizualiza împreună se poate utiliza **cbind**. Comanda $merge arată, pentru fiecare etapă a algoritmului de grupare, perechea de elemente adăugate, conform metodei de agregare alese. Valorile precedate de ”-” indică elementul unic (singular), în timp ce valorile pozitive reprezintă clusterele formate în etapele anterioare.
Astfel, în etapa 1, primul cluster va fi format pe baza perechii de elemente (13,16), corespunzătoare modelelor Finlandei și Suediei, în timp ce clusterul trei (la etapa 10) va fi format din elementele clusterului 2 (Grecia, Italia) plus elementul 1 (Franța). Câmpul $height arată distanța cosiderată pentru agregarea între elemente/ grupuri.

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
```

Pentru a secționa arborele, se utilizează comanda cutree, pentru parametrul k vom alege punctul în care distanța de agragare intră pe un trend orizontal:

groups <- cutree(hc_s, k=4)

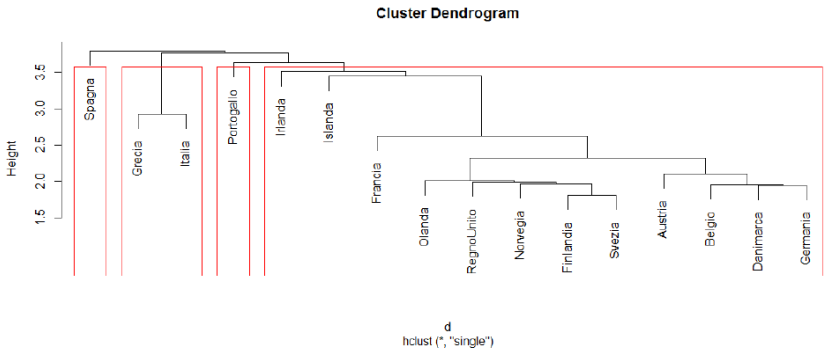
Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

plot(hc_s)

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

Dendograma va fi:



Putem spune că acest tip de agregare nu este una bună, deoarce există clustere care conțin un singur element și un cluster care este mult prea omogen în interior.

În continuare, procedăm la aplicarea celorlalte metode de agregare, într-un mod similar.

Agregarea completă:

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

Screeplot aferent distanțelor de agregare pentru Full bond:

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

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)

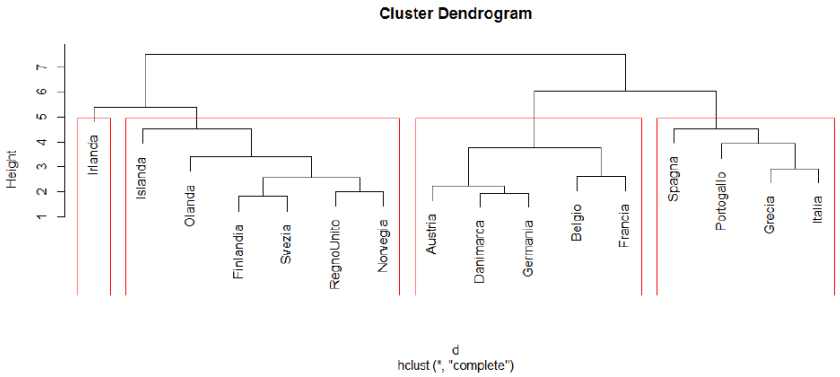
cbind(hc_c$merge,hc_c$height)

Secționarea graficului de tip arbore pentru agregare completă, unde pentru k vom atribui valoarea corespunzătoare screeplot-ului distanțelor de agregare:

groups <- cutree(hc_c, k=4)

plot(hc_c)

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



Agregare medie:

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

summary(hc_a)

plot(hc_a)

n<-nrow(X)

n_clus<-seq(n-1,1)

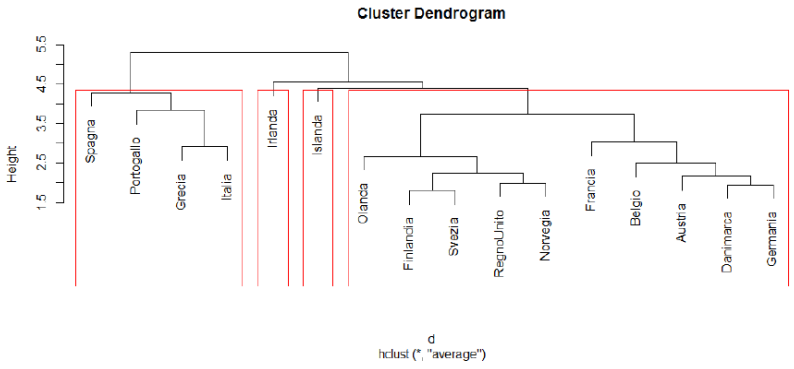
hc_a$merge

hc_a$height

d_fus_a<-hc_a$height

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	Screeplot aferent distanțelor de agregare pentru agregare medie: **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)** **cbind(hc_a$merge,hc_a$height)** Secționarea arborelui pentru agregarea medie, pentru k vom atribui valoarea conform screeplot aferent distanțelor de agregare: **groups <- cutree(hc_a, k=4)** **plot(hc_a)** **rect.hclust(hc_a, k=4, border="red")** 
Auto-evaluare (întrebări cu răspuns multiplu și răspunsuri)	1. Matricea de distanțe: A) **Are pe diagonala principală toate valorile 0** B) Are pe diagonala principală toate valorile 1 C) Are pe diagonala principală distanțele dintre i și j 2. Care din aceste distanțe este mai robustă sau mai putin senzitivă la valorile extreme? A) Indicele Jaccard B) City block

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	C) **Distanța Euclidiană** 3. Standardizarea va face posibilă: A) **Eliminarea frecvențelor ridicate** B) Eliminarea efectelor determinate de unitățile de măsură C) Acordarea de ponderi diferite variabilelor
Resurse (video, link-uri)	
Materiale adiționale	
PPT	
Bibliografie	Johnson, S. C. (1967). Hierarchical clustering schemes, Psychometrika, 32, 241-254. Pollice, A. (2013). Statistica multivariata, http://www.uniba.it/ricerca/dipartimenti/dse/dipartimento/personale/personale-docente/pollice/stat_mult/disp10.pdf Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function, Journal of American Statistical Association, 58, 236-244.
Realizat de	[Unisalento]

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

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.