https://datascience-project.eu/

Training Fiche Template

Title	Analisi in Cluster
Keywords (meta tag)	Unità statistiche, cluster, intra-cluster, inter-cluster, indice di dissimilarità, distanza di fusione, dendogramma.
Language	Italiano
Objectives / Goals / Learnig outcomes	**Obiettivo di questo modulo è introdurre e spiegare la tecnica dell'Analisi in Cluster.** **Alla fine di questo modulo sarai in grado di:** - **Conoscere la logica dell'Analisi in Cluster** - **Conoscere i requisiti** - **Condurre un analisi in cluster**

Training course:	
Data Science Literacy	
Data Visualisation and Visual Analytics Module	X
Introduction to Data science for Human & Social Sciences	
Data Science for good	
Data Journalism and Storytelling	

Description	In questo modulo formativo verrà presentata la tecnica di analisi multidimensionale denominata Analisi in Cluster, detta anche analisi automatica dei gruppi. Le cluster analysis sono utilizzate per raggruppare unità statistiche che hanno caratteristiche in comune ed assegnarle a categorie non definite a priori. I gruppi che si formano devono essere il più possibili omogenei all'interno (intra-cluster) ed eterogenei all'esterno (inter-cluster). L'applicazione di questo tipo di analisi è molteplice: informatica, medicina, biologia, marketing. L'ultima parte del modulo sarà dedicata all'applicazione dell'analisi in cluster con il software R.
Contents arranged in 3 levels	**1. INTRODUZIONE** Le cluster analysis sono utilizzate per raggruppare unità statistiche che hanno caratteristiche in comune ed assegnarle a categorie non definite

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.

a priori. I gruppi che si formano devono essere il più possibili omogenei all'interno (intra-cluster) ed eterogenei all'esterno (inter-cluster).

Le cluster analysis sono delle procedure che si compongono essenzialmente di quattro fasi:

- Scelta delle variabili
- Rilevazione dei dati
- Elaborazione dei dati
- Verifica e utilizzo dei risultati

2. REQUISITI DELL'ANALISI IN CLUSTER

Nell'analisi in cluster possono essere utilizzate diversi tipi di variabili:

- Variabili descrittive (esempio: demografiche, socio-economiche, geografiche)
- Variabili comportamentali (ossia quelle variabili che rispondono alle domande: cosa, quando, dove, come e perchè)

Quindi parliamo di variabili sia di tipo qualitativo che quantitativo.

Il campione a disposizione per l'analisi in cluster deve essere sufficientemente numeroso, identificabile, abbastanza stabile, facilmente raggiungibile e sufficientemente redditizio.

3. Come condurre Analisi in Cluster

3.1 Matrice di Dissimilarità (o matrice delle Distanze), D

Partiamo dalla nostra matrice dei dati **X**, con dimensioni nxp e la trasformiamo in una **matrice di dissimilarità D**, con dimensioni nxn. Quest'ultima è utile per sapere quante unità statistiche sono diverse tra loro e quindi utile per scegliere quali variabili devono essere considerate nell'analisi.

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

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.

Come possiamo vedere la matrice **D** è una matrice simmetrica che lungo la diagonale maggiore ha tutti 0, in quanto la distanza di un punto con se stesso è nulla.

Per calcolare le distanze tra i punti si utilizza l'indice $d_{i,j}$, ossia la misura del grado di similarità tra i e j.

Ci sono diversi indici per poter calcolare tali distanze, a seconda del tipo di variabile che si sta utilizzando.

3.2 Distanze

- Quando si utilizzano **variabili quantitative** si fa riferimento al **grado di dissimilarità**, ci sono diversi modi per poterlo calcolare:
  Distanza Euclidea:
  Essa si rifà al teorea di Pitagora, risulta essere sensibile ai valori anomali. Si calcola:

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

  Distanza di Manhattan:
  Detta anche City Block, risulta essere più robusta della distanza Euclidea e dunque quando possibile si preferisce utilizzare questa. Si calcola:

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

Nel calcolo delle distanze si tiene sempre conto delle unità di misura delle variabili, si può eliminare l'effetto della misura attraverso la standardizzazione della matrice **X** nella matrice **Z**, che sarà data da:

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

Una volta standardizzata la matrice, ovviamente, la utilizzeremo per calcolare l'indice di dissimilarità. La distanza di Manhattan sarà:

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.

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

Dove $\frac{1}{S_k}$ è la ponderazione.

NB: La standardizzazione si esegue se vogliamo date a tutte le variabili lo stesso peso; se invece si ritiene opportuno che una variabile debba avere un peso maggiore alle altre allora non si procederà con la standardizzazione.

- Quando si utilizzano **variabili di tipo Binario,** cioè variabili che presentano due sole modalità (quando parliamo di modalità significa che la variabili a nostra disposizione sono **variabili qualitative**). Alle modalità delle variabili Binarie viene assegnato lo stato 0 e 1. Con questo tipo di variabili si **calcola il grado di similatità,** ossia la somiglianza tra i e j.
  Le variabili binarie si distinguono in:
  Variabili Binarie Simmetriche, **BS**: qui i due stati (0 e 1) hanno la stessa importanza.
  Variabili Binarie Asimmetriche, BA: qui, invece, si da più importanza allo stato 1 rispetto allo stato 0.

  Indice di Zubin:
  Viene utilizzato per le varaibili **binarie simmetriche,** si calcola sommando le frequenze di concordanza e le frequenze di discordanza, poi si divide per il totale.

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

  Indice di Jaccard:
  Viene utilizzato per le variabili **binarie asimmetriche**, si calcola dividendo la frequenza di concordanza per la differenza tra il totale e la frequenza di discordanza.

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

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 Tipi di Cluster

Ci sono diversi tipi di cluster a seconda dell'approccio che si vuole utilizzare nella creazione dei gruppi.

Gli algoritmi gerarchici realizzano fusioni o divisioni successive dei dati, una volta che un oggetto è entrato a far parte di un cluster la sua assegnazione è irrevocabile.

- **Cluster Agglomerativi o aggregativi (bottom-up)**:
  Obiettivo è quello di raggruppare i molti cluster ed ottenere un unico claster che contenga tutti quelli presenti dall'inzio.

- **Cluster Divisi o scissori (top-down):**
  In questo caso si parte da un unico cluster e l'obiettivo finale è quello di dividerlo in tanti cluster.

3.4) Tipi di Legami tra le Unità Statistiche

I cluster si possono formare attraverso diversi tipi di legami:

- **Legame singolo** o semplice (simple linkage)
- **Legame completo** (complete linkage)
- **Legame medio** o del centroide (average linkage)

Il **legame singolo** utilizza la tecnica "del confinante più vicino", il grado di vicinanza tra due gruppi viene stabilito prendendo in considerazione la minima distanza minima tra i punti. In altre parole, si prendono in considerazione le unità che sono più vicine tra loro. Questo legame però, nonostante sia il più veloce da realizzare a livello computazione, crea gruppi troppo omogenei tra loro.

Il **legame completo** utilizza, invece, la tecnica del "confinante più lontano", considera le similarità/distanze fra i gruppi più lontani (quindi quelli meno simili tra loro). In pratica prende il considerazione la minima distanza massima tra i punti. Questo legame, nonostante sia il più lento da un punto di vista computazione, crea gruppi molto eterogenei all'esterno ed omogenei all'interno.

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.

Il **legame medio** nella creazione dei cluster utilizza la minima distanza media, in pratica prima si calcola la distanza media tra tutte osservazioni e poi si prende in considerazione quella più piccola. Anche questo legame è lento sotto al profilo computazionale ma è robusto, risulta essere meno sensibile a valori anomali.

Il **legame di Ward** può essere utilizzato con i dati quantitativi. Questa tecnica minimizza la varianza all'interno dei gruppi omogeneizzandoli, in pratica questo metodo massimizza l'omogeneità interna (o minimizza l'eterogeneità interna) e massimizza l'eterogeneità esterna.

3.5 Dendogramma e Distanza di Fusione

Una volta che è stato scelto il legame che meglio rappresenta i dati in nostro possesso si otterrà il **dendogramma,** cioè possiamo visualizzare attraverso un **grafico ad albero** come sono state distribuite le unità statistiche. Ad ogni passo la distanza tra i cluster tende ad aumentare e quindi è necessario scegliere una **regola di stop** che ci permettere di scegliere il numero di gruppi che vogliamo ottenere. Si può utilizzare la tecnica del taglio dell'albero attraverso il grafico delle **distanze di fusione** (o altezze), che indica il punto in cui si creano i cluster. Graficamente si osserva il punto in cui ci registra una maggiore impennata. Questa parte sarà successivamente ripresa nella parte del modulo dedicato al software R.

4. Esempio con il software R

L'analisi in Cluster ha come obiettivo quello di individuare la migliore ripartizione possibile, in termini di numero e composizione, di un insieme di elementi in gruppi in modo che questi risultino: il più possibile omogenei al loro interno e il più possibile differenti gli uni dagli altri. Tali costruzioni possono essere effettuate sia in funzione della scelta delle strategie di raggruppamento, che in relazione al criterio scelto per la misura della somiglianza/dissimilarità.

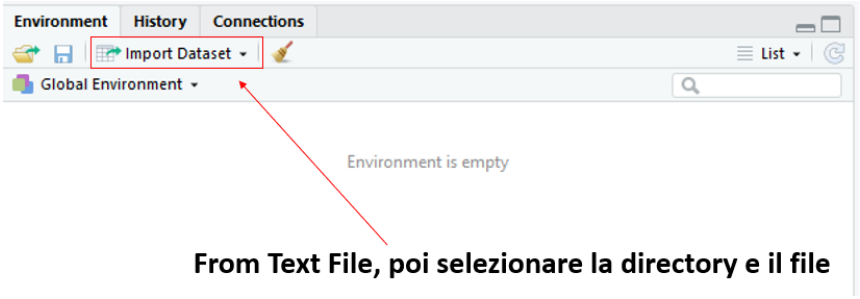
Dataset:

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

Importiamo il dataset:



From Text File, poi selezionare la directory e il file

Nel campo ***row names*** selezionare la dicitura: "***use first column***" in modo da avere sui grafici le etichette sia degli individui sia delle variabili.

Nel campo ***decimal*** selezioniamo la dicitura "***comma***".

Con il comando:

X<-as.matrix(nome_del_dataset)

Attribuiamo ad **X**, come oggetto, il dataset utilizzato nell'analisi.

Standardizziamo la matrice **X**:

Z<-scale(X)

Successivamente calcoliamo la distanza tra gli elementi, possiamo utilizzare o la distanza euclidea o la distanza di Manhattan. Rispettivamente i comandi sono:

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: il comando **round** ci permette di arrotondare alla cifra significativa che preferiamo, in questo caso alla seconda.

Successivamente si passa alla scelta del legame tra gli elementi.

Partiamo con **il legame singolo**:

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

Possiamo visualizzare una **sintesi dei risultati** del legame singolo con il comando:

summary(hc_s)

Possiamo visualizzare il **dendogramma** con la funzione plot:

plot(hc_s)

Per decidere dove tagliare l'albero si utilizza il comando **cutree,** la scelta di quanti gruppi ottenere visualizzando il punto di fusione attraverso lo **scree-plot** del legame della distanza di fusione. I comandi da eseguire sono:

n<-nrow(X)

n_clus<-seq(n-1,1)

hc_s$merge

hc_s$height

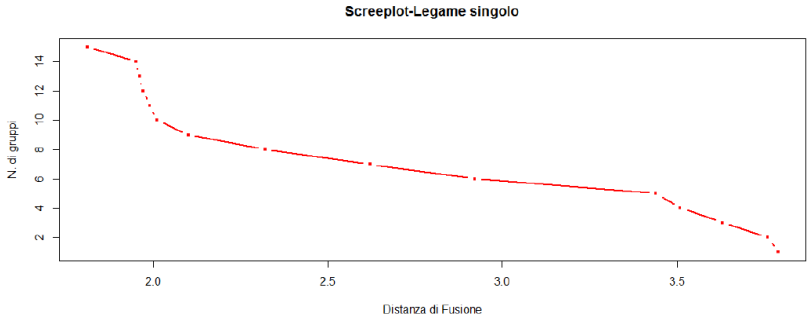
d_fus_s<-hc_s$height

plot(d_fus_s,n_clus,"b", main="Screeplot Legame singolo", xlab="Distanza di Fusione", ylab="N. di gruppi",cex=0.6, col="red",lwd=2.5)

Graficamente:

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.



Ora si vogliono vedere i punti di fusione (hc_s$merge) e le altezze (hc_s$height), per poterle visualizzare insieme si fa ricorso a **cbind**. Il comando $merge riporta, per ogni passo, dell'algoritmo di raggruppamento, la coppia di elementi accorpati a seconda del legame scelto. I valori preceduti da "-" indicano il singolo elemento, mentre i valori positivi rappresentano i cluster formati nei passi precedenti. Così, al primo passo, si avrà la formazione del primo cluster composto dalla coppia (13, 16), corrispondente ai modelli Finlandia e Svezia, mentre il terzo cluster (passo 10) sarà formato dagli elementi del cluster 2 (Grecia, Italia) più l'elemento 1 (Francia). Il campo $height riporta la distanza considerata per la fusione tra elementi/gruppi.

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
```

Per il taglio dell'albero utilizziamo il comando cutree, a k mettiamo il punto in cui la distanza di fusione prende un andamento orizzontale:

groups <- cutree(hc_s, k=4)

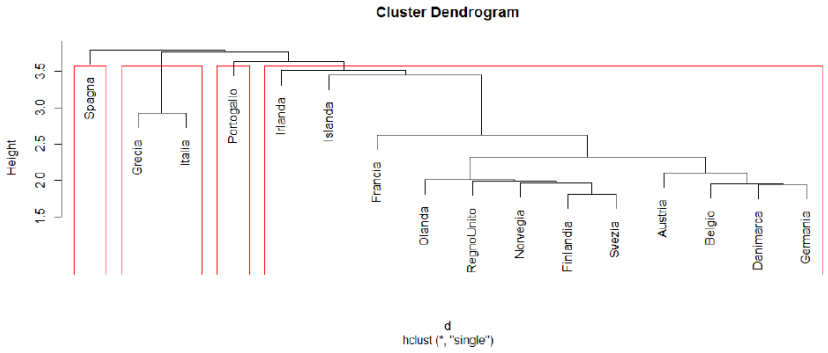
plot(hc_s)

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.

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

Il dendogramma sarà:



Possiamo dire che questo tipo di legame non va bene, perché ci sono cluster che contengono singolo elementi e un cluster che invece è troppo omogeneo al suo interno.

Si procede con gli altri legami allo stesso modo.

Legame 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

Screeplot delle distanze di fusione per il legame Completo:

plot(d_fus_c,n_clus,"b", main="Screeplot Legame completo", xlab="Distanza di Fusione", ylab="N. di gruppi",cex=0.6, col="red",lwd=2.5)

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.

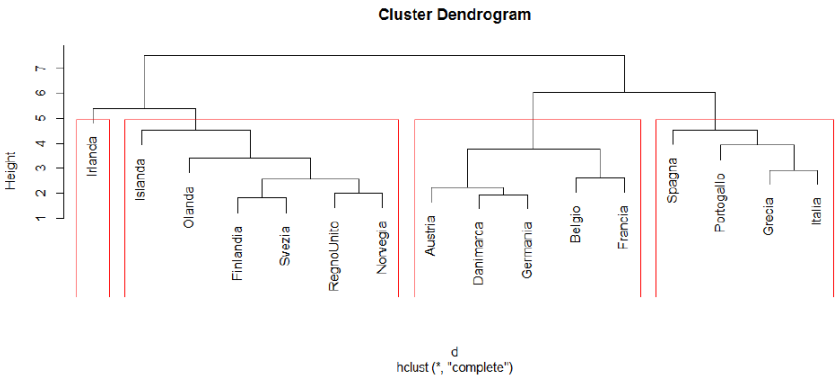
cbind(hc_c$merge,hc_c$height)

Taglio dell'albero per il legame Completo, a k attribuiremo la cifra in base allo screeplot delle distanze di fusione:

groups <- cutree(hc_c, k=4)

plot(hc_c)

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



Legame Medio

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

Screeplot delle distanze di fusione per il legame Medio:

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_a,n_clus,"b", main="Screeplot Legame medio", xlab="Distanza di Fusione", ylab="N. di gruppi",cex=0.6, col="red",lwd=2.5)

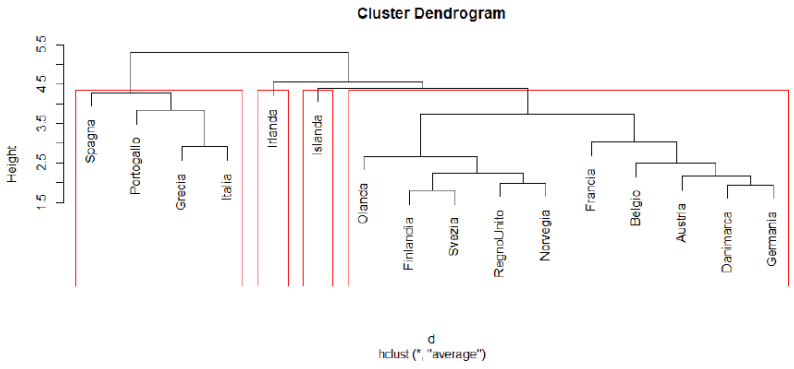
cbind(hc_a$merge,hc_a$height)

Taglio dell'albero per il legame Medio, a k attribuiremo la cifra in base allo screeplot delle distanze di fusione:

groups <- cutree(hc_a, k=4)

plot(hc_a)

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



Self-assessment (multiple choice queries and answers)

1. La matrice delle distanze:

   A) Ha sulla diagonale maggiore tutti 0
   B) Ha sulla diagonale maggiore tutti 1
   C) Ha sulla diagonale maggiore le distanze tra i e j

2. Quale tra queste distanze è più robusta, o insensibile ai valori estremi?

   A) Indice di 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) Distanza Euclidea 3. La standardizzazione consente di: A) Eliminare le frequenze più elevate B) Eliminare l'effetto dell'unità di misura C) Dare peso diverso alle variabili
Resources (videos, reference link)	
Related material	
Related PPT	
Bibliography	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.
Provided by	[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.