Business Analytics
Metodi statistici per il Decision Making
Modelli esplorativi

Questo libro è stato aggiornato per il software KNIME® Analytics Platform (Versione 4.5.2 e superiori), R (Versione 4.2.0 e superiori).

Indice

- **Regole di associazione**
- **Metodi di segmentazione non supervisionata**
- **Riduzione delle variabili**

Indice

- **Regole di associazione**
  - Market Basket Analysis
  - Misure di qualità e affidabilità delle regole
- **Metodi di segmentazione non supervisionata**
  - Cluster Analysis
    - Gerarchica
    - Non gerarchica
- **Riduzione delle variabili**
  - Analisi delle componenti principali (Principal Component Analysis, PCA)
  - Analisi delle corrispondenze (Correspondence Analysis, CA)

Market Basket Analysis

La Market Basket Analysis (MBA) è un tipo di analisi usata dai professionisti del marketing per conoscere il comportamento di acquisto dei clienti.

Viene usata per scopi di **cross/up selling** influendo sulle promozioni di vendita, i programmi di fidelizzazione, la disposizione dei prodotti negli scaffali e i piani di sconto.

Si basa sulla tecnica delle **regole di associazione**[1] per scoprire **affinità tra prodotti** singoli o gruppi.

Può essere applicata a qualsiasi settore che vende prodotti diversificati come supermercati, operatori telefonici, istituti di credito e i nuovi canali di vendita, in particolare Internet (***Amazon, Netflix***, …)

[1] R. Agrawal; T. Imielinski; A. Swami: *Mining Association Rules* Between Sets of Items in Large Databases", SIGMOD Conference 1993: 207-216 https://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf

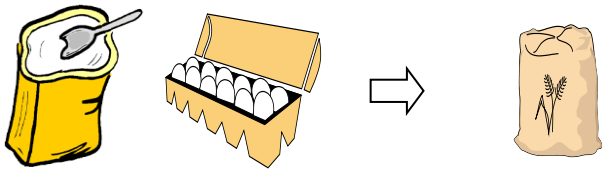
Market Basket Analysis

Dato un insieme di elementi ("*item*"), e un insieme di transazioni che comprendono quegli item, una regola associativa del tipo

$$\boldsymbol{X \Rightarrow Y} \qquad \textit{(X implica Y)}$$

si può enunciare come: ***Chi compra X, compra anche Y***

Ad esempio: se un cliente compra zucchero e uova, allora il 86% delle volte compra anche farina.

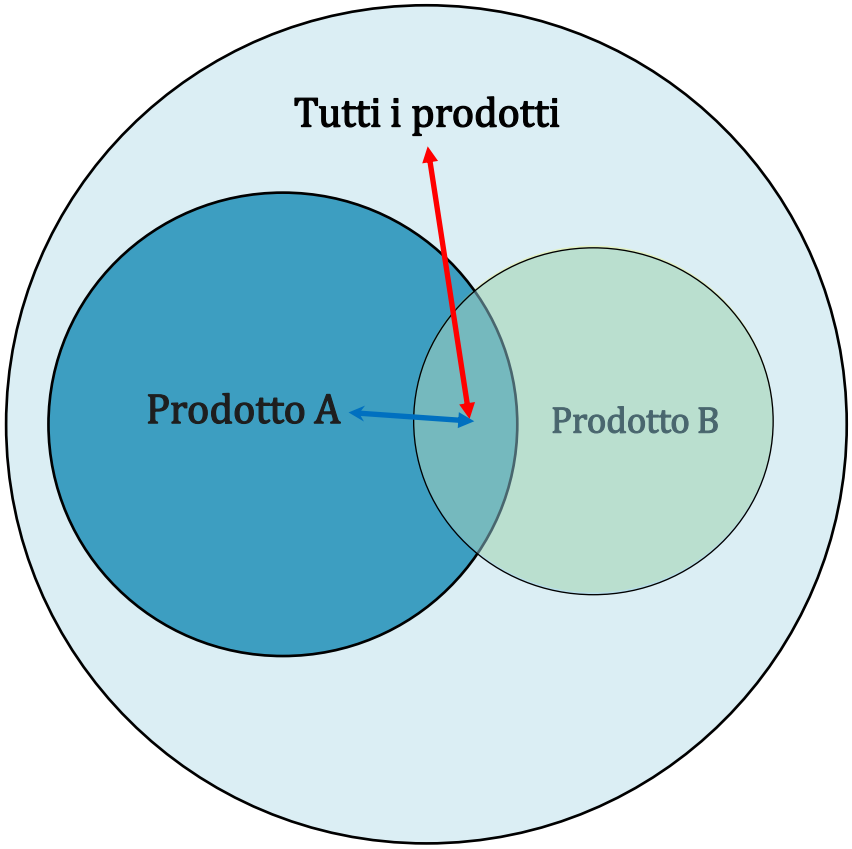


Questa regola può quindi orientare un rivenditore al posizionamento ottimale dei prodotti da disporre sugli scaffali o **suggerire a un cliente ulteriori prodotti da acquistare** in base ai suoi comportamenti d'acquisto abituali (p. e. promuovere un'offerta sulle uova a chi acquista farina).

- Misure di qualità e affidabilità

$$\textit{Supporto}\ (A \rightarrow B) = \frac{\textit{Prodotto } A \cap B}{\textit{Tutti i prodotti}}$$

$$\textit{Confidenza}\ (A \rightarrow B) = \frac{\textit{Prodotto } A \cap B}{\textit{Prodotto } A}$$



Misure di qualità e affidabilità

Supporto

Misura la **frequenza** della regola. E' la percentuale delle transazioni che contengono sia X sia Y

$$Supporto\ (X \Rightarrow Y) = \frac{\#\ transazioni\ che\ contengono\ X\ e\ Y}{Totale\ transazioni}$$

Confidenza

Misura l'**affidabilità** (la "*forza*") della regola. E' la percentuale di transazioni che contengono Y se queste contengono X.

$$Confidenza\ (X \Rightarrow Y) = \frac{\#\ transazioni\ che\ contengono\ X\ e\ Y}{\#\ transazioni\ che\ contengono\ X}$$

Confidenza attesa

È' la misura della confidenza nell'assunzione che gli item siano tra loro indipendenti (non c'è correlazione).

$$Confidenza\ attesa\ (X \Rightarrow Y) = \ Supporto\ (Y)\ = \frac{\#\ transazioni\ che\ contengono\ Y}{Totale\ transazioni}$$

Misure di qualità e affidabilità

Lift (Interest)

È dato dalla confidenza diviso la confidenza attesa. Può essere interpretato come una misura generale di associazione. **Se vicino o uguale a 1 c'è indipendenza, se ≠1 c'è correlazione**.

$$Lift = \frac{Confidenza\,(X \Rightarrow Y)}{Confidenza\,attesa\,(X \Rightarrow Y)} = \frac{Supporto\,(X \Rightarrow Y)}{Supporto\,(X)\;Supporto\,(Y)}$$

Leverage

È una variante del Lift, ma più 'robusto'. **Se vicino o uguale a 0 c'è indipendenza, se diverso da 0 c'è correlazione**.

$$Leverage = Supporto(X \Rightarrow Y) - Supporto(X)\;Supporto(Y)$$

lift > 1 *o* ***leverage > 0*** *significa che acquistando l'antecedente (X) aumenta le probabilità che in una transazione si acquisti anche il conseguente (Y).*

lift < 1 *o* ***leverage < 0*** *significa che acquistando l'antecedente riduce le probabilità che nella stessa transazione si acquisti anche il conseguente (in pratica, gli item sono visti come alternativi).*

lift =1 *o* ***leverage = 0*** *significa che la probabilità di acquisto del conseguente non è condizionata dall'acquisto dell'antecedente e quindi la regola non ha alcuna utilità.*

- **Esempi di misure di qualità e affidabilità**

Regola	Supporto	Confidenza	Conf. Attesa	LIFT
A ⇒ D	**2/5 (40%)**	**2/3 (66,7%)**	**3/5 (60%)**	**0,66/0,6=1,11**
C ⇒ A	2/5 (40%)	2/4 (50%)	3/5 (60%)	0,50/0.6=0,83
A ⇒ C	2/5 (40%)	2/3 (66,7%)	4/5 (80%)	0,66/0,8=0,83
B&C ⇒ D	1/5 (20%)	1/3 (33,3%)	3/5 (60%)	0,33/0,6=0,56

La regola **A ⇒ D**, ad esempio, da queste misure, può essere così tradotta in linguaggio naturale:

"Ogni volta che qualcuno acquista A è probabile che acquisti anche D con una probabilità del 66,7%, che è 1,11 volte più verosimile della sola probabilità di D che è del 60%. Questa combinazione si trova nel 40% di tutti i carrelli."

Esempi di misure di qualità e affidabilità

In base alla tabella a lato si calcolano le misure della regola
Libretto Risparmio (LR) ⇒ Conto Corrente (CC)

Supporto (LR ⇒ CC) = 2000/5000 = 40%
Confidenza (LR ⇒ CC) = 2000/3000 = 66,7%
Confidenza Attesa (LR ⇒ CC) = 3750/5000 = 75%

Lift (LR ⇒ CC) = 0,667/0,75 = **0,889 (<1)**
Leverage (LR ⇒ CC) = 0,4-(0,6*0,75) = **-0,05 (<0)**

		Conto Corrente: No	Conto Corrente: Sì	
Libretto Risparmio	No	250	1750	2000
	Sì	1000	2000	3000
		1250	3750	5000

In base a queste misure, questa regola potrebbe essere considerata "forte" (40% di supporto e 66,7% di confidenza), se non che i due item sono sostitutivi avendo Lift <1 e Leverage <0 (in pratica, chi ha un libretto tende a non avere un conto).

Infatti, considerando la regola **LR ⇒ non CC**, si hanno bassi valori di supporto e confidenza (20% e 33,3%) ma con lift >1 e leverage >0:

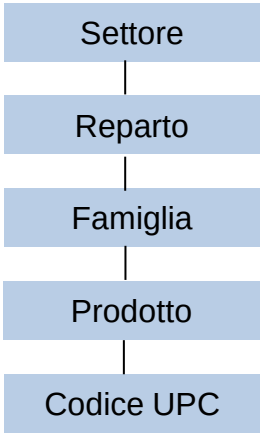
Lift (LR ⇒ *non* CC) = 0,333/0,25 = **1,333 (>1)**
Leverage (LR ⇒ non CC) = 0,2–(0,6*0,25) = **0,05 (>0)**

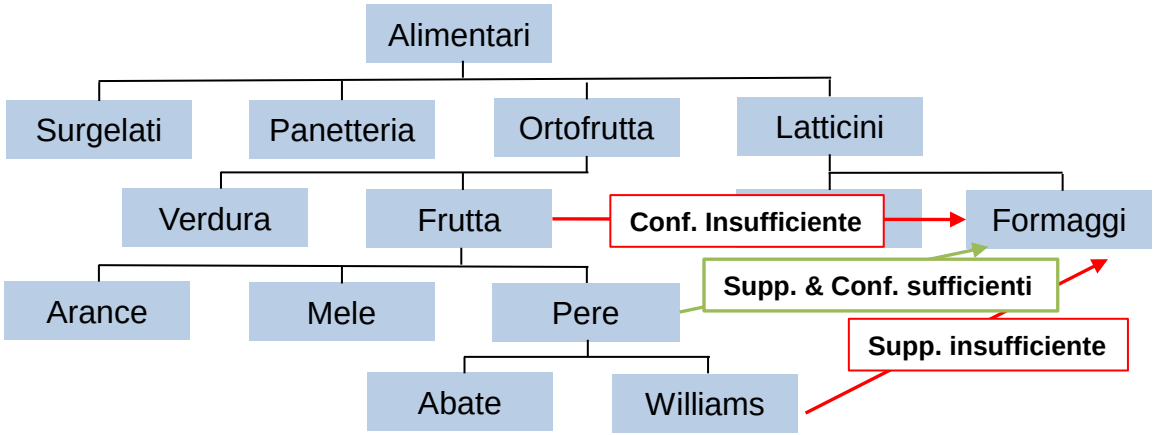
Limitazioni

Per evitare la complessità dovuta al numero elevato di combinazioni, si fissano delle soglie per il supporto e la confidenza (rispettivamente al 5% e 10%) con il rischio però di perdere **regole "interessanti"** aventi un **basso supporto e un'alta confidenza**.

Tassonomia

Le regole potrebbero però non essere interessanti a livello di massimo dettaglio (p.e. il singolo codice a barre o UPC), in questi casi è consigliabile combinare gli item a livelli diversi di gerarchia.





Utilizzo del software

Si supponga di avere un dataset strutturato come illustrato a lato (colonne ***Cliente*** e ***Prodotto***), dove ogni riga rappresenta il singolo prodotto acquistato da un certo cliente.

Cliente	Prodotto
1	Zucchero
1	Uova
1	Latte
1	Farina
1	Lievito
2	Zucchero
2	Uova
2	Olio
…	…
9	Olio
9	Latte
9	Farina
9	Lievito
10	Zucchero
10	Uova
10	Latte
10	Farina
(lista parziale)	

Si vogliono elaborare le regole di associazioni utilizzando due software open-source[1]:

- **Analytics Platform**
- **Project**

Al fine di ottenere regole credibili, i risultati devono superare per il supporto la soglia del 5% e per la confidenza la soglia del 10%.

[1] Per installazione vedi Appendice.

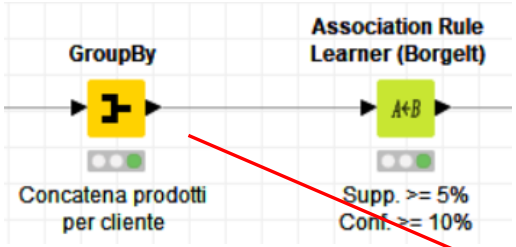
Utilizzo del software Knime - Modulo2_Esempio1

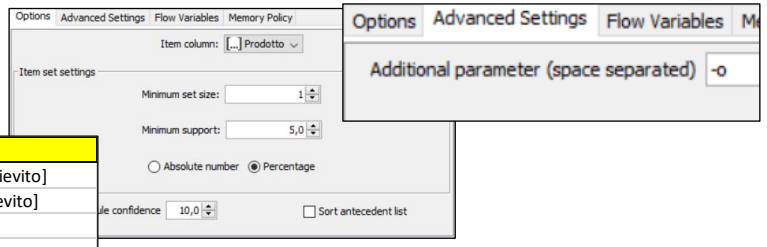
Il workflow di Knime

Il nodo che in Knime utilizza le regole di associazione è l'**Association Rule Learner (Borgelt)** che però deve avere in ingresso i dati strutturati come sotto illustrato dove la colonna Prodotto deve essere di tipo **Lista** (una collezione di valori).

Nella scheda Advanced Settings inserire nel campo *Additional parameter* il valore **-o** ("original definition of the support of a rule")

Per ottenere questo bisogna usare il nodo **GroupBy** selezionando *Cliente* come colonna di raggruppamento nella scheda Groups e "***List***" come metodo di aggregazione per la colonna *Prodotto* nella scheda Manual Aggregation.



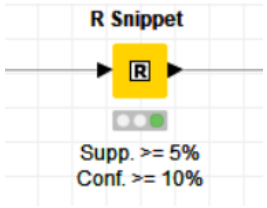


Cliente	Prodotto
1	[Zucchero,Uova,Latte,Farina,Lievito]
2	[Zucchero,Uova,Olio,Farina,Lievito]
3	[Uova,Olio,Latte,Farina]
4	[Zucchero,Uova,Latte,Lievito]
5	[Zucchero,Uova,Olio,Latte,Farina,Lievito]
6	[Zucchero,Uova,Farina]
7	[Zucchero,Uova,Olio,Farina,Lievito]
8	[Uova,Olio]
9	[Uova,Olio,Latte,Farina,Lievito]
10	[Zucchero,Uova,Latte,Farina]

Utilizzo del software R all'interno di Knime - Modulo2_Esempio1

Lo script di R

Codifica del linguaggio R nel nodo **R Snippet** dove viene usata la funzione **apriori** del package **arules**.



R Script

```
1  folder <- "C:\\Program files\\R\\Packages"
2  library(arules, lib.loc=folder)
3
4  lst <- split(knime.in[,"Prodotto"] ,knime.in[,"Cliente"])
5  trx <- as(lst, "transactions")
6
7  rules <- apriori(trx,
8           parameter  = list(supp = 0.05,
9                             conf = 0.1,
10                            minlen=1,
11                            maxlen=4,
12                            maxtime=0,
13                            target="rules"))
14
15 im <- interestMeasure(rules,
16      c("leverage","chiSquared") , significance=TRUE,  transaction=trx)
17
18 knime.out <- cbind(as(rules, "data.frame"), im)
19
```

Suddivide colonne per gruppi creando oggetto di tipo ***list***

Crea oggetto di tipo ***transaction***

Statistiche **leverage e chi-quadro**

Utilizzo del software R e Knime - Modulo2_Esempio1

Estrazione di alcune regole:

Regola	Supporto	Confidenza	Lift	Leverage
{Uova} => {Farina}	0,8	0,8	1,0	0
{Farina} => {Uova}	0,8	1	1,0	0
{Olio,Farina} => {Lievito}	0,4	0,8	1,3	0,1
{Zucchero,Uova} => {Farina}	**0,6**	**0,86**	**1,07**	**0,04**
{Zucchero,Uova,Latte} => {Lievito}	0,3	0,75	1,25	0,06

Una regola credibile ha un valore alto di confidenza e di supporto e un valore di Lift maggiore di 1 ***[4]***. Le regole che hanno un livello alto di confidenza, ma che hanno scarso supporto, devono essere interpretate con cautela **[3,5]**, di nessuna utilità se hanno Lift = 1 o Leverage = 0 ***[1,2].***

Utilizzo del software R e Knime - Modulo2_Esempio1

Esempio di Tassonomia

Livello 1	Livello 2
Aceto	CONDIMENTI
Olio	CONDIMENTI
Uova	LATTE E DERIVATI, UOVA
Latte	LATTE E DERIVATI, UOVA
Farina	PASTA, CEREALI E FARINE
Zucchero	PRODOTTI DOLCIARI
Lievito	PRODOTTI DOLCIARI

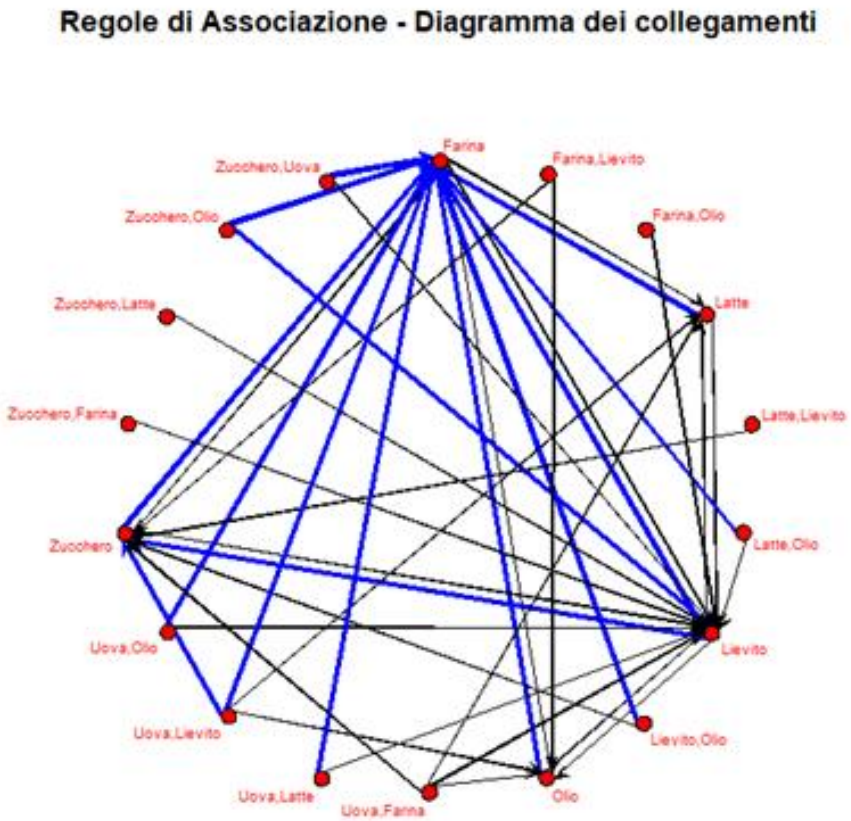
Regola	Supporto	Confidenza	Lift	Leverage
{PRODOTTI DOLCIARI} => {Farina}	0,7	0,875	1,094	0.06
{Zucchero} => {CONDIMENTI}	0,3	0,439	0,714	-0,12
{CONDIMENTI} => {PRODOTTI DOLCIARI}	0,4	0,667	0,833	-0,08

Utilizzo del software R all'interno di Knime - Modulo2_Esempio1

La visualizzazione grafica delle regole permette di cogliere velocemente quelle più interessanti.

Di fianco viene rappresentata la misura della **Confidenza con Lift > 1** (in questo caso se è >80% il collegamento è più marcato e colorato).

Questo grafico è ottenuto attraverso uno script di R nel nodo **R View** dove viene usata la funzione **network** del package **network**.



Utilizzo del software Knime - Modulo2_Esercizio1

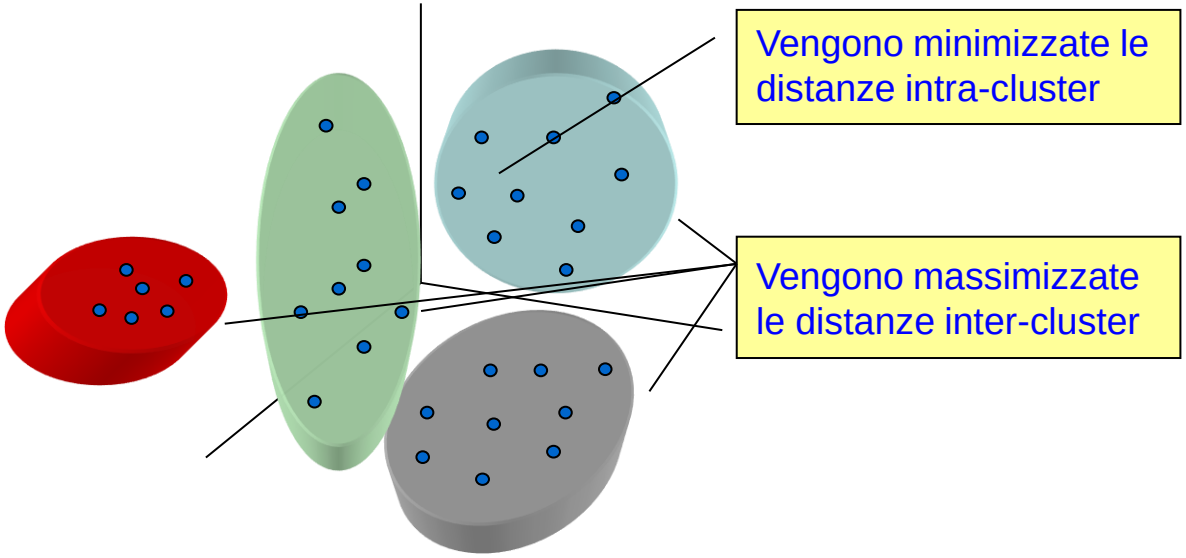
- Importare con il nodo **Excel Reader** dalla cartella ***Dati*** nella chiavetta Usb la tabella ***Banca.xlsx***;
- raggruppare con il nodo **GroupBy** per la colonna *Codice_Cliente* e cambiando la tipologia della colonna *Prodotto* da Stringa a Lista;
- utilizzare il nodo **Association Rule Learner (Borgelt)** impostando il supporto minimo al 10% e la confidenza minima al 25% e, nelle impostazioni avanzate, il parametro ***-o***;
- selezionare con il nodo **Column Filter** le colonne *Consequent, Antecedent, RuleLift, RelativeItemSetSupport%, RuleConfidence%*;
- mettere come prima colonna Antecedent con il nodo **Column Resorter**;
- selezionare con il nodo **Rule-based Row Filter** solo le righe che hanno *RuleLift* > 1 oppure *RuleLift* < 1;
- ordinare con il nodo **Sorter** la tabella per i valori decrescenti della colonna *RuleLift;*
- interpretare in risultati.

Indice

- Regole di associazione
  - Market Basket Analysis
  - Misure di qualità e affidabilità delle regole
- **Metodi di segmentazione non supervisionata**
  - Cluster Analysis
    - Gerarchica
    - Non gerarchica
- Riduzione delle variabili
  - Analisi delle componenti principali (Principal Component Analysis, PCA)
  - Analisi delle corrispondenze (Correspondence Analysis, CA)

La Segmentazione

La segmentazione serve all'individuazione di gruppi (detti ***segmenti*** o ***cluster***) di soggetti **simili al loro interno** per determinate caratteristiche[1], rispetto ai soggetti presenti in gruppi diversi.



[1] Si possono raggruppare, ad esempio, in ambito bancario, i clienti per comportamento inteso come l'insieme delle attività che il cliente intrattiene con la banca (per cui si parla di "segmentazione comportamentale").

Preparazione (1/2)

Standardizzazione

Molto spesso le variabili utilizzate per l'analisi dei gruppi (o **cluster analysis**) non hanno la stessa unità di misura. Se una o più variabili hanno una scala molto più grande delle altre, la metrica utilizzata per misurare le distanze ne sarà influenzata.

Un altro vantaggio della standardizzazione è rendere più simili le metriche utilizzate.

Una delle standardizzazioni più utilizzate è quella del metodo **Z-score** che "standardizza" una variabile avente media uguale a 0 e varianza pari a 1 ottenuta sottraendo alla variabile la sua media e dividendo il tutto per la deviazione standard.

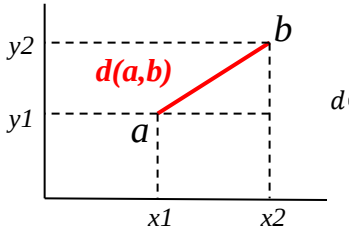
$$z = \frac{x - \mu}{\sigma}$$

Preparazione (2/2)

- **Metrica**
  Per decidere quali cluster combinare insieme è necessaria una **misura di distanza** tra i diversi insiemi di osservazioni. Tra le misure di distanza più utilizzate rientrano:

  - La distanza euclidea $dE(p,q) = \sqrt{(q_1 - p_1)^2 + \cdots + (q_n - p_n)^2}$

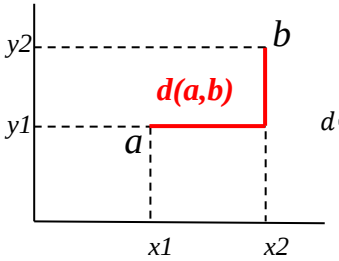
    In coordinate cartesiane dove *a=(x1,y1)* e *b=(x2,y2)* equivale al teorema di Pitagora

    $d(a,b) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

  - La distanza di Manhattan $dM(p,q) = |q_1 - p_1| + \ldots + |q_n - p_n|$

    In coordinate cartesiane è la somma delle differenze (assolute) dei due punti.

    $d(a,b) = |x_2 - x_1| + \ldots + |y_2 - y_1|$

Metodi di segmentazione

Metodi per la costruzione di gruppi

I metodi statistici utilizzati per la cluster analysis si dividono in

- **Gerarchici**

  Partono dagli n soggetti presi singolarmente e li aggregano in gruppi creando una gerarchia ad albero rappresentabile graficamente (dendrogramma).

- **Non gerarchici**

  prevedono la decisione a priori del numero di cluster che si vogliono ottenere e una procedura iterativa che assegna i vari soggetti ai gruppi.

Cluster Analysis Gerarchica

Metodi

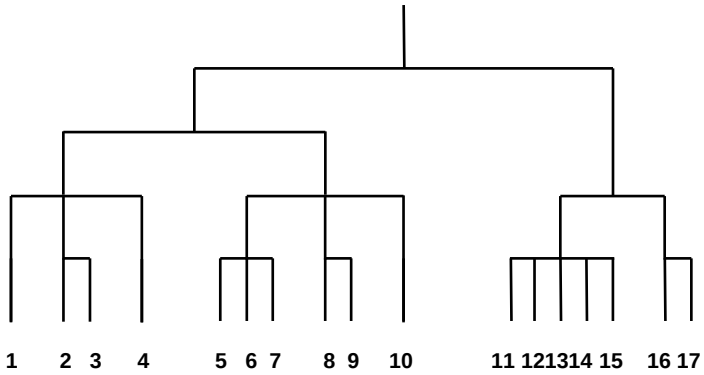
Per misurare la differenza tra due gruppi di osservazioni, sono stati sviluppati diversi metodi. I più comuni sono:

- **Complete linkage** tende a produrre gruppi di uguale diametro, **sensibile però agli outlier**
- **Average linkage** tende a produrre gruppi con la **stessa varianza**
- **Single linkage** tende a produrre gruppi con **allungati e irregolari**
- **Centroid** più robusto rispetto agli outlier, meno performante dei metodi Average e Ward
- **Ward** tende a produrre gruppi con la **stessa numerosità**

Cluster Analysis Gerarchica

Il principio d'azione di un algoritmo gerarchico, di tipo agglomerativo, è rappresentato sotto forma di diagramma ad albero (*dendrogramma*):

1. Ogni caso, all'inizio, è un cluster.
2. Si individua la coppia di cluster, secondo il metodo scelto, più vicini.
3. La coppia viene fusa in un unico cluster.
4. Si ripete dal punto 2 fino a ottenere un unico cluster comprensivo di tutti i casi.



Una volta terminato il processo di aggregazione, bisogna scegliere il livello di "taglio" della gerarchia per determinare il **numero di cluster da tenere**.

Determinazione del numero di cluster – Modulo2_Esempio2

Determinare il numero ottimale di cluster è soggettivo. Si possono però utilizzare dei metodi che aiutano nella scelta

Elbow method (metodo del gomito)

Si rappresenta graficamente la somma dei quadrati **all'interno dei cluster** (*WSS*); il punto in cui la *curva scende in modo più consistente* è un indicatore del numero appropriato di cluster.

pseudo-F $\text{pseudo}-\text{F} = \frac{BSS/(k-1)}{WSS/(n-k)}$

Calcola il rapporto tra la variazione tra i cluster e la variazione all'interno dei cluster, dove BSS è la somma dei quadrati **tra i cluster**, k è il numero dei cluster e n è il numero delle osservazioni.
Il numeratore, che è la varianza tra i cluster, misura quanto i cluster sono separati gli uni dagli altri.
Più il valore di pseudo-F è alto, più i cluster sono distinti.

Silhouette

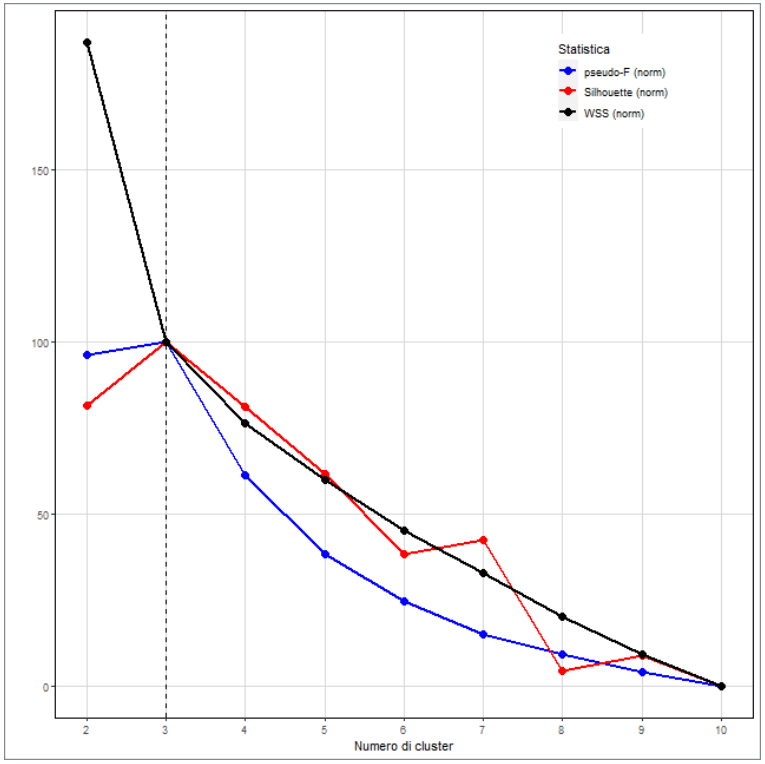
Viene usata per valutare la bontà di assegnazione delle osservazioni ai diversi cluster. I valori che si avvicinano a 1 evidenziano un raggruppamento ben strutturato.

Determinazione del numero di cluster – Modulo2_Esempio2

Si vuole fare una cluster analysis basata sul metodo k-means dai risultati di un'analisi chimica dei vini di una stessa regione d'Italia ottenuti da 3 diverse cultivar. L'analisi ha determinato le quantità di 13 componenti presenti in ciascuna delle 3 tipologie di vini tra le quali il grado alcolico, l'acido malico, il magnesio, ecc.

Si vuole confrontare il numero ottimale di cluster utilizzando i metodi illustrati prima e confrontarli con la classe della cultivar.

Tutti i metodi indicano 3 come numero ottimale.



Utilizzo del software Knime

I nodi che si utilizzano in Knime per la Cluster Analysis sono:

• **Normalizer**	Normalizzazione delle variabili
• **Denormalizer**	Ritorno ai valori originali dopo la normalizzazione
• **k-Means**	Calcolo dei cluster (metodo non gerarchico)
• **Distance Matrix Calculate**	Calcolo delle distanze (metodo gerarchico)
• **Hierarchical Clustering (DistMatrix)**	Calcolo dei cluster (metodo gerarchico)
• **Hierarchical Clustering Assigner**	Assegnazione dei cluster (metodo gerarchico)
• **R Snippet**	Contiene script di R per il calcolo dei cluster
• **Entropy Scorer**	Valutazione della bontà del cluster in termini di entropia

Utilizzo del software Knime - Modulo2_Esempio3

Una Banca vuole identificare gruppi di clienti in base al tipo di prodotti posseduti. Questo permetterà di adeguare l'offerta di prodotti e servizi in base alle loro attese.

La tabella dei dati contiene l'indicazione del possesso o meno di un certo prodotto (1=Sì,0=No).

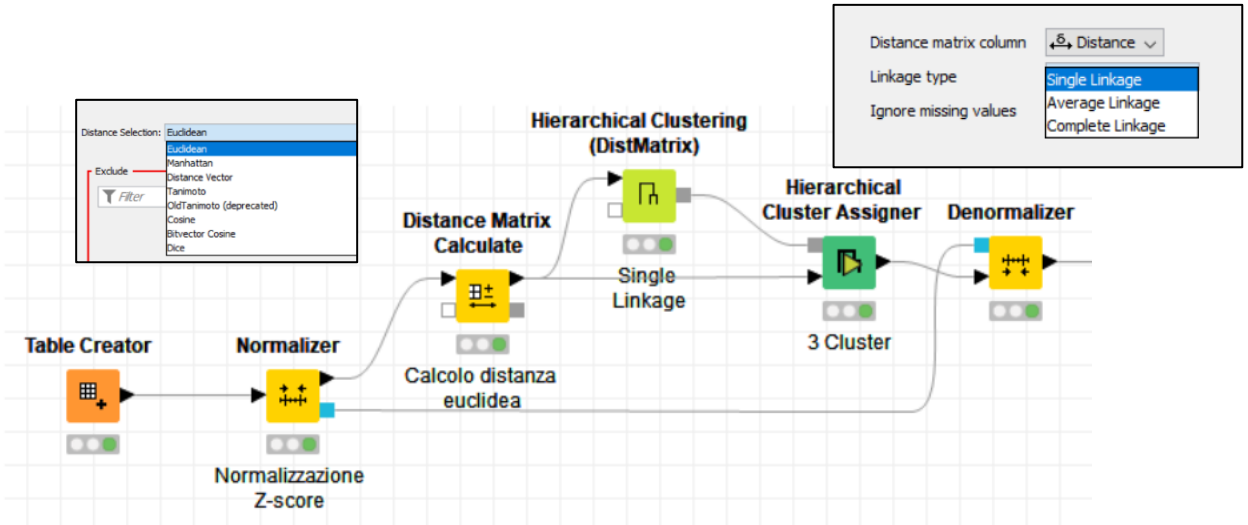
ID Cliente	Conti Correnti	Libretti	Mutui	Prestiti Personali	Altri Prestiti	Obbligazioni	Fondi	Certificati Deposito
1	1	1	0	0	0	2	0	0
2	1	1	0	0	0	0	1	0
3	1	0	1	1	1	1	1	0
4	1	0	1	1	0	0	0	0
5	1	0	0	1	0	0	0	0
6	0	0	1	0	1	0	0	0
7	0	1	0	0	0	0	0	0
8	1	0	0	0	0	1	1	0
9	0	0	0	1	0	0	0	1
10	1	0	0	0	0	1	0	0
11	1	0	0	0	0	1	1	0
12	1	0	1	1	0	0	0	0
13	1	0	1	1	0	0	0	0
14	1	1	0	0	0	0	1	0
15	0	1	0	0	0	0	0	0
16	0	0	1	0	1	0	0	0
17	0	0	0	0	0	1	0	0
18	0	1	0	1	0	0	0	0
19	1	1	0	0	0	1	0	1

(segue)

Utilizzo del software Knime - Modulo2_Esempio3

Il workflow di Knime

1. Si normalizzano i valori delle colonne con il metodo Z-Score
2. Si calcola la distanza tra tutte le coppie di righe della tabella con la funzione euclidea
3. Si utilizza il cluster gerarchico con il metodo "*Single Linkage*" e si scelgono 3 cluster
5. Si riportano i valori delle colonne ai loro valori originari



Utilizzo del software R all'interno di Knime - Modulo2_Esempio3

Lo script di R

Codifica del linguaggio R nel nodo **R Snippet** dove vengono usate le funzioni **dist**, ***hclust*** e **cutree**.

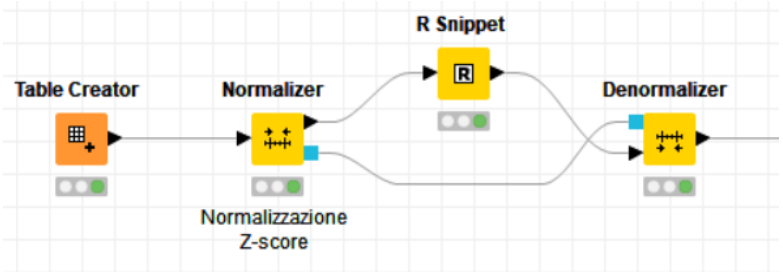
R Script

```
1 df <- knime.in[, -1]
2
3 df <- dist(df, method = "euclidean")
4 hc <- hclust(df,  method = "ward.D2")
5
6 cluster <- cutree(hc, k = 3)
7
8 knime.out <- cbind(cluster,knime.in)
9
```

Calcolo della distanza

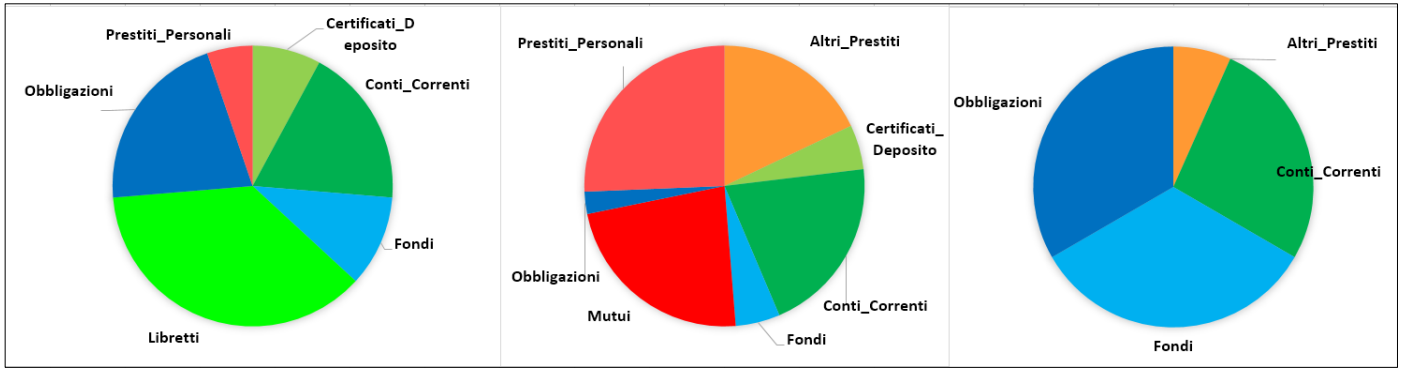
Cluster analysis con il metodo Ward

Indica il numero di cluster desiderato



Utilizzo del software R all'interno di Knime - Modulo2_Esempio3

Dall'analisi svolta sui 3 cluster si sono ottenuti questi risultati. Per una loro interpretazione viene visualizzato un diagramma circolare (*grafico a torta*), dove ogni colore evidenzia il prodotto posseduto dai clienti nel cluster e l'ampiezza dello spicchio è proporzionale alla sua frequenza relativa:



- Il **Cluster A,** "***Risparmiatori Tradizionali***". È caratterizzato da ***libretti, obbligazioni e fondi***.

- Il **Cluster B,** ***"Indebitati"***. È caratterizzato da ***mutui e prestiti***.

- Il **Cluster C,** "***Investitori***". È caratterizzato in gran parte da ***fondi e obbligazioni***.

Metodi di segmentazione

Metodi per la costruzione di gruppi

I metodi statistici utilizzati per l'analisi dei gruppi (*cluster analysis*) si dividono in

- **Gerarchici**

  Partono dagli n soggetti presi singolarmente e li aggregano in gruppi creando una gerarchia ad albero rappresentabile graficamente (dendrogramma).

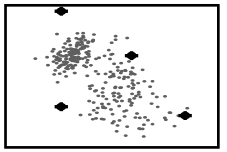
- **Non gerarchici**

  prevedono la decisione a priori del numero di cluster che si vogliono ottenere e una procedura iterativa che assegna i vari soggetti ai gruppi.

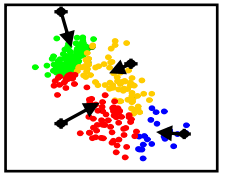
Metodo delle k-medie

Il più diffuso algoritmo non gerarchico è il k-Means method (metodo delle k-medie) basato sulle **distanze euclidee**, del quale di seguito viene illustrata brevemente la procedura:

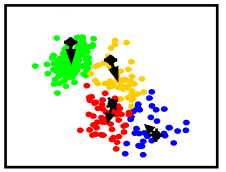
1. Si **stabilisce a priori il numero k dei gruppi** e, per ciascuno, viene individuato un centroide casuale.



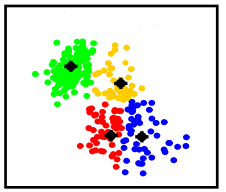
2. Ogni unità viene quindi assegnata al centroide più vicino in modo da minimizzare la varianza totale intra-cluster.



3. Viene ricalcolato il centroide di ogni cluster.



4. Si assegnano nuovamente le unità al centroide più vicino.



5. Le iterazioni proseguono finché l'algoritmo non converge, ossia non ci sono più trasferimenti di unità da un cluster all'altro.

Utilizzo del software Knime - Modulo2_Esempio4

Si vogliono segmentare le zone di una certa area geografica per capire **dove aprire nuove agenzie** bancarie. L'insieme delle variabili utilizzate[1] riguardano la struttura demografica della popolazione, il sistema produttivo e l'occupazione, i servizi e le infrastrutture, gli indicatori di benessere e il capitale umano.

SSLL	t_att_tot	Tasso_AGR	T_IND	T_SERV	Sport. BANC	DEPOSITI/RESIDENT	IMPONIB/RESIDENT	IMPONIB/CONTRIE
BUSTO ARSIZIO1	45.9	0.49	55.6	43.9	62	24.0	23.6	26.4
GALLARATE	47.5	0.58	60.5	39.0	85	21.1	13.8	24.9
LUINO	44.1	2.04	47.9	50.1	14			
SESTO CALENDE	44.7	1.55	56.2	42.3	42			
VARESE1	45.4	1.07	49.6	49.4	104			
MILANO1	45.7	0.65	48.9	50.5	35			
VARESE2	45.6	1.29	58.8	39.9	4			
BELLAGIO1	44.0	5.19	41.4	53.4	3			
CAMPIONE D'ITALIA	43.9	3.97	35.0	61.0	7			
COMO	45.9	1.13	49.8	49.0	207			
MENAGGIO	42.5	4.43	45.9	49.6	18			
PORLEZZA	42.8	3.78	47.0	49.2	7			
MORBEGNO1	44.4	8.39	48.5	43.2	1			
DESIO1	46.6	1.03	60.7	38.3	24			
BORMIO	42.3	2.11	32.9	65.0	14			
CHIAVENNA	42.3	4.71	45.8	49.5	11			
CHIESA IN VALMALEI	41.3	3.38	47.2	49.4	4			
MORBEGNO2	42.3	4.10	48.5	47.4	23			
SONDALO	42.7	3.24	38.9	57.8	7			
SONDRIO	42.7	4.75	32.5	62.7	27			
TIRANO	43.5	7.32	31.4	61.3	7	21.6	15.2	22.2
EDOLO1	44.4	2.51	20.9	76.6	2	25.5	17.2	17.1
BUSTO ARSIZIO2	46.3	0.76	56.2	43.0	103	18.9	13.8	25.3
LECCO1	44.2	1.57	54.9	43.5	2	12.4	5.6	26.1
DESIO2	46.9	0.55	53.6	45.9	172	18.9	13.1	24.4
MILANO2	46.5	0.46	34.8	64.7	1499	32.2	65.0	30.9
SANT'ANGELO LODIG	39.7	3.98	34.3	61.8	3	28.3	12.7	23.8

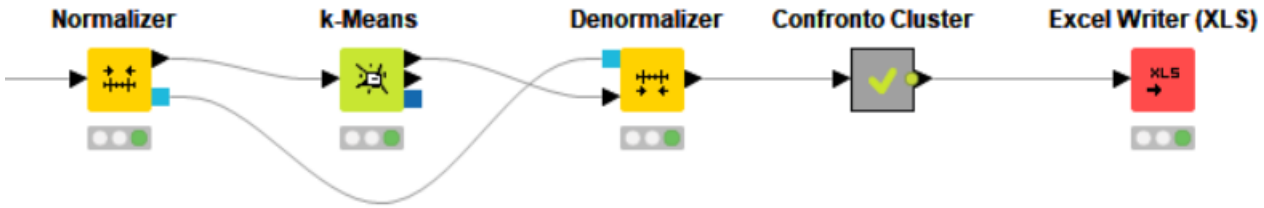
- Popolazione attiva sulla popolazione residente, in % (CP 1991)
- Quota di occupati nell'agricoltura sul totale occupati, in % (CP 1991)
- Quota di occupati nell'industria sul totale occupati, in % (CP 1991)
- Quota di occupati nei servizi sul totale occupati, in % (CP 1991)
- Numero di sportelli bancari (1996)
- Ammontare medio dei depositi bancari per abitante, in mil di lire (1996)
- Imponibile medio per residente, in mil di lire (1994)
- Imponibile medio per contribuente, in mil di lire (1994)

[1] I dati, relativi al censimento del 1991 e al censimento intermedio dell'industria e servizi del 1996, riguardano il **Sistema Locale di Lavoro (SSLL)**, cioè aggregati di comuni contigui, non necessariamente della stessa provincia, della Regione Lombardia.

Utilizzo del software Knime - Modulo2_Esempio4

Il workflow di Knime

1. Si normalizzano i valori delle colonne con il metodo Z-Score
2. Si utilizza il cluster non gerarchico con il metodo "*k-Means*" e si scelgono 5 cluster
3. Si riportano i valori delle colonne ai loro valori originari
4. Si traspongono i risultati per una migliore lettura semantica
5. I risultati vengono riportati in un foglio di lavoro di tipo Excel



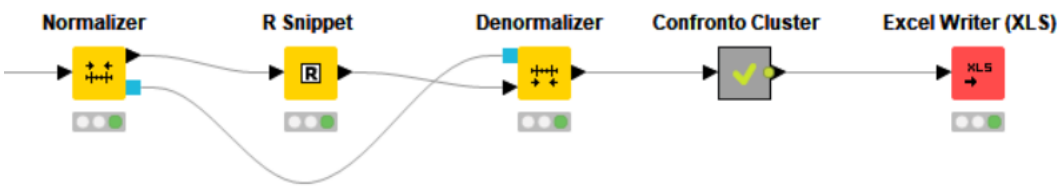
Utilizzo del software R all'interno di Knime - Modulo2_Esempio4

Lo script di R

Codifica del linguaggio R nel nodo **R Snippet** dove viene usata la funzione **kmeans**.

R Script

```
1 knime.out <- knime.in
2
3 kmeans <- kmeans(knime.out[,2:9], 5,  iter.max = 100 )
4
5 knime.out$Cluster <- as.factor(kmeans$cluster)
6
```



Utilizzo del software Knime - Modulo2_Esempio4

Si confronta **la media di ogni variabile riferita a un cluster specifico** con la **media riferita alla totalità dei SSLL** considerati.

Variabile	Statistica	TOTALE	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
	Numero	88	12	20	12	28	16
	%	100	13	21	13	31	22
DEPOSITI/RESIDENTI	Media	21.1	22.4	23.0	25.5	16.9	21.9
IMPONIB/CONTRIB	Media	22.0	26.0	22.4	22.4	21.1	20.3
IMPONIB/RESIDENTI	Media	13.4	19.1	17.5	19.5	8.2	10.0
SPORT. BANC	Media	31.7	119.8	25.0	33.9	13.2	10.9
T_IND	Media	48.6	51.3	58.6	33.4	49.2	45.6
T_SERV	Media	44.7	47.2	35.9	61.3	46.3	39.4
TASSO_AGR	Media	6.7	1.5	5.5	5.3	4.5	15.0
T_ATT_TOT	Media	43.9	45.8	45.7	42.8	43.2	42.8

Dal confronto delle medie di ciascun segmento con le rispettive medie della popolazione risulta di particolare interesse il **cluster 5** che ha:

- un valor medio degli ***occupati nell'agricoltura superiore*** *al valor medio* della popolazione
- un valor medio del ***numero di sportelli bancari inferiore*** *al valor medio* della popolazione
- un valor medio dei ***depositi bancari per abitante superiore*** *al valor medio* della popolazione

I comuni appartenenti al **cluster 5** sono dunque buoni candidati per **l'apertura di nuove agenzie.**

Utilizzo del software Knime - Modulo2_Esercizio2

- Importare dalla cartella ***Dati*** nella chiavetta Usb con il nodo **Excel Readel (XLS)** la tabella ***Acque_Minerali.xlsx***;
- normalizzare con il nodo **Normalizer** le colonne con il metodo z-score;
- produrre con il nodo **k-Means** 4 cluster;
- riportare con il nodo **Denormalizer** (unendo la porta del modello in uscita del nodo Normalizer) i valori normalizzati a quelli originari;
- l'uscita va collegata a 2 nodi di **GroupBy**; il primo, senza raggruppamenti, per calcolare le medie delle sole colonne numeriche; il secondo, raggruppando per la colonna *Cluster*, per calcolare le medie di tutte le colonne numeriche e concatenando i valori della colonna *Marca*;
- accodare la seconda tabella in uscita alla prima con il nodo **Concatenate**; Riordinare le colonne con il nodo **Column Resorter** mettendo *Cluster* e *Marca* all'inizio;
- interpretare i risultati.

Indice

- Regole di associazione
  - Market Basket Analysis
  - Misure di qualità e affidabilità delle regole
- Metodi di segmentazione non supervisionata
  - Cluster Analysis
    - Gerarchica
    - Non gerarchica
- **Riduzione delle variabili**
  - Analisi delle componenti principali (*Principal Component Analysis, PCA*)
  - Analisi delle corrispondenze (*Correspondence Analysis, CA*)

Riduzione delle variabili con l'analisi delle componenti principali (PCA)

Spesso per ridurre la complessità del problema in termini di numero di variabili è possibile eseguire un'analisi delle componenti principali (*Principal Component Analysis* o *PCA*) che **sostituisce,** con una limitata perdita di informazione, **le variabili osservate con un numero ridotto, dette componenti** o variabili "*latenti*".

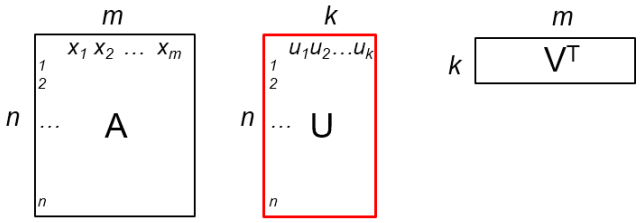
I vantaggi principali di questo approccio sono:

- Una **migliore interpretazione** dei gruppi dovuta alle componenti in luogo delle variabili originarie;
- La possibilità di **rappresentare graficamente** i gruppi su un grafico di dispersione (scatter plot).

Analisi delle componenti principali (PCA)

Con la PCA si ottiene una matrice U di dimensioni ridotte ($k \leq m$) rispetto alla matrice originale A attraverso una trasformazione lineare tale che $A = UV^T$.

Si prende come esempio lo stato di salute di alcune importanti società[1] per ridurre i 6 indicatori di performance aziendali a 3 variabili di sintesi:



La matrice **A** dei dati originari

Azienda	Profitto Econom.	Cash Flow sul Fatt. (%)	Costo Lavoro sul Valore Aggiunto (%)	ROE - Utile Netto sul Patrim. (%)	Indebitam. su Cap. Proprio	Fatturato
Barilla	-25,4	7,39	59,54	4,2	0,83	2867
Eridania	-141	4	68,99	4,2	0,83	1693
Ferrero	65,8	9,61	53,7	21,12	-0,02	3031
Galbani	-71,9	8,4	56,32	2,66	-0,02	2136
Kraft	-32	5,88	72,11	3,2	0,35	1563
Lavazza	-28,9	4,96	39,08	5,29	-0,05	1117
Nestlé	-98,8	2,72	81,25	0	1,69	3463
Parmalat	-145,1	5,96	38,51	2,23	2,91	1664
Plasmon	31,7	27,76	31,35	24,6	1,35	858
Star	2,4	6,47	62,49	10,6	0	811

La matrice **U** delle ***nuove variabili***

Azienda	C1	C2	C3
Barilla	-0,54672	0,460739	0,729316
Eridania	-1,40883	-0,32649	-0,5464
Ferrero	1,476889	1,706276	1,221071
Galbani	-0,44279	0,389704	-0,41893
Kraft	-0,61107	0,706074	-0,62381
Lavazza	0,507905	0,035891	-1,35646
Nestlé	-2,46727	0,216682	1,339776
Parmalat	-0,93717	-2,72256	0,233681
Plasmon	3,789719	-1,22557	0,584723
Star	0,639339	0,759245	-1,16297

La matrice **V** dei coefficienti (***factor loadings***)

	C1	C2	C3
Profitto Economico	0,776854	0,561187	0,128569
Cash Flow su Fatt. (%)	0,870951	-0,23819	0,240295
Costo Lavoro sul Valore Aggiunto (%)	-0,70768	0,524198	0,095379
ROE - Utile Netto sul Patrim. (%)	0,916376	0,128641	0,259253
Indebitam. su Cap. Proprio	-0,24234	-0,83509	0,444691
Fatturato	-0,49179	0,361539	0,755359

[1] Le 5000 società leader, supplemento a Milano Finanza, 1988.

Scelta e interpretazione delle componenti

Un criterio per decidere quante componenti sintesi tenere è quello della percentuale di variabilità spiegata[1] che, per ogni singola componente, non dovrebbe essere sotto il 5-10% o, cumulate, sotto il 90-95%; un altro criterio è che l'autovalore sia > 1.

In questo esempio si scelgono 3 componenti che spiegano il 90% della variabilità complessiva.

Autovalore	% di varianza	% di varianza
2,7029	50,1	50,1
1,3420	24,9	74,9
0,8270	15,3	90,2
0,3066	5,7	95,9
0,1347	2,5	98,4
0,0869	1,6	100,0

L'interpretazione delle componenti viene effettuata in base ai loro coefficienti o ***factor loadings*** (sono i pesi con i quali le variabili originarie contribuiscono alla definizione di ogni singola componente attraverso la loro combinazione lineare).

*In questo esempio il Cash Flow e il ROE sono fortemente correlati positivamente con la C1 che si può interpretare come "**redditività**"; il Profitto Economico e l'Indebitamento sono correlati, uno negativamente e l'altro positivamente con la C2 che si può interpretare come "**Indebitamento**".*

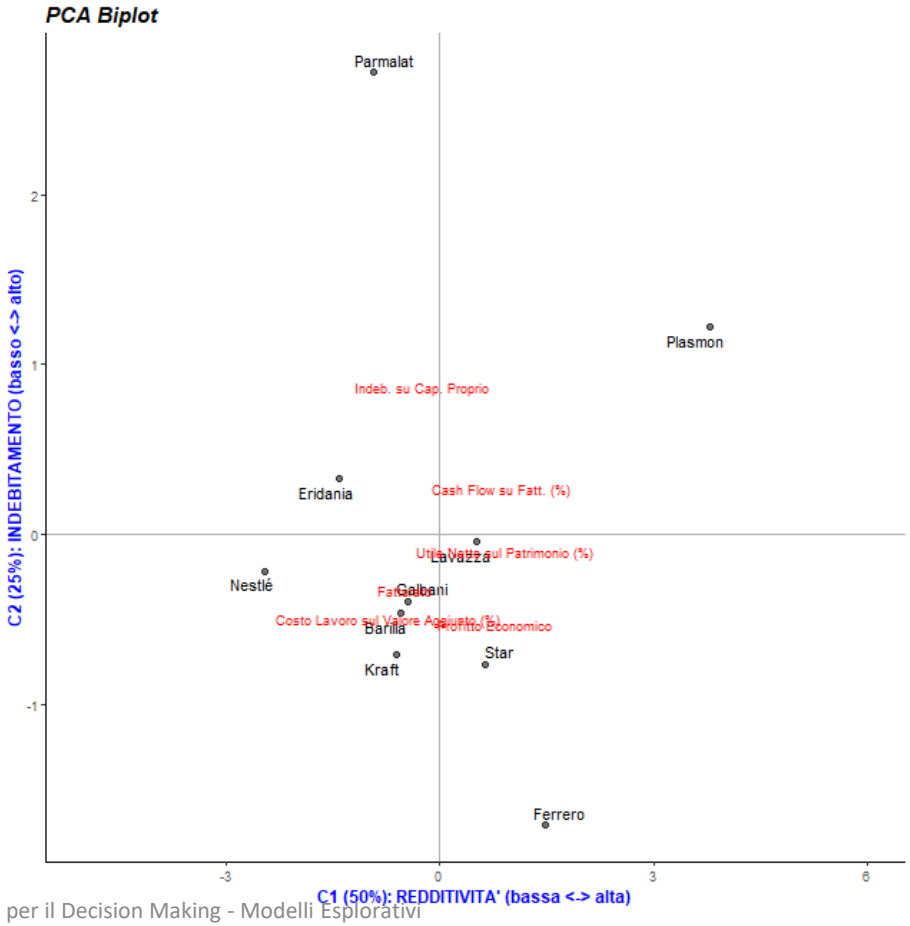
	C1	C2	C3
Profitto Economico	0,776854	-0,56119	0,128569
Cash Flow su Fatt. (%)	0,870951	0,238189	0,240295
Costo Lavoro sul Valore Aggiunto (%)	-0,70768	-0,5242	0,095379
ROE - Utile Netto sul Patrim. (%)	0,916376	-0,12864	0,259253
Indebitam. su Cap. Proprio	-0,24234	0,835086	0,444691
Fatturato	-0,49179	-0,36154	0,755359

[1] È il rapporto tra l'autovalore della componente di interesse e la somma di tutti gli autovalori della matrice di correlazione.

Rappresentazione grafica delle prime due componenti

Il **grafico di dispersione ("*biplot*")** delle prime due componenti, insieme ai factor loadings, permette di fare qualche osservazione:

- La *Plasmon* presenta elevatissimi valori di redditività (C1) e un indebitamento sopra la media (C2).
- La *Parmalat* presenta scarsi valori di redditività (C1) e un fortissimo indebitamento sopra la media (C2).
- Le aziende vicino al centro degli assi presentano redditività e indebitamento nella media.



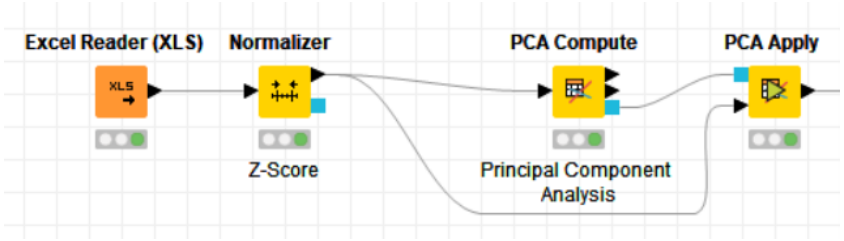
Utilizzo del software Knime - Modulo2_Esempio6

Si vogliono creare dei gruppi di modelli di automobili in base alle loro caratteristiche tecniche:

Modello	Rendimento (Km/litro)	Numero di cilindri	Cilindrata (cm cubici)	Cavalli vapore	Rapporto assale posteriore	Peso (kg)	Lentezza (0,5 km / 1 sec)	Tipo di motore (0=V, 1=in linea)	Transmissione (0=automatica, 1=manuale)	Numero marce	Numero carburatori
Mazda RX4	3,80	6	2622	110	3,9	1188	26,49	0	1	4	4
Mazda RX4 Wa	3,80	6	2622	110	3,9	1304	27,39	0	1	4	4
Datsun 710	4,10	4	1770	93	3,85	1052	29,95	1	1	4	1
Hornet 4 Drive	3,90	6	4228	110	3,08	1458	31,29	1	0	3	1
Hornet Sportal	3,40	8	5899	175	3,15	1560	27,39	0	0	3	2
Valiant	3,30	6	3687	105	2,76	1569	32,54	1	0	3	1
Duster 360	2,60	8	5899	245	3,21	1619	25,49	0	0	3	4
Merc 240D	4,40	4	2404	62	3,69	1447	32,19	1	0	4	2
Merc 230	4,10	4	2307	95	3,92	1429	36,85	1	0	4	2
Merc 280	3,50	6	2746	123	3,92	1560	29,45	1	0	4	4
Merc 280C	3,20	6	2746	123	3,92	1560	30,42	1	0	4	4
Merc 450SE	3,00	8	4520	180	3,07	1846	28	0	0	3	3
Merc 450SL	3,10	8	4520	180	3,07	1692	28,32	0	0	3	3
Merc 450SLC	2,70	8	4520	180	3,07	1715	28,97	0	0	3	3

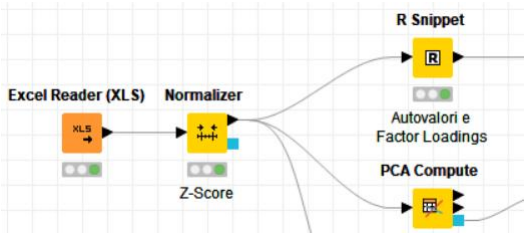
(segue)

Si procede con la normalizzazione delle variabili, poi si applica il nodo **PCA Compute** per effettuare l'analisi delle componenti principali e infine il nodo **PCA Apply** per portare in uscita le componenti scelte.



Utilizzo del software R all'interno di Knime - Modulo2_Esempio6

Per il calcolo dei factor loadings si utilizza il linguaggio R attraverso il nodo **R snippet**, in particolare la funzione **PCA** del pakage FactoMiner.



R Script

```
folder <- "C:\\Program files\\R\\Packages"
library(FactoMineR, lib.loc=folder)

df <- knime.in[,2:ncol(knime.in)]

pca <- PCA(df, scale.unit=FALSE, ncp=3, graph=FALSE)

pca$eig

pca$var$cor
```

con la quale si ottengono gli autovalori, la percentuale di varianza e i factor loadings.

Secondo il criterio della percentuale di varianza si dovrebbero tenere almeno 3 componenti che, insieme, spiegano circa il 90% della variabilità complessiva.

	Autovalore	% varianza	% varianza cumulata
Comp.1	6.3991	60.0	60.0
Comp.2	2.5676	24.1	84.1
Comp.3	0.6072	5.7	89.8
Comp.4	0.2612	2.5	92.3
Comp.5	0.2171	2.0	94.3
Comp.6	0.2055	1.9	96.3
Comp.7	0.1312	1.2	97.5
Comp.8	0.1195	1.1	98.6
Comp.9	0.0757	0.7	99.3
Comp.10	0.0507	0.5	99.8
Comp.11	0.0214	0.2	100.0

Utilizzo del software R all'interno di Knime - Modulo2_Esempio6

Per interpretare le componenti bisogna identificare le variabili che hanno, in valore assoluto, degli alti valori di factor loading. Di solito un valore di factor loading è considerato "alto" quando, in valore assoluto, supera lo 0,3. In questa tabella sono stati nascosti i valori di factor loading superiori in valore assoluto a 0,3.

La prima componente (**asse orizzontale**) risulta essere correlata positivamente con le variabili numero cilindri e cilindrata e negativamente con il rendimento, il tipo motore (quello a V) e il rapporto assale posteriore (coppia). Quest'asse potrebbe essere interpretato come "***Fascia***" (bassa vs alta).

La seconda componente (**asse verticale**) risulta essere correlata positivamente con le variabili numero di marce, trasmissione e carburatori e negativamente con la lentezza. Quest'asse potrebbe essere interpretato come "***Prestazioni***" (basse vs alte).

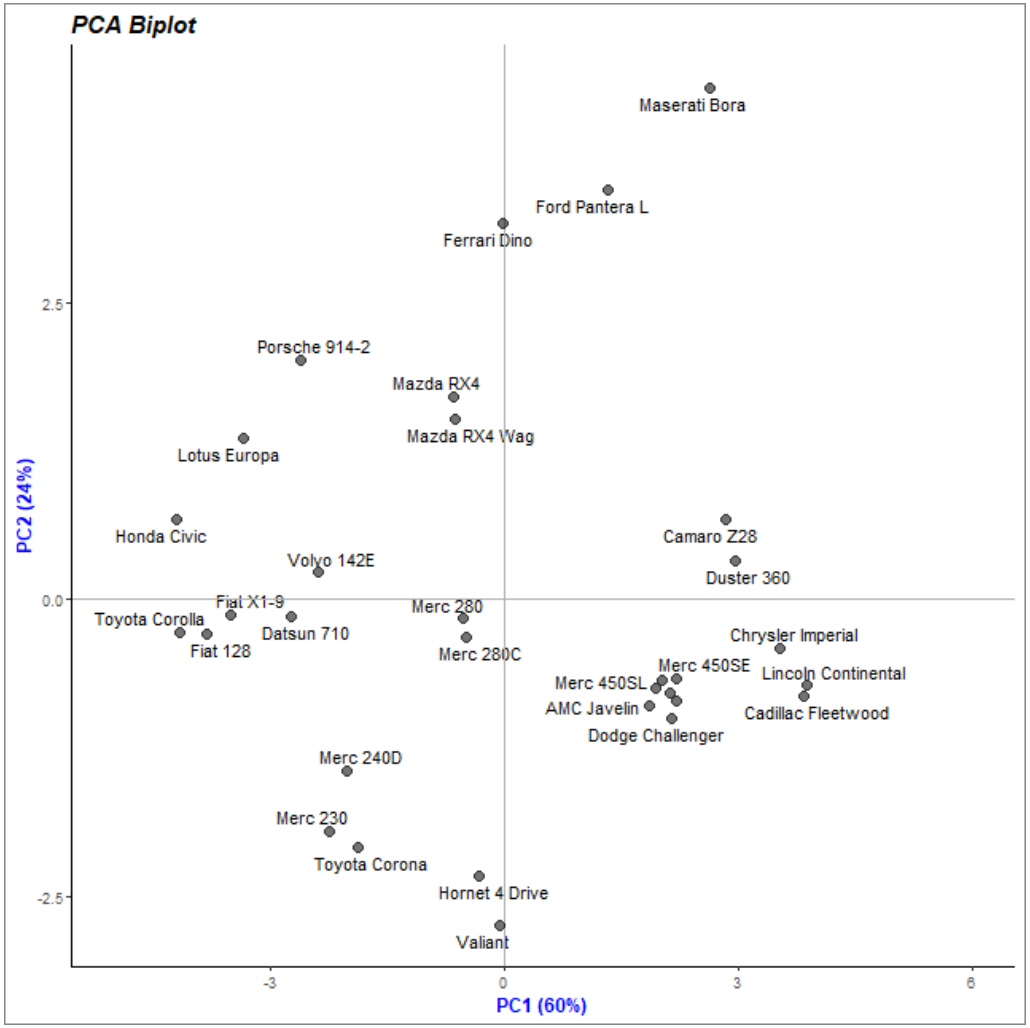
Caratteristica tecnica	Comp.1	Comp.2	Comp.3
Consumo (Km/litro)	**-0,930**		
Numero di cilindri	**0,961**		
Cilindrata (cm cubici)	**0,946**		
Cavalli vapore	0,848	0,405	
Rapporto assale posteriore	**-0,756**	0,447	
Peso (kg)	**0,889**	-0,232	
Tempo per 400 metri (secondi)	-0,515	**-0,755**	0,319
Tipo di motore (0=V, 1=in linea)	**-0,788**	-0,378	0,339
Trasmissione (0=autom., 1=manuale)	-0,604	**0,699**	
Numero marce	-0,532	**0,752**	
Numero carburatori	0,550	**0,674**	**0,419**

Utilizzo del software R all'interno Knime - Modulo2_Esempio6

Biplot ottenuto dalle prime due componenti principali.

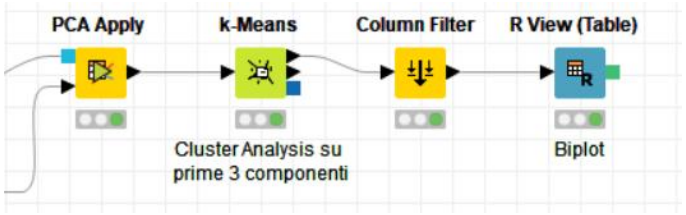
La Maserati Bora, che si trova nell'angolo in alto a destra del biplot è un'auto sportiva e potente.

Al contrario, una Toyota Corona, posizionata nell'angolo inferiore sinistro del biplot, è un'auto economica e lenta.



Utilizzo del software Knime - Modulo2_Esempio6

Si sono infine calcolati 4 cluster utilizzando il metodo delle k-means sulle tre componenti come input.



Cluster 0: "***City cars***", auto leggere, con cilindrata ridotta e basso consumo di carburante.

Cluster 1: "***Personal luxury cars***", auto pesanti, con molti cilindri, alto consumo di canburante.

Cluster 2: "***Supercars***", auto pesanti, con motori di grande cilindrata.

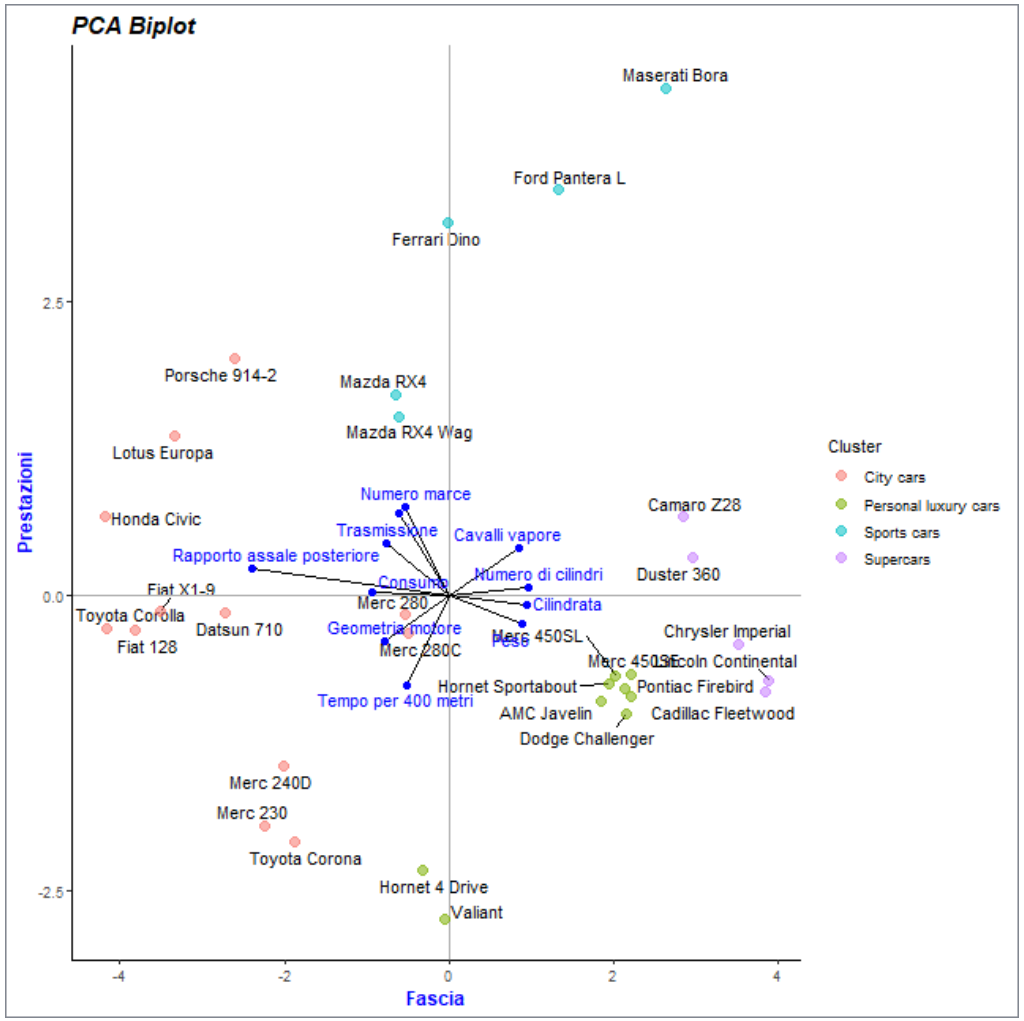
Cluster 3: "***Sports cars***", auto veloci, con molti cilindri, più carburatori e più marce.

Colonna	Statistica	Totale	cluster_0	cluster_1	cluster_2	cluster_3
	N		13	9	5	5
	%		40,6	28,1	15,6	15,6
Cavalli vapore	Media	146,7	88,8	156,1	228,0	**198,8**
Cilindrata (cm cubici)	Media	3780,9	1880,3	4902,4	**6823,4**	3661,0
Consumo (Km/litro)	Media	3,6	**4,6**	3,2	2,3	3,4
Geometria motore (0=V, 1=in linea)	Media	0,4	0,9	0,2	0,0	0,0
Numero carburatori	Media	2,8	1,9	2,1	4,0	**5,2**
Numero di cilindri	Media	6,2	4,3	7,6	**8,0**	6,8
Numero marce	Media	3,7	4,1	3,0	3,0	**4,6**
Peso (kg)	Media	1459,2	1117,3	1637,7	**2125,2**	1361,0
Rapporto assale posteriore	Media	3,6	4,0	3,0	3,2	3,8
Tempo per 400 metri (secondi)	Media	28,7	30,7	28,8	27,2	**25,1**
Trasmissione (0=automatica, 1=manu	Media	0,4	0,6	0,0	0,0	1,0

Riduzione delle variabili

- **Utilizzo del software R all'interno di Knime - Modulo2_Esempio6**

Biplot delle prime due component con evidenziati i cluster di appartenenza e i Factor Loadings



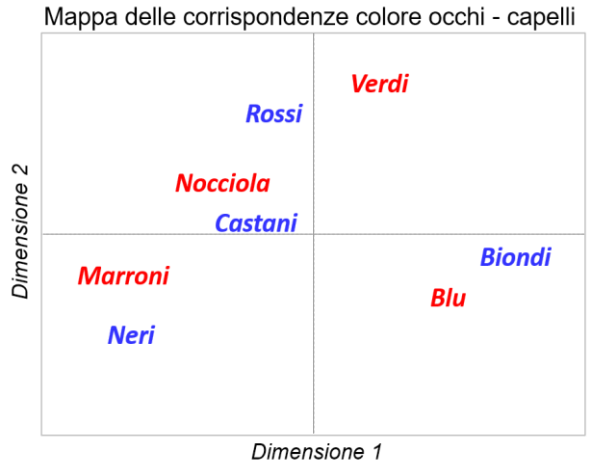
Analisi delle corrispondenze (CA)

Quando le variabili da analizzare sono di **tipo categorico**, si ricorre all'**analisi delle corrispondenze** (*Correspondence Analysis*), semplici (*CA*) o multiple (*MCA*).

La CA è una tecnica che consente di ottenere una **rappresentazione grafica a 2 dimensioni (biplot)** delle righe e delle colonne di una tabella di contingenza[1] che spiegano buona parte della variabilità osservata.

In questo modo si ottiene così una "mappa" sulla quale le modalità delle righe e delle colonne della tabella vengono a essere rappresentate dalle proiezioni dei loro punti rappresentativi.

	Neri	Castani	Rossi	Biondi
Marroni	68	119	26	7
Blu	20	84	17	94
Nocciola	15	54	14	10
Verdi	5	29	14	16



L'interpretazione delle **prossimità delle proiezioni** sulla mappa conduce l'analista a riconoscere i **legami tra le modalità** delle righe e delle colonne.

[1] Le tabelle di contingenza sono un particolare tipo di tabelle a doppia entrata in cui vengono incrociate due variabili *qualitative* per analizzare le relazioni tra due o più variabili in base alle **frequenze** delle combinazioni delle loro categorie

Le variabili di riga, di colonna e supplementari

Data una tabella di dati, come quella sotto riportata, si devono indicare all'inizio quali sono le **variabili di riga e di colonna**.

Professione	Sesso	Area	Soddisfazione
Studente	Maschio	Sud	2: Bassa
Studente	Femmina	Sud	1: Molto bassa
Studente	Femmina	Nord	3: Media
Studente	Femmina	Centro	3: Media
Studente	Femmina	Centro	2: Bassa
Pensionato	Maschio	Sud	2: Bassa
Pensionato	Maschio	Nord	4: Alta
Pensionato	Maschio	Centro	3: Media
Pensionato	Femmina	Sud	3: Media
Pensionato	Femmina	Nord	2: Bassa
Pensionato	Femmina	Centro	2: Bassa
Impiegato	Maschio	Sud	5: Molto alta
Impiegato	Maschio	Nord	5: Molto alta
Impiegato	Maschio	Nord	4: Alta

(segue)

Ad esempio, scegliendo *Soddisfazione* come variabile di riga e *Area* e *Professione* come variabili di colonna, si ottiene la seguente **tabella di contingenza**:

Categoria	Area=Centro	Area=Nord	Area=Sud	Professione=Altro	Professione=Impiegato	Professione=Pensionato	Professione=Studente
1: Molto bassa	1	0	3	3	0	0	1
2: Bassa	4	3	2	3	1	3	2
3: Media	3	3	1	1	2	2	2
4: Alta	0	2	1	0	2	1	0
5: Molto alta	0	1	1	0	2	0	0

Si possono anche indicare delle **variabili supplementari** che **non vengono usate nell'analisi** (per esempio la variabile *Sesso*). Le loro coordinate possono però essere calcolate con il modello ottenuto per migliorare l'interpretazione dei risultati.

Categoria	Sesso=Femmina	Sesso=Maschio
1: Molto bassa	3	1
2: Bassa	5	4
3: Media	4	3
4: Alta	1	2
5: Molto alta	0	2

Riduzione delle variabili

Le variabilità spiegata e le statistiche di riepilogo

- La **percentuale di variabilità spiegata** (o inerzia)

  In questo esempio, le 2 dimensioni spiegano il 96% circa della variabilità.

Dettagli

Valore singolare	Inerzia	Chi-quadrato	Percentuale	Percentuale cumulativa	20 40 60 80
0,55026	0,30279	15,140	63,10	63,10	
0,39685	0,15749	7,874	32,82	95,92	
0,13110	0,01719	0,859	3,58	99,50	
0,04890	0,00239	0,120	0,50	100,00	

- Le **statistiche di riepilogo** e le **coordinate** di riga (OBS) e di colonna (VAR)

  SqCos indicano l'**importanza che ha ciascuna dim. sul singolo item**; la somma in orizzontale dà 1.

  Contr sono i contributi che ogni item dà alla variabilità per quella dimensione: **più i valori sono alti, più l'item è importante per quella dimensione**; la somma in verticale di ogni contributo per le righe e per le colonne dà 1.

		Quality	Mass	Inertia	Dim1	Dim2	Contr1	Contr2	SqCos1	SqCos2
OBS	1: Molto bas	**0,999**	0,160	0,182	**-0,873**	0,611	**0,403**	**0,380**	**0,670**	**0,329**
OBS	2: Bassa	0,945	0,360	0,042	-0,205	-0,259	0,050	0,153	0,364	0,581
OBS	3: Media	0,851	0,280	0,036	0,123	-0,310	0,014	0,170	0,115	**0,735**
OBS	4: Alta	0,920	0,120	0,098	0,842	0,203	**0,281**	0,031	**0,869**	0,050
OBS	5: Molto alta	**0,969**	0,080	0,122	**0,977**	**0,722**	**0,252**	**0,265**	**0,626**	**0,342**
SUP	Sesso: M				0,350	0,104				
SUP	Sesso:F				-0,323	-0,096				
VAR	Prof.: Pen	0,842	0,120	0,039	0,143	-0,501	0,008	**0,191**	0,063	0,778
VAR	Prof.: Imp	**0,986**	0,140	0,147	**0,955**	0,350	**0,422**	0,109	**0,869**	0,117
VAR	Prof.: Stu	0,800	0,100	0,027	-0,377	-0,265	0,047	0,045	0,535	0,264
VAR	Prof.: Alt/Dis	**0,992**	0,140	0,102	**-0,808**	0,269	**0,302**	0,064	**0,893**	0,099
VAR	Area: N	**0,972**	0,180	0,049	**0,487**	-0,162	0,141	0,030	**0,876**	0,097
VAR	Area: C	0,926	0,160	0,047	-0,301	-0,426	0,048	**0,185**	0,308	0,618
VAR	Area: S	**0,994**	0,160	0,069	-0,247	**0,608**	0,032	**0,376**	0,141	**0,853**

Quality è la somma degli SqCos: indica quanto il punto identificato dalle 2 coordinate è rappresentato nel piano: se il valore è alto, è identificato bene; se il valore è basso, 2 dimensioni non bastano.

Dim sono le coordinate dello spazio bidimensionale; gli item che hanno un valore assoluto alto di coordinata aiutano notevolmente all'interpretazione della dimensione.

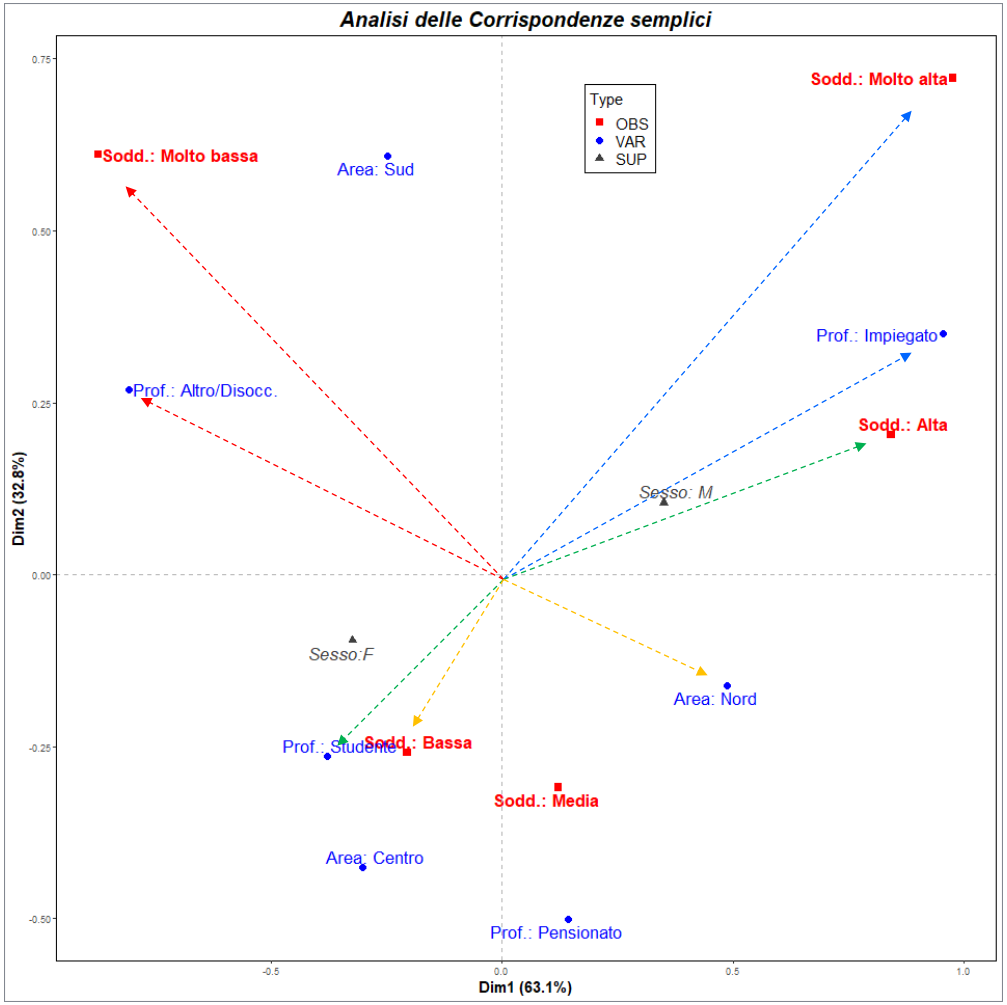
Interpretazione dei risultati

1. Si guardano le lunghezze delle linee che collegano la modalità di riga all'origine: le **linee più lunghe indicano che c'è un'alta associazione** con le modalità di colonna; le linee più corte indicano anch'esse che c'è associazione, ma più debole.

2. Si fa lo stesso procedimento con le modalità di colonna.

3. Si osserva l'angolo formato da queste 2 linee:

- piccoli angoli indicano **associazione positiva** (*Sodd.:Molto bassa* e *Prof.:Altro/Disocc.*, *Sodd.:Molto alta* e *Prof.: Impiegato*)
- angoli a 90° **nessuna associazione** (*Sodd.:Bassa* e *Area: Nord*);
- angoli vicini a 180° indicano **associazione negativa** (*Sodd.:Alta* e *Prof.:Studente*)



Utilizzo del software R all'interno di Knime - Modulo2_Esempio8

I dati di un **questionario sugli ascolti radiofonici**, come illustrato a lato, sono stati trasformati nella sottostante tabella. La preferenza viene usata come variabile di riga; sesso, età e ora come variabili di colonna:

QUESTIONARIO ASCOLTATORI

1. Qual è la tua età?
2. Sesso?
3. Mediamente, nei giorni FERIALI, quante ore ascolti la radio?
4. Mediamente, nei giorni FESTIVI, quante ore ascolti la radio?

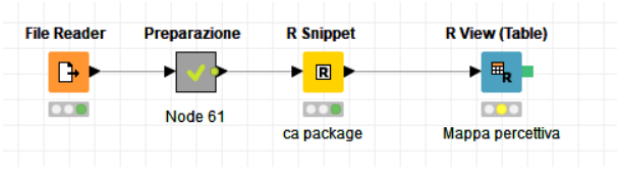
Codici di risposta

0 Non ascolto a quell'ora	
1 Rock	**5 Classica**
2 Top 40	**6 Easy Listening**
3 Country	**7 News/Info/Discuss**
4 Jazz	**8 Altro**

In un tipico giorno FERIALE, che genere di musica ascolti
6. dalle 6 alle 9?
7. dalle 9 a mezzogiorno?
8. da mezzogiorno alle 13?
9. dalle 13 alle 16?
10. dalle 16 alle 18?
11. dalle 18 alle 20?
12. dalle 20 alle 2?

In un tipico giorno FESTIVO, che genere di musica ascolti
13. dalle 6 alle 9?
14. dalle 9 a mezzogiorno?
15. da mezzogiorno alle 13?
16. dalle 13 alle 16?
17. dalle 16 alle 18?
18. dalle 18 alle 20?
19. dalle 20 alle 2?

Utilizzo del software R all'interno di Knime - Modulo2_Esempio8



Le risposte sono state inserite in un foglio elettronico e poi importate in KNIME,

	A	B	C	D	E	F	G	H	I
1	Età	Sesso	Orario:06-09	Orario:09-12	Orario:12-13	Orario:13-16	Orario:16-18	Orario:18-20	Orario:20-02
2	32	f	7	5	5	5	7	0	0
3	30	f	5	0	0	0	5	0	0
4	39	f	1	0	0	0	1	0	0
5	40	f	7	5	0	5	7	0	0
6	37	m	1	5	5	5	5	4	4
7	35	f	7	0	0	0	7	0	0
8	42	m	7	0	0	0	7	5	4
9	39	f	7	0	0	0	7	7	0
10	28	m	7	0	0	0	0	0	0
11	28	f	1	0	0	0	0	1	1
12	24	f	2	0	0	0	0	0	0

Categorie delle variabili di colonna

e poi opportunamente trasposte e ricodificate:

	Sesso.F	Sesso.M	Età.<21	Età.21.25	Età.25.30	Età.30.35	Età.>35	Orario.06.09	Orario.09.12	Orario.12.13	Orario.13.16	Orario.16.18	Orario.18.20	Orario.20.02
News/Info/Discuss	113	88	1	17	43	61	79	96	3	9	2	69	20	2
Classica	81	72	4	21	34	34	60	19	33	18	30	22	20	11
Non ascolto a quell'c	854	370	64	141	388	303	328	34	214	238	216	56	202	264
Rock	246	127	20	74	144	84	51	98	39	43	38	93	38	24
Jazz	34	39	1	3	26	17	26	6	6	3	7	19	15	17
Top 40	123	34	16	17	88	23	13	45	20	12	22	37	16	5
Altro	42	13	4	10	19	13	9	17	4	4	3	18	7	2
Easy Listening	70	8	2	9	20	19	28	14	12	6	13	16	9	8
Country	19	19	0	9	8	6	15	7	5	3	5	6	9	3

Categorie della variabile di riga

Utilizzo del software R all'interno di Knime - Modulo2_Esempio8

Si è proceduto all'analisi delle corrispondenze utilizzando la funzione ca() del package *ca* e le funzioni del package *factoextra* all'interno di un nodo R Snippet. Come variabile di riga viene assunta *Preferenza* e come variabili di colonna *Fascia_Età* e *Orario*; *Sesso* come variabile supplementare.

```
library(ca); library(factoextra)
...
ca <- ca(ct, nd=2, supcol=1:2)
summary(ca, scree=T, rows=F, columns=F)
```

Dallo standard output del nodo R Snippet si legge che la prima dimensione spiega il 70,6% della variabilità, il secondo il 14,8% (insieme l'90,5%).

```
Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.162563  75.6  75.6   ******************
 2      0.031893  14.8  90.5   ****
 3      0.008446   3.9  94.4   *
 4      0.005909   2.7  97.2   *
 5      0.004421   2.1  99.2   *
 6      0.001276   0.6  99.8
 7      0.000334   0.2 100.0
 8      7.8e-050   0.0 100.0
        -------- -----
 Total: 0.214921 100.0
```

Utilizzo del software R all'interno di Knime - Modulo2_Esempio8

La **DIMENSIONE 1** ha il valore positivo più alto in corrispondenza della modalità di riga *News/Info/Discuss* e negativo per la modalità *Non ascolto a quell'ora*; positivo più alto per la modalità di colonna *6-9, 16-18* e negativo per le modalità *12-13, 20-2*.

Graficamente, quindi, nell'asse orizzontale, si concentrano a destra gli ***orari di punta con preferenza ai notiziari*** e a sinistra ***quelli dei pasti o notturni***.

La ***DIMENSIONE 2*** ha valore positivo in corrispondenza delle modalità di riga Jazz, News/Info/Discuss e negativo per la modalità *Top 40*; positivo per le modalità di colonna *>35* e negativo per le modalità *<21*.

Graficamente, quindi, nell'asse verticale, si concentra in alto il ***pubblico adulto che ascolta notiziari, musica jazz e popolare*** e in basso ***quello dei giovani che ascoltano musica rock e le hit parade***.

	Dim1	Dim2
News/Info/Discuss	**0,847804**	**0,331977**
Classica	-0,01709	0,181648
Non ascolto a quell'ora	**-0,34148**	0,009718
Rock	0,369471	-0,19076
Jazz	0,041679	0,201475
Top 40	0,395253	**-0,40847**
Altro	0,558614	-0,13154
Easy Listening	0,148053	0,167853
Country	0,13152	0,218539

	Dim1	Dim2
Età:<21	-0,10607	**-0,43293**
Età:21-25	0,067504	-0,13888
Età:25-30	0,022793	-0,22798
Età:30-35	-0,00281	0,10167
Età:>35	-0,04009	**0,343016**
Orario:06-09	**1,005286**	0,023798
Orario:09-12	-0,32379	-0,04582
Orario:12-13	**-0,3667**	-0,04655
Orario:13-16	-0,33432	-0,0622
Orario:16-18	0,752844	0,000829
Orario:18-20	-0,18417	0,067159
Orario:20-02	**-0,54915**	0,062785

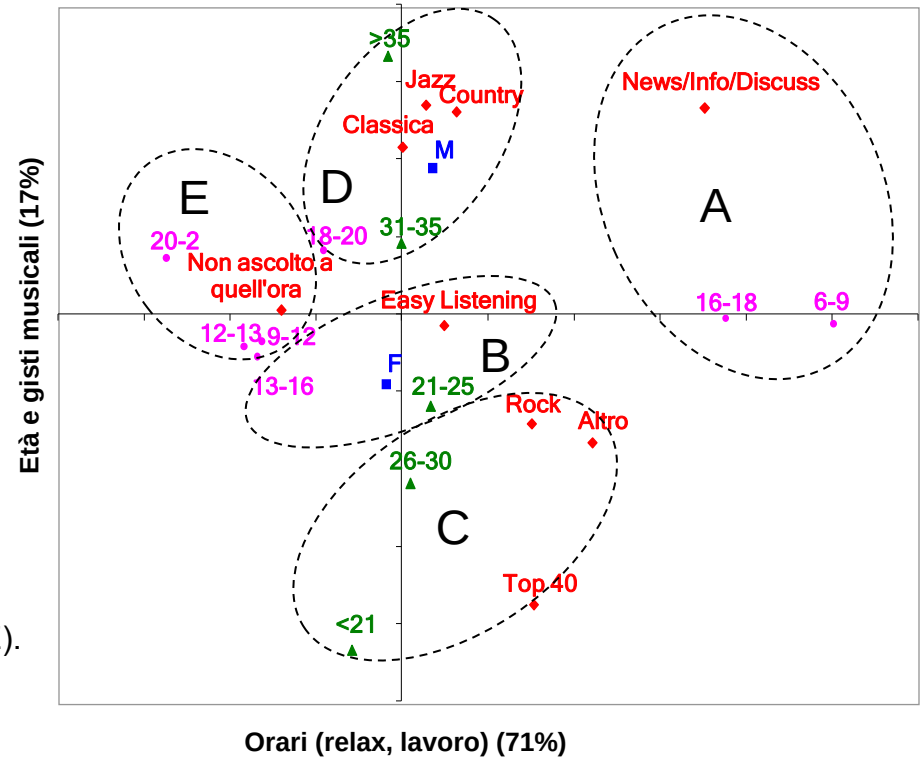
Utilizzo del software R all'interno di Knime - Modulo2_Esempio8

Dalle coordinate ottenute è possibile creare una “mappa” delle relazioni tra le modalità delle diverse variabili per **riassumere le preferenze degli intervistati** in base alle loro risposte.

.

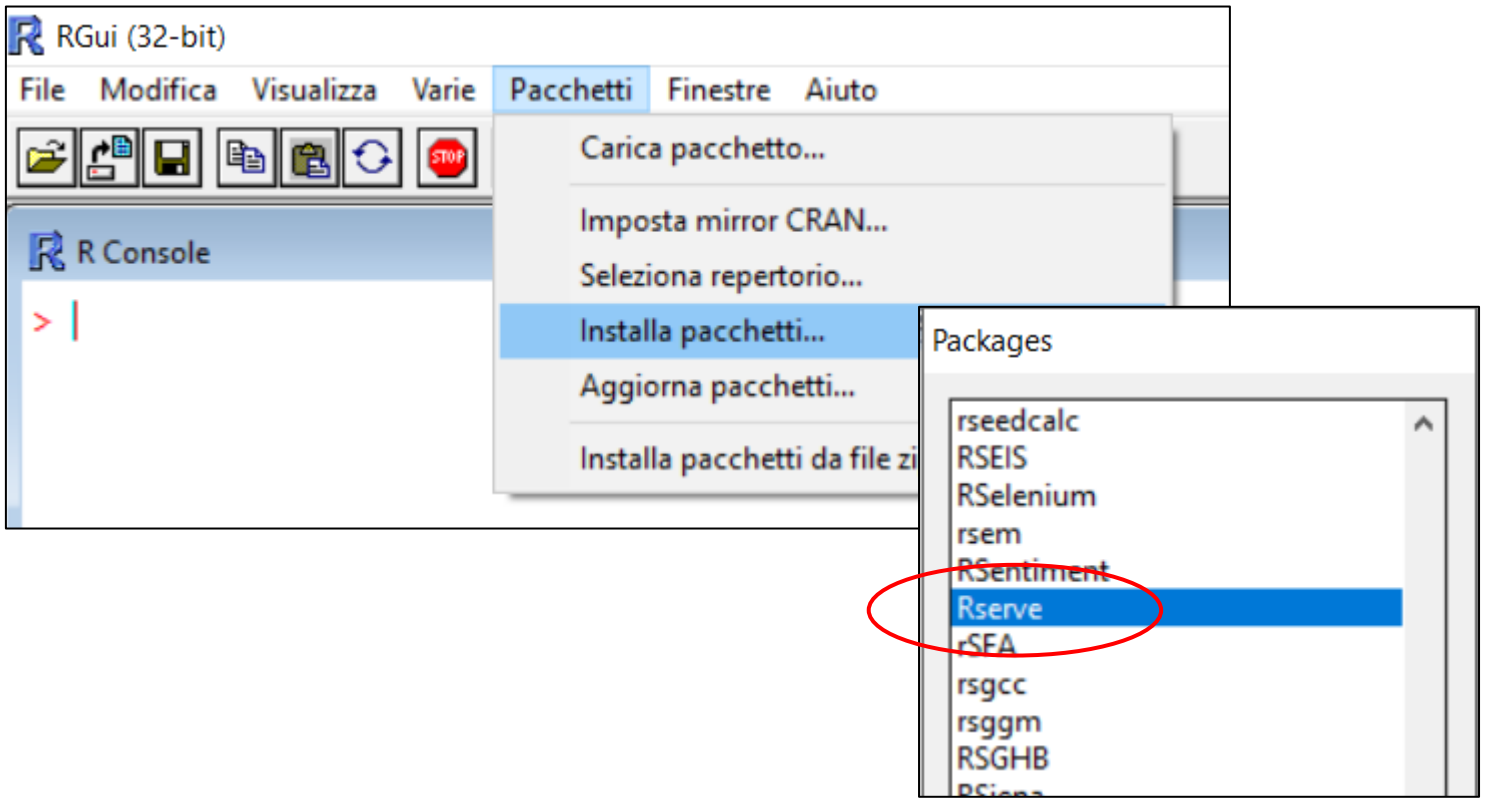
Proiettando i punti delle modalità su un sistema di assi a 2 dimensioni si possono **visualizzare graficamente le associazioni** tra le diverse risposte. Si possono quindi distinguere cinque tipologie di ascoltatori:

- Pubblico adulto che ascolta i notiziari nelle ore di punta (A)
- Pubblico maschile che ascolta musica “impegnata” nella fascia serale (D).
- Pubblico femminile giovane che ascolta musica piacevole e poco impegnativa (B).
- Pubblico under 30 che ascolta musica rock e le canzoni più vendute e popolari del momento (C).
- “Non pubblico” durante la notte e orari di lavoro (E).



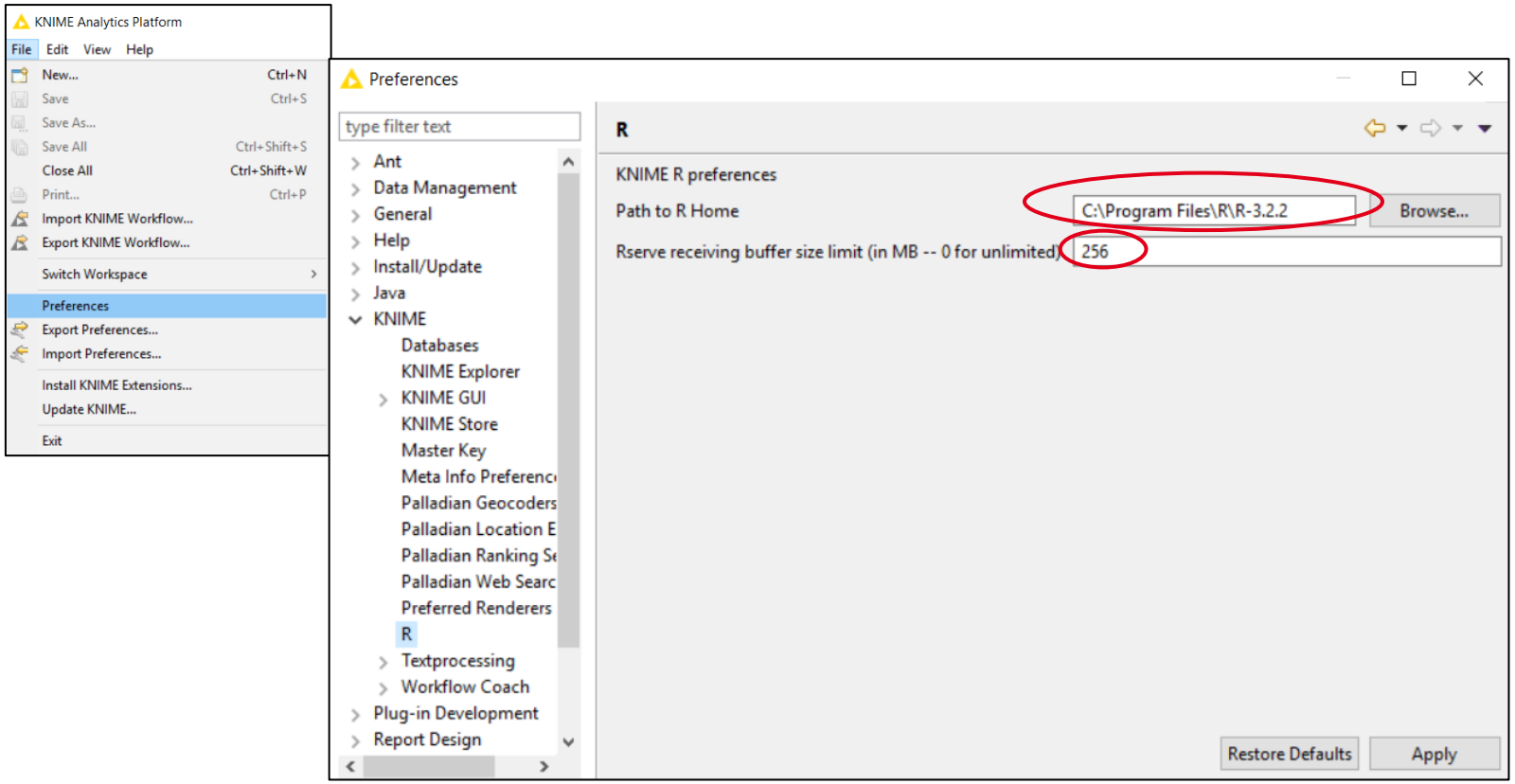
Integrazione con R - 1/2

Installazione in R del package **Rserve**



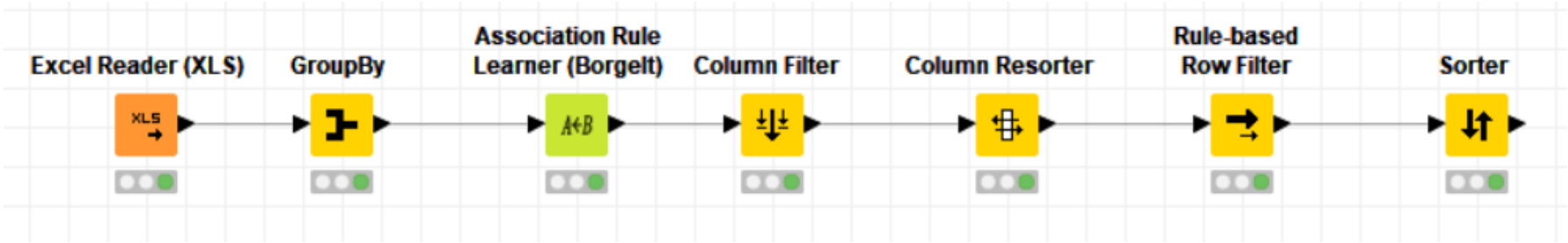
Integrazione con R - 2/2

Indicare nelle preferenze il percorso dell'installazione di R



Appendice

- **Modulo2_Esercizio1**



Appendice

- Modulo2_Esercizio2