



**StatSoft®**

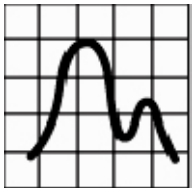
*data analysis • data mining • quality control • web-based analytics*

# **Tecniche di Raggruppamento**

**e**

# **STATISTICA**

## **Case Study: Definizione di Gruppi di Clienti di un Centro Commerciale**



### **STATISTICA**

**Soluzioni per Business Intelligence,  
Data Mining, Quality Control, e  
Web-based Analytics**

## Tabella dei Contenuti

**TECNICHE DI RAGGRUPPAMENTO E STATISTICA..... 3**

Panoramica sugli algoritmi di raggruppamento .....3

    Unione (Raggruppamento Gerarchico) .....4

    Raggruppamento K-Means.....4

    Raggruppamento EM (Expectation Maximization) .....5

**CASE STUDY: DEFINIRE GRUPPI DI VISITATORI DI UN CENTRO COMMERCIALE ..... 6**

**ANALISI DEI DATI CON STATISTICA ..... 9**

Risultati dell’Analisi dei Gruppi *k*-Means ..... 9

Risultati dell’Analisi dei Gruppi EM ..... 10

    Risultati Selezione di Caratteristiche e Screening di Variabili ..... 11

**CONCLUSIONI..... 12**

## Tecniche di Raggruppamento e **STATISTICA**

Il termine analisi dei gruppi o di raggruppamento (usato per la prima volta da Tryon, 1939) in effetti racchiude un numero elevato di algoritmi di classificazione differenti. Un problema generale che incontrano i ricercatori di molte aree di ricerca è quello di organizzare i dati osservati in una struttura con un significato, cioè sviluppare tassonomie. Hartigan (1975) fornisce un sommario eccellente dei tanti studi pubblicati che riportano risultati di analisi dei gruppi. Per esempio, nel campo medico, il raggruppamento delle patologie, delle cure per le malattie o dei sintomi delle malattie può portare a tassonomie estremamente utili. Nel campo psichiatrico, la diagnosi corretta di gruppi di sintomi quali la paranoia, la schizofrenia, ecc. sono essenziali per produrre una terapia utile. In archeologia, i ricercatori hanno cercato di stabilire tassonomie su strumenti di pietra, oggetti funerari, ecc. applicando tecniche di raggruppamento analitiche. In generale, ogni volta in cui si devono classificare "montagne" di informazioni in pile significative e gestibili, l'analisi dei gruppi risulta di grande utilità.

Gli algoritmi di raggruppamento funzionano bene con tutti i tipi di dati, inclusi quelli categoriali, numerici, e testuali. In questo caso non sarà necessario identificare gli input e gli output in principio di analisi. Solitamente la sola decisione che l'utente dovrà prendere è quella di richiedere un numero specificato di gruppi candidati. Altre caratteristiche avanzate disponibili in **STATISTICA** consentiranno all'utente anche di determinare il numero ottimale di gruppi in base alle caratteristiche del file dati di riferimento. L'algoritmo di raggruppamento individuerà il miglior partizionamento di tutti i record di clienti (disponibili nel ns. file dati d'esempio) e fornirà descrizioni circa "medie e centroidi" di ogni gruppo. In molti casi, tali gruppi saranno soggetti ad un'interpretazione tale da fornire suggerimenti circa la segmentazione "naturale" dei clienti che visitano i negozi.

Le tecniche di raggruppamento generalmente appartengono al gruppo di strumenti di data mining indiretto; tali tecniche alcune volte vengono chiamate tecniche di "apprendimento non supervisionato", in quanto non vi è alcuna particolare variabile dipendente o di output i cui valori possano essere previsti (e da usare per "supervisionare" il processo di apprendimento, allo scopo di portare alla produzione del miglior modello predittivo). Invece, l'obiettivo del data mining indiretto o dell'apprendimento non supervisionato è individuare la struttura dei dati. Non vi è alcuna variabile target da prevedere; e quindi non deve esser fatta alcuna distinzione tra variabili indipendenti e dipendenti.

## Panoramica sugli algoritmi di raggruppamento

Le tecniche di raggruppamento vengono usate per organizzare i casi o le osservazioni in gruppi che soddisfino due criteri principali:

1. Ogni gruppo è omogeneo al suo interno; le osservazioni (i casi) che appartengono al medesimo gruppo sono simili tra loro.
2. Ogni gruppo dovrebbe essere differente dagli altri gruppi, ovvero sia le osservazioni appartenenti ad un gruppo dovrebbero essere differenti da quelli contenuti negli altri gruppi.

I metodi di raggruppamento sono comunemente distinti in tre categorie.

1. Unione (Raggruppamento Gerarchico)
2. Raggruppamento *K-means*

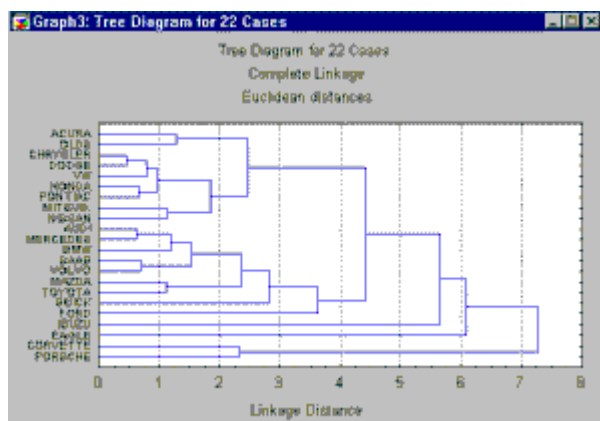
### 3. Raggruppamento *EM* (Expectation Maximization)

Seguono brevi spiegazioni di questi tre metodi.

## Unione (Raggruppamento Gerarchico)

Lo scopo di questo algoritmo è quello di riunire oggetti (per es., animali) in gruppi sempre più grandi, utilizzando una qualche misura di somiglianza o distanza. Un risultato tipico di questo tipo di raggruppamento è il dendrogramma o diagramma ad albero.

Si consideri un *Grafico a Dendrogramma Orizzontale*; A sinistra nel grafico si inizia con ogni oggetto che appartiene ad una classe a sé stante. Ora si immagina che, per passi molto piccoli, si "rilassi" il criterio per cui definire (o meno) unico (separato dagli altri) un oggetto. Detta in altro modo: si abbassa sempre più la soglia relativa alla decisione di dichiarare due o più oggetti come membri di uno stesso gruppo.



Come risultato si riuniranno sempre più oggetti e si aggregeranno (amalgameranno) gruppi sempre più ampi di elementi sempre più dissimili tra loro. Infine, all'ultimo passo, tutti gli oggetti saranno riuniti insieme. In questi grafici, l'asse orizzontale indica la distanza di unione (nei Dendrogrammi Verticali, disponibili, sarà l'asse verticale a indicare le distanze di unione). Quindi, per ogni nodo nel grafico (dove si forma un nuovo gruppo) è possibile leggere il criterio di distanza per cui i rispettivi elementi sono stati riuniti tra loro in un gruppo unico. Quando i dati contengono una chiara "struttura" in termini di gruppi di oggetti simili tra loro, allora questa struttura si rifletterà spesso nel dendrogramma come ramificazioni distinte. Come risultato di una analisi ben riuscita con il metodo d'unione (gerarchico), si dovrebbe essere in grado di individuare gruppi (ramificazioni) ed anche di interpretarli.

## Raggruppamento K-Means

Questo metodo di raggruppamento è molto differente da quello dell'Unione (Raggruppamento Gerarchico). Si supponga di avere già formulato ipotesi concernenti il numero di gruppi nei casi o nelle variabili. Si può per esempio "dire" al computer di formare esattamente 3 gruppi che siano più distinti possibile tra loro. Questo è il tipico problema di ricerca che può essere risolto dall'algoritmo di raggruppamento k-means. In generale, il metodo k-means produrrà esattamente  $k$  gruppi differenti più distinti possibile.

Si supponga che un ricercatore clinico "sospetti", per esperienza clinica acquisita, che i suoi pazienti con patologie cardiache formino fondamentalmente tre categorie differenti rispetto all'idoneità

fisica Egli potrebbe rallegrarsi se la sua intuizione potesse essere quantificata, cioè, se una analisi di raggruppamento *k*-means delle misure di idoneità fisica potesse produrre i tre gruppi di pazienti attesi. Se così fosse, le medie delle differenti misure di idoneità fisica per ogni gruppo rappresenterebbero un percorso quantitativo per esprimere l'ipotesi o l'intuizione del ricercatore (cioè, i pazienti nel gruppo 1 hanno misura 1 elevata, misura 2 bassa, ecc.).

A livello computazionale, si può pensare a questo metodo come ad una analisi della varianza (ANOVA) "invertita". Il programma inizierà con *k* gruppi casuali e quindi sposterà oggetti all'interno di questi gruppi con l'obiettivo di (1) minimizzare la variabilità all'interno dei gruppi e (2) massimizzare la variabilità tra gruppi. Questa operazione è analoga ad una "ANOVA invertita", poiché il test di significatività dell'ANOVA valuta la variabilità tra gruppi rispetto alla variabilità entro gruppi per verificare l'ipotesi che le medie nei gruppi siano differenti tra loro. Nel raggruppamento *k*-means, il programma cerca di spostare gli oggetti (per es., i casi) dentro e fuori dai gruppi per ottenere i risultati ANOVA più significativi.

**Definizione delle "distanze" tra ed all'interno dei gruppi, e raggruppamento "corretto".** Un aspetto importante da notare è che la classificazione prodotta con tutti gli strumenti di raggruppamento (il metodo *k*-means così come il metodo di unione) dipende in modo determinante dalla particolare metrica utilizzata per la determinazione delle "distanze" tra gli oggetti (osservazioni) ed i gruppi. Si ricordi che tutti i metodi di raggruppamento hanno come scopo quello di organizzare le osservazioni in modo che gli item simili tra loro vengano assegnati al medesimo gruppo, mentre gli item dissimili tra loro vengano assegnati a gruppi diversi. Dato che vi sono molti modi di definire le "distanze" (distanze Euclidee, distanze euclidee al quadrato, percentuali discordanze, e così via), non esisterà un'unica soluzione per il raggiungimento della classificazione corretta, sebbene vi sia sempre lo spazio per tentativi diretti alla definizione delle soluzioni 'ottimali'.

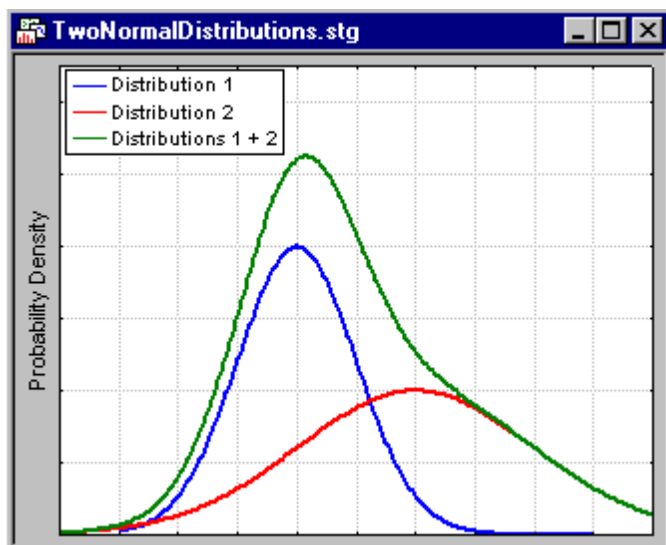
## Raggruppamento EM (Expectation Maximization)

La tecnica di Raggruppamento EM (Expectation Maximization) è un altro tool disponibile in *STATISTICA*. L'obiettivo generale di tali tecniche è quello di individuare i gruppi di osservazioni (o di variabili) e di assegnare le osservazioni ai gruppi. L'algoritmo *EM* (expectation maximization) estende l'approccio di raggruppamento *k*-means in due modi importanti:

1. Invece di assegnare i casi o le osservazioni ai gruppi con l'obiettivo di massimizzare le differenze tra le medie delle variabili continue (od il grado di associazione tra variabili categoriali), l'algoritmo di raggruppamento *EM* calcola le probabilità di appartenenza ai gruppi sulla base di una o più distribuzioni di probabilità. L'obiettivo dell'algoritmo di raggruppamento è quindi di massimizzare la probabilità generale o la verosimiglianza dei dati, considerati i gruppi (finali).
2. Diversamente dalla classica implementazione del raggruppamento *k*-means, l'algoritmo *EM* generale può venire applicato sia a variabili continue che categoriali (sebbene in *STATISTICA* il classico algoritmo *k*-means sia stato modificato in modo da poter trattare anche le variabili categoriali, attraverso la definizione di appropriate misure di "distanza").

L'algoritmo di raggruppamento *EM* è descritto in dettaglio in Witten and Frank (2001). L'approccio e la logica alla base di tale metodo di raggruppamento sono i seguenti. Si supponga che si sia misurata una singola variabile continua in corrispondenza di un grande campione di osservazioni. Inoltre, si supponga che il campione consista di due gruppi di osservazioni con medie differenti (e

forse differenti deviazioni standard); all'interno di ogni campione, la distribuzione dei valori (nella popolazione) potrebbe apparire come segue:



**Misture e distribuzioni.** La precedente illustrazione riporta due distribuzioni normali aventi differenti medie e deviazioni standard, e la somma delle due distribuzioni. Verrà osservata solo la mistura (somma) delle due distribuzioni normali. L'obiettivo del raggruppamento *EM* è stimare le medie e le deviazioni standard per ogni gruppo al fine di massimizzare la verosimiglianza dei dati osservati (distribuzione). In altre parole, l'algoritmo *EM* prova ad approssimare le distribuzioni osservate dei valori sulla base delle misture delle diverse distribuzioni nei differenti gruppi.

**Variabili categoriali.** L'algoritmo *EM* consente di trattare anche variabili categoriali. Il programma dapprima assegnerà casualmente differenti probabilità (pesi, per essere precisi) ad ogni classe o categoria, in corrispondenza di ogni gruppo. Nelle iterazioni successive, tali probabilità verranno ridefinite (aggiustate) in modo da massimizzare la verosimiglianza dei dati dato il numero specificato di gruppi.

**Probabilità di classificazione e non classificazioni.** I risultati del raggruppamento *EM* sono differenti da quelli calcolati dal raggruppamento *k*-means. Quest'ultimo algoritmo assegnerà le osservazioni ai gruppi in modo da massimizzare le distanze tra i gruppi. L'algoritmo *EM* non calcola le reali assegnazioni delle osservazioni ai gruppi, bensì le *probabilità* di classificazione. In altre parole, ogni osservazione appartiene ad ogni gruppo con una certa probabilità. Certamente, come risultato sarà possibile visualizzare i dettagli circa l'assegnazione di osservazioni ai gruppi sulla base della (maggiore) probabilità di classificazione.

## Case Study: Definire Gruppi di Visitatori di un Centro Commerciale

Questo Case Study è basato sull'analisi delle risposte a 502 domande facenti parte di un questionario sottoposto ad un campione di negozianti di un noto centro commerciale di San Francisco (si veda Hastie, Tibshirani, Friedman, 2001). L'obiettivo generale di questo studio è identificare gruppi "tipici" e omogenei di visitatori del centro commerciale. Una volta identificati tali gruppi "prototipi" di clienti, potranno essere progettate speciali campagne di marketing, servizi,

o consigli per gli acquisti specifiche per gruppo al fine di migliorare la qualità dell’offerta e dell’esperienza di shopping.

In questo esempio verrà analizzato un grande numero di clienti indifferenziati per osservare se questi cadono entro gruppi naturali. Questo è un puro esempio di “data mining indiretto” o di “apprendimento non-supervisionato” in cui l’analista non ha una chiara conoscenza a priori dei “tipi” di clientela che visita il centro commerciale e spera che lo strumento di data mining riveli qualche struttura significativa attraverso l’identificazione di gruppi di acquirenti relativamente omogenei tra loro.

**1. ENTRATE FAMILIARI ANNUALI (ENTRATE PERSONALI SE SINGLE)**

1 – Meno di \$10,000	4 – Tra \$20,000 e \$24,999	7 – Tra \$40,000 e \$49,999
2 – Tra \$10,000 e \$14,999	5 – Tra \$25,000 e \$29,999	8 – Tra \$50,000 e \$74,999
3 – Tra \$15,000 e \$19,999	6 – Tra \$30,000 e \$39,999	9 – Maggiore o uguale a \$75,000

**2. SESSO**

1. Maschio	2. Femmina
------------	------------

**3. STATO CIVILE**

1. Sposato	3. Divorziato o separato	5. Single, mai sposato
2. Convivente	4. Vedovo	

**4. ETA’**

1. da 14 a 17	5. da 45 a 54
2. da 18 a 24	6. da 55 a 64
3. da 25 a 34	7. da 65 in poi
4. da 35 a 44	

**5. ISTRUZIONE**

1. Fino al grado 8	3. Diplomato	5. Laureato
2. Tra il grado 9 all’11	4. tra il primo ed il terzo anno di college	6. Studi post-laurea

**6. OCCUPAZIONE**

1. Professionista/Amministrativo	4. Religioso	7. Militare
2. Venditore	5. Costruttore	8. Ritirato
3. Operaio	6. Studente	9. Disoccupato

**7. DURATA DELLA PERMANENZA NELL'AREA DI SAN FRANCISCO /OAKLAND/SAN JOSE?**

1. Meno di un anno	3. Da quattro a sei anni	5. Più di dieci anni
2. Da uno a tre anni	4. Da sette a dieci anni	

**8. SECONDO REDDITO (SE SPOSATO)**

1. Non sposato	3. No
2. Sì	

**9. COMPONENTI DELLA FAMIGLIA**

1. Uno	4. Quattro	7. Sette
2. Due	5. Cinque	8. Otto
3. Tre	6. Sei	9. Nove o più

**10. COMPONENTI DELLA FAMIGLIA SOTTO I 18 ANNI**

1. Uno	4. Quattro	7. Sette
2. Due	5. Cinque	8. Otto
3. Tre	6. Sei	9. Nove o più

**11. STATO DELL'ALLOGGIO**

1. Di proprietà	3. Presso i genitori/famiglia
2. In affitto	

**12. TIPO DI ALLOGGIO**

1. Casa	3. Appartamento	5. Altro
2. Condominio	4. Casa Mobile	

**13. ETNIA**

1. Indiano-Americana	4. Asiatico-Indiana	7. Bianca
----------------------	---------------------	-----------

2. Asiatica	5. Ispanica	8. Altro
3. Di Colore	6. Pacifica	

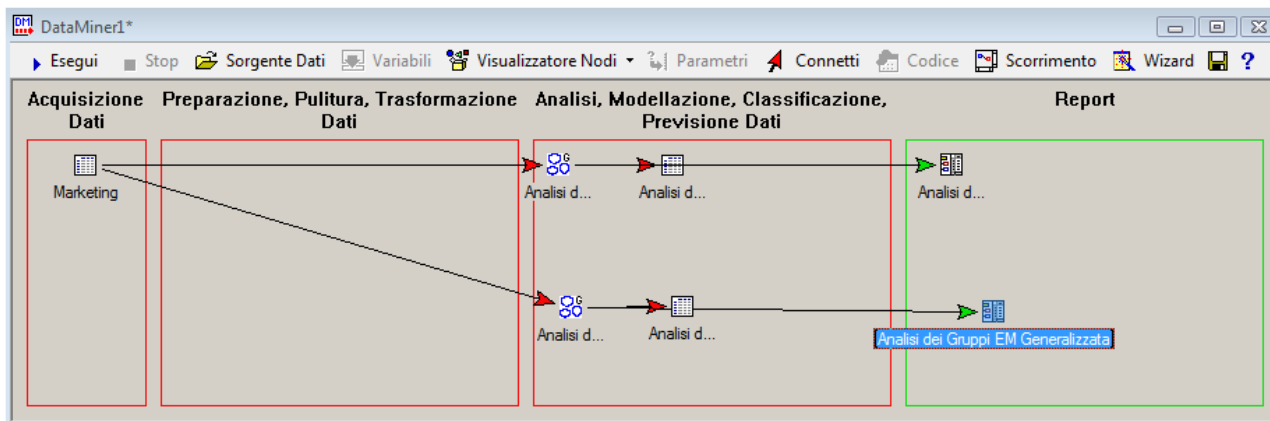
14 LINGUA PARLATA PIU' SPESSO IN FAMIGLIA

1. Inglese	2. Spagnolo	3. Altro
------------	-------------	----------

## Analisi dei Dati con STATISTICA

La preparazione dei dati è un passaggio cruciale di ogni progetto di data mining. In questo caso, tuttavia, la preparazione dei dati è già stata completata. Si è infatti già verificata l'accuratezza e la ragionevolezza dei dati; vi sono dati mancanti, ecc.

La seguente illustrazione rappresenta lo spazio di lavoro di *STATISTICA Data Miner*, in cui viene utilizzato lo strumento di Analisi dei Gruppi K-Means Generalizzata e di Analisi dei Gruppi EM Generalizzata per raggruppare i dati contenuti nell'insieme di dati *Marketing*. Una volta individuati i gruppi, lo strumento di Selezione delle Caratteristiche verrà impiegato per stabilire quale siano le variabili che più fortemente determinano l'appartenenza di gruppi.



## Risultati dell'Analisi dei Gruppi k-Means

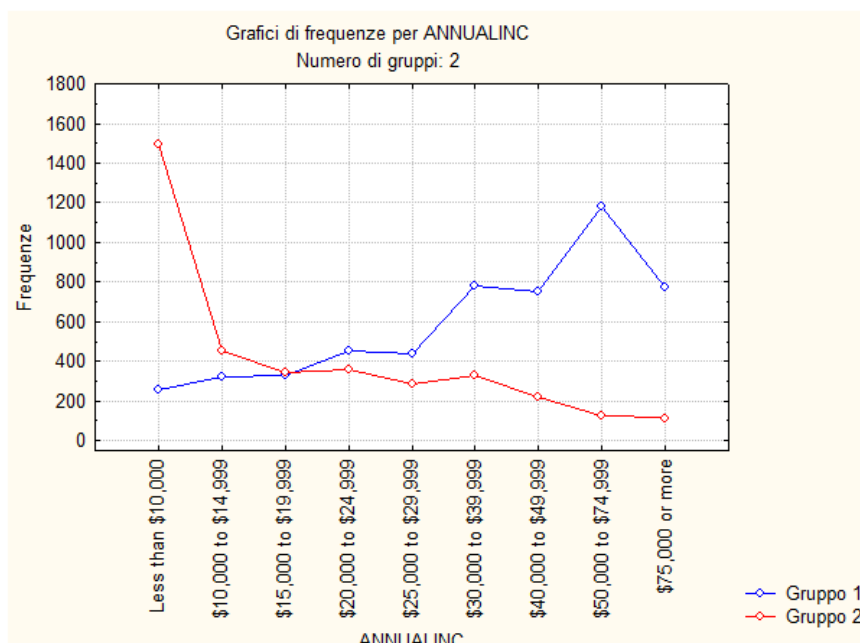
Nei risultati seguenti, sono riportati la *Classificazione Finale* e la *Distanza dal Centroide* quale output dell'Analisi dei Gruppi k-Means. Questo output può essere utilizzato per confrontare i casi appartenenti ai diversi gruppi previsti e la forza della rispettiva appartenenza di gruppo.

Dati: Elementi di gruppo (Marketing)\*

Elementi di gruppo (Marketing)  
 Numero di gruppi: 2  
 Numero totale casi di addestramento: 8993

Caso N.	Classificazione finale	ANNUALINC	SEX	MARSTATUS	Distanza dal centroide	AGE	ET
1	1	\$75,000 or more	Female	Married	2,449490	45 thru 54	1 to :
2	1	\$75,000 or more	Male	Married	2,828427	45 thru 54	
3	1	\$75,000 or more	Female	Married	2,449490	25 thru 34	
4	2	Less than \$10,000	Female	Single, never married	2,645751	14 thru 17	
5	2	Less than \$10,000	Female	Single, never married	2,828427	14 thru 17	
6	1	\$50,000 to \$74,999	Male	Married	2,000000	55 thru 64	1 to :
7	2	Less than \$10,000	Male	Single, never married	1,732051	18 thru 24	Gradi
8	2	\$30,000 to \$39,999	Male	Divorced or separated	2,236068	25 thru 34	1 to :
9	2	\$10,000 to \$14,999	Male	Married	2,449490	55 thru 64	Gradi
10	1	\$20,000 to \$24,999	Male	Married	2,828427	65 and Over	1 to :

Secondo l’algoritmo *k*-Means una variabile importante nella determinazione dei gruppi è il Reddito Annuale (Annual Income). Nel grafico successivo sono riportate le frequenze relative ad ogni categoria di reddito per i due gruppi. I soggetti caratterizzati da un reddito inferiore sono più spesso classificati come appartenenti al Gruppo 2, mentre le persone con un reddito più alto rientrano nel Gruppo 1.



## Risultati dell’Analisi dei Gruppi EM

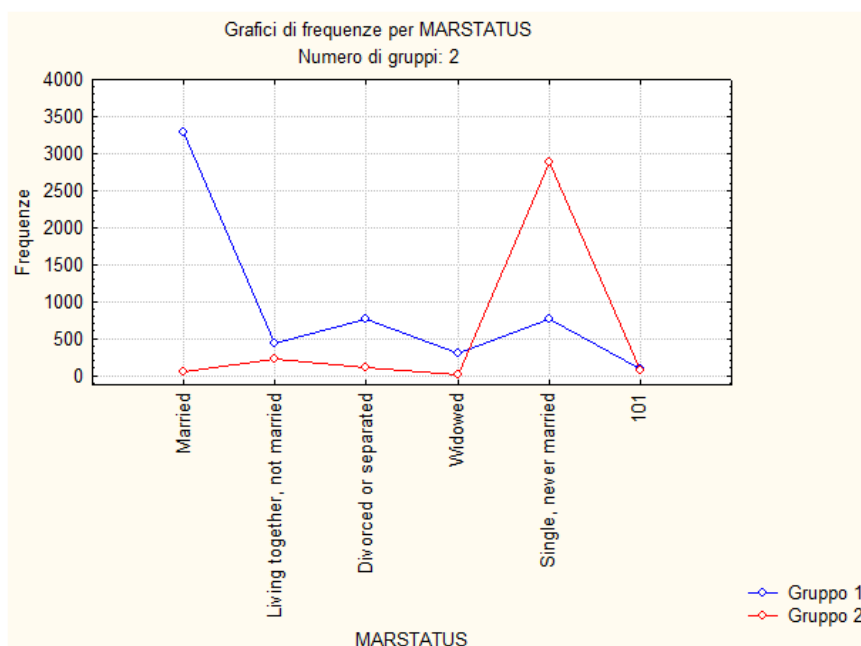
L’output dell’Analisi dei Gruppi EM sarà piuttosto simile a quello ottenuto tramite l’Analisi dei Gruppi *k*-Means. Assieme alla classificazione finale dei gruppi, l’Analisi dei Gruppi EM riporta le probabilità di appartenenza ai gruppi. Nello spreadsheet di output riportato sotto sono riportati i primi casi del dataset.

Dati: Probabilità di classificazione (pesi) per raggruppamento EM (Marketing)\*

Probabilità di classificazione (pesi) per raggruppamento EM (Marketing)  
 Numero di gruppi: 2  
 Numero totale casi di addestramento: 8993

	Gruppo 1	Gruppo 2	Classificazione finale
1	1,000000	0,000000	1
2	1,000000	0,000000	1
3	0,999993	0,000007	1
4	0,000000	1,000000	2
5	0,000000	1,000000	2
6	1,000000	0,000000	1
7	0,000134	0,999866	2
8	0,953011	0,046989	1
9	0,999997	0,000003	1
10	1,000000	0,000000	1

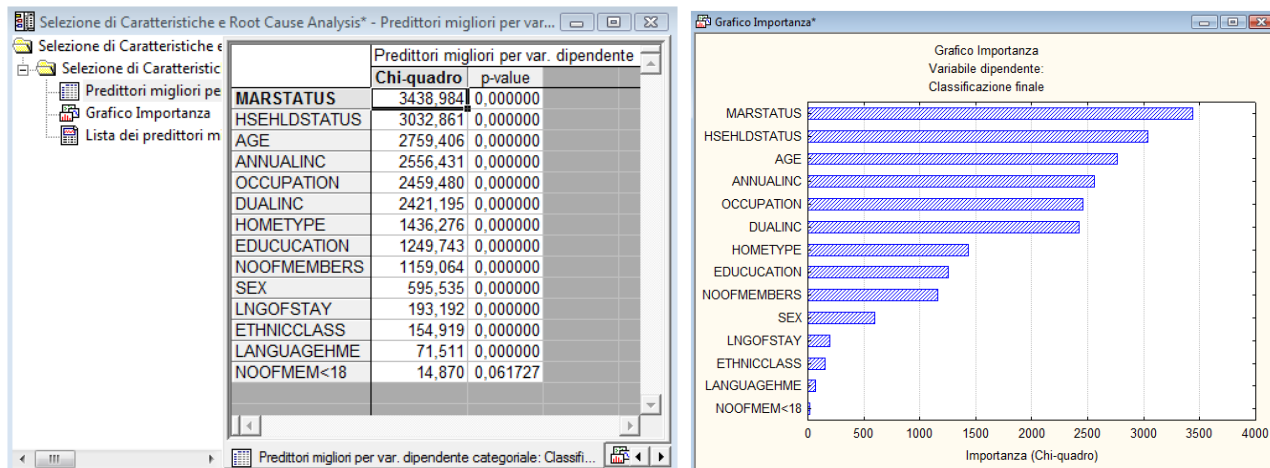
Come per i risultati *k*-Means, è possibile ad esempio analizzare le frequenze delle osservazioni classificate nei 2 gruppi in base allo Stato Civile (Marital Status). Le persone sposate rientrano nel Gruppo 1. I single e le persone che non si sono mai sposate ricadono più frequentemente nel Gruppo 2. Dal seguente grafico, è facile osservare come lo Stato Civile sia una variabile importante per la determinazione dell'appartenenza di classe.



## Risultati Selezione di Caratteristiche e Screening di Variabili

Una volta individuate le assegnazioni di gruppo nel nostro dataset, sarà di grande importanza trovare (generalmente in presenza di un gran numero di variabili) le variabili predittrici più fortemente legate alla variabile di classificazione finale

Nello spreadsheet sono elencate in modo abbastanza chiaro le variabili “Migliori Predittori Categoriali” dalla più importante alla meno importante, sulla base del rispettivo valore di chi-quadro. L’istogramma, che riporta la stessa informazione in formato grafico, indica chiaramente come la variabile MARITALSTAT sia il predittore più influente nel processo di raggruppamento, seguita da HSEHLDSTATUS, AGE. ecc.



## Conclusioni

Riassumendo i risultati delle analisi dei migliori predittori (estratti attraverso l’uso dello strumento di Selezione delle Caratteristiche) in una semplice tabella, avremo quanto segue. Si noti che tale tabella indica la rispettiva classe o categoria “dominante” per ognuna delle variabili riportate sotto, ovvero sia la classe o la categoria che viene osservata con la maggiore frequenza all’interno del rispettivo gruppo.

	Gruppo 1	Gruppo 2
<b>Stato Civile</b>	Sposato/a	Single, Mai Sposato/a
<b>Proprietà Abitazione</b>	Propria	Affitto
<b>Età</b>	25 - 34	18 - 24
<b>Reddito Annuo</b>	\$50,000 - \$74,999	Meno di \$10,000
<b>Occupazione</b>	Professionista/Manager	Studente

L’informazione riportata sopra indica chiaramente come la tecniche di raggruppamento abbia consentito ad identificare due gruppi significativi distinti di acquirenti.

Tale informazione potrebbe essere sfruttata per meglio soddisfare i bisogni dei clienti del proprio esercizio, e quindi per migliorare le vendite e per individuare la via da seguire per il conseguimento di una più alta profittabilità. Per esempio, sulla base di una migliore comprensione della tipologia di clienti, sarà possibile progettare speciali campagne di marketing (promozioni speciali per gli studenti, ecc.); tale informazioni potrebbero essere usate per raccogliere in particolari zone del proprio esercizio insiemi omogenei di beni che possano soddisfare un gruppo ben caratterizzato di clienti. In genere, meglio si conosce e si “comprende” la propria clientela, meglio l’ esercente sarà preparato a servirla, e quindi ad assicurare un successo d’impresa.