Analisi Esplorativa dei Dati

MISURE DI ETEROGENEITÀ

Dato un carattere qualitativo rilevato su scala nominale con k modalità diverse

Riassumendo

X	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
x_1	n_1	f_1	p_1
x_2	n_2	f_2	p_2
x_j	n_j	f_j	p_j
x_k	n_k	f_k	p_k
Totale	N	1	100%

Si definisce eterogeneità la diversificazione esistente fra le modalità di un carattere qualitativo.

Casi estremi

1. eterogeneità nulla ovvero max omogeneità

Tutte le unità del collettivo sono concentrate in una modalità

f_j =1 per un certo j

f_j =0 per ogni altro j

Si definisce eterogeneità la diversificazione esistente fra le modalità di un carattere qualitativo.

Casi estremi

2. eterogeneità massima ovvero minima omogeneità

f_j =1/k per j=1,2,…,k

Gli indici di eterogeneità permettono di valutare dove si posiziona una distribuzione di frequenze rispetto ai casi estremi.

Indice di eterogeneità di Gini

In generale, l'indice di eterogeneità di Gini è così definito:

$$G = 1 - \sum_{j=1}^{k} f_j^2$$

dove f_j è la frequenza relativa dell'j-esima modalità di una variabile qualitativa con k modalità. G assume tutti i valori compresi tra zero (massima omogeneità) e (k-1)/k (massima eterogeneità).

Indice relativo

$G' = G/G_{max}$

Indice di eterogeneità di Gini

In generale, l'indice di eterogeneità di Gini è così definito:

$$G = 1 - \sum_{j=1}^{k} f_j^2$$

G assume tutti i valori compresi tra zero (massima omogeneità) e (k-1)/k (massima eterogeneità).

G_{max}=$1 - \Sigma 1/k^2$=1-1/k= (k-1)/k

G_{min}=$1 - 1 = 0$

Indice relativo

G'=G/G_{max} = G/(k-1)/k = G k/(k-1)

Indice di eterogeneità Entropia di Shannon

$$H = \sum_{j=1}^{k} f_j log\ 1/f_j = -\sum_{j=1}^{k} f_j log\ f_j$$

dove f_j è la frequenza relativa dell'j-esima modalità di una variabile qualitativa con k modalità. H assume tutti i valori compresi tra zero (massima omogeneità) e log k (massima eterogeneità).

Indice relativo

$H' = H/H_{max}$

Indice di eterogeneità Entropia di Shannon

$H = \sum_{j=1}^{k} f_j log\ 1/f_j = -\sum_{j=1}^{k} f_j log\ f_j$

H assume tutti i valori compresi tra zero (massima omogeneità) e log k (massima eterogeneità).(logaritmo in base 2)

$H_{max} = -\Sigma 1/k \log 1/k = \Sigma 1/k \log k = k\ 1/k \log k$ OSS. (log 1/k)= -log k

$H_{min} = 0$

Indice relativo

$H' = H/H_{max} = H/\log k$

Esempio

Da un sondaggio condotto da un giornale sportivo in due regioni sul tifo per le principali squadre di calcio risulta

Piemonte		Lombardia	
squadra		squadra	
Inter	331	Torino	591
Milan	450	Juventus	721
Torino	675	Inter	2125
Juventus	2354	Milan	3374

In quale delle due regioni c'è più omogeneità?

$$G = 1 - \sum_{j=1}^{k} f^2_j$$

$$H = \sum_{j=1}^{k} f_j \log 1/f_j = -\sum_{j=1}^{k} f_j \log f_j$$

$G' = G/G_{max} = G/(k-1)/k = G\ k/(k-1)$

$H' = H/H_{max} = H/\log k$

piemonte	n_j	f_j	f_j^2	$\log f_j$	$f_j \log f_j$
Inter	331	0,086877	0,007548	-3,52489	-0,30623
Milan	450	0,11811	0,01395	-3,08179	-0,36399
Torino	675	0,177165	0,031388	-2,49683	-0,44235
Juventus	2354	0,617848	0,381736	-0,69468	-0,4292
	3810	1	0,434621		-1,54178

G	0,56537		H	1,54177
Grel	0,75383		Hrel	0,77088

lombardia	n_j	f_j	f_j^2	$\log f_j$	$f_j \log f_j$
Torino	591	0,086771	0,007529	-3,52664	-0,30601
Juventus	721	0,105858	0,011206	-3,2398	-0,34296
Inter	2125	0,311995	0,097341	-1,6804	-0,52428
Milan	3374	0,495375	0,245397	-1,01341	-0,50202
	6811	1	0,361473		-1,67526
G	0,638527		H	1,675265	
Grel	0,85137		Hrel	0,837632	

In quale delle due regioni c’è più omogeneità?

Piemonte

G	0,56537		H	1,54177
Grel	0,75383		Hrel	0,77088

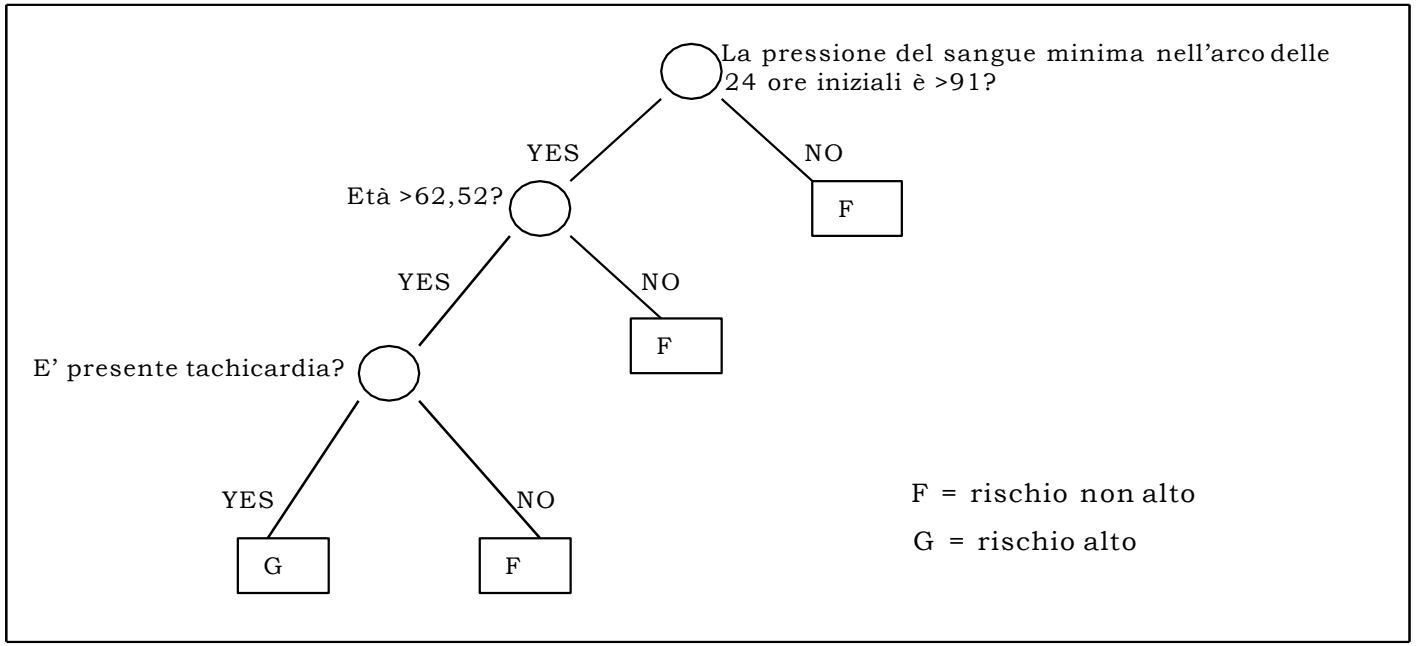
Lombardia

G	0,638527		H	1,675265
Grel	0,85137		Hrel	0,837632

Esempio Alberi Decisionali

Ormai da diversi anni, al “San Diego Medical Center” dell’Università della California, quando arriva un paziente con problemi di cuore, vengono misurate nelle prime ventiquattro ore di degenza diciannove variabili tra cui la pressione del sangue, l’età, il ritmo cardiaco ed altre sedici sia binarie che ordinali.

Questo permette di ottenere la storia medica e la situazione fisiologica dei pazienti in modo da identificare quelli ad alto rischio, ossia che non sopravvivranno ad un minimo di trenta giorni, sulla base dei dati delle ventiquattro ore iniziali. Per fare ciò, viene costruita una struttura ad albero che viene usata come regola decisionale per classificare ogni paziente (Breiman L. *et al.*, 1984).



Esempio di albero decisionale usato in ambito medico.

(Fonte: Breiman L. *et al.*, 1984)

Un altro campo in cui frequentemente comparivano, oltre a quello medico in genere, era la botanica, che li utilizzava per classificare i diversi tipi di specie di piante in base a determinate caratteristiche.

Un importante uso degli alberi decisionali è quello fatto dagli enti assicurativi e di credito perché essi permettono di valutare il rischio potenziale di un credito e quindi il suo grado di solvibilità a seconda delle caratteristiche del cliente/assicurato. Si parla in questo caso di Credit Scoring.

Un semplice esempio di Credit Scoring è rappresentato nella seguente tabella (esempio ripreso da Zani Cerioli 2007)

Es. alberi di classificazione

Matrice dei dati 8 clienti di un Istituto di Credito con il corrispondente rischio di credito

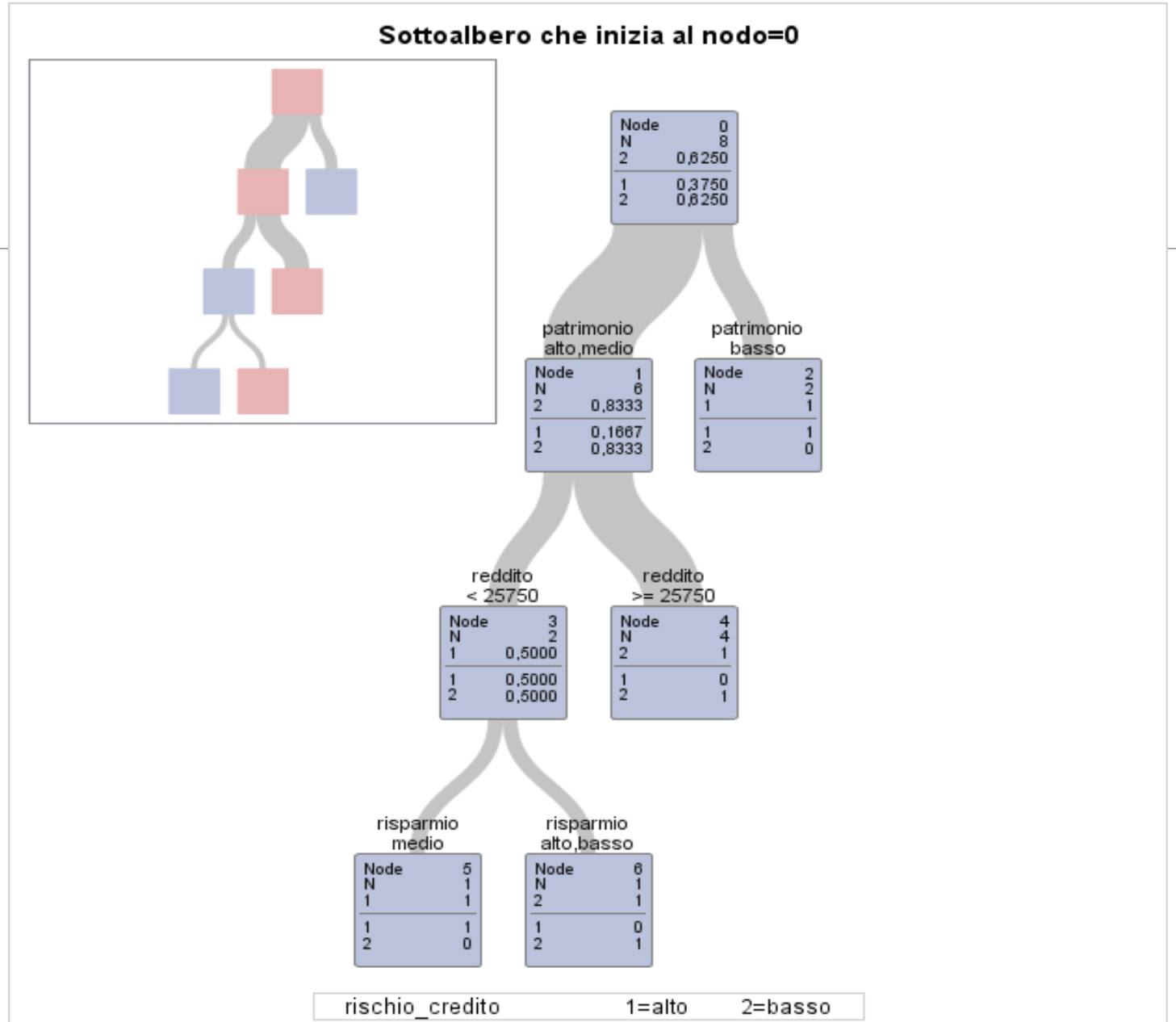
cliente	risparmio	patrimonio	reddito	rischio_credito	
A	medio	alto	75000	basso	
B	basso	basso	50000	alto	
C	alto	medio	25000	alto	
D	medio	medio	50000	basso	
E	basso	medio	100000	basso	
F	alto	alto	25000	basso	
G	basso	basso	25000	alto	
H	medio	medio	75000	basso	

Oss. Ciascun cliente è classificato sulla base dell'esito del finanziamento ricevuto: basso = solvente alto = insolvente

OBIETTIVO: Costruire una regola di decisione che, in base ai dati disponibili, consenta di assegnare un nuovo cliente ad una delle due classi.

SAS ALBERI (Gini)

```
 model rischio_credito = risparmio patrimonio reddito;
    grow Gini;
    prune off;
    output out=sasuser.alberi_scoring1; /* salvo l'output */
run;
```


Sottoalbero che inizia al nodo=0
Node 0
N 8
2 0,6250
1 0,3750
2 0,6250
patrimonio alto,medio
patrimonio basso
Node 1
N 6
2 0,8333
1 0,1667
2 0,8333
Node 2
N 2
1 1
1 1
2 0
reddito < 25750
reddito >= 25750
Node 3
N 2
1 0,5000
1 0,5000
2 0,5000
Node 4
N 4
2 1
1 0
2 1
risparmio medio
risparmio alto,basso
Node 5
N 1
1 1
1 1
2 0
Node 6
N 1
2 1
1 0
2 1
rischio_credito 1=alto 2=basso

Indice di eterogeneità di Gini

In generale, l'indice di eterogeneità di Gini è così definito:

$$G = 1 - \sum_{i=1}^{r} f_i^2$$

dove è la frequenza relativa dell'i-esima modalità di una variabile qualitativa con r modalità. G assume tutti i valori compresi tra zero (massima omogeneità) e (massima eterogeneità).

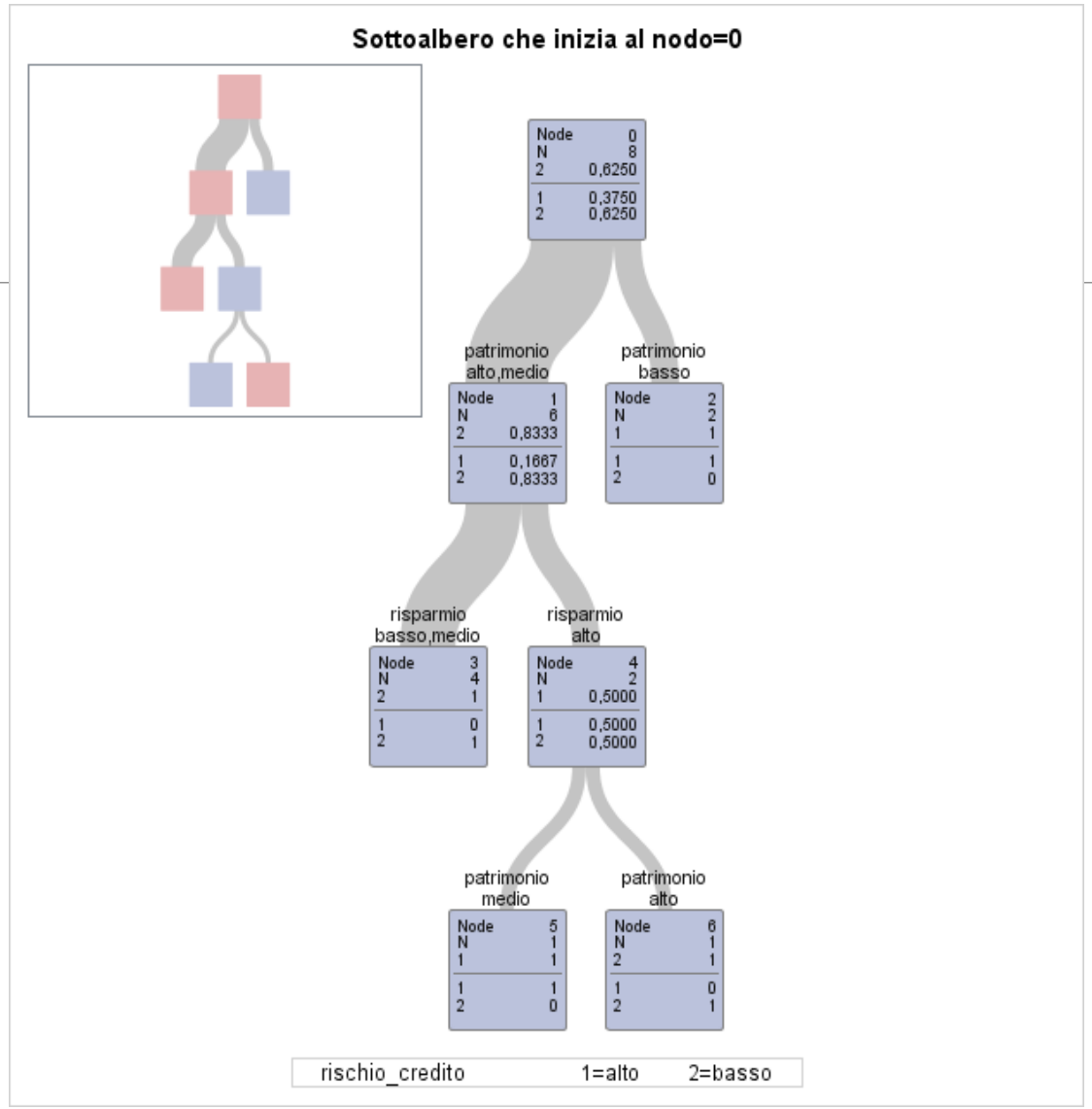
Ad es. considerando i dati da Zani Cerioli 2007,

nodo 0 G=0,468765

nodo 1 G=0 nodo 2 G=0,2782.

SAS ALBERI (Entropia)

```
proc hpsplit data=sasuser.scoring maxdepth=4;
class rischio_credito risparmio patrimonio; /* var qualitative */
   model rischio_credito = risparmio patrimonio reddito;
   prune off; /* off o none? */
output out=sasuser.alberi_scoring;
run;
```


Sottoalbero che inizia al nodo=0
Node 0
N 8
2 0,6250
1 0,3750
2 0,6250
patrimonio alto,medio
patrimonio basso
Node 1
N 6
2 0,8333
1 0,1667
2 0,8333
Node 2
N 2
1 1
1 1
2 0
risparmio basso,medio
risparmio alto
Node 3
N 4
2 1
1 0
2 1
Node 4
N 2
1 0,5000
1 0,5000
2 0,5000
patrimonio medio
patrimonio alto
Node 5
N 1
1 1
1 1
2 0
Node 6
N 1
2 1
1 0
2 1
rischio_credito 1=alto 2=basso