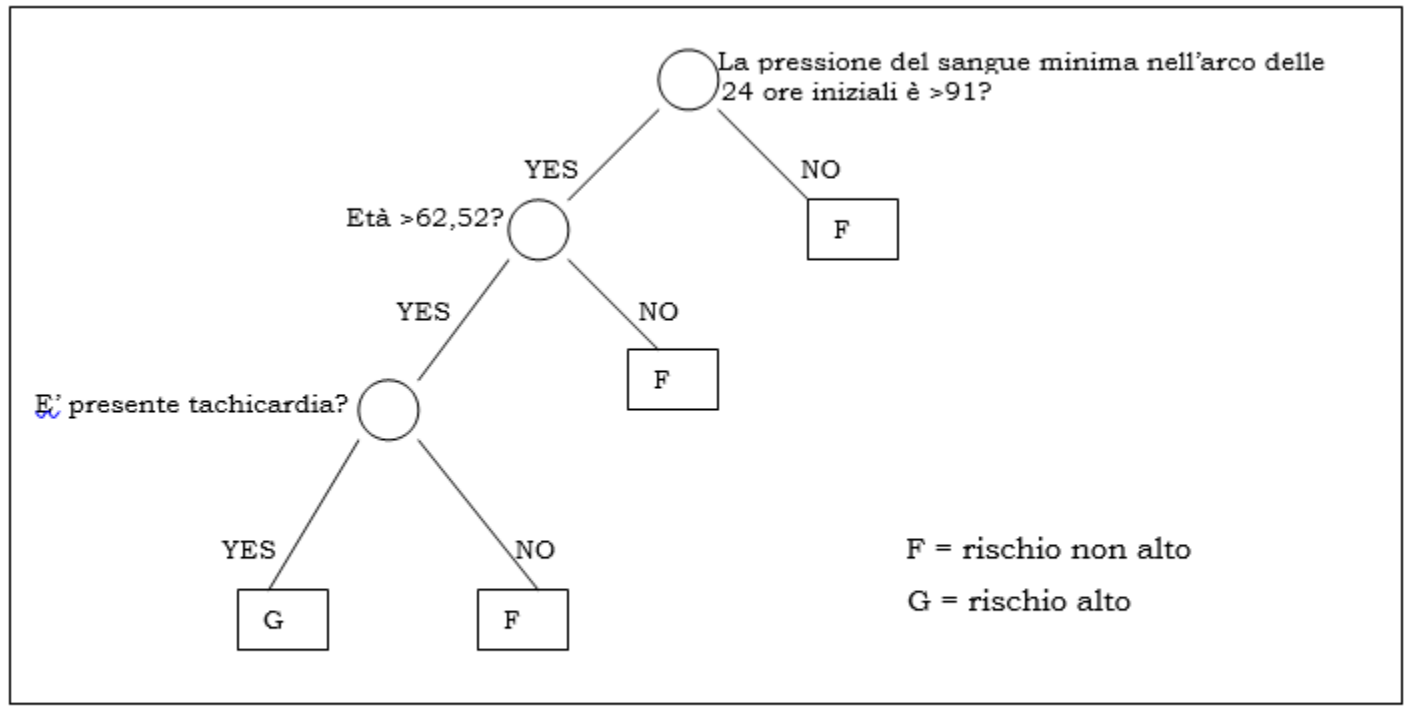
GLI ALBERI DECISIONALI

Sviluppi nell'uso degli 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 venivano utilizzati, oltre a quello medico in genere, era la botanica, che li usava per classificare i diversi tipi di specie di piante in base a determinate caratteristiche

- Gli alberi decisionali fanno parte delle tecniche di classificazione o segmentazione gerarchica
- Tali tecniche hanno lo scopo di “smistare” unità statistiche di elevata numerosità nelle varie classi di una variabile dipendente in base ai valori di una o più variabili esplicative
- Viene individuata una regola che, graficamente, ha una struttura ad albero e che viene successivamente usata per classificare nuove unità non rientranti tra quelle usate inizialmente per la sua creazione.

- L’input in un problema di classificazione è un dataset di *training records*, chiamato *training dataset*.
- In ciascun record sono presenti diversi attributi: quelli il cui dominio è numerico sono chiamati *numerical attributes* gli altri sono detti *categorical attributes*.
- Uno di questi attributi è la variabile dipendente gli altri sono i predittori.
- **Scopo degli alberi decisionali** è costruire un conciso modello di distribuzione dell’attributo dipendente in base ai predittori. Il modello risultante viene poi usato per classificare valori di un database in cui il valore della variabile dipendente è noto ma quello delle variabili predittive no.

OSSERVAZIONE

Questa tecnica può sembrare molto simile alla *Cluster Analysis*, poiché anche quest'ultima fornisce come risultato una partizione delle unità statistiche, ma ci sono delle sostanziali differenze.

- Una tecnica di segmentazione richiede la conoscenza a priori della classe di appartenenza delle unità, mentre la *Cluster Analysis* costruisce i gruppi a partire da un insieme indistinto. Gli alberi decisionali appartengono quindi alle tecniche di classificazione supervisionata mentre la cluster analysis a quelle non supervisionate.
- Per fare la segmentazione si utilizza una sola variabile selezionata tra tutte quelle a disposizione, mentre la formazione dei *cluster* avviene per mezzo di alcune misure di distanza fra le unità calcolate usando tutte le variabili a disposizione.
- La regola di classificazione ottenuta attraverso la segmentazione viene utilizzata a fini previsionali su nuove unità statistiche e questo non è previsto nella *Cluster Analysis*.

La tecnica degli alberi decisionali è particolarmente interessante per il Data Mining per diversi motivi.

- Primo, perché, grazie ad una rappresentazione intuitiva, lo schema di classificazione generato è facile da interpretare e capire.
- Secondo, perché sono dei modelli non parametrici e, quindi, sono particolarmente indicati per la *Exploratory Knowledge Discovery*.
- Terzo, perché possono essere costruiti piuttosto velocemente
- Quarto perché hanno una buona accuratezza nelle regole di classificazione generate.

Alcuni esempi di utilizzo degli alberi decisionali

Da un mero utilizzo a fini grafico-rappresentativi fatto all'inizio da medici e botanici, si è passati ad un pesante ricorso agli alberi decisionali come strumento di classificazione ma anche di previsione e di supporto alle decisioni fatto in numerosissimi campi.

- Nel Marketing, per esempio, essi sono diventati strumenti indispensabili per costruire profili dei consumatori, degli utenti, dei clienti e per progettare i servizi forniti sulla base dei bisogni del pubblico; inoltre, vengono usati per scoprire quali gruppi risponderanno favorevolmente alle offerte promozionali e per capire come raggiungere più facilmente gli utenti.
- Nella programmazione aziendale sono adottati allo scopo di individuare gli elementi cruciali da inserire nella programmazione dell'assistenza sanitaria, sociale e di servizio.
- Nel controllo di gestione servono per analizzare le relazioni esistenti tra centri di costo e fattori.
- Etc…..

- Altro 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)

- 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

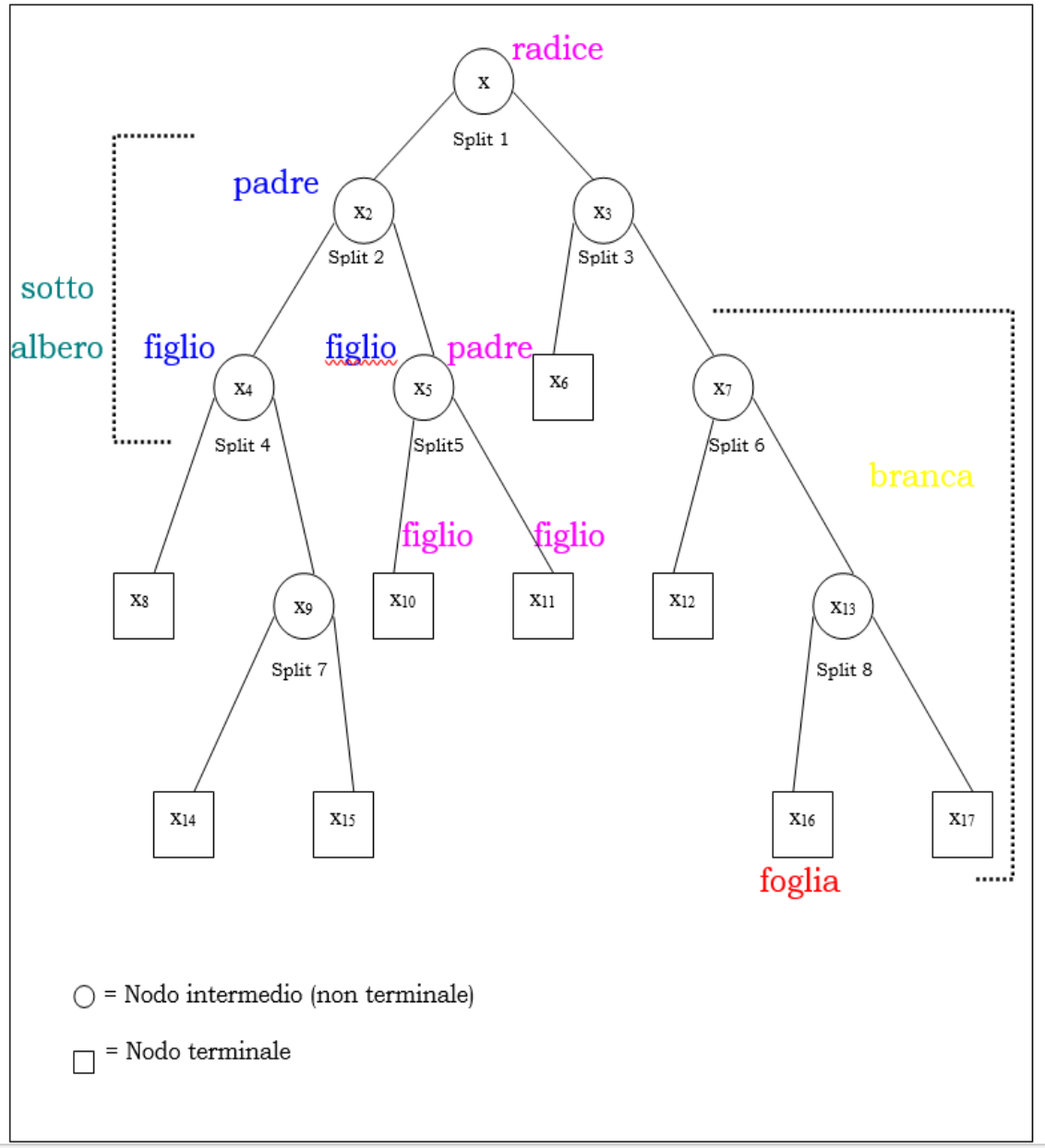
- OBIETTIVO: Costruire una regola di decisione che, in base ai dati disponibili, consenta di assegnare un nuovo cliente ad una delle due classi.

Dal punto di vista formale e grafico, un albero è formato da un insieme finito di elementi detti “nodi”.

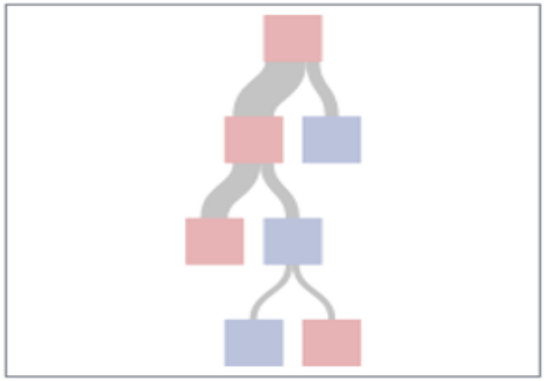
- Quello principale, ossia il primo, è detto “radice”, talvolta viene indicato con il termine di “nodo 0” o con la lettera R; da esso si diramano i successivi nodi. In genere, la radice è posizionata in alto nella rappresentazione grafica, quindi l’albero si sviluppa in discesa.
- L’insieme dei nodi, ad esclusione della radice, può essere diviso in *h* insiemi distinti $S_1,\ldots S_h$ detti “sottoalberi” del nodo radice.

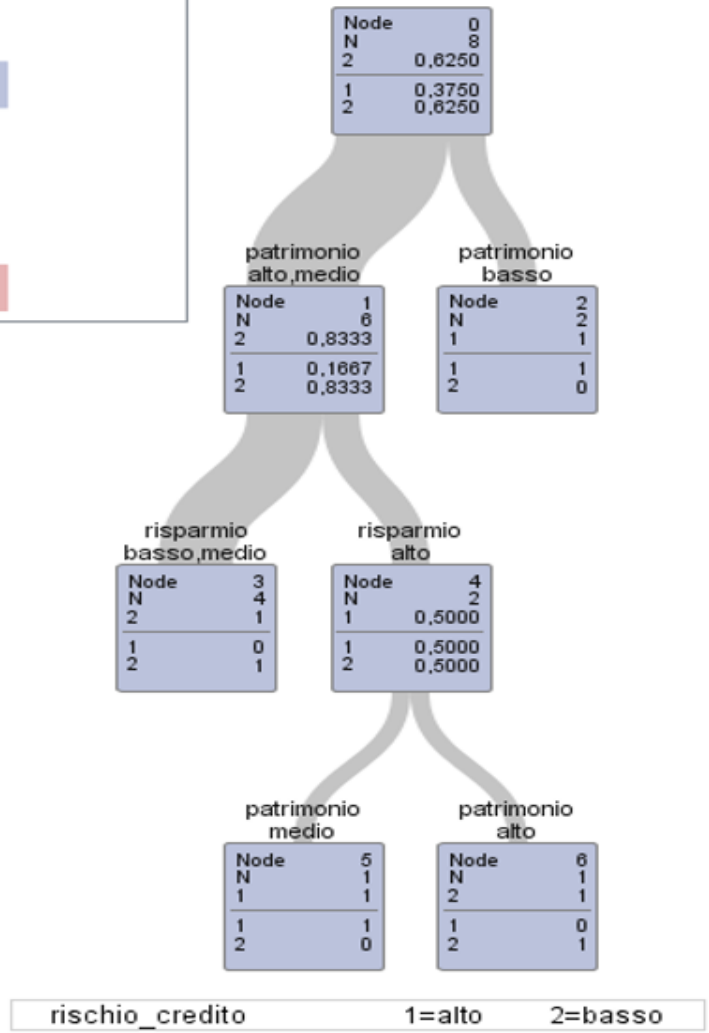
L’insieme dei nodi discendenti da un determinato nodo intermedio è chiamato “branca” o “ramo”.

- Un nodo può essere “padre” se genera dei nodi che stanno sotto di lui ed essi sono detti “figli”. Un nodo padre può avere più figli ma non viceversa.
- I nodi terminali, ossia quelli senza diramazioni, sono denominati “foglie”.
- Per ottenere le diramazioni vengono usati dei valori soglia di una variabile detti “split”, che dividono le unità di un determinato nodo padre nei suoi due o più nodi figli.



Logica di costruzione di un albero decisionale





Alberi di classificazione e di regressione

Ci sono due tipi di variabili o attributi: numeriche e categoriche quindi anche la variabile dipendente su cui si basa la costruzione di un albero essere di due tipi e, di conseguenza, si avranno due tipi di alberi decisionali:

- se la variabile dipendente è di tipo categorico o qualitativo si parla di “alberi di classificazione”;
- se, invece, essa è di tipo numerico o quantitativo si parla di “alberi di regressione”.

Es. alberi di classificazione

La variabile dipendente in questo caso è: rischio di credito

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

Logica di costruzione di un albero decisionale: Definizioni e procedura di classificazione

- La variabile dipendente (target) viene indicata con Y. Essa presenta J modalità, se qualitativa, oppure è suddivisa in J classi, se quantitativa.
- Oltre alla variabile dipendente o target si considerano *p* variabili esplicative o predittive $X_1,...,X_p$, sia qualitative che quantitative, rilevate sulle *n* unità statistiche che compongono il training database. Si indichi con $\mathbf{x}_i=[x_{i1},...,x_{is},...,x_{ip}]$ il vettore che contiene le informazioni sulle variabili esplicative relative all'*i*-esima unità statistica (valori numerici per le variabili quantitative o codici per quelle qualitative).

- Sia X, sottospazio di R^p ($X \subseteq R^p$) lo spazio dei valori che possono assumere le *p* variabili esplicative, tale insieme è detto spazio degli attributi (feature space).
- Per creare una regola di classificazione si considera $d(\mathbf{x}),\ \mathbf{x} \in X,$ una regola che associa ad ogni elemento del sottospazio X (vale a dire ad ogni possibile vettore p dimensionale contenente le informazioni sulle variabili esplicative $X_1,\ldots,X_p,$) un numero intero compreso tra 1 e J:

$$d(\mathbf{x}): \mathbf{x} \Rightarrow \text{j con j} \in \{1,\ldots,J\}$$

- Definizione

Si indichino con A_j (j=1,....,J) gli elementi di una partizione del sottospazio X in J sottinsiemi.

Si definisce "regola di classificazione" una qualsiasi partizione di X in J sottoinsiemi $A_1,\ldots,A_J$ tale che per ogni $\mathbf{x} \in A_j$, la classe prevista è j; cioè

$$A_j = \{ x ; d(x)=j \}.$$

ESEMPIO

- Vogliamo classificare delle persone in base al Rischio di Credito (basso, alto) J=2

utilizzando l'albero della figura precedente,

poiché la variabile reddito annuo non compare nella costruzione dell'albero stesso (si veda la fig.) **x** è un vettore bidimensionale nello spazio degli attributi X (*p*=2 invece di *p*=3)

- La regola utilizzata è:

"alle foglie dell'albero è assegnata la classe corrispondente

alla modalità più frequente all'interno della foglia".

Secondo tale criterio le unità della prima foglia a destra (Nodo 2) sono assegnate

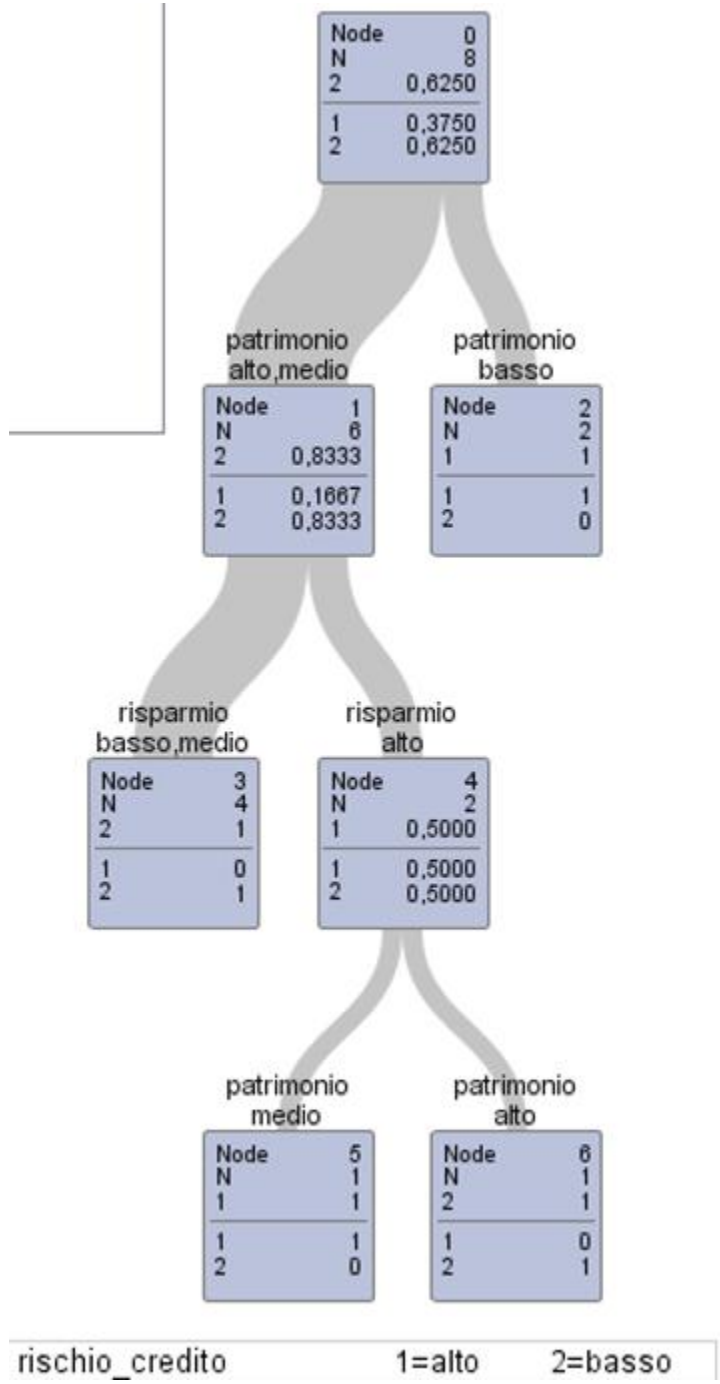
alla classe 1 mentre quelle appartenenti alla seconda foglia (Nodo 3)

sono assegnate alla classe 2.

Le altre due foglie discendono dalla suddivisione del nodo 4:

La foglia a destra (Nodo 6) è assegnata alla classe 2 mentre quella a sinistra

(Nodo 5) è assegnata alla Classe 1

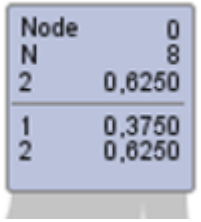


ESEMPIO in dettaglio

Vogliamo classificare delle persone in base al Rischio di Credito (basso, alto) J=2 utilizzando l'albero della figura precedente,

La regola utilizzata è:"alle foglie dell'albero è assegnata la classe corrispondente alla modalità più frequente all'interno della foglia".

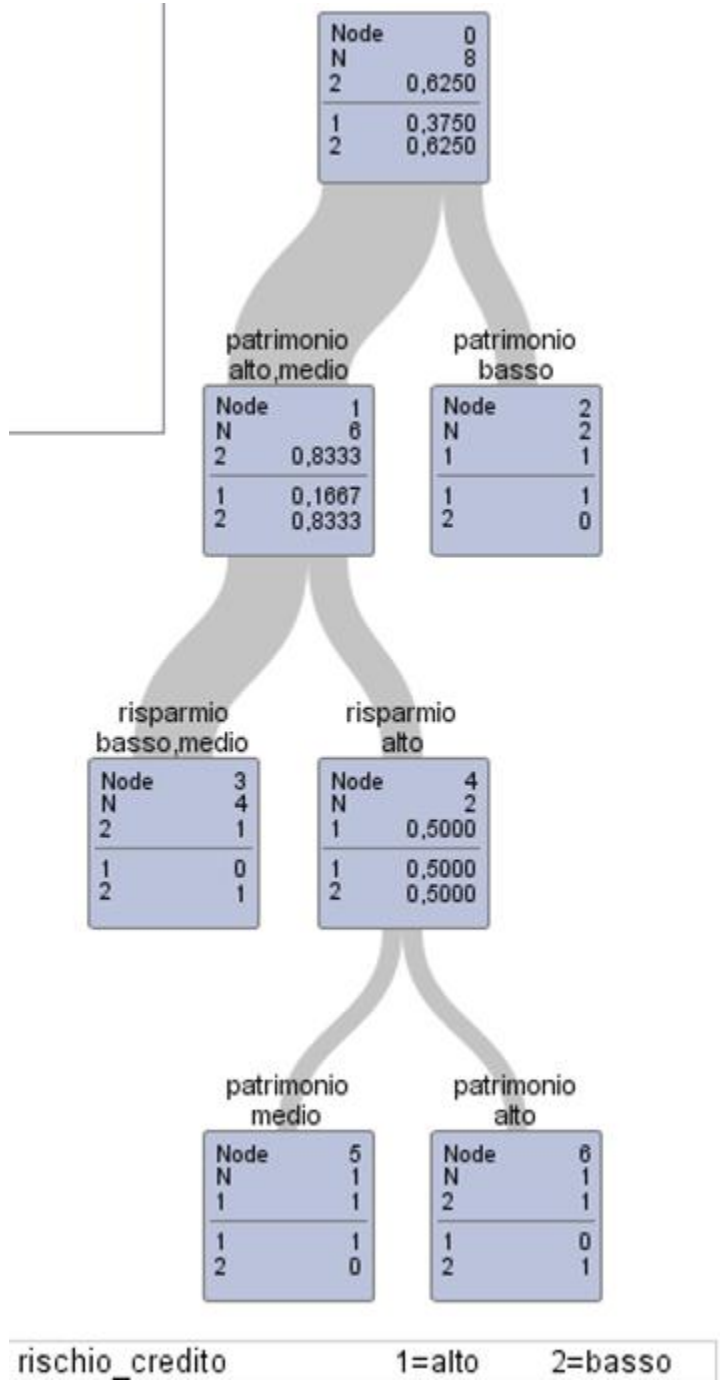
- Nodo 0 : 8 unità 5 in classe 2 (62.5 %) e 3 in classe 1 (37.5 %)



La regola decisionale assegna tutte le unità alla classe 2 .

La probabilità di errore è 0.375 .

- I due nodi figli della radice corrispondono alla classificazione riportata nella tabella di contingenza delle slide seguente



ESEMPIO in dettaglio

- I due nodi figli della radice Nodo 1 e Nodo 2 corrispondono alla classificazione riportata nella tabella di contingenza
- La regola assegna clienti con basso livello patrimoniale alla classe rischio alto (Nodo 2) (massima freq. nella prima riga della tabella) e clienti con livello patrimoniale

medio alto alla classe rischio basso (Nodo 1)

(massima freq. nella seconda riga della tabella)

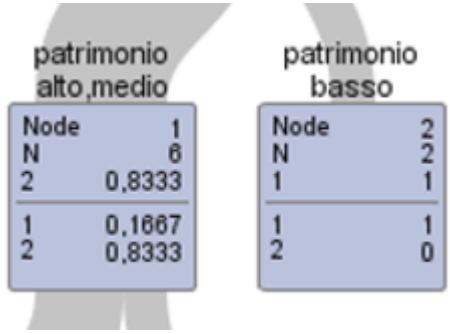
La probabilità di errore associata alla regola è

(1/6)+ (0/2) = 0.167.

Si migliora rispetto alla

situazione non strutturata

del nodo radice



La procedura FREQ

Frequenza Percentuale Pct riga Pct col	Tabella di patrimonio1 per rischio_credito		
	rischio_credito		
patrimonio1	alto	basso	Totale
basso	2 25.00 100.00 66.67	0 0.00 0.00 0.00	2 25.00
medal	1 12.50 16.67 33.33	5 62.50 83.33 100.00	6 75.00
Totale	3 37.50	5 62.50	8 100.00

ESEMPIO in dettaglio

- La previsione del rischio di credito del Nodo 2

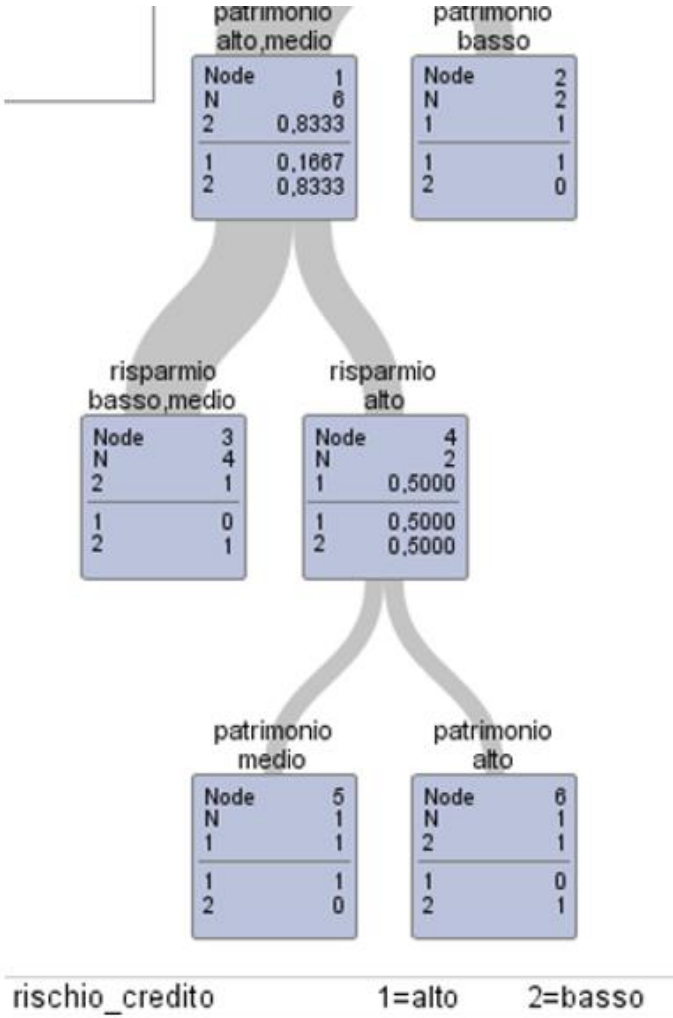
(formato solo da clienti con alto rischio di credito)
non si può migliorare Nodo 2 è Nodo terminale/ foglia.

- Il Nodo 1 contiene clienti appartenenti ad entrambe

le classi di rischio, si può quindi segmentare:

La regola assegna clienti con livello patrimoniale medio alto e risparmio medio basso alla classe rischio basso (Nodo 3).Essendo il Nodo 3 composto solo da clienti con basso rischio esso è Nodo terminale/ foglia.
I clienti del Nodo 4 con livello patrimoniale medio alto e risparmio alto vengono assegnati in ugual misura ad entrambe le classi, si deve segmentare:

La regola assegna clienti con risparmio alto e patrimonio medio alla classe 1 (Nodo 5) e clienti con risparmio alto e patrimonio alto alla classe 2 (Nodo 6), entrambi i nodi sono foglie.

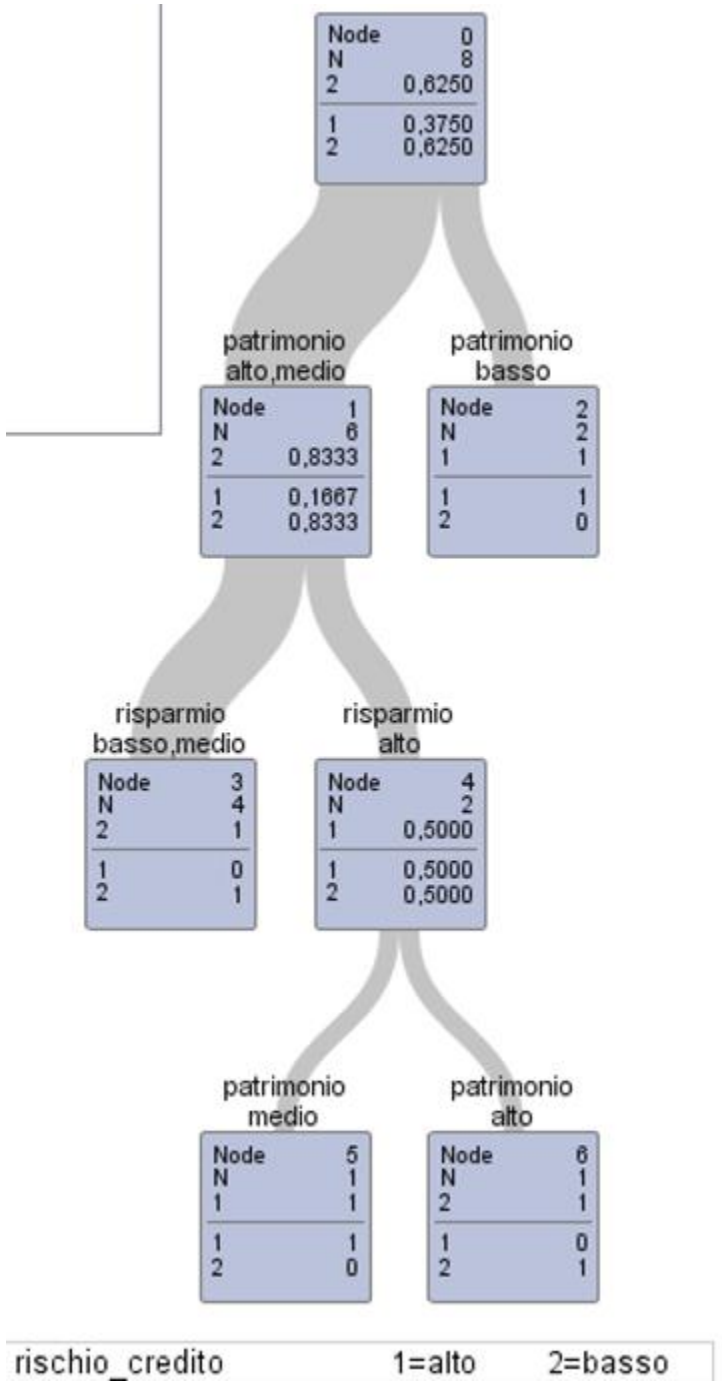


ESEMPIO

- A_2 ={**x**:*d*(**x**)=*2*}= { **x** =[patrimonio>basso, risparmio≤ medio]
 e **x** =[patrimonio>medio, risparmio> medio] }
- A_1 ={**x**:*d*(**x**)=*1*}= {patrimonio≤basso, risparmio qualunque]
e **x** =[patrimonio≤medio, risparmio> medio] }

Dove:
2= classe di clienti con basso rischio di credito
1= classe di clienti con alto rischio di credito

Riassumendo:
alto rischio di credito: nodo 2, nodo 5
basso rischio di credito: nodo 3, nodo 6



- Problema: instabilità di un albero decisionale cioè piccole variazioni nella matrice dei dati possono portare a differenze nella regola di classificazione. (Hastie, Tibshirani, and Friedman 2009; Kuhn and Johnson 2013).
- Per limitare questo problema si dovrebbe rinunciare a fare crescere l'albero fino al livello massimo possibile ma fissare un limite minimo e/ potando in modo opportuno l'albero di dimensione massima.

- Un approccio più sofisticato è il bagging (Hastie et al. 2001). Esso consiste nel ripetere più volte la procedura di segmentazione gerarchica su un campione casuale di n unità estratto con ripetizione dall'insieme di dati di partenza e nel calcolare la media aritmetica delle diverse previsioni ottenute per ciascun vettore **x.**
- Bagging, Random forest e boosting usano gli alberi come blocchi per costruire modelli più potenti. (PER APPROFONDIMENTI si veda ad es. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani: Introduzione all'Apprendimento Statistico con Applicazioni in R. Piccin (2021))

- Problema: instabilità di un albero decisionale cioè piccole variazioni nella matrice dei dati possono portare a differenze nella regola di classificazione.
- Matrice dei dati 8 clienti di un Istituto di Credito con il corrispondente rischio di credito
- MODIFICHIAMO DA alto A medio IL LIVELLO DI RISPARMIO DEL CLIENTE C

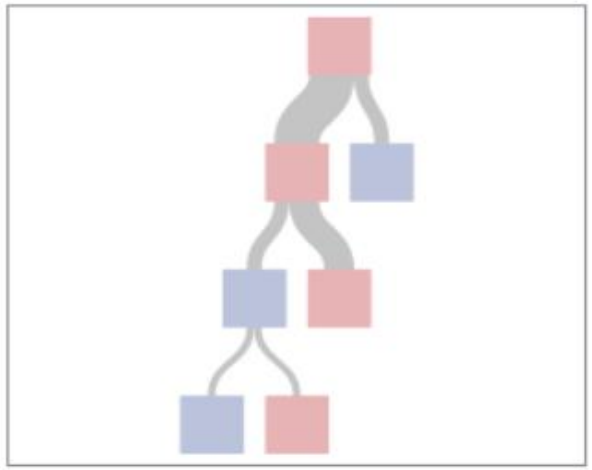
cliente	risparmio	patrimonio	reddito	rischio credito
A	medio	alto	75000	basso
B	medio	basso	50000	alto
C	alto	medio	25000	alto
D	medio	medio	50000	basso
E	basso	medio	100000	basso
F	alto	alto	25000	basso
G	basso	basso	25000	alto
H	medio	medio	75000	basso

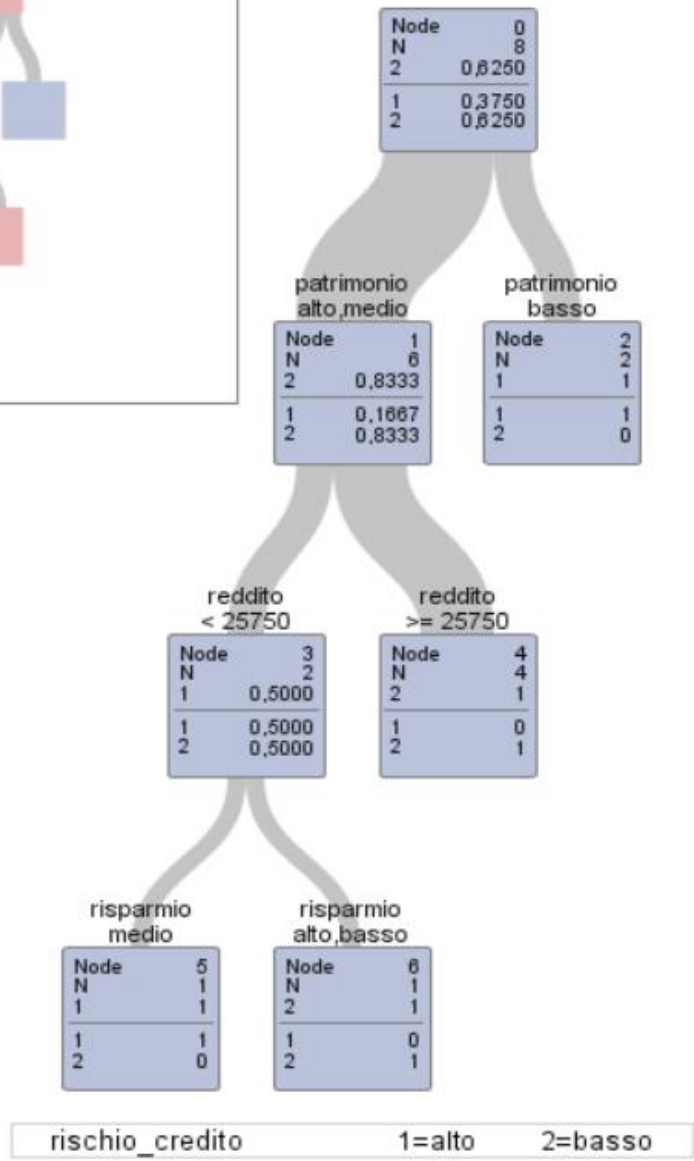
Oss. Ciascun cliente è classificato sulla base dell'esito del finanziamento ricevuto:

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.

Sottoalbero che inizia al nodo=0





Fasi di costruzione di un albero decisionale

- dicotomizzazione delle variabili esplicative;
- scelta del criterio di suddivisione di ogni nodo padre nei nodi figli;
- definizione di un criterio d'arresto nella costruzione dell'albero;
- scelta della regola per assegnare una delle J modalità della variabile dipendente ad ogni foglia;
- definizione, in base al passo precedente, della regola di classificazione $d(x)$ da applicare ai nuovi casi;
- stima del tasso di errata classificazione.

Riassumendo:

- Tramite il processo di segmentazione gerarchica usato nella costruzione degli alberi decisionali, le *n* unità statistiche vengono suddivise progressivamente, rispettando un criterio di ottimizzazione, in un numero finito di sottogruppi disgiunti tra loro in modo da garantire un'omogeneità interna superiore a quella dell'insieme iniziale ed una eterogeneità elevata tra i sottogruppi ma allo stesso tempo decrescente all'interno di ciascuno di essi ad ogni passo del processo.
- Una volta terminata la segmentazione, abbiamo a disposizione una regola che permette di classificare in qualsiasi momento altre unità non appartenenti al dataset originario e di cui non si conosce la classe di appartenenza.

Dicotomizzazione delle variabili esplicative

- Il punto principale nella creazione di una regola di classificazione sta nell'individuare la miglior partizione o suddivisione dello spazio X in base alle variabili esplicative al fine di prevedere quali siano le varie classi della variabile dipendente.
- Per fare ciò, bisogna prima ottenere tutte le possibili partizioni e poi scegliere quella ottimale.
- Le suddivisioni sono molteplici perché dipendono dalla natura dei predittori che possono essere sia quantitativi che qualitativi.

Le variabili esplicative possono essere:

- quantitative continue o discrete(fatturato, numero di dipendenti, ecc...);
- qualitative ordinali (giudizio di preferenza in una scelta che può essere “ottimo”, ”buono”, ”discreto”, ecc...)
- qualitative nominali (forma giuridica delle aziende, ecc...);
- qualitative dicotomiche (“paga mensile” e “paga settimanale”, ”commercio estero” e “commercio nazionale”, ecc...)

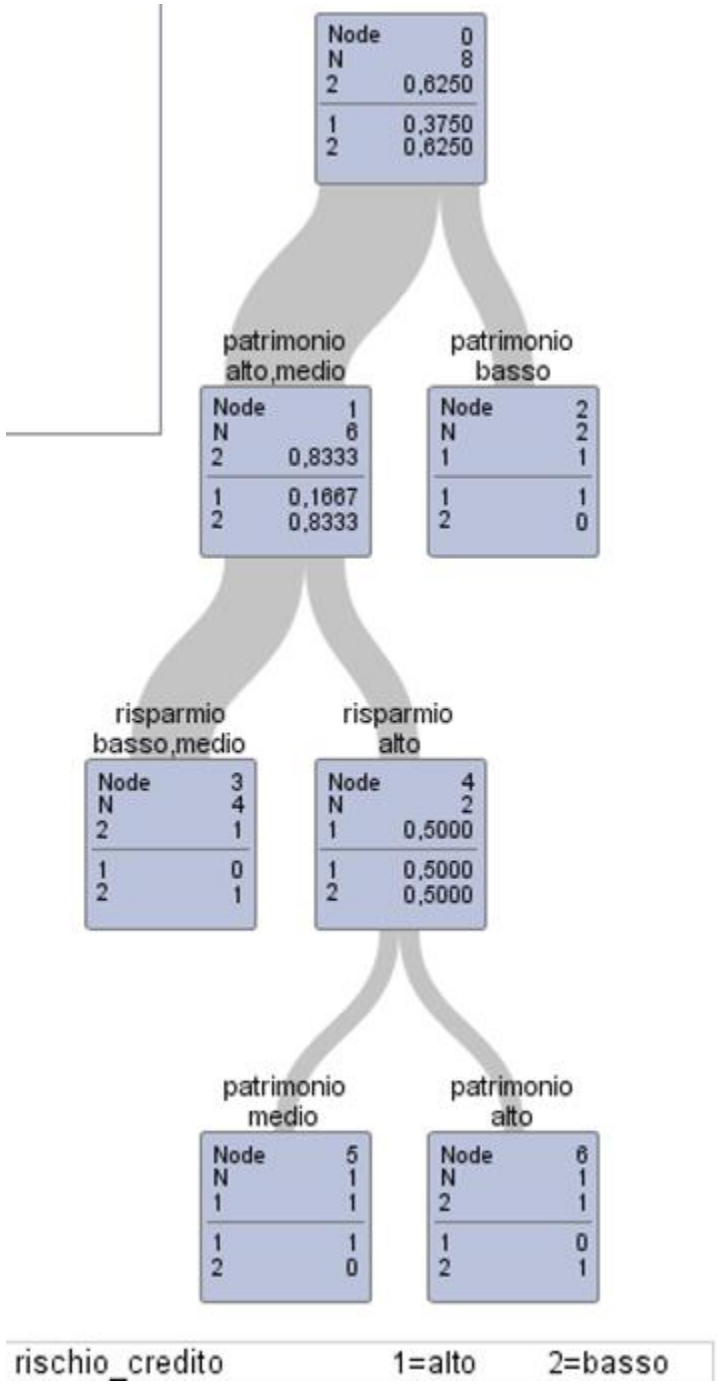
- Se una variabile esplicativa X_s è quantitativa, la dicotomizzazione delle unità statistiche viene attuata individuando un valore che svolga il ruolo di soglia di ripartizione; tramite questa soglia si creano due sottoinsiemi di cui uno contiene tutti i valori inferiori o uguali al valore considerato e l'altro tutti i valori superiori. Il valore soglia corrisponde a uno degli n valori distinti che la variabile Xs può assumere, ad esclusione dell'ultimo e viene scelto nella serie dei valori ordinati in senso non decrescente. Il numero di split possibili, in relazione a quella variabile, è pari a *n-1*.
- Se la variabile X_s è di tipo ordinale con *m* modalità, il numero di possibili suddivisioni in due gruppi (split) è m-1.
- Nel caso in cui la variabile esplicativa sia una variabile nominale con *m* modalità, ci si trova nella situazione più complessa perché non è possibile stabilire un ordinamento. Le suddivisioni possibili sono $2^{m-1}-1$ e, al crescere di *m*, tale valore cresce più che proporzionalmente e può, quindi, diventare molto elevato.
- Quando, infine, la variabile è dicotomica, ci si riconduce chiaramente al caso particolare di una variabile nominale con *m*=2. La suddivisione, quindi, è univoca poiché corrisponde alle due modalità che X_s può assumere.

Criterio di suddivisione di un nodo padre

- Questa è la fase centrale nell'intero processo di segmentazione perché comporta la scelta del criterio in base al quale effettuare la ripartizione delle unità appartenenti al nodo padre nei suoi nodi figli.
- E' proprio questa fase che distingue le varie tecniche di classificazione nonché i diversi programmi informatici che permettono di costruire gli alberi di classificazione e di regressione.
- Il criterio di suddivisione o split consiste nel calcolo di un indice statistico, precedentemente scelto, che permetta di individuare la miglior partizione tra tutte quelle che sono state create e che corrispondono ad ogni singolo predittore.
- In genere, gli split sono nella forma $X_j \leq t$ e $X_j > t$, dove X_j è la variabile e t è un valore dell'indice statistico.

- Fra tutti i predittori viene scelto, poi, quello migliore, cioè quello che riduce maggiormente l'eterogeneità in base ad un determinato criterio di riduzione. Viene, inoltre, valutato il grado di 'omogeneità' interna ai nodi figli ed il grado di 'eterogeneità' tra gli stessi
- Questi valori permettono di selezionare il predittore più efficiente e, di conseguenza, la partizione ottimale.
- Vengono utilizzati degli algoritmi di segmentazione che sono in grado di ricercare lo split migliore analizzando tutte le *p* variabili esplicative.
- Generalmente tutti gli algoritmi di costruzione degli alberi decisionali usano una struttura di tipo "top-down". Questo vuol dire che, a partire dal nodo radice, si esamina il *training database* per selezionare un criterio di split per tale nodo e, poi, in maniera ricorsiva si considera il nuovo nodo intermedio (ossia né radice né foglia) creato e si cerca un nuovo criterio di split per quel nodo.

Generalmente tutti gli algoritmi di costruzione
degli alberi decisionali usano una struttura
di tipo “top-down”.
Questo vuol dire che, a partire dal nodo radice,
si esamina il *training database* per selezionare
un criterio di split per tale nodo
e, poi, in maniera ricorsiva si considera
il nuovo nodo intermedio (ossia né radice né foglia)
creato e si cerca un nuovo criterio di split
per quel nodo.



Criterio di arresto

- La costruzione di un albero decisionale tramite segmentazione è una tecnica “ricorsiva”,
- cioè una tecnica che può andare avanti, se non all’infinito, sicuramente fino a che non viene fermata tramite un intervento esterno.
- Altrimenti si arresta quando i nodi terminali contengono solo una unità statistica o solo casi appartenenti alla stessa classe della variabile dipendente, senza, perciò, dare un significativo apporto alla conoscenza.

- Per questo è richiesta la definizione di una o più regole di arresto, al verificarsi delle quali il processo si ferma.
- Tale regola deve avere della proprietà tra le quali due di fondamentale importanza:

la semplicità: tra due criteri di arresto si sceglie quello che determina l'albero di ampiezza minore e, quindi, è più facilmente interpretabile e leggibile per quanto riguarda i risultati;

il potere discriminatorio, una regola di arresto deve permettere di distinguere efficacemente unità statistiche appartenenti a classi diverse.

- I criteri di arresto più noti si basano sul concetto di numerosità minima dei nodi terminali o su livelli massimi di crescita dell'albero.
- Un metodo che propone una tecnica diversa è il CART o *Classification And Regression Tree* (Breiman L. *et al*., 1984): viene dapprima costruito l'albero di massima dimensione, cioè quello in cui ogni nodo contiene un solo elemento oppure elementi appartenenti alla stessa classe e, successivamente, lo si "pota" (in inglese *pruning*) sulla base di una regola che minimizza la complessità a parità di potere discriminatorio.

Regola per l'assegnazione delle classi alle foglie e per la classificazione di nuovi casi

Nel passo precedente viene portato a termine il vero e proprio processo di costruzione dell'albero.

Ora è necessario stabilire la corrispondenza tra classi ed ogni nodo terminale (foglia).

- Si possono riscontrare tre casi:

a) la foglia comprende unità statistiche appartenenti ad una sola classe e, quindi, alla foglia viene assegnata proprio la classe corrispondente a quella della unità che ne fanno parte (regola dell'unanimità);

b) la foglia comprende unità statistiche di classi diverse di cui una, però, con frequenza superiore alle altre e, quindi, alla foglia viene assegnata la classe corrispondente a quella con frequenza massima (regola della maggioranza o *plurality rule*); (Breiman et al. 1984)

c) la foglia comprende unità statistiche di classi diverse ma con medesima frequenza e, in questo caso, si cade in una zona di indecisione; nelle tecniche che usano il *pruning* (CART) questo avviene molto raramente.

- Una volta terminata l’assegnazione delle classi alle foglie dell’albero, si può procedere alla classificazione di nuove unità non appartenenti al campione utilizzato per la costruzione di tale albero. Ogni singolo caso finirà in un nodo terminale e sarà classificato in base alla classe assegnata al nodo corrispondente.

OSS.

La previsione della classe di appartenenza per nuove unità statistiche costituisce l’obiettivo fondamentale in diversi ambiti tra cui il data mining, la pattern recognition e il machine learning (Bishop 2006). In questi ambiti risulta fondamentale la scoperta di regolarità nei dati utilizzabili a fini classificatori e previsivi.

Stima del tasso di errata classificazione

- Il “tasso di errata classificazione”, indicato con $R(d)$, dove d è la regola di classificazione a cui è associato, serve a valutare la bontà di una classificazione.

- Questo significa che, a parità di semplicità nella rappresentazione di diversi alberi, misurata in termini di numero di foglie, verrà selezionata la regola che consente di classificare correttamente la percentuale più alta di unità statistiche.

- Sia *S* il campione di unità statistiche sulla cui base viene costruita la regola di classificazione *d* ed *Ω* un insieme artificiale di unità statistiche, molto numeroso, ipoteticamente infinito ed avente le medesime caratteristiche di S.
- Dal confronto tra la reale classificazione delle unità in *Ω* e la classificazione derivata dalla regola *d* dovrebbe scaturire il tasso di errata classificazione. Ma questo è vero solo in teoria perché, nella pratica avendo a disposizione *S*, è necessario usare una stima di *R*(*d*), che sarà indicata con $\hat{R}(d)$
- Esistono diversi metodi per il calcolo ne vengono esposti tre.

1)*Stima basata sul campione S (Resubstitution Estimate)*

- Definendo con $C_j(i)$ la classe di effettiva appartenenza della *i*-esima unità statistica, con $d(x_i)$ la classe assegnata dalla regola *d* all'*i*-esima unità e con $I(\cdot)$ una funzione indicatore di evento la stima per risostituzione del tasso di errata classificazione è:

$$\hat{R}(d) = \frac{1}{n}\sum_{i=1}^{n} I[d(x_i) \neq C_j(i)]$$

- E' stato constatato che questo metodo, pur essendo computazionalmente semplice, fornisce una buona stima

2)*Stima basata su un campione test (Test Sample Estimate)*

Si suddivide casualmente il campione S in due sottocampioni S_1 e S_2 rispettivamente di numerosità n_1 e n_2, tali che $S_1 \cup S_2 = S$ e $S_1 \cap S_2 = \emptyset$.

S_1 si chiama "campione di apprendimento" (*learning sample*) ed S_2 si chiama "campione test" (*testing sample*). La regola di stima che ne deriva è:

$$\hat{R}_{ts}(d) = \frac{1}{n_2} \sum_{i \in S_2} I[d(x_i) \neq C_j(i)]$$

è la stima sul campione test, dato che la regola *d* viene costruita tramite il campione S_1 e viene testata stimando $R(d)$ su S_2.

- La stima del tasso di errata classificazione ottenuta con questo metodo è più affidabile rispetto alle altre perché vengono usati anche dati esterni oltre a quelli impiegati per la determinazione della regola di classificazione.

3)*Stima basata sulla cross-validation (V-fold Cross-Validation Estimate)*

Il campione S viene suddiviso casualmente in V>2 sottocampioni $S_1,...,S_v,...,S_V$ di dimensione in più possibile simile fra loro. Si costruisce una regola di classificazione $d^{(v)}(x)$ sul campione S - S_v (v=1,...,V) e si fa una stima di $R(d^{(v)})$ basata sul campione test

$$\hat{R}_{ts}(d^{(v)}) = \frac{1}{n_v} \sum_{i \in S_v} I[d^{(v)}(x_i) \neq C_j(i)] \qquad \text{con } v=1,...,V$$

Se V è sufficientemente elevato, ogni classificatore $d^{(v)}(x)$ viene costruito usando un campione di apprendimento di dimensione *n(1-1/V)* ossia prossima alla dimensione di S. La base portante del metodo basato sulla *cross-validation* è quindi la stabilità, perché ogni classificatore $d^{(v)}(x)$ ha un tasso di errata classificazione $R(d^{(v)})$ molto prossimo a $R(d)$.

- La stima del tasso di errata classificazione calcolato con questo metodo viene indicato con $\hat{R}_{cv}(d)$

$$\hat{R}_{cv}(d) = \frac{1}{V}\sum_{v=1}^{V} \hat{R}_{ts}(d^{(v)}).$$