Classificazione

Introduzione

- I modelli di classificazione si collocano tra i metodi di apprendimento supervisionato e si rivolgono alla predizione di un attributo target categorico.
- A partire da un insieme di osservazioni riferite al passato, per le quali è nota la classe di appartenenza, si genera un modello che consente di effettuare predizioni.
- Vedremo alcuni esempi, i principali indicatori di valutazione e alcuni metodi.

Problemi di classificazione

- Si dispone di un dataset D contenete m osservazioni (*sample*), costituite da n attributi esplicativi e da un attributo target categorico (*label*).
  - Gli attributi esplicativi possono essere di qualsiasi tipo.
  - L'attributo target viene indicato anche come classe.
- Si vogliono individuare legami ricorrenti tra le variabili esplicative di osservazioni appartenenti alla stessa classe.
- I legami vengono tradotti in un modello che viene utilizzato per predire la classe di osservazioni di cui non è nota la label.

Credit card promotion database

Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19

- Una società di gestione di carte di credito inserisce offerte promozionali per polizze vita nell'estratto conto mensile.
- Vuole costruire un modello di previsione per Life Insurance Promotion.
- Lo scopo è proporre l'offerta solo ai clienti che verosimilmente accetteranno.

Sviluppo di un modello

- **Training**. L'algoritmo di classificazione viene applicato agli esempi appartenenti ad un sottoinsieme $T \subset D$ per ricavare il modello.
- **Test**. Il modello viene impiegato per classificare le osservazioni V=D – T. La classe di appartenenza viene confrontata con quella predetta.
  - Per evitare sovrastime, T e V devono essere disgiunti.
- **Predizione**. Il modello viene utilizzato per predire la classe di nuove osservazioni per cui non è nota l'appartenenza.

Tassonomia dei modelli

- **Modelli euristici**. Procedure basate su schemi semplici e intuitivi.
  - Nearest neighbor, alberi di classificazione,…
- **Modelli di separazione**. Si ricavano regioni disgiunte dello spazio che permettono di separare le osservazioni appartenenti a classi diverse.
  - Analisi discriminante, reti neurali, support vector machine
- **Modelli di regressione**. Si ipotizza una forma funzionale per la probabilità che una osservazione venga assegnata dal supervisore ad una classe target.
  - Regressione logistica.
- **Modelli probabilistici**. Si ipotizza una forma funzionale per la probabilità condizionate delle osservazioni data la classe di appartenenza.
  - Reti bayesiane.

Valutazione dei modelli

- *Accuratezza di f su $(\mathbf{x}_i, y_i)$:*

$$L(y_i, f(\mathbf{x_i})) = \{ \begin{array}{lll} 0 & se & y_i = f(\mathbf{x}_i) \\ 1 & se & y_i \neq f(\mathbf{x}_i) \end{array} ,$$

$$acc_A(V) = 1 - \frac{1}{v} \sum_{i=1}^{v} L(y_i, f(\mathbf{x}_i)),$$

- *Velocità*
- *Robustezza*
- *Scalabilità*
- *Interpretabilità*

Cross validadation

- Per valutare l'accuratezza di un metodo si suddivide l'insieme di train in k sottoinsiemi disgiunti (*fold*) $A_1, \ldots, A_k$ e prevede k iterazioni.
- In corrispondenza della r-esima iterazione si sceglie come insieme di test A_r e come train l'unione dei rimanenti A_i, $i \neq r$.
- L'accuratezza complessiva è valutata come media delle singole accuratezze.
- Per $r = m$, si ottiene il *leave one out*.

Matrice di confusione

- L'elemento sulla riga i e sulla colonna j è il numero di casi della classe "vera" i che il classificatore ha classificato nella classe j.
- Sulla diagonale ci sono i casi classificati correttamente.
  - Gli altri sono errori.

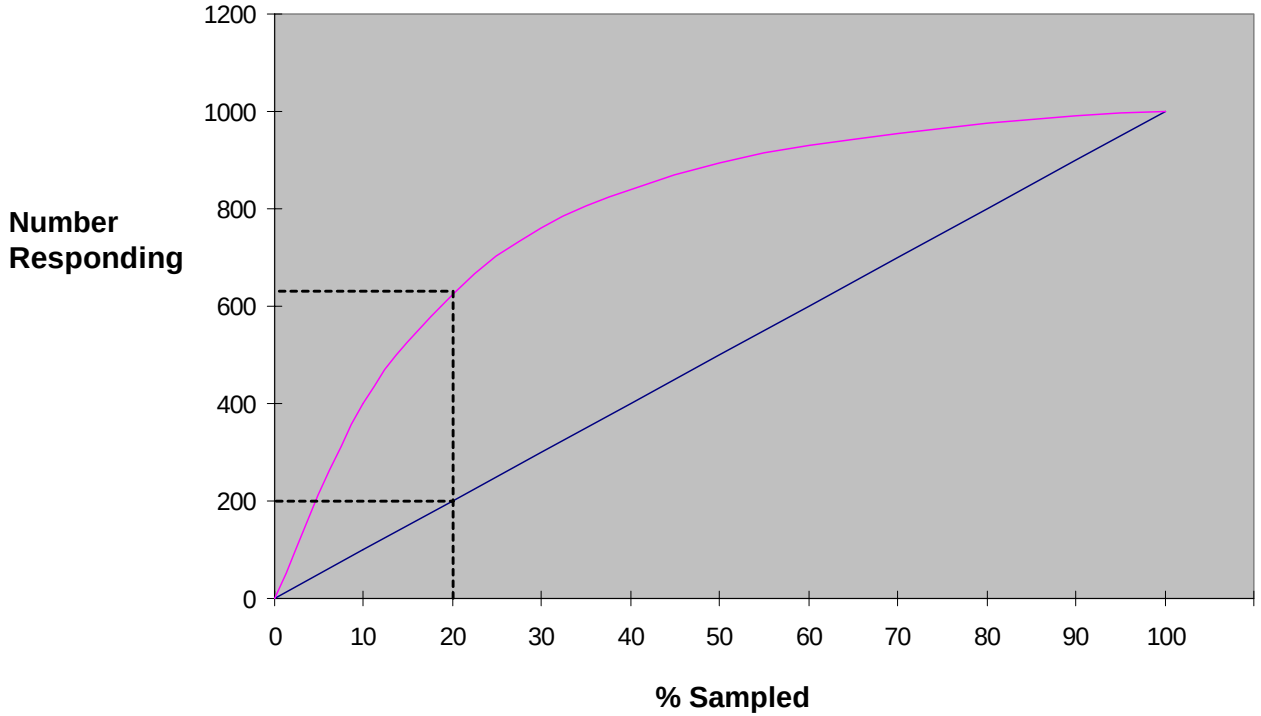
				Classe calcolata						
				A	**B**	**C**		**Totale**		
Classe vera		**A**		60	14	13		**87**		**69.0%**
		B		15	34	11		**60**		**56.7%**
		C		11	0	42		**53**		**79.2%**
		Totale		**86**	**48**	**66**		**200**		**68.0%**

Grafici di Lift

- La misura di lift corrisponde all'idea di valutare l'accuratezza di un classificatore in base alla densità di osservazioni positive presenti nel gruppo identificato in base alle predizioni del modello.
- Il lift è il rapporto tra la performance del modello scelto e quella del modello casuale:

$$lift = \frac{b / campione}{a / popolazione}$$

Grafici di lift



- Supponiamo che ci siano 100.000 clienti con media di successi dell'1% su campagne analoghe.
- La retta descrive la performance attesa del modello casuale.
- La curva superiore descrive un modello che sui 20.000 clienti con i migliori score coglie 625 successi, e ha quindi lift(0,20) = 625 / 200 = 3,125.

Campagna di promozione

Nessun modello	Classe Accetta	Classe Rifiuta	Modello ideale	Classe Accetta	Classe Rifiuta
Accetta	1.000	0	Accetta	1.000	0
Rifiuta	99.000	0	Rifiuta	0	99.000

- Con nessun modello, si invia a tutti i 100.000 clienti e si hanno 1.000 successi. Il lift è 1, perché campione e popolazione coincidono.
- Con il modello ideale si scelgono proprio i 1.000 che accetteranno. Il lift è 100 / 1 = 100.
- Il modello ideale è 100 volte migliore in senso conoscitivo di quello che classifica tutte come Accetta.
- In senso economico dipende dai costi di rifiuto e dai profitti di accettazione.

Due matrici di confusione per modelli alternativi con lift uguale a 2.25

Modello X	Classe Accetta	Classe Rifiuta	Modello Y	Classe Accetta	Classe Rifiuta
Accetta	540	460	Accetta	450	550
Rifiuta	23.460	75.540	Rifiuta	19.550	79.450

$$\text{Lift(Modello X)} = \frac{540/24000}{1000/100000} = 2{,}25$$

$$\text{Lift(Modello Y)} = \frac{450/20000}{1000/100000} = 2{,}25$$

Il modello *X* porta a spedire 24.000 offerte, con 540 successi.
Il modello *Y* porta a spedire 20.000 offerte, con 450 successi.
Il modello *Y* è migliore se le 4.000 spedizioni risparmiate superano in valore le 90 vendite perse. Altrimenti è migliore *X*.

Alberi di decisione

- Gli alberi di classificazione rappresentano uno dei modelli meglio noti ed utilizzati nelle applicazioni di data mining.
  - Concettualmente semplici, facili da utilizzare, robusti per dati mancanti ed outlier.
- Permettono di ricavare regole per separare le osservazioni appartenenti a classi differenti.

Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19

Attributi e target

- Ci sono 4 attributi esplicativi: Income range, Credit card insurance, Sex, Age e il target *Life insurance promotion d*i tipo categorico: due classi, Yes e No.
- Ogni attributo ha un potere discriminante, cerchiamo quello che lo ha massimo, cioè l'attributo che meglio differenzia le due classi di risposta.
- L'ideale è un attributo che per ogni valore individua in modo univoco la risposta.

Determinazione delle regole

- Se si classifica un nuovo cliente in maniera casuale rispetto a *Life*, conviene metterlo nella classe Yes.
  - Predizione su un *individuo tipico*, non su quel cliente.
  - Per la classe Yes 9/15 è una probabilità a priori: *p(Life = Yes)* = 9/15.
- Guardando i dati si può calcolare le probabilità a posteriori relativa all'attributo *Sex*.
- Se il nuovo cliente è maschio, l'attributo *Life* vale *Yes* 3 volte su 8. La stima è *p(Life = Yes | Sex = Male)* = 3/8.
- Se invece è femmina la frequenza è 6 volte su 7. *p(Life = Yes | Sex = Female)* = 6/7.
  - Disaggregando si impara qualcosa e si può predire meglio.

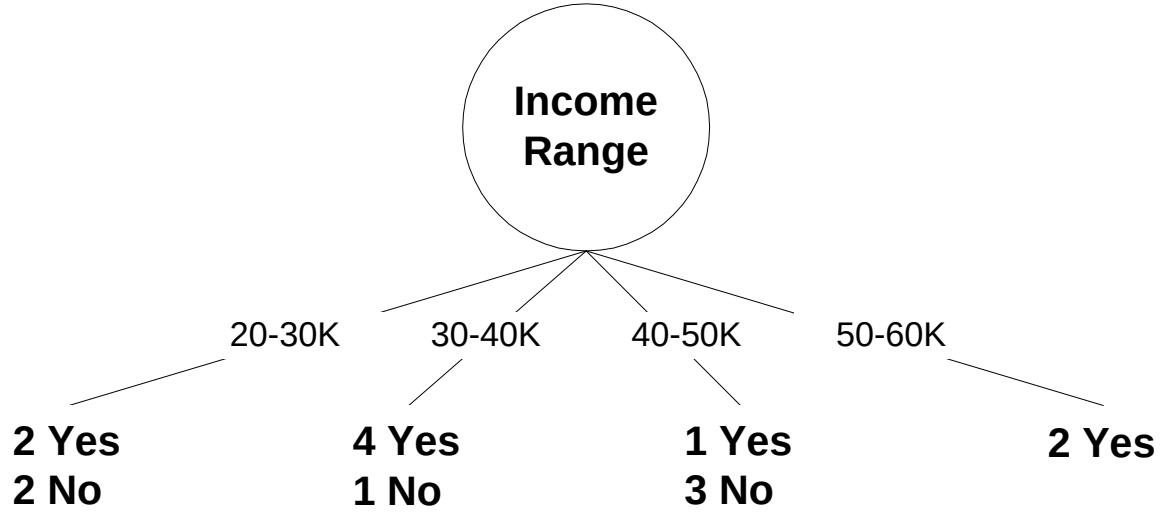
Attributo Sex

- Se il campione nel database è rappresentativo della popolazione, la regola *predici Yes* indovina 9 volte su 15 (accuratezza 60%).
- Usando Sex si può definire la regola:

  se è maschio predici No, se è femmina predici Yes
- Questa regola:
- in 8/15 maschi, predice No e indovina 5/8 dei casi;
- in 7/15 femmine, predice Yes e indovina 6/7 dei casi.

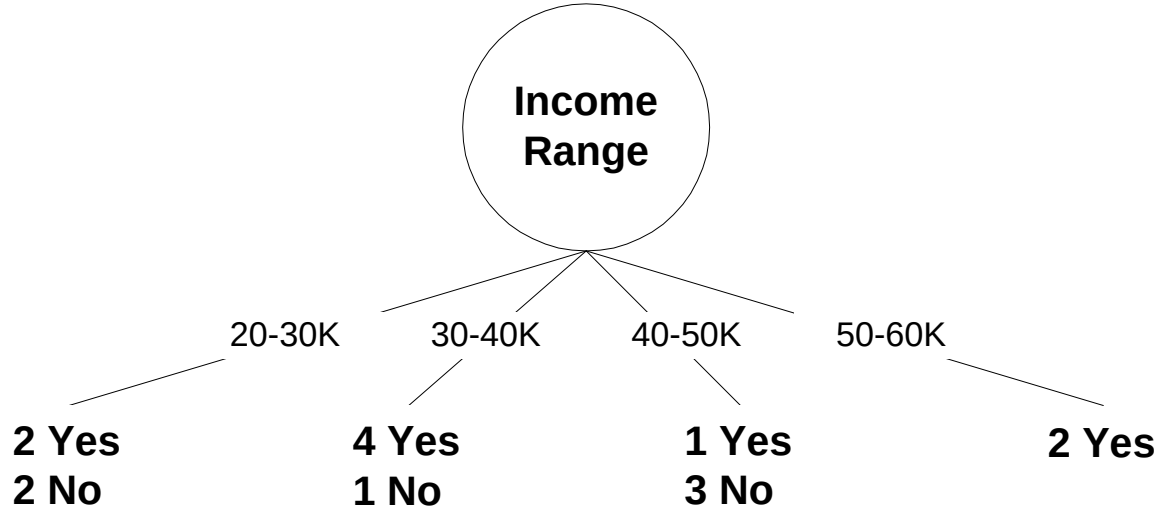
  ➢ La percentuale di successi è del 73.3%.
- I risultati sui problemi ristretti sono migliori che sul problema generale.

Attributo Income range



- Il range ha 4 valori, quindi i casi sono divisi in 4 classi.
- Il valore del reddito fra 50.000 e 60.000 $ è sufficiente a predire senza errori la risposta (sui dati di training).
- C'è una incertezza residua sulle prime 3 classi che sono *impure,* cioè miste.

Attributo Income range



Se ha reddito in 20-30K allora prevediamo risposta Yes.

Se ha reddito in 30-40K allora prevediamo risposta Yes.

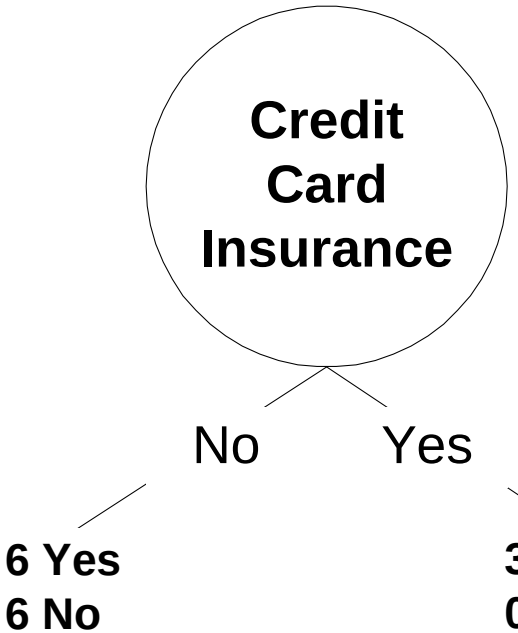
Se ha reddito in 40-50K allora prevediamo risposta No.

Se ha reddito in 50-60K allora prevediamo risposta Yes.

- Si è scelta la risposta più frequente nella classe nel caso di parità si è scelta la risposta più frequente.
- La regola minimizza gli errori sul training set e classifica 11 casi su 15.

Attributo Credit Card Insurance

Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19



L'attributo *Credit card insurance* classifica correttamente 9 casi su 15.
Accuratezza = 0.6
Numero rami = 2
Rapporto = 0.3
Rispetto a *Income range* riduce meno l'incertezza, ma generalizza meglio perché le foglie sono meno numerose.

Attributo Age

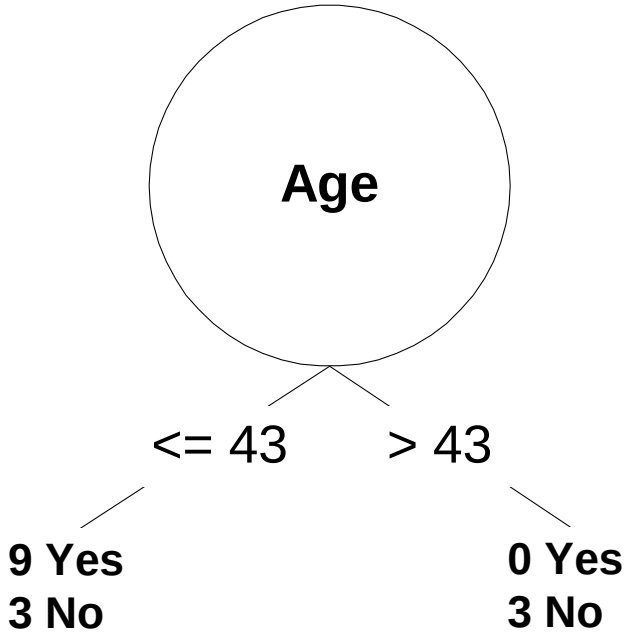
- Un modo comune di trattare le variabili numeriche è dividere il loro campo di variazione in due parti, con una soglia.

19	27	29	35	38	39	40	41	42	43	43	43	45	55	55
Y	N	Y	Y	Y	Y	Y	Y	N	Y	Y	N	N	N	N

- L'algoritmo prova tutti i tagli possibili
  - Tra 19 e 27, tra 27 e 29,…, tra 45 e 55.
- Si sceglie il taglio con il massimo rapporto accuratezza-numero rami.
  - In questo caso è tra 43 e 45.

The Credit Card Promotion Database

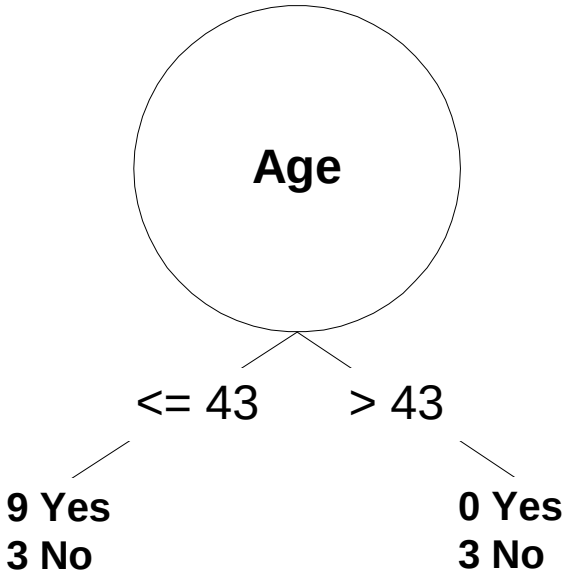
Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19



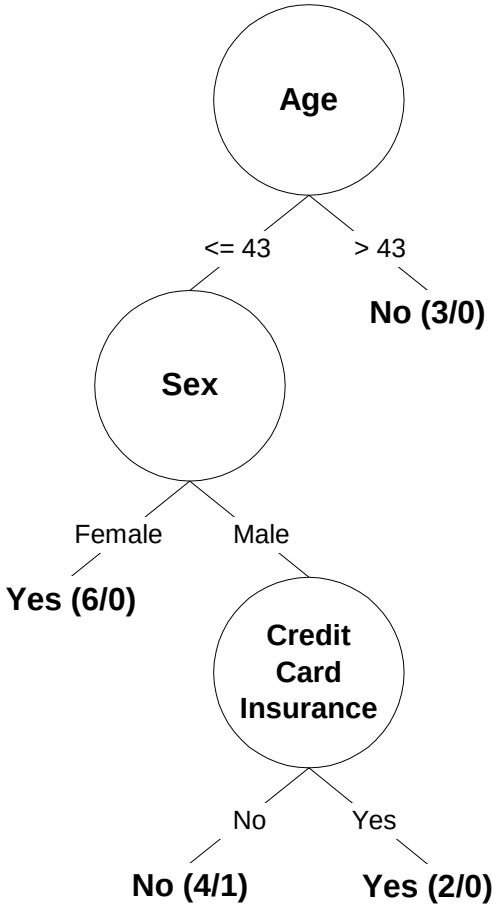
L'attributo *Age* classifica correttamente 12 casi su 15 (con il taglio ottimo).
Accuratezza = 0.8
Numero rami = 2
Rapporto = 0.4
Riduce l'incertezza meglio più delle altre due, e generalizza come *Credit card insurance*.
Sembra preferibile alle altre due.

Scelta dell'attributo

- Se si sceglie Sex il rapporto è 0.367.
- Classifica correttamente 11 casi su 15 come Income range, ma con 2 rami invece di 4, e quindi il suo rapporto è doppio.
- **Conclusione**: l'attributo che combina meglio capacità di discriminazione e generalizzazione è Age.
- Partendo dall'albero provvisorio con *Age* come radice si vede se è il caso di espandere le foglie allo stesso modo.



The Credit Card Promotion Database

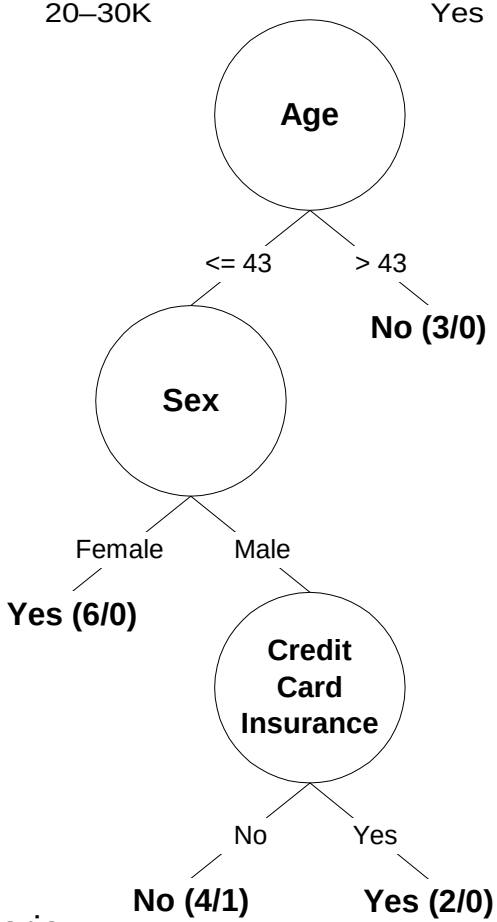


Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19

La seconda foglia (*Age* > 43) è già pura: 3 casi *No* e 0 casi *Yes*.
La prima foglia ha 6 casi *Yes* e 6 casi *No*.
Si espandie con il migliore attributo, che risulta essere *Sex*.
Il ramo *Sex* = *Female* ha 6 casi *Yes* e 0 casi *No* Il ramo *Sex* = *Male* ha 3 casi *Yes* e 3 casi *No*.
Si espande la foglia con *Credit card insurance.*
Il ramo *Yes* è puro.
Il ramo *No* è ancora impuro, contiene un caso *Yes*.

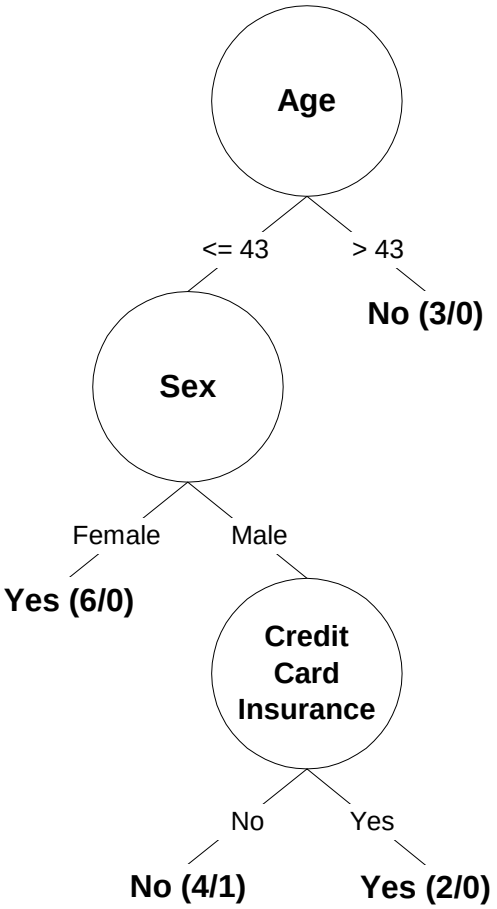
Ramo impuro

Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	42
20–30K	No	No	Male	27
30–40K	No	No	Male	43
20–30K	Yes	No	Male	29



Non c'è più modo di proseguire.
Non si riesce a caratterizzare il caso *Yes* utilizzando *Income range*.
Si potrebbe utilizzare di nuovo *Age* suddividendo nei rami *Age* <= 27 e *Age* > 27.
Creando nuovi nodi si classificherebbe correttamente tutti i casi di training.
Però i nodi aumentano e si in capacità di generalizzazione.
Conviene rinunciare a proseguire e accettare l'impurità.

Percorsi e regole



- Ogni percorso sull'albero corrisponde a una regola di decisione sul target.
- Una delle 4 regole è

IF Age <= 43
and Sex = Male
and Credit card insurance = No
THEN Life insurance promotion = No

- La regola si applica a 4 casi su 15 e classifica correttamente 3 casi su 4.
- Tra gli indici utilizzati per generare le regole ci sono l'entropia e Gini.