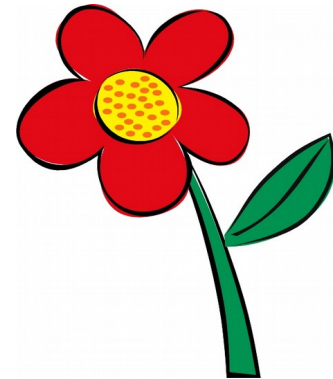


Classificazione: alberi di decisione

Materiale parzialmente tratto dalle slide associate al libro: Introduction to Data Mining di Tan, Steinbach, e Kumar

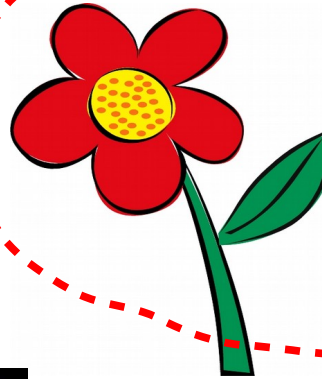
Il problema



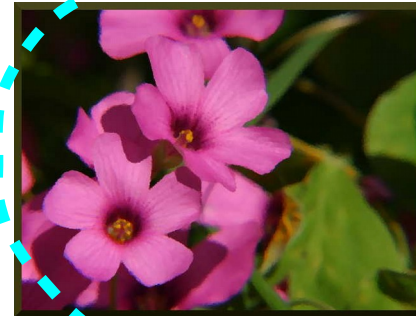
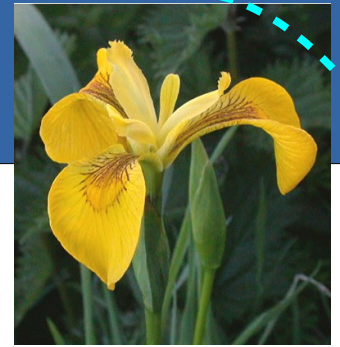
Il problema

- Le classi non sono insite negli esempi (o istanze)
- Non si tratta di raggruppamenti naturali inequivocabili
- Sono predefinite e dipendono dal fine per il quale si vuole costruire il predittore

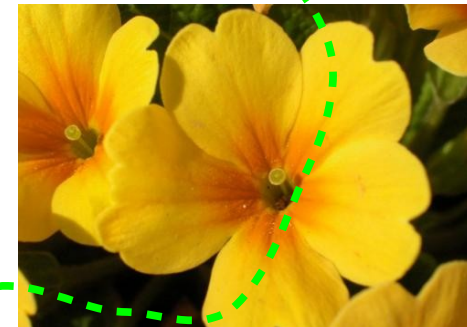
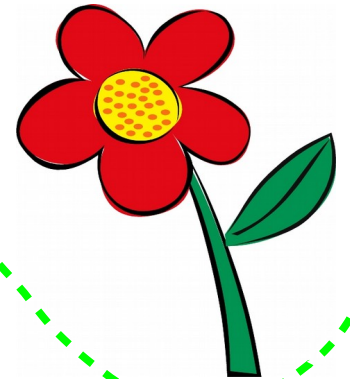
Raggruppamento per colore



Raggruppamento per forma



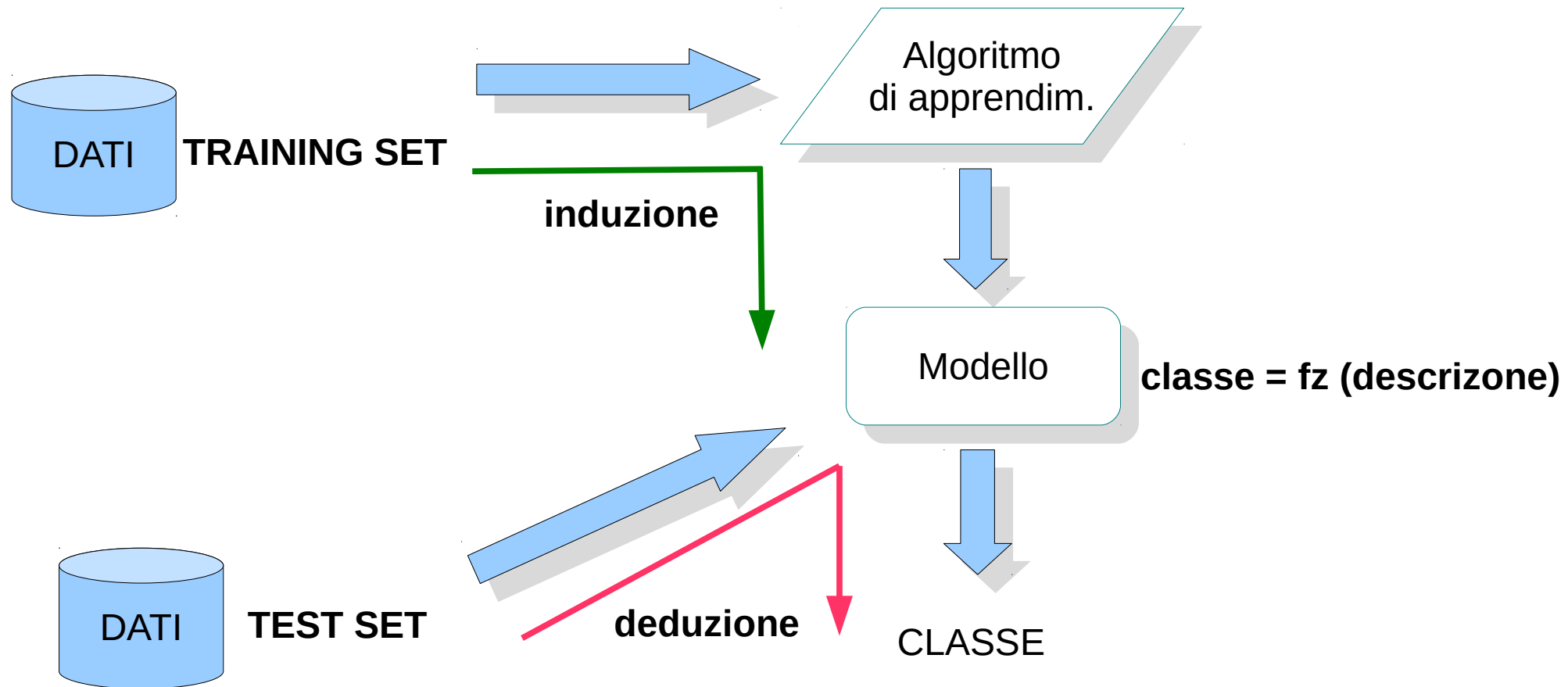
Fiori finti e fiori veri



Il problema

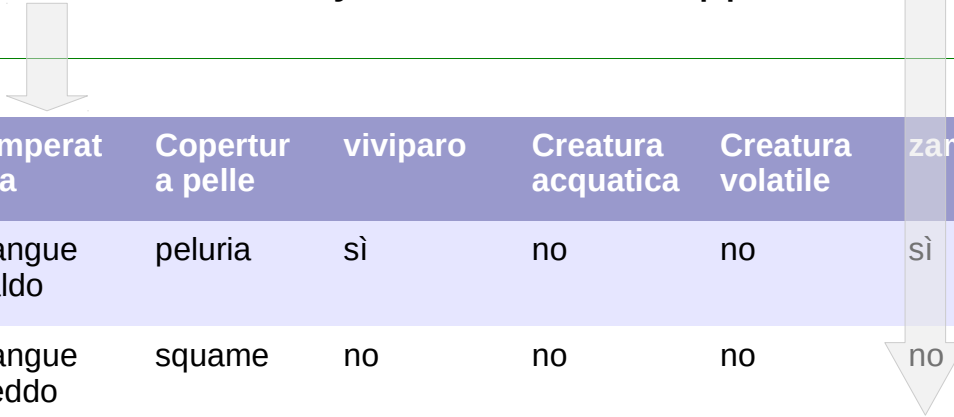
- **Dati:**
 - **Esempi** (fiori)
 - Categorie o classi (fiori finti/fiori veri, fiori rosa/gialli/bianchi/rossi, ...)
 - **Costruire:**
 - Una rappresentazione astratta (**modello**) che permetta di associare in modo corretto nuove istanze alla classe (o alle classi) di appartenenza
 - **Apprendimento supervisionato:** gli esempi dal quale astrarre le definizioni delle classi hanno associata la classe a cui appartengono
- 1) Problema di rappresentazione dei dati
 - 2) Problema di analisi dei dati e costruzione delle definizioni
 - 3) Problema di utilizzo della conoscenza acquisita

Schema generale

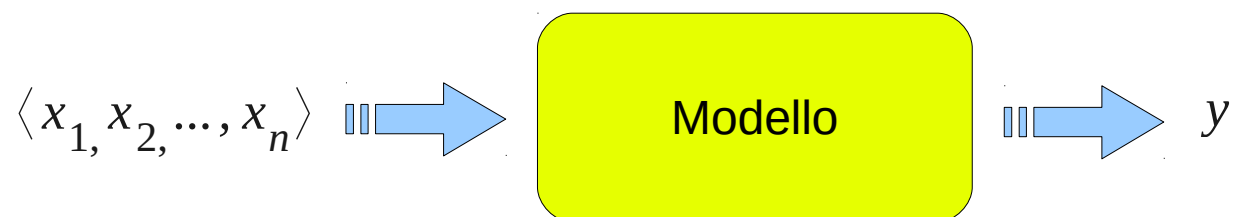


Learning (o training) set

Per **learning** (o training) **set** si intende la collezione di dati usati per svolgere il compito di apprendimento. I dati sono divisi in **istanze** (o record o esempi). Ogni esempio è rappresentato da una **tupla** (\bar{x}, y) dove \bar{x} è a sua volta una tupla di valori di **attributi descrittivi** e y è la **classe** di appartenenza dell'istanza



nome	temperatura	Copertura pelle	viviparo	Creatura acquatica	Creatura volatile	zampe	letargo	CLASSE
uomo	Sangue caldo	peluria	sì	no	no	sì	no	mammifero
pitone	Sangue freddo	squame	no	no	no	no	sì	rettile
salmone	Sangue freddo	squame	no	sì	no	no	no	pesce
balena	Sangue caldo	peluria(?)	sì	sì	no	no	no	mammifero
rana	Sangue freddo	nessuna	no	semi	no	sì	sì	anfibia
pinguino	Sangue caldo	piumaggio	no	semi	no	sì	no	uccello
piccione	Sangue caldo	piumaggio	no	no	sì	sì	no	uccello



Nota:

Come classe si usano attributi binari o categorie nominali

Sì no
1 0
Vero falso

Etichette:
es. Mammifero, pesce, ...

Usi dei Modelli appresi

Modello predittivo

Viene usato per predire la classe di appartenenza di istanze ignote in fase di apprendimento

Es. data la descrizione di una salamandra (sangue freddo, nessuna, no, semi, no, sì, sì) si usano le regole apprese per decidere a quale classe appartiene

Modello descrittivo

Viene usato come strumento esplicativo che permette di evidenziare quali caratteristiche distinguono le diverse categorie

Esprime in maniera sintetica delle descrizioni evitando di ragionare direttamente sugli esempi

Es. i mammiferi hanno il sangue caldo e solitamente non sono esseri acquatici

Test set

Qual è la bontà dei modelli appresi?

Valutazione sperimentale: il modello viene usato per classificare le istanze di un *test set*. La valutazione della bontà è fatta sulla base del comportamento di classificazione corretto/sbagliato su questi dati

Matrice di confusione:

		<i>classe predetta</i>	
		Classe 1	Classe 2
<i>classe reale</i>	Classe 1	f11	f12
	Classe 2	f21	f22

Errori !!

Predizioni corrette

Accuratezza ed error rate

$$\text{Accuratezza} = \frac{f_{11} + f_{22}}{f_{11} + f_{22} + f_{12} + f_{21}}$$

Predizioni
corrette

Predizioni
totali

$$\text{Error rate} = \frac{f_{12} + f_{21}}{f_{11} + f_{22} + f_{12} + f_{21}}$$

Predizioni
sbagliate

Predizioni
totali

Matrice di confusione: esempio

	<i>classe predetta</i>	
	<i>lattina</i>	<i>altro ogg.</i>
<i>lattina</i>	18	2
<i>altro ogg.</i>	1	19

Accuratezza: $(18 + 19) / 40 = 92.5\%$

Error rate: $3 / 40 = 7.5\%$

Matrice dei costi

		<i>classe predetta</i>	
		Classe 1	Classe 2
<i>classe reale</i>	Classe 1	c1	c2
	Classe 2	c3	c4

Costi degli errori !!

Costi delle valutazioni corrette

Matrice dei costi: esempio

		<i>classe predetta</i>	
		Classe 1	Classe 2
<i>classe reale</i>	Classe 1	-1	50
	Classe 2	10	0

Ogni istanza di classe 1 correttamente classificata riduce il costo complessivo; sbagliare a classificare gli oggetti di classe 1 è molto costoso; sbagliare a classificare gli oggetti di classe 2 è meno grave

Esempio: è più grave dire a un malato che è sano o dire a una persona sana che è malata?

Matrice dei costi: esempio d'uso

COSTI		classe predetta	
		Classe 1	Classe 2
classe reale	Classe 1	-1	50
	Classe 2	10	0

Mat. Conf.		classe predetta	
		lattina	altro ogg.
classe reale	lattina	18	2
	altro ogg.	1	19

$$\text{Costo} = -1 \times 18 + 50 \times 2 + 1 \times 10 + 0 \times 19 = 92$$

Attenti ai numeri ...

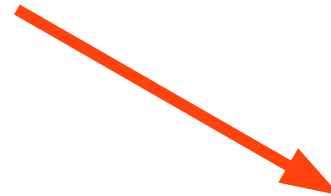
Istanze di classe A: 9990

Istanze di classe B: 10

Supponiamo che il nostro classificatore dica sempre che l'istanza è di classe A



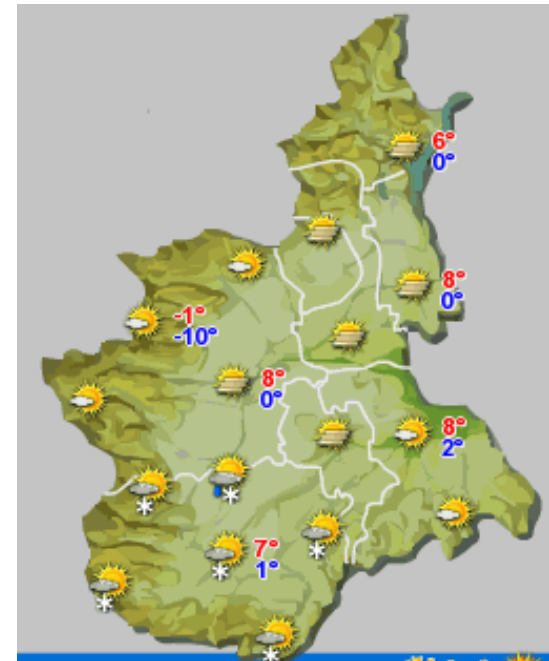
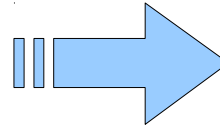
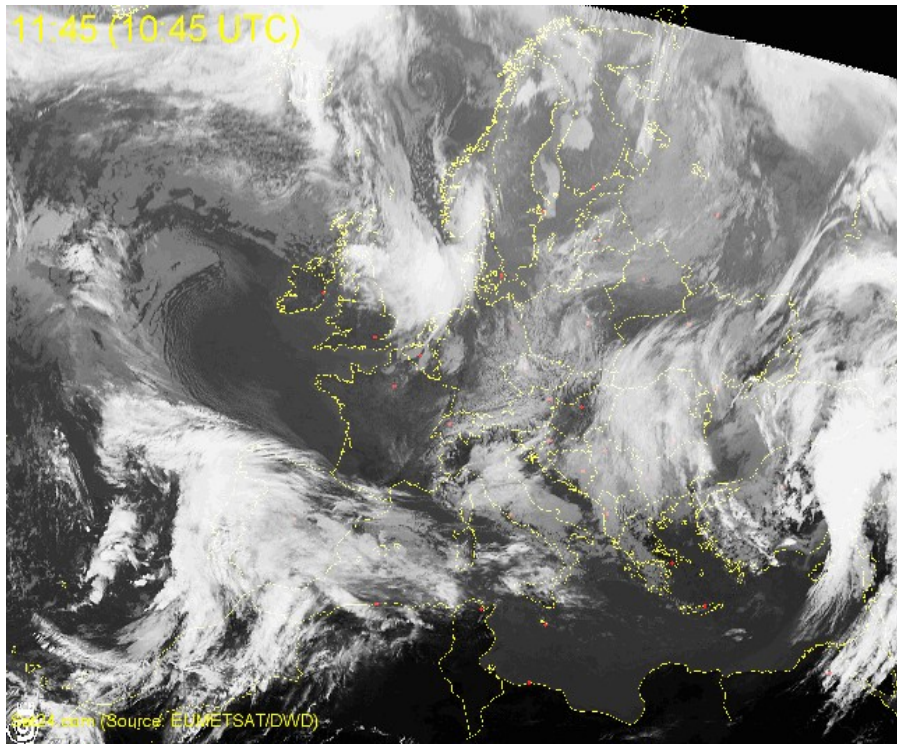
Accuratezza: $9990/10000 = 0.999$!!



Error rate: $10 / 10000 = 0.001$



Usi della classificazione (esempio)



Identificazione degli stati meteorologici

Usi della classificazione (esempio)



Classificazione della struttura
secondaria delle proteine
(alpha-helix, beta sheet, random
coil)

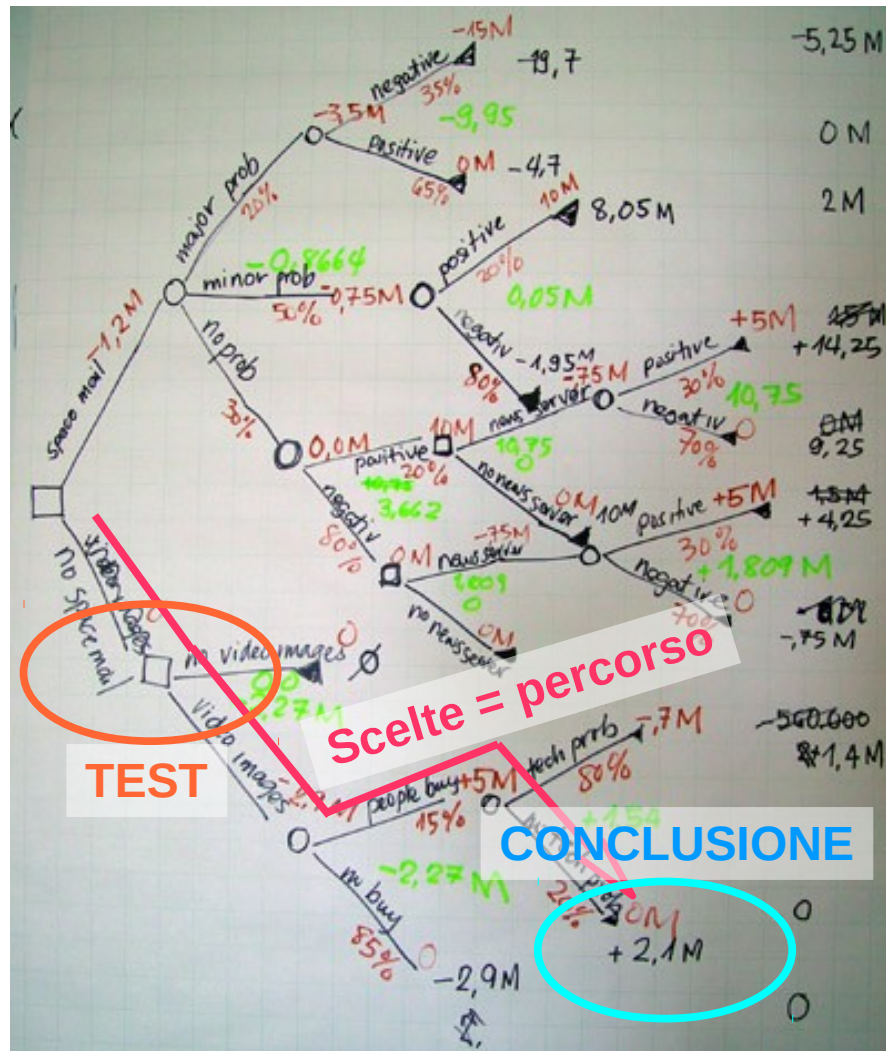
Usi della classificazione (esempio)



Uso lecito o fraudolento
in transazioni on-line con
carta di credito

Induzione di alberi di decisione

Alberi di decisione



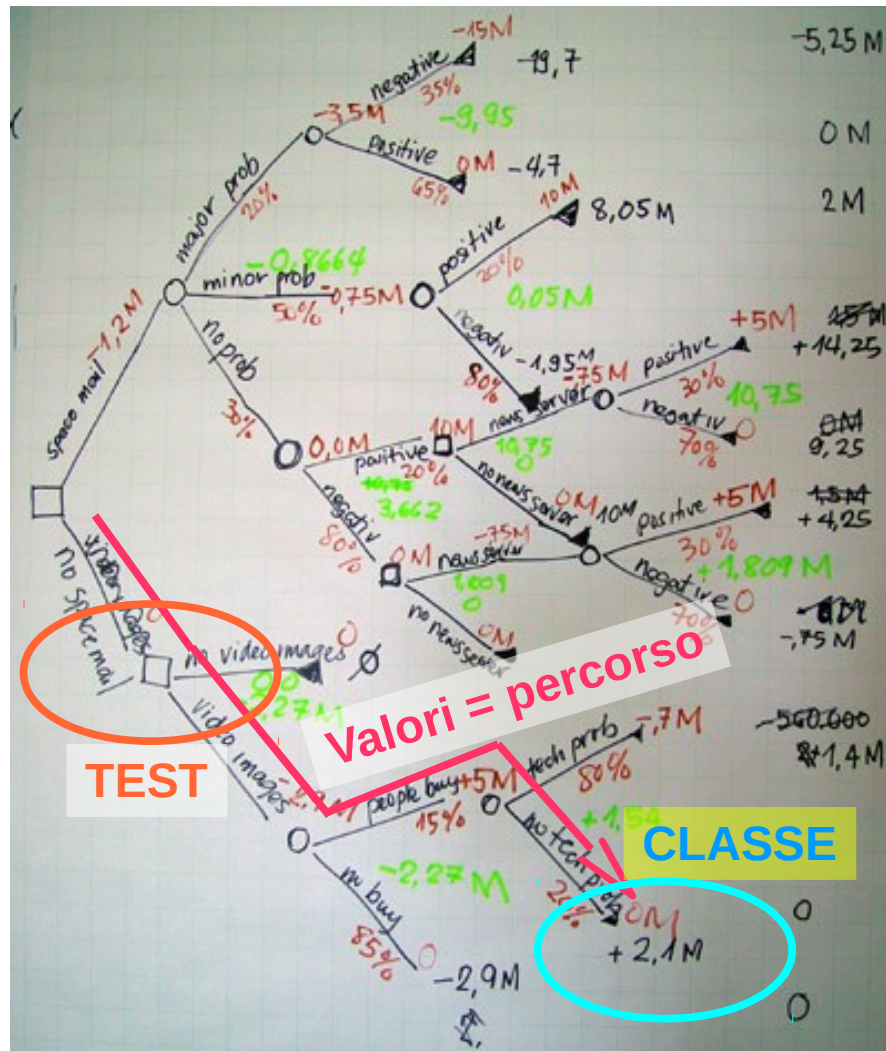
Sono strumenti di supporto alle decisioni che usano modelli strutturati ad albero, comunemente utilizzati per esempio per la definizione di strategie mirate al conseguimento di un goal

Esistono da tempi immemorabili

Hanno moltissime applicazioni. Per esempio la struttura di navigazione di molti negozi on-line implementa un albero di decisione (ti interessa un libro, un film o un gioco?)

Fonte: wikipedia

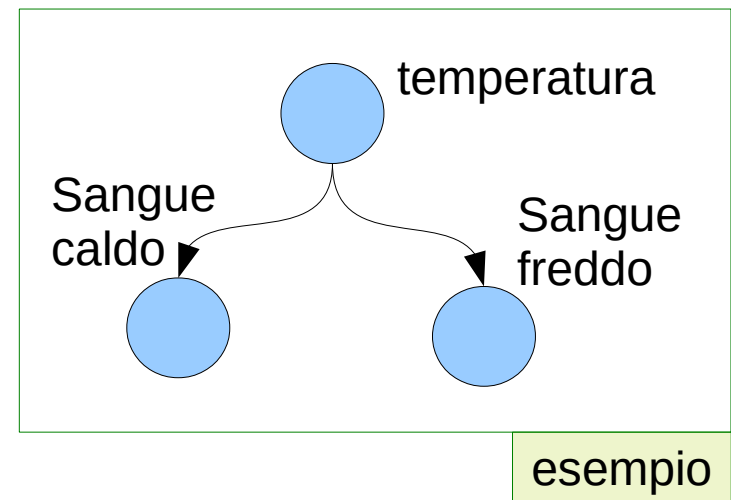
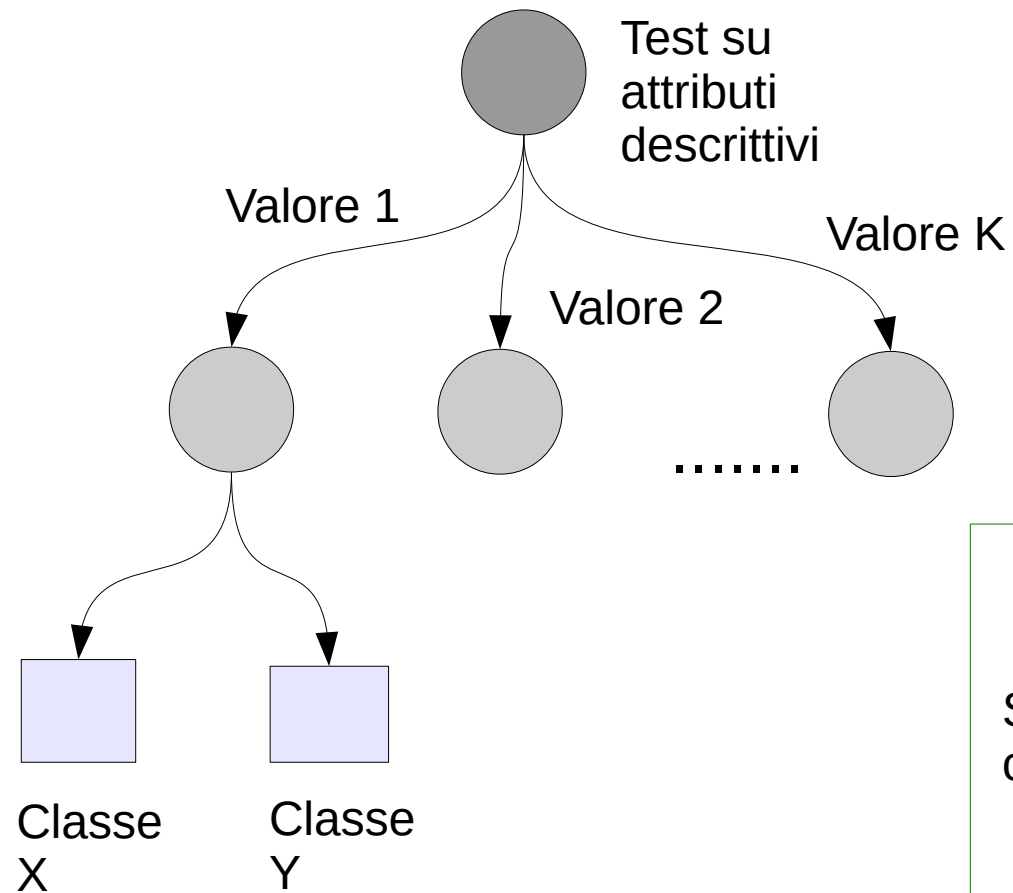
Alberi di decisione



Noi studieremo dei metodi che consentono di indurre da esempi degli alberi di decisione da utilizzare in compiti di classificazione

Fonte: wikipedia

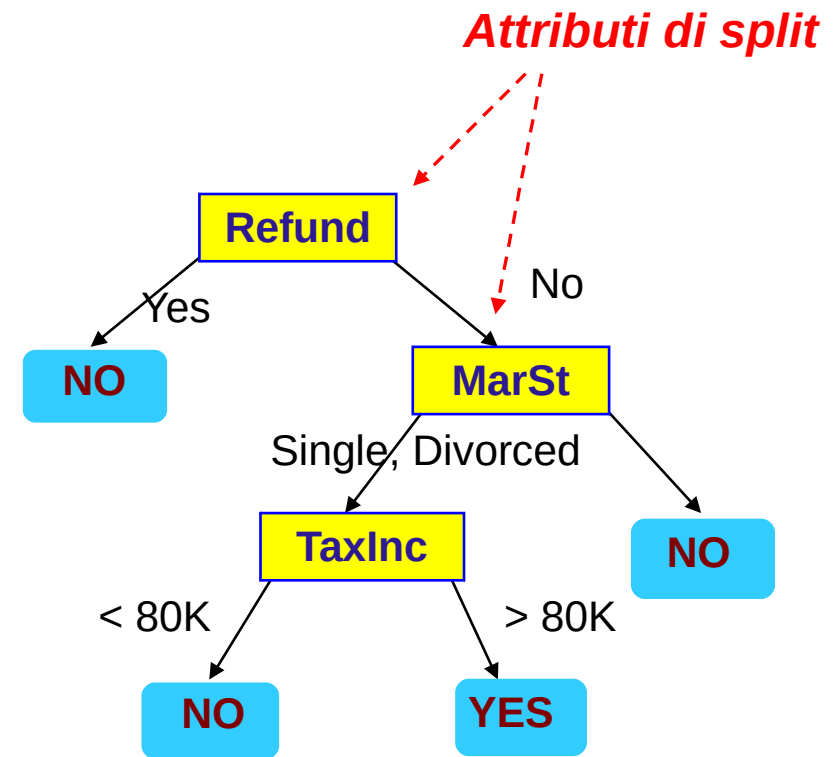
Struttura dei nostri alberi di decisione



Indurre alberi

<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Learning set



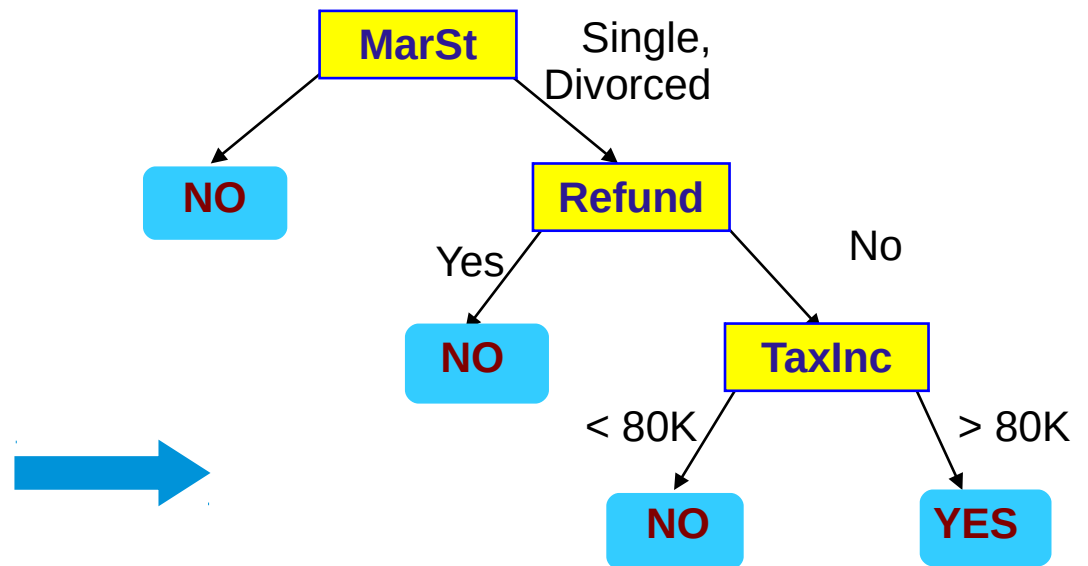
Modello: Albero di Decisione

Esistono più alberi per gli stessi dati !!

categorico
categorico
continuo
classe

<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

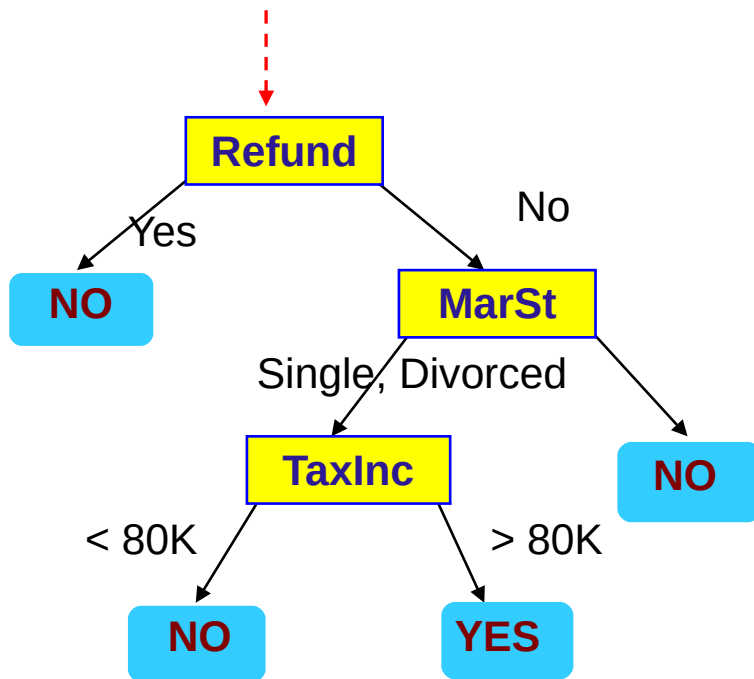
Stesso Learning
set di prima



Altro Modello!
Altro albero di Decisione

Uso di un albero appreso: predizione

Si inizia dalla radice



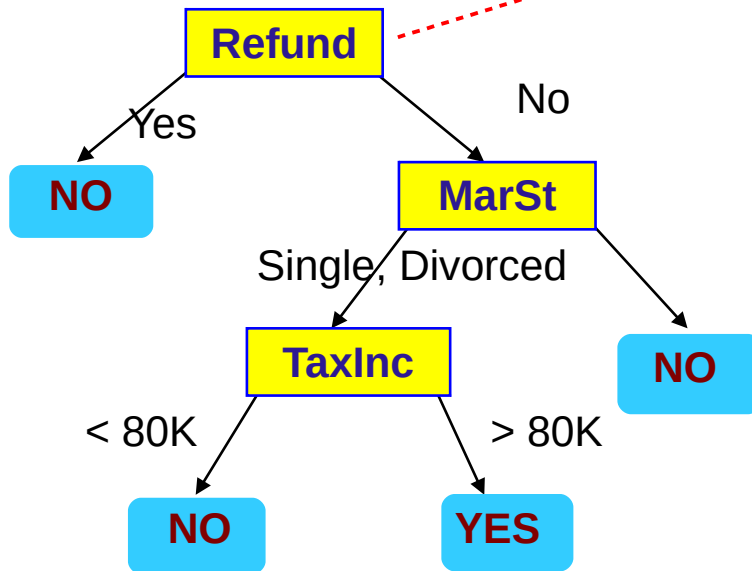
Test Set

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Uso di un albero appreso: predizione

Test Set

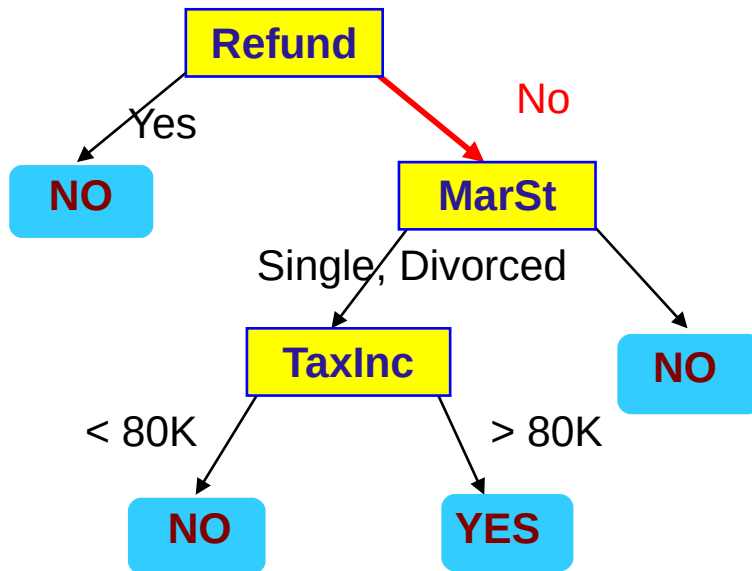
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Uso di un albero appreso: predizione

Test Set

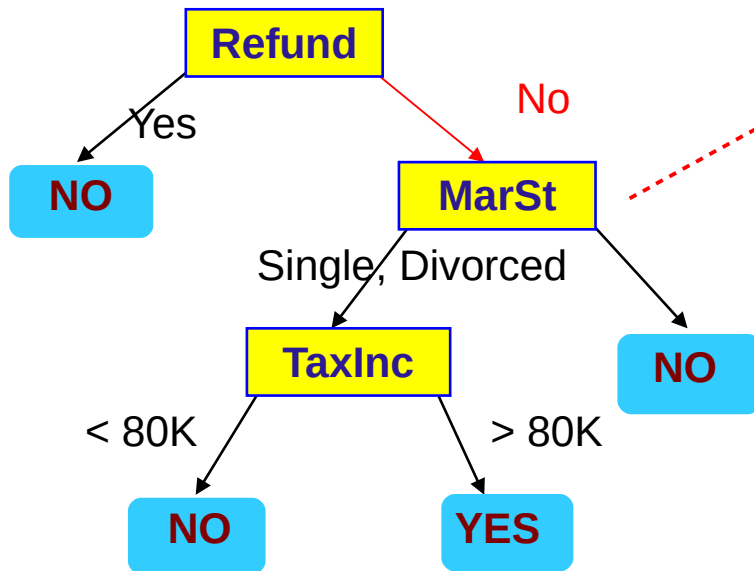
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Uso di un albero appreso: predizione

Test Set

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Uso di un albero appreso: predizione

Test Set

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

