Project work

Data Mining e IT

Introduzione al Data Mining

Nicola Barbieri

Obiettivi

- Illustrare i **campi d'applicazione** e le **potenzialità** del data mining

- Inquadrare il data mining nel contesto delle attività necessarie ad **estrarre conoscenza dai dati**.

- Comprendere caratteristiche e finalità di tutte le **fasi** necessarie per l'estrazione di conoscenza dai dati.

- Analizzare e sperimentare alcune delle principali classi di **algoritmi** utilizzati nel data mining.

Definizione di Data Mining

Attività **automatica** o semi-automatica che prevede **l'estrazione** di informazioni da **grandi quantità** di dati, allo scopo di ricercare relazioni ricorrenti (pattern) **non note** a priori.

Image source: deepclimate.org

Principali problemi affrontati dal Data Mining

- **Analisi del carrello**: analizzare le scelte dei clienti per suggerire loro nuovi prodotti a cui, molto probabilmente, saranno interessati.
- **Ricerca di anomalie**: identificare possibili comportamenti fraudolenti (carte di credito, assegni, account web, cellulari, ecc.) o pericoli per il business.
- **Churn Analysis**: identificare clienti propensi a passare alla concorrenza.
- **Segmentazione della clientela**: raggruppare i clienti in base ai loro comportamenti, dedicando a ciascun gruppo una strategia di marketing.
- **Pubblicità mirate**: identificare i potenziali clienti più promettenti per raggiungerli con campagne pubblicitarie mirate.
- **Previsioni**: analizzare i dati passati per prevedere come si evolverà il mercato, ma anche un virus o un uragano.
- **Ottimizzazione della produzione**: identificare i processi produttivi e logistici che possono essere migliorati in efficienza, costi e rendimento.

Citazioni

"If you're not prepared to be wrong,
you will not come up with anything original.“
Ken Robinson

“You can have data without information,
but you cannot have information without data.”
Daniel Keys Moran

“If you torture the data long enough,
it will confess.”
Ronald Coase

“In God we trust.
All others must bring data.”
William Edwards Deming

“If we have data, let’s look at data.
If all we have are opinions, let’s go with mine.”
Jim Barksdale

KDD e Data Mining

Il Data Mining è parte del processo Knowledge Discovery in Databases (KDD), che mira a **scoprire nuova conoscenza** analizzando le informazioni presenti nelle banche dati informatiche.

KDD prevede le seguenti fasi:

- **Selezione dei dati**: scartare dati inutili, acquisire dati mancanti.
- **Pre-elaborazione**: pulire, integrare, completare i dati, ed eventualmente estrarne un campione ideale da analizzare.
- **Trasformazione**: effettuare approssimazioni, conversioni di dati e valute, normalizzazioni di valori richieste dall'analisi.
- **Data Mining**: scegliere l'algoritmo da applicare, costruire un *modello* e verificarne l'efficacia.

  Serie di regole che permettono di processare nuovi dati alla luce della conoscenza attuale.

- **Interpretazione dei risultati**: valutare qualità e utilità dei risultati.

KDD e Data Mining

Talvolta si parla di Data Mining per indicare l'intero processo di KDD.

Il DM è il motore del processo KDD, ma senza le altre fasi non dà risultati.

È importante definire gli obiettivi della ricerca
e conoscere bene il business da analizzare.

- **Selezione dei dati**
- **Pre-elaborazione**
- **Trasformazione**
- **Data Mining**
- **Interpretazione dei risultati**

Tecniche, Algoritmi e Modelli

Tecnica: approccio fondato su una struttura matematica o logica, utilizzata da algoritmi per calcolare modelli e rispondere a una o più esigenze di ricerca.
Esempi: Alberi decisionali, clustering, reti neurali…

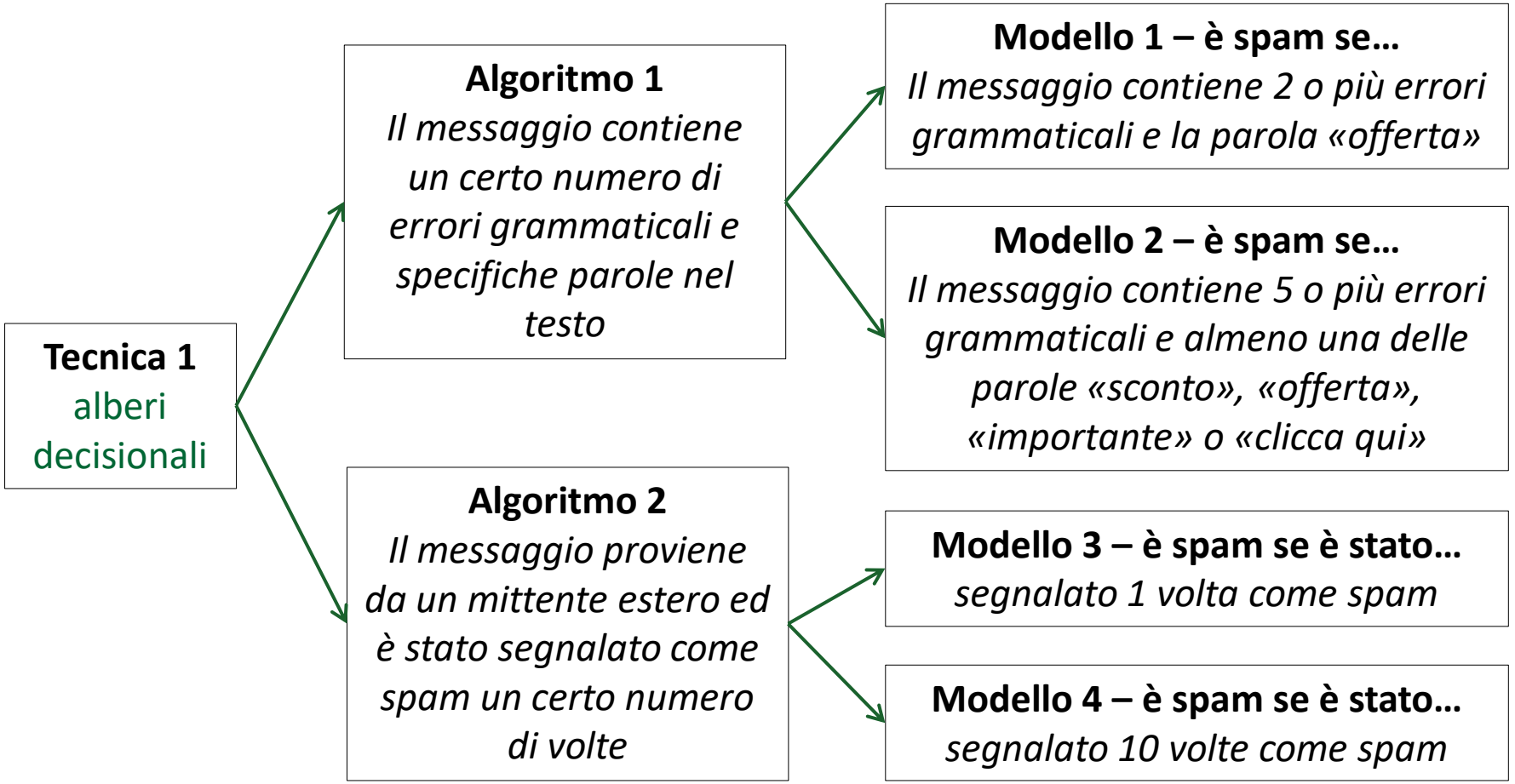
Algoritmo: Metodo che consente di calcolare diversi possibili modelli, cambiando parametri e soglie.

Esempi: K-means, ID3, CHAID, MARS…

Modello: Serie di regole che permettono di processare nuovi dati alla luce della conoscenza attuale.

Tecniche, Algoritmi e Modelli

Esempio: **riconoscere le e-mail indesiderate** (problema di classificazione)
Potranno esistere varie **tecniche** adatte a risolvere il problema. Per ciascuna ci saranno vari **algoritmi** da tarare, che daranno origine a diversi **modelli** più o meno validi (che restituiranno più o meno falsi positivi e falsi negativi).



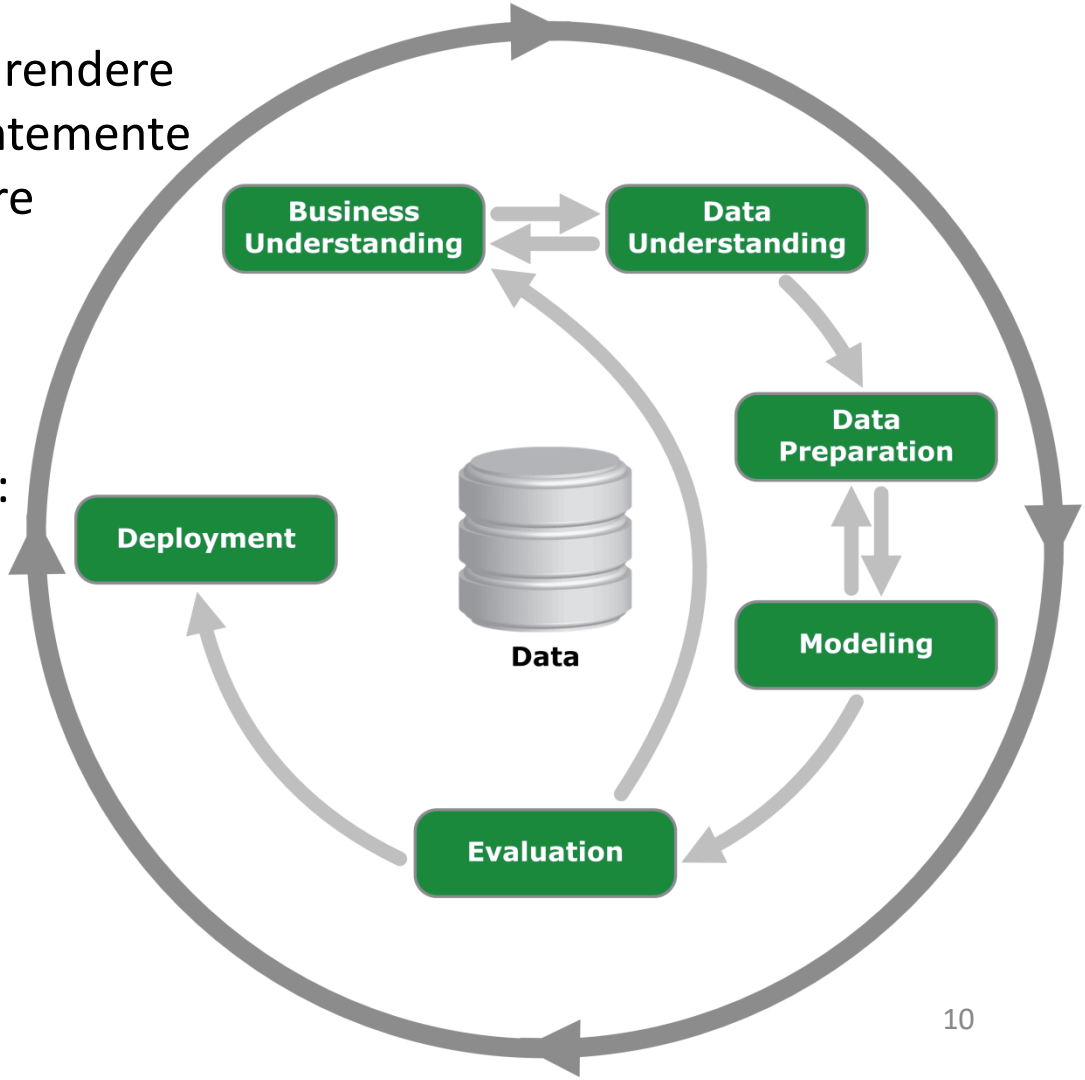
CRISP-DM: uno standard per il Data Mining

CRISP-DM (CRoss Industry Standard Process for Data Mining) nasce nel 1999 da un consorzio di aziende, con finanziamento dell'Unione Europea.

Obiettivo: definire **linee guida** per rendere il processo di Data Mining sufficientemente **affidabile** e **robusto** da poter essere utilizzato da persone con poche competenze tecniche (ma con una elevata conoscenza del business).

CRISP-DM prevede le seguenti fasi:

- Comprensione del business
- Comprensione dei dati
- Preparazione dei dati
- Costruzione del modello
- Valutazione del modello
- Attuazione del modello

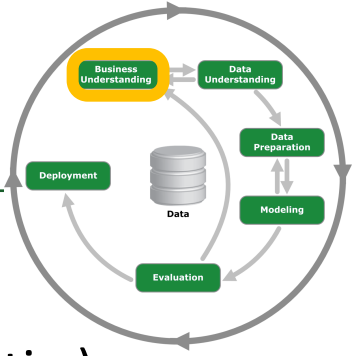


Nota sulla diapositiva precedente

Come tutte le raccolte di linee guida, per quanto standardizzato, il CRISP-DM solitamente non viene seguito al 100%. Rappresenta un modello di riferimento a cui tendere, ma la variabilità dei dati e delle situazioni reali di business richiedono una certa flessibilità e comportano l'allontanamento dal percorso teoricamente suggerito.

Già il comprendere e mettere in pratica l'utilità della revisione ciclica dell'analisi è un modo per ispirarsi al CRISP-DM.

Comprensione del business (1)



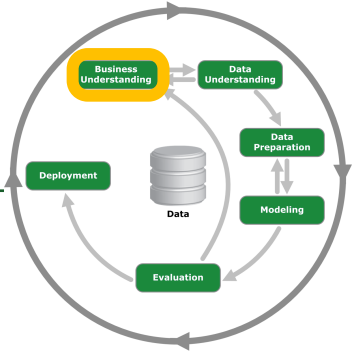
- **Che obiettivi ha il business?**
  - Analisi obiettivi aziendali (profitto, espansione, investimenti…)

Es: *Fidelizzare i clienti.*

- **Com'è la situazione attuale?**
  - Inventario delle risorse
  - Requisiti, presupposti, vincoli
  - Rischi e imprevisti
  - Analisi di costi e benefici

Es: *Entro un anno dal primo acquisto, solo il 15% dei clienti torna ad acquistare.*
Vogliamo portare il dato al 30%. Si stima che i ricavi salirebbero di X €/anno.
Possiamo stanziare per il progetto un budget di B € e 3 persone per 2 mesi.
Le campagne via e-mail sembrano infastidire gli utenti.
Un concorrente potrebbe lanciare una campagna aggressiva entro l'anno…

Comprensione del business (2)

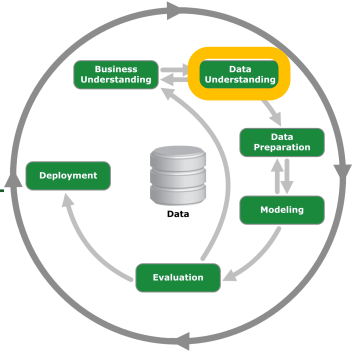


- **Cosa cerchiamo tramite il DM?**
  - Determinazione degli aspetti da esplorare con il DM

Es: *Cosa attira i clienti che fanno il primo acquisto?*
Come si può stimolare l'interesse verso altri articoli?
Che caratteristiche hanno i clienti fidelizzati e i loro acquisti?

- **Come si svilupperà il progetto?**
  - Creazione di un piano di progetto che guidi tutta l'attività di ricerca

Comprensione dei dati



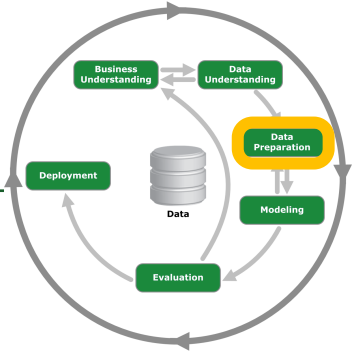
- **Identifichiamo, recuperiamo e descriviamo i dati**
  - Individuazione dei dati rilevanti per creare il modello
  - Creazione di report evidenziando fonti di dati e criteri di scelta

Es: *I dati di interesse vanno estratti dal database principale, dalla piattaforma web analytics e dalle statistiche trimestrali sull'andamento del nostro settore di mercato…*

- **I dati sono di qualità?**
  - Identificazione dati mancanti o inutili
  - Identificazione anomalie (outliers)

Es: I dati delle carte di credito non servono.
Dobbiamo acquisire i dati storici dei prezzi della concorrenza.
I clienti che hanno acquistato solo tramite buono regalo non si considerano…
NB: il vero utente da curare è quello che ha regalato il buono regalo.

Preparazione dei dati



- **Selezioniamo puntualmente i dati utili**
  - Scelta delle singole tabelle e dei campi di interesse
- **Puliamo i dati e integriamoli**
  - Pulizia dei dati e integrazione con dati esterni in base alle verifiche di qualità
  - Normalizzazione e uniformazione dei dati

Es: *Eliminiamo dati duplicati, convertiamo le valute in € e le date in formato ISO. Aggiungiamo dati mancanti anche acquistandoli da banche dati specializzate. Troviamo un buon valore di default per i campi NULL.*

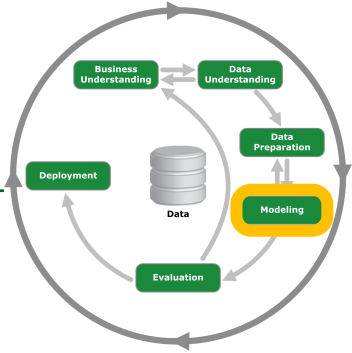
- **Ricaviamo dati per l'analisi**
  - Costruzione di dati derivati da quelli disponibili

Es: *Dividere i clienti per regioni di residenza (dovremo reperire le associazioni comune-regione).*

Nota sulla diapositiva precedente

I passi di preparazione dei dati possono spesso essere automatizzati e codificati in un data warehouse.

Costruzione del modello



- **I dati sono pronti per l'algoritmo scelto?**
  - Eventuale preparazione supplementare dei dati secondo i requisiti dell'algoritmo scelto
- **Come configuriamo i parametri di base dell'algoritmo?**
- **Quali dati scegliamo come *training set* e *test set*?**
  - Training set: ampio insieme di dati per scegliere e calibrare l'algoritmo
  - Test set: insieme di dati su cui si metterà alla prova il modello
- **Partenza della fase di training dell'algoritmo**
  - L'algoritmo analizza le relazioni nascoste nei dati

Gli insiemi di training e di test

Esempio: vogliamo cercare relazioni tra gli acquisti degli articoli 1, 2, 3 e l'acquisto dell'articolo 4.

Sappiamo cosa ha acquistato ciascun cliente.

Prendiamo a caso due insiemi di dati:

Training set (2/3 dei dati totali)

Test set (1/3 tot., metà Training set)

Dati noti (1=vero; 0=falso)				
Cliente	Ha acquistato			
	art. 1	art. 2	art. 3	art. 4
C1	1	1	1	1
C2	0	1	0	1
C3	1	0	1	1
C4	0	0	0	0
C5	1	1	1	0
C6	1	1	0	1
C7	1	1	1	1
C8	0	1	0	0
C9	1	0	1	1

Il Training set verrà usato per addestrare gli algoritmi candidati a creare il modello.

Il Test set servirà a verificare l'efficacia dei modelli candidati, per scegliere il migliore.

Nota sulla diapositiva precedente

Utilizzare gli eventi di acquisto degli articoli 1, 2, 3, come indicatori della probabilità che venga acquistato l'articolo 4 è fatto a scopo esemplificativo.

Anziché l'acquisto di specifici articoli, potremmo avere altri fattori come l'età, il sesso, la regione di residenza.

Gli insiemi di training e di test

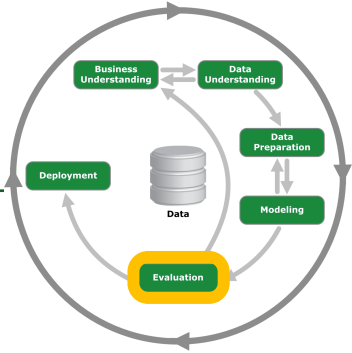
Training set con i risultati previsti dai modelli.

Dati noti (1=vero; 0=falso)					Previsioni	
Cliente	Ha acquistato				Modello A	Modello B
	art. 1	art. 2	art. 3	art. 4	art. 4	art. 4
C1	1	1	1	1	1	0
C2	0	1	0	1	1	1
C3	1	0	1	1	1	1
C4	0	0	0	0	0	0
C5	1	1	1	0	1	0
C6	1	1	0	1	1	1
				Errori	1 su 6	1 su 6

Test set per mettere alla prova i modelli.

Dati noti (1=vero; 0=falso)					Previsioni	
Cliente	Ha acquistato				Modello A	Modello B
	art. 1	art. 2	art. 3	art. 4	art. 4	art. 4
C7	1	1	1	1	1	0
C8	0	1	0	0	1	1
C9	1	0	1	1	1	1
				Errori	1 su 3	2 su 3

Valutazione del modello

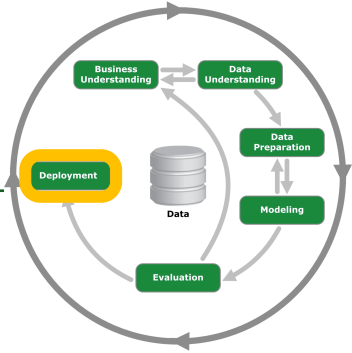


- **Siamo soddisfatti dei risultati?**
  - Il modello perfetto non esiste
    Si sceglie il migliore tra tutti quelli testati

- **Ci siamo dimenticati qualcosa?**
  - Revisione di tutte le fasi del processo

- **“La accendiamo?”**
  - Decisione finale sull’attuazione del modello

Attuazione del modello



- **Mettiamo in atto il modello**
  - Si effettuano le azioni suggerite dal modello
  - Si integra il modello con i sistemi esistenti

Es: *Progettiamo una nuova campagna di offerte ad personam.*
Automatizziamo l'emissione della campagna configurando i sistemi in uso.

- **Come faremo le verifiche e la manutenzione?**
  - Pianificazione delle verifiche periodiche
  - Pianificazione e assegnazione dei compiti di manutenzione
- **Revisione e documentazione**
  - Redazione di documentazione dell'intero processo
  - Revisione con la partecipazione degli utenti

Alcune strutture e algoritmi per il Data Mining

Seguono alcuni esempi di tecniche usate nel data mining.

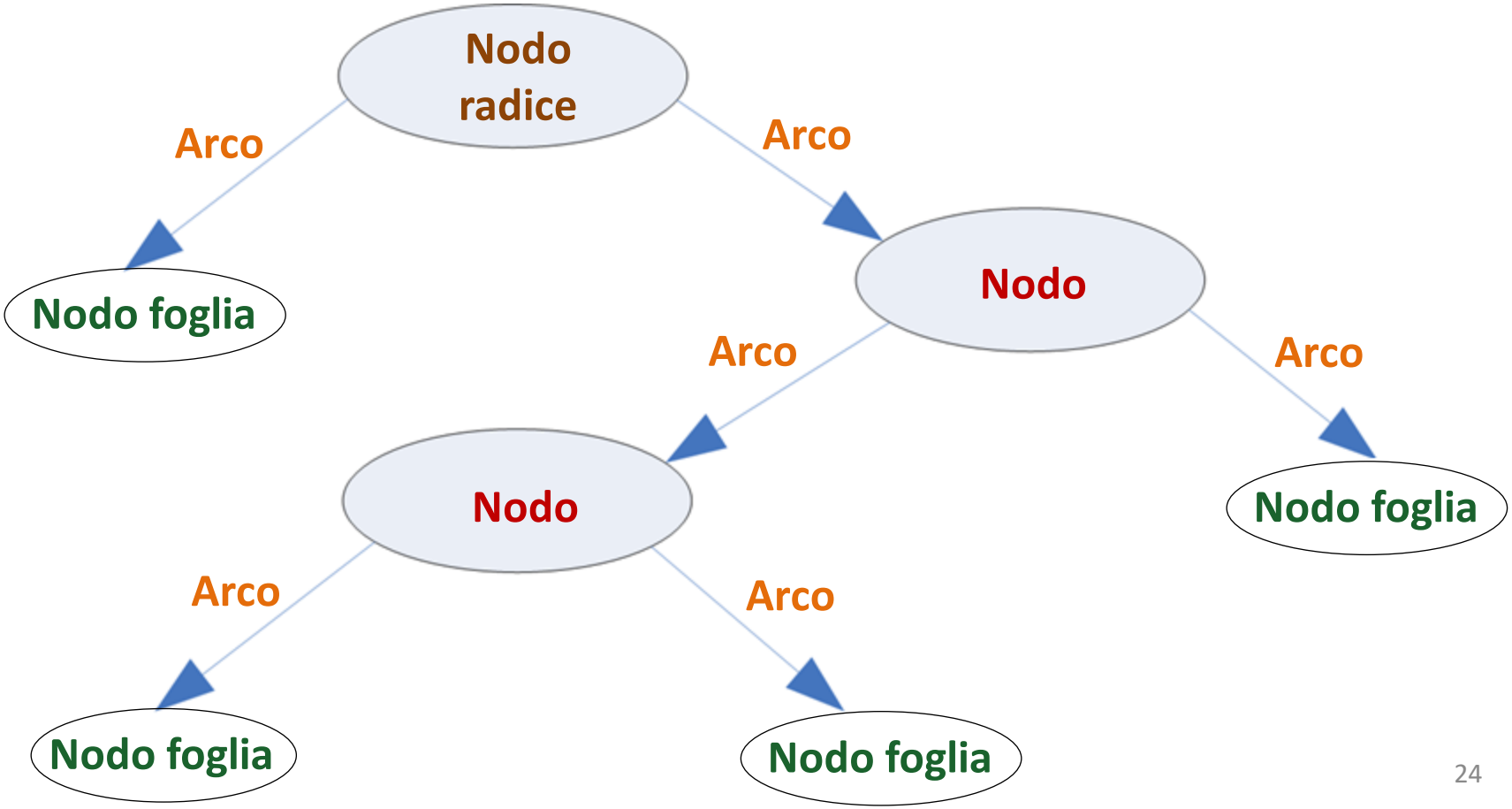
Ciascuna costituisce l'ossatura di un modello.

Una tecnica può essere usata da molti algoritmi, ciascuno dei quali costruirà un modello alternativo.

Alberi decisionali

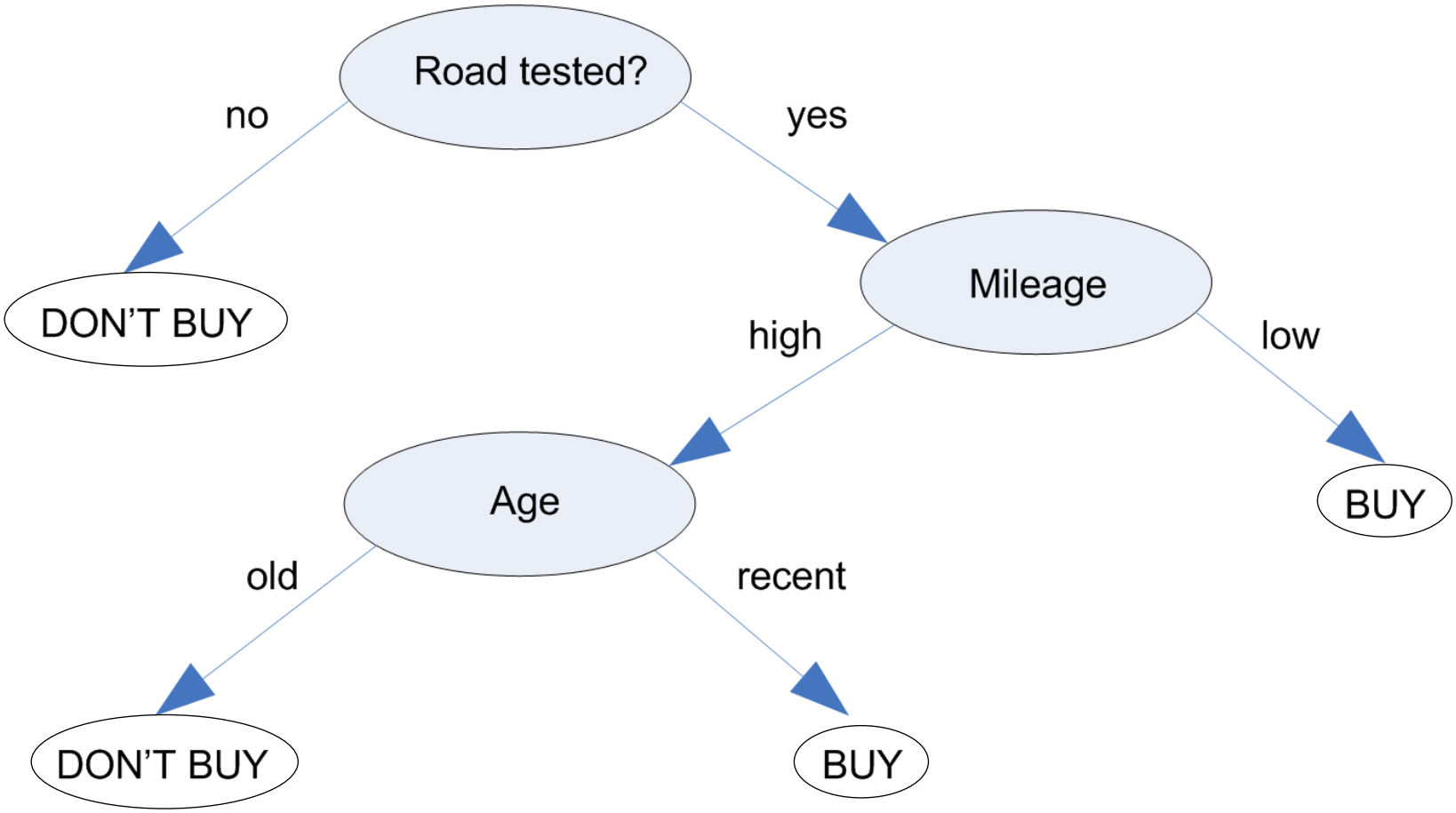
Cos'è un albero?

Un albero è un insieme di nodi e archi, in cui due nodi qualsiasi sono uniti da un solo cammino.



Alberi decisionali

Quest'albero rappresenta un modello semplice ma valido per **decidere se acquistare un'auto usata**.
Si inizia dalle condizioni più selettive, per decidere ponendosi il minor numero di domande.



Alberi decisionali

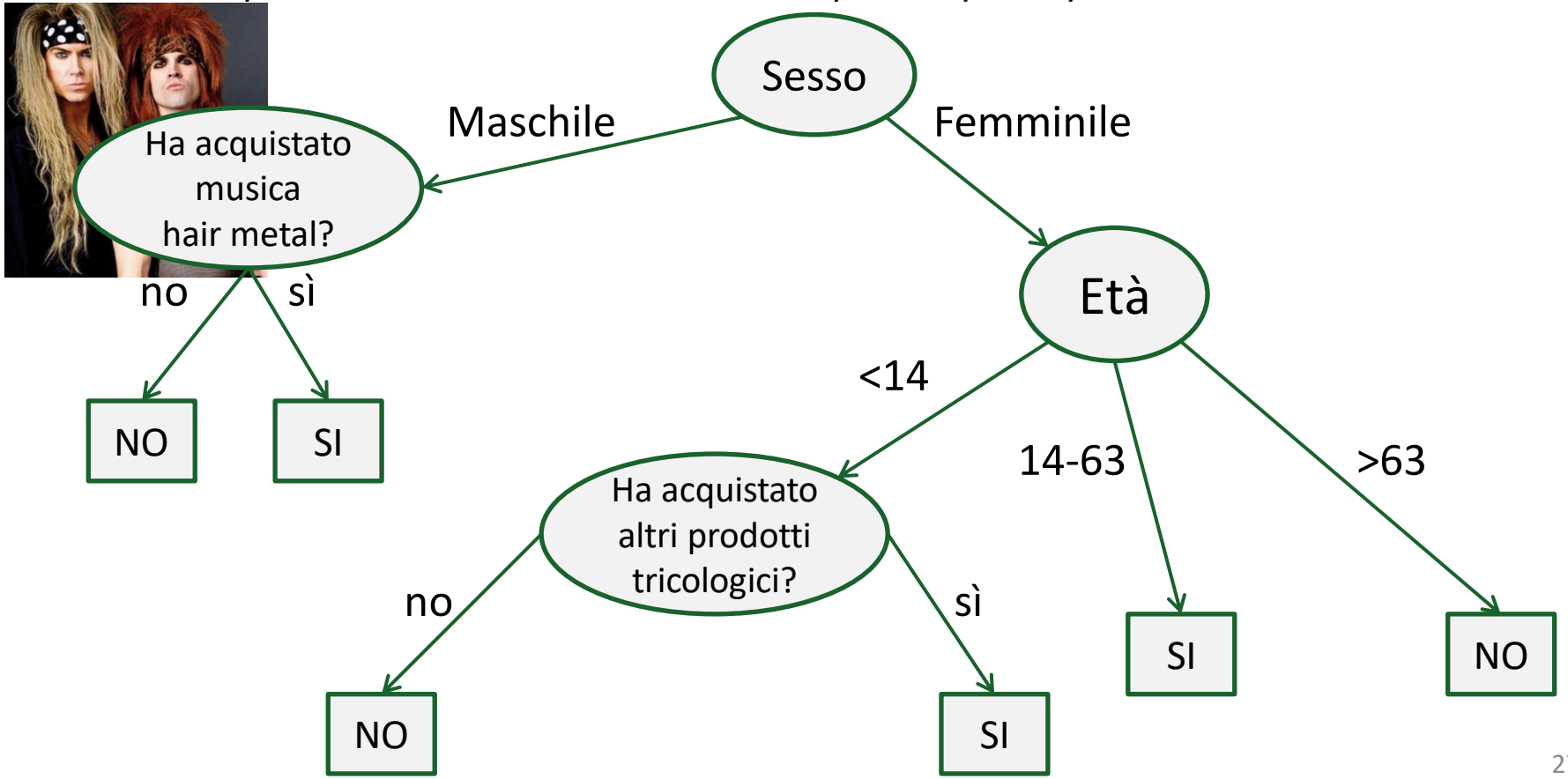
Principali campi di applicazione

- Ricerche di mercato
- Segmentazioni
- Pianificazione aziendale
- Modellazione degli investimenti
- Valutazione delle categorie di rischio
- Valutazioni di casi clinici
- Modelli epidemiologici
- Andamento della borsa
- Decisioni make or buy

Alberi decisionali di classificazione

Permettono di predire comportamenti e azioni.

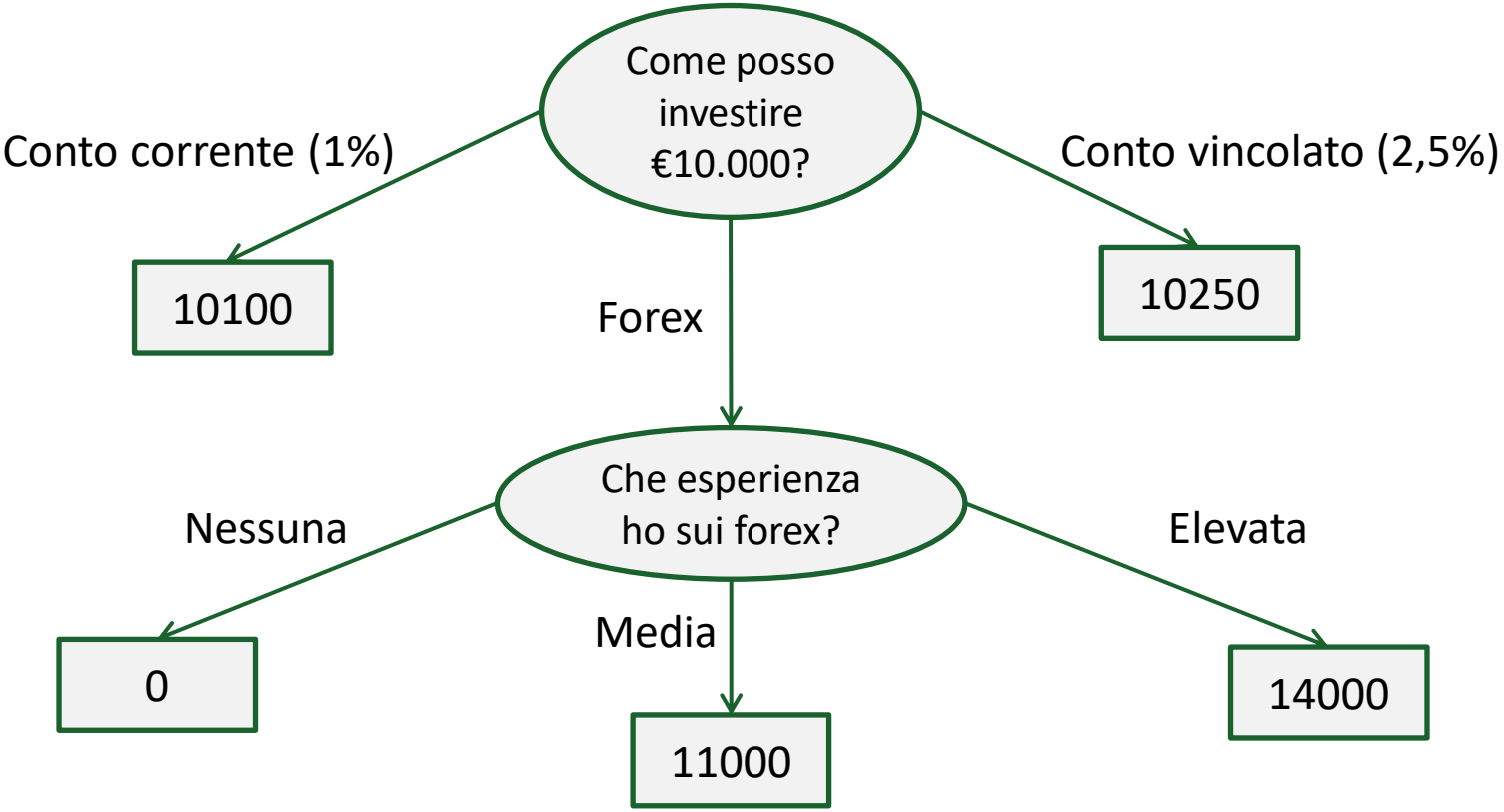
Quali possibili malattie indicano i sintomi di un paziente?
Un debitore sarà in grado di pagare?
Un cliente potrebbe essere interessato a una piastra per capelli?



Alberi decisionali di regressione

Permettono di predire il valore di una variabile quantitativa.

Quanto costa licenziare una persona?
Che valore ha un'auto usata?
Che guadagni possiamo ottenere in un anno investendo 10.000 €?



Costruzione di alberi decisionali

L'idea: creiamo un algoritmo che costruisca un albero decisionale a partire da una serie di dati conosciuti (training set).

L'obiettivo: creare un albero delle giuste dimensioni:
troppo grande => non generalizza abbastanza
troppo piccolo => approssima troppo

I passi:

1. Individuare le variabili e i rispettivi valori soglia che permettono di partizionare il data set (splitting dei nodi).
   Es: Età {<14; 14-63; >63}
2. Determinare se dividere un nodo o renderlo terminale (foglia).
3. Assegnare etichette ad ogni nodo foglia: cosa fare se i dati mi portano a quel nodo? (unanimità, maggioranza, indecisione)

Alcuni algoritmi per alberi decisionali

CART (Classification And Regression Trees) è l'algoritmo più diffuso per costruire alberi di classificazione e regressione.

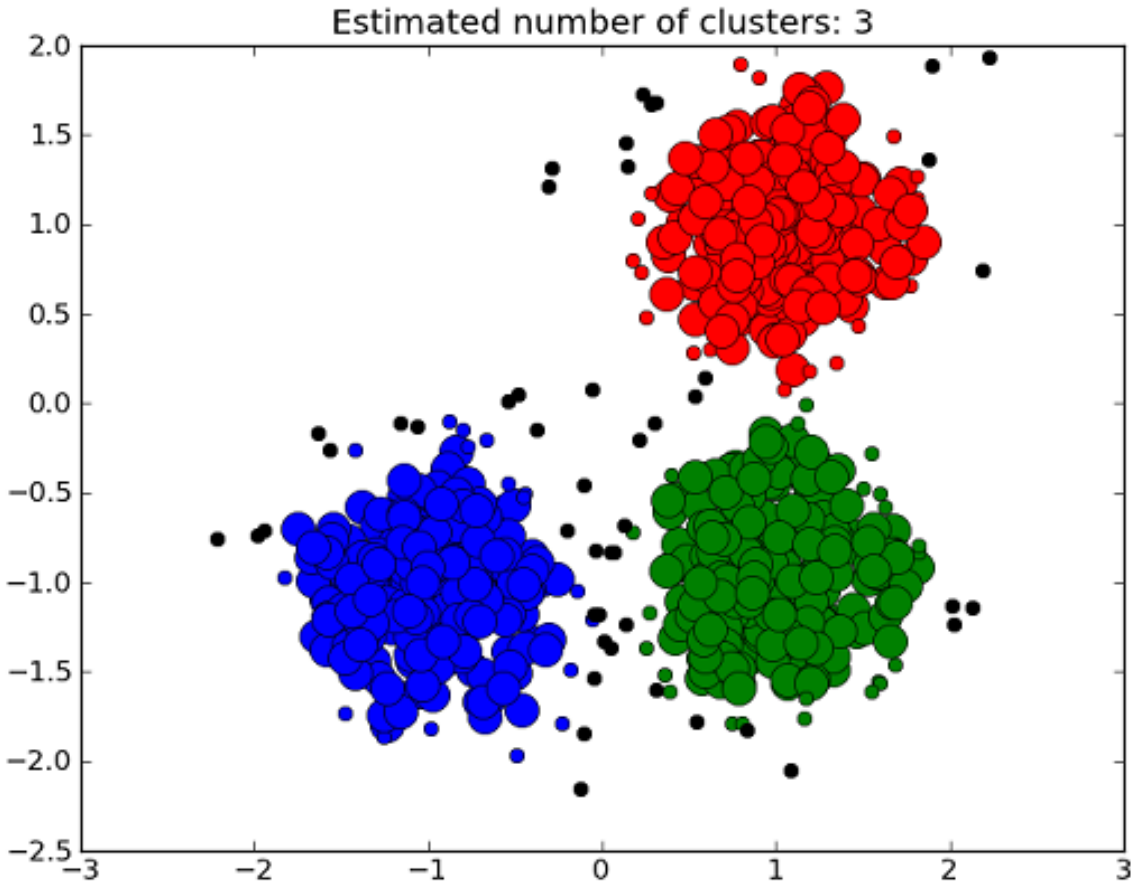
Funziona con alberi binari (ogni nodo ha massimo 2 archi uscenti).

Altri algoritmi:

- ID3
- C4.5 (Successore di ID3)
- CHAID
- MARS

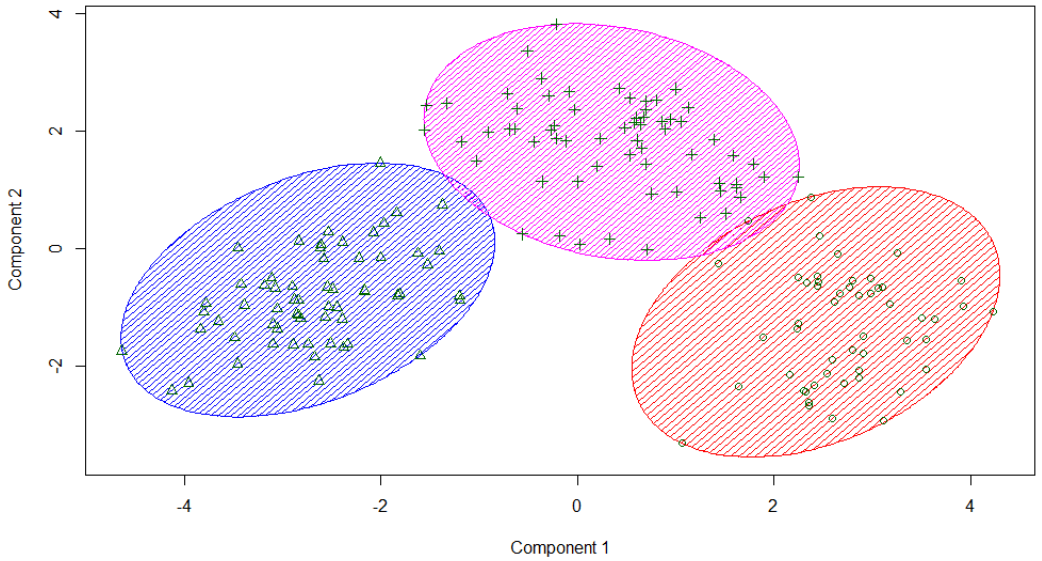
Clustering: analisi dei gruppi

La tecnica del clustering prevede di classificare elementi raggruppandoli a seconda di caratteristiche comuni.



Clustering

- Analisi multivariata (2+ variabili)
- Elementi raggruppati in base alla distanza reciproca



Due approcci:

- Aggregativo (bottom-up)

Si uniscono cluster vicini.

- Divisivo (top-down)

Si dividono elementi lontani dello stesso cluster.

Clustering

Principali campi di applicazione

- Segmentazioni
- Posizionamento prodotti
- Sviluppo di nuovi prodotti
- Riconoscimento prodotti o news uguali sul web
- Riconoscimento gruppi su social network
- Raggruppamento risultati utili su motore di ricerca
- Regolazione visibilità marcatori su mappe digitali
- e ancora…

  Biologia, Medicina, Scienze naturali, Informatica, ecc.

Alcuni algoritmi di clustering

K-means è uno dei più noti algoritmi per l'individuazione di gruppi.

K è il numero di gruppi che si intende creare, ed è un parametro che va fornito all'algoritmo. La qualità dei risultati dipende fortemente dalla scelta di K, che può essere fatta a mano o con degli algoritmi.

Altri algoritmi di clustering

- K-medoids
- Fuzzy
- Quantum

https://en.wikipedia.org/wiki/Category:Cluster_analysis_algorithms

Costruzione del modello

- **Quale tecnica applicare?** Alcuni esempi:

Problema	Esempio	Tecnica
Stima attributo discreto	*Stimare se il destinatario di una campagna di mailing diretta acquisterà un prodotto, sulla base di vari dati comportamentali e anagrafici*	**Alberi decisionali** **Clustering** Classificatori Bayesiani Reti neurali
Stima attributo continuo	*Stimare le vendite del prossimo anno*	Serie storiche Reti neurali
Ricerca gruppi di elementi comuni nelle transazioni	*Suggerire a un cliente prodotti da acquistare tramite analisi di mercato*	**Alberi decisionali** Regole di associazione
Ricerca di gruppi di elementi simili	*Segmentare i clienti per comportamenti di acquisto simili*	**Clustering**
Ricerche di anomalie nei dati	*Individuare usi fraudolenti di strumenti di pagamento*	**Clustering**

Software per il Data Mining

Esistono decine di software, gratuiti o a pagamento, per ricerche di data mining.

- **Accettano dati** (e possono spesso dividerli in insiemi di training e test).
- **Incorporano algoritmi**, che si possono configurare ed eseguire sui dati forniti per creare modelli.
- **Mostrano i risultati** della creazione dei modelli, con più o meno ricchezza grafica.

I software più completi (e spesso più costosi) offrono strumenti di statistica avanzati, algoritmi migliorati e più personalizzabili, strumenti per gestire l'intero processo KDD e data warehouse.

I principali software gratuiti:
https://www.predictiveanalyticstoday.com/top-free-data-mining-software/

Altri software:
https://financesonline.com/top-15-data-mining-software-systems/