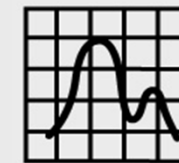


Breve Corso sul Data Mining



StatSoft®

STATISTICA

data analysis • data mining • quality control • web-based analytics

StatSoft Italia • via Parenzo 3 • 35010, Vigonza (PD) • Tel : 049 8934654 • Fax: 049 8932897 • info@statsoft.it • www.statsoft.it

Australia: StatSoft Pacific Pty Ltd.

Brazil: StatSoft South America

Bulgaria: StatSoft Bulgaria Ltd.

Czech Rep.: StatSoft Czech Rep. s.r.o.

China: StatSoft China

France: StatSoft France

Germany: StatSoft GmbH

Hungary: StatSoft Hungary Ltd.

India: StatSoft India Pvt. Ltd.

Israel: StatSoft Israel Ltd.

Italy: StatSoft Italia srl

Japan: StatSoft Japan Inc.

Korea: StatSoft Korea

Netherlands: StatSoft Benelux BV

Norway: StatSoft Norway AS

Poland: StatSoft Polska Sp. z o.o.

Portugal: StatSoft Ibérica Lda

Russia: StatSoft Russia

Spain: StatSoft Ibérica Lda

S. Africa: StatSoft S. Africa (Pty) Ltd.

Sweden: StatSoft Scandinavia AB

Taiwan: StatSoft Taiwan

UK: StatSoft Ltd.

Sommario

Panoramica Introduttiva di Data Mining

- Cos'è il Data Mining?
- Modelli per il Data Mining
- Fasi del Data Mining
- Panoramica sulle tecniche di Data Mining
- Punti da Ricordare

Cos'è il Data Mining?

- Il ricorso al Data Mining diviene necessario quando costosi problemi di business (produttivi, tecnici, ecc.) non hanno soluzioni ovvie
 - Ottimizzazione di un processo produttivo o di un prodotto.
 - Individuazione di transazioni fraudolente.
 - Valutazione del rischio.
 - Segmentazione della clientela.

Deve essere trovata una soluzione:

- Nega che il problema esista.
- Consulta un astrologo.
- Usa il data mining.



Nota: Noi raccomandiamo questo approccio ...

Cos'è il Data Mining?

- **Data mining** è un processo analitico, progettato per consentire l'esplorazione di grandi moli di dati alla ricerca di tendenze e/o relazioni sistematiche tra le variabili, e, in seguito, per convalidare le scoperte tramite l'applicazione dei comportamenti rilevati su nuovi insiemi di dati.
- **Data mining** è un processo chiave per massimizzare il valore dei dati raccolti dall'azienda.
- **Data mining** viene utilizzato per
 - Individuare tendenze nelle transazioni fraudolente, nelle richieste assicurative, ecc.
 - Individuare eventi e comportamenti ripetitivi
 - Modellare i comportamenti della clientela per effettuare operazioni di marketing
 - Ottimizzare la qualità dei prodotti e i processi produttivi
 - **Il data mining può essere utilizzato da qualsiasi azienda che desideri utilizzare le possibili informazioni nascoste nei dati, quando queste conoscenze possono far aumentare il valore del proprio business.**

Cos'è il Data Mining?

Gli obiettivi più comuni dei progetti di data mining sono:

- Identificare **gruppi, strati o dimensioni** presenti in un insieme di dati che ha mostra una struttura non immediatamente comprensibile,
- Individuare fattori che sono collegati ad un particolare fenomeno di interesse (**root-cause analysis**)
- **Prevedere con precisione** delle variabili risposta di interesse (in termini di nuova clientela, nuovi candidati, ecc.; questa tipologia è nota di solito con il nome di data mining predittivo)

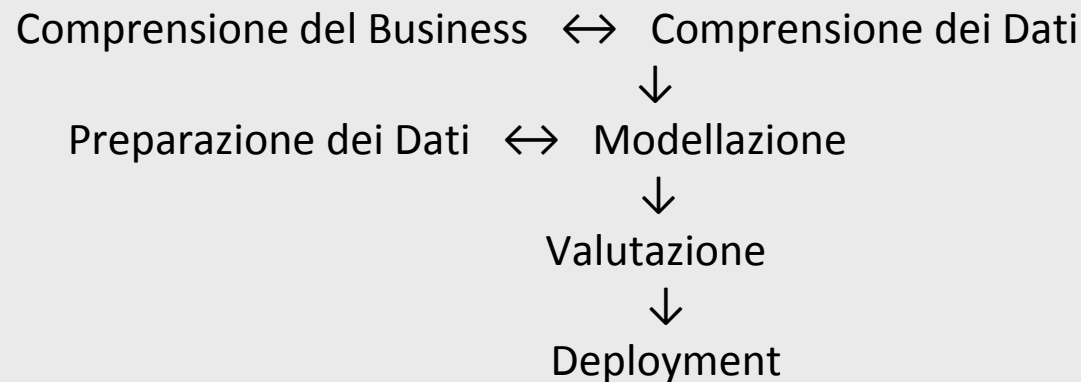
Cos'è il Data Mining?

- Il data mining è uno strumento, non una bacchetta magica.
- Il data mining non trova automaticamente delle soluzioni senza una guida.
- Il data mining non risiede all'interno dei vostri database e non invia un'email ogni volta che scopre qualcosa di interessante.
- Il data mining potrebbe trovare delle tendenze interessanti, ma non ne comunica immediatamente il valore.
- Il data mining non trova delle relazioni di causalità.
 - Per esempio, si potrebbe scoprire che gli uomini con un certo reddito e che fanno esercizio regolarmente sono dei buoni consumatori di un determinato prodotto. Tuttavia la conoscenza di questi fattori, non indica che l'acquisto di tale prodotto sia sistematico, ma solo che esiste questa relazione.

Modelli per il Data Mining

La letteratura sul data mining propone diverse “strutture generali” per l’organizzazione del processo di raccolta dei dati, di analisi dei dati, di diffusione dei risultati, di implementazione dei risultati e di controllo dei miglioramenti.

- **CRISP**: progettato a metà degli anni 90 da un consorzio di aziende europee, serve da modello standard non-proprietario dei processi per il data mining.



- **DMAIC** Metodologia *Six Sigma* – metodologia guidata-dai-dati per l’eliminazione dei difetti, degli sprechi o dei problemi di qualità di tutti i tipi.

Definisci → Misura → Analizza → Migliora → Controlla

- **SEMMA** (SAS Institute) – focalizzato su più aspetti tecnici del data mining.
Campiona → Esplora → Modifica → Modella → Verifica

Fasi del Data Mining

Fase 0: Definizione del problema.

- Prima di aprire il software ed eseguire l'analisi, l'analista deve sempre aver chiari gli interrogativi a cui desidera rispondere. Se non si dispone di una formulazione precisa del problema che si sta provando a risolvere, è sempre meglio non sprecare tempo e denaro.

Fase 1: Esplorazione iniziale.

- Questa fase inizia con la preparazione dei dati, che coinvolge la "pulitura" dei dati (ad es., per l'identificazione e la rimozione di dati codificati in modo errato, ecc.), la trasformazione dei dati, la selezione di sottoinsiemi di record, e, nel caso di insiemi di dati con un grande numero di variabili ("campi"), l'esecuzione preliminare di una selezione delle caratteristiche. La descrizione e la visualizzazione dei dati sono componenti chiave di questa fase (ad es., statistiche descrittive, correlazioni, scatterplot, box plot, ecc.).

Fase 2: Costruzione e convalida dei modelli.

- Questa fase prevede la valutazione di diversi modelli e la scelta del migliore sulla base delle rispettive performance predittive.

Fase 3: Deployment.

- Quando l'obiettivo del progetto di data mining è prevedere o classificare nuovi casi (ad es., per prevedere la solvibilità dei singoli richiedenti un prestito), la fase finale in genere prevede l'applicazione del modello o dei modelli migliori (individuati nella fase precedente) con l'obiettivo di generare previsioni.

Fase 1: Esplorazione iniziale

- **“Pulitura” dei dati,**
 - Identificazione e rimozione di dati codificati in modo scorretto, Maschio=Sì, Incinta=Sì

- **Trasformazione dei dati,**

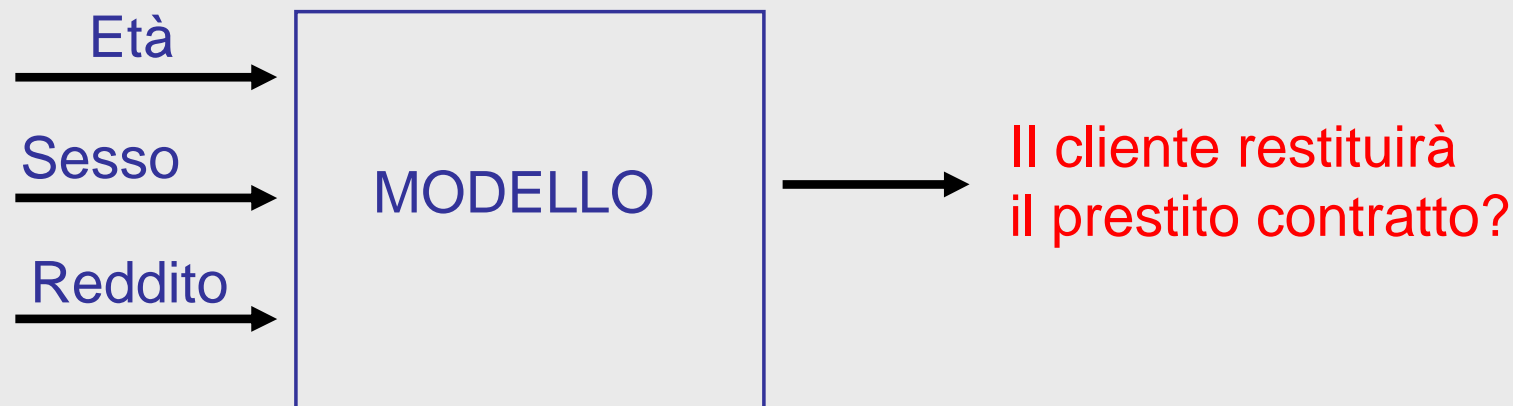
I dati potrebbero essere asimmetrici (ossia, potrebbero essere presenti outlier in una direzione o nell'altra). Trasformazioni Logaritmiche, Box-Cox, ecc.

- **Riduzione dei dati,** Selezione di sottoinsiemi di record, e, nel caso di dataset con un grande numero di variabili (“campi”), esecuzione preliminare di una selezione delle caratteristiche.

- **Descrizione e visualizzazione dei dati** sono componenti chiave di questa fase (ad es., statistiche descrittive, correlazioni, scatterplot, box plot, strumenti di brushing, ecc.)
 - La descrizione dei dati consente di ottenere una fotografia delle caratteristiche importanti dei dati (come ad esempio la tendenza centrale e le misure di dispersione).
 - Le tendenze sono spesso più facili da individuare visivamente che attraverso liste e tabelle numeriche.

Fase 2: Costruzione e validazione dei modelli

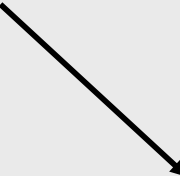
- Il Data Mining prevede la creazione di modelli della realtà
- Un modello riceverà uno o più input e produrrà uno o più output



- Un modello può essere “trasparente”, ovvero sia composto da una serie di considerazioni “se/allora” in cui la struttura è facilmente derivabile. Viceversa un modello può essere visto come una scatola nera, quando ad esempio si ricorre alle reti neurali, la cui struttura o le regole che governano le previsioni sono impossibili da comprendere totalmente.

Fase 2: Costruzione e validazione dei modelli

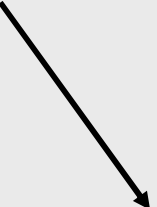
- Un modello può solitamente essere valutato secondo 2 criteri:
 - Accuratezza
 - Comprensibilità
- Questi aspetti talvolta sono in conflitto tra loro.
- Gli alberi decisionali ed i modelli di regressione lineare sono meno complicati e più semplici di modelli quali le reti neurali, gli alberi boosted, ecc., e più facili da comprendere, anche se si potrebbe essere costretti a rinunciare ad una maggiore precisione predittiva.
- Si ricordi di non confondere il modello di data mining con la realtà (la mappa di una strada non è mai la perfetta rappresentazione della strada), nonostante possa essere utilizzato come utile guida.



Generalizzazione è la capacità di un modello di creare previsioni accurate in presenza di dati non estratti dall'insieme di dati di addestramento originale .

Fase 2: Costruzione e validazione dei modelli

- La validazione richiede l'addestramento del modello su un insieme di dati e la sua successiva valutazione su un insieme di dati diverso.
- Esistono due i metodi di validazione principali
 - Suddivisione dei dati in sottoinsiemi di addestramento/test (75% vs 25%)
 - Convalida incrociata v-fold, soprattutto nel caso in cui non si disponga di grandi quantità di dati.



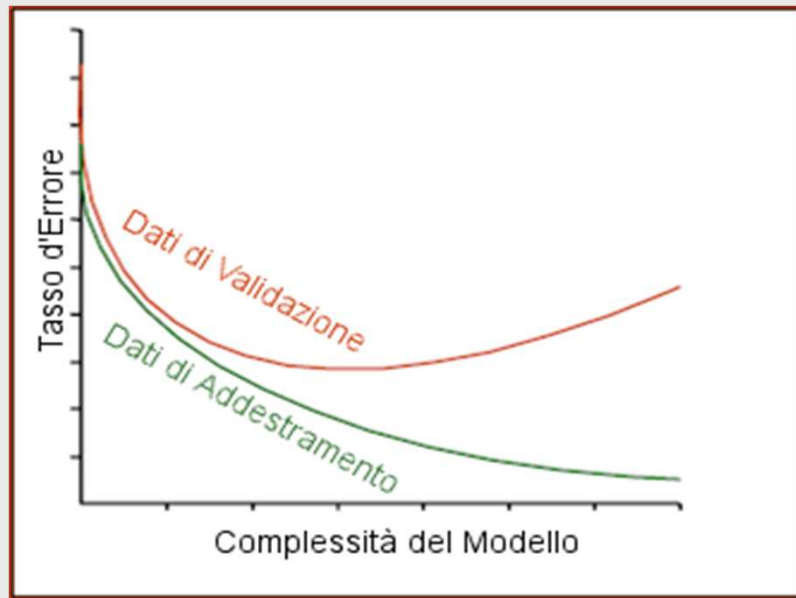
Nella **convalida incrociata v-fold** si estraggono v campioni casuali dai dati dell'analisi, e il rispettivo modello viene utilizzato per calcolare i valori previsti, le classificazioni, ecc. In genere, gli indici riassuntivi della precisione della previsione sono calcolati per tutte le v repliche; quindi questa tecnica consente all'analista di valutare l'accuratezza generale del rispettivo modello predittivo nei campioni casuali ripetutamente estratti.

Fase 2: Costruzione e validazione dei modelli

- Se non si dispone di un numero sufficiente di dati, occorre ricorrere alla convalida incrociata v-fold.



Fase 2: Costruzione e validazione dei modelli



In generale, il termine **sovra-adattamento** fa riferimento alla condizione in cui un particolare modello predittivo è così "specifico" da riuscire a riprodurre le diverse variazioni casuali che caratterizzano i dati dal quale sono stati stimati i parametri; quindi, tali modelli potrebbero non riuscire più a prevedere accuratamente nuove osservazioni (ad es., durante il deployment di un progetto di data mining predittivo). Per evitare il **sovra-adattamento** vengono solitamente applicate tecniche quali la convalida incrociata e la convalida incrociata v-fold.

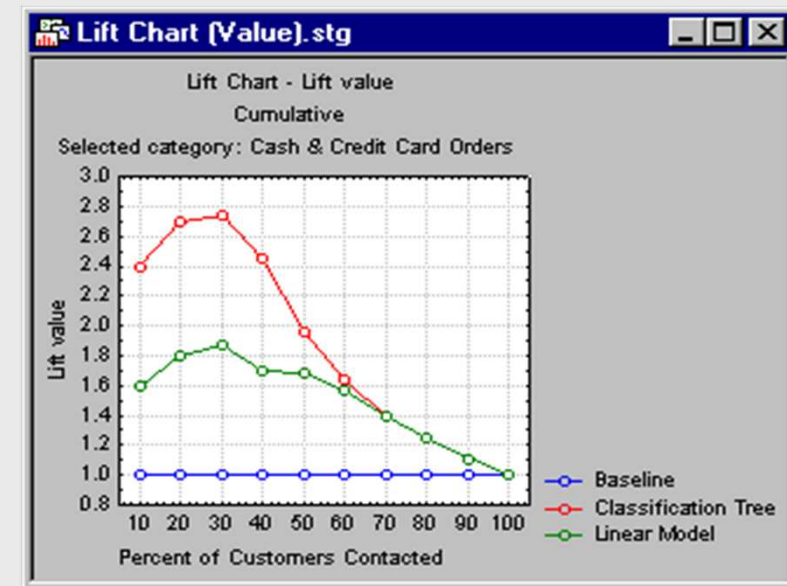
- Quanto è predittivo il modello?
 - Due misure sono date dalle somme degli errori al quadrato (regressione) o dalla matrice di confusione (classificazione)
- L'errore dato dai dati di addestramento non è un buon indicatore della prestazione su dati futuri
 - I nuovi dati probabilmente non saranno esattamente uguali a quelli utilizzati durante l'addestramento!
- **Sovra-adattamento:** Adattare il modello troppo ai dati di addestramento, solitamente, porta a prestazioni predittive scarse per dati diversi.

Fase 2: Costruzione e validazione dei modelli

Misurare la Qualità del Modello

- Possibili misure di validazione
 - Accuratezza della classificazione
 - Costi/benefici totali – quando differenti tipi di errore comportano costi differenti
 - Curve Lift e Gain
 - Errore delle previsioni numeriche

- Tassi di errore
 - Proporzione degli errori sull'intero insieme d'istanze
 - Tasso d'errore sull'insieme di addestramento: approccio oltremodo ottimistico!
 - È possibile individuare tendenze anche nei dati casuali



La **lift chart** fornisce un riassunto grafico sull'utilità dell'informazione fornita da uno o più modelli statistici per la previsione di una variabile dipendente binomiale (categoriale); per le variabili dipendenti multinomiali (a categorie multiple), le lift chart possono essere prodotte per ogni categoria. In particolare, il grafico riassume l'utilità che un utente potrebbe attendersi dall'impiego dei rispettivi modelli predittivi rispetto alla sola informazione di base.

Fase 3: Deployment

- Il modello viene costruito una sola volta, ma può essere utilizzato in più di un'occasione.
- Si dovrebbe effettuare il deployment del modello in modo semplice.
 - La regressione lineare è un tipico esempio di modello semplice, in quanto è facile calcolare i coefficienti...
 - Per esempio, in presenza di un nuovo vettore di dati osservati {x1, x2, x3}, basterà semplicemente inserire tale vettore nell'equazione lineare, in modo da ottenere il valore previsto,

$$\text{Previsione} = B0 + B1 * X1 + B2 * X2 + B3 * X3$$

- Anche gli Alberi di Classificazione e di Regressione sono considerati modelli semplici sui quali effettuare il deployment: sono costituiti da una serie di If/Then/Else ...

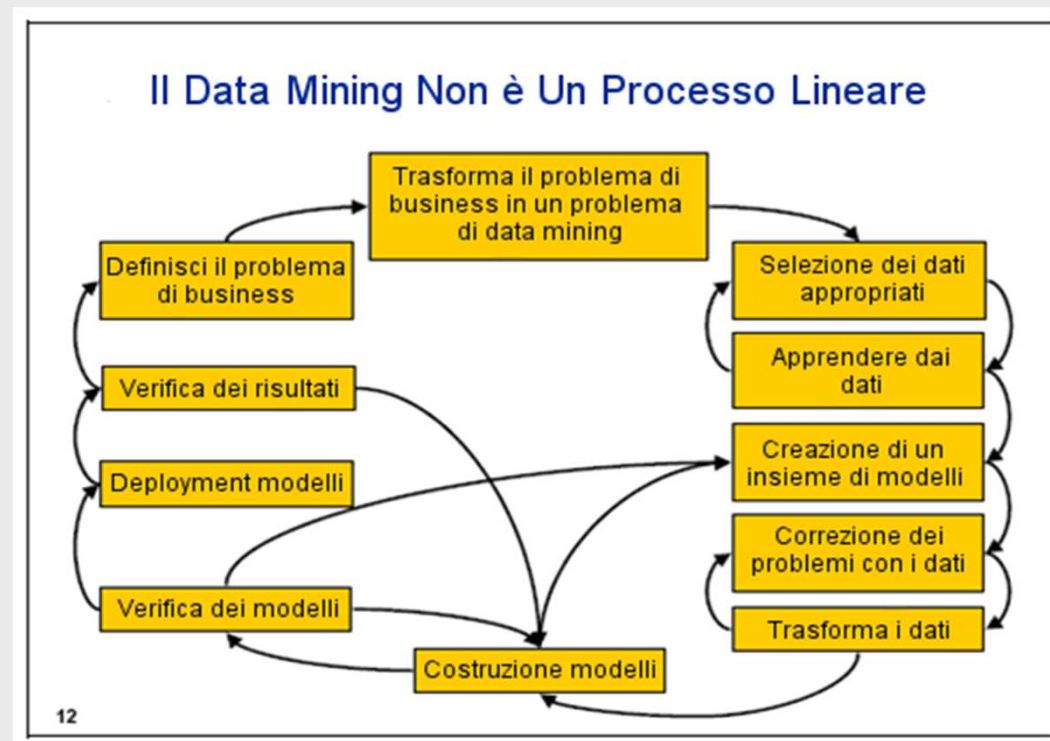
Fasi del Data Mining

Fase 0: Definizione precisa del problema.

Fase 1: Esplorazione iniziale.

Fase 2: Costruzione e convalida del modello.

Fase 3: Deployment.



Panoramica sulle tecniche di Data Mining

- Apprendimento supervisionato
 - **Classificazione:** risposta categoriale
 - **Regressione:** risposta continua
 - **Serie Storiche:** analisi temporale delle osservazioni
 - **Ottimizzazione:** minimizzare o massimizzare alcune caratteristiche
- Apprendimento non-supervisionato
 - **Analisi delle Componenti Principali:** riduzione della dimensione del problema
 - **Analisi dei Gruppi:** raggruppamento di oggetti simili tra di loro
 - **Analisi di Associazioni e Collegamenti:** approccio descrittivo all'esplorazione dei dati, in modo da identificare delle relazioni tra i valori di un database, come ad esempio nella Market basket analysis: i clienti che acquistano i martelli acquistano anche i chiodi. Esaminare le probabilità condizionate.

Apprendimento Supervisionato:
Categoria di metodi di data mining, nei quali si utilizza un insieme di dati composto da variabili dipendenti ed indipendenti, con l'obiettivo di stimare un modello che verrà impiegato in seguito per il deployment.

Apprendimento non-supervisionato:
Metodo di data mining basato su dati dove non sono presenti delle variabili dipendenti.

Panoramica sulle tecniche di Data Mining

Le prossime diapositive si concentreranno sulle tecniche utilizzate più comunemente, senza soffermarsi sui metodi più complicati, in modo da fornire un panoramica globale sul data mining.

- Statistiche Descrittive
- Regressione Lineare e Logistica
- Analisi della Varianza (ANOVA)
- Analisi Discriminante
- Alberi Decisionali
- Tecniche di Analisi dei Gruppi (K Means & EM)
- Reti Neurali
- Analisi delle Associazioni e dei Collegamenti
- MSPC (Controllo Statistico di Processo Multivariato)

Svantaggi dei modelli non parametrici

- Alcuni algoritmi hanno bisogno di grandi quantità di dati...

Maledizione della dimensionalità è un'espressione coniata da Richard Bellman, e indica il problema derivante dal rapido incremento dello spazio (matematico), associato all'aggiunta di dimensioni.

Leo Breiman riporta come esempio il fatto che 100 osservazioni possano coprire piuttosto bene l'intervallo uni-dimensionale $[0,1]$ sulla retta reale. Si potrebbe quindi tracciare un istogramma dei risultati, e fare dell'inferenza su tali dati. Se si considera invece un ipercubo di 10 dimensioni, le 100 osservazioni saranno ora dei punti isolati, in uno spazio vasto e vuoto. Per ottenere una copertura simile a quella osservata sullo spazio unidimensionale sarebbero necessarie 10^{20} osservazioni: una quantità difficilmente ottenibile.

Il termine **maledizione della dimensionalità** (Bellman, 1961, Bishop, 1995) di solito si riferisce alle difficoltà incontrate durante la stima di modelli su degli spazi a multi-dimensionali. All'aumentare della dimensione dei dati di input (cioè, del numero di predittori), diventa esponenzialmente più difficile trovare i punti di ottimo globale per la stima dei parametri. Quindi, risulta necessario eseguire una preselezione delle variabili di input (predittori) sulla base della loro capacità predittiva, in modo da ridurre e contenere la dimensione del problema.

La maledizione della dimensionalità rappresenta un ostacolo significativo nei problemi che coinvolgono pochi dati ma numerose dimensioni (caratteristiche).

Si veda anche :

http://en.wikipedia.org/wiki/Curse_of_dimensionality

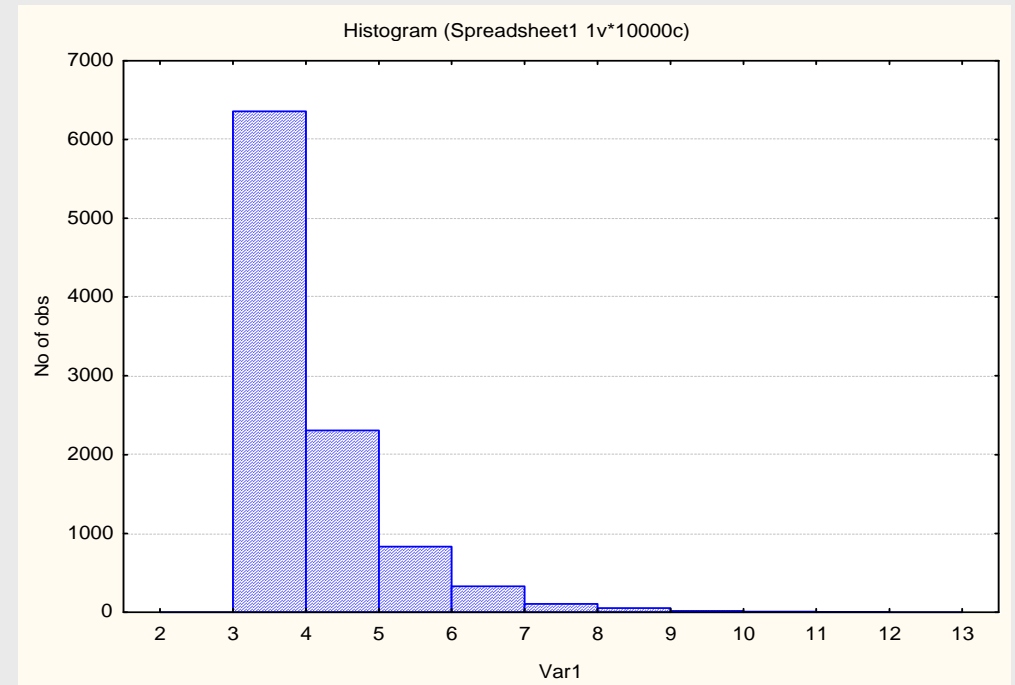
Statistiche Descrittive, ecc.

- Solitamente la quantità di dati, presente in un database, è tale da dover ricorrere a degli indici riassuntivi per ottenere delle indicazioni sul fenomeno in esame. La prima fase del Data Mining richiede la descrizione ed il riepilogo dei dati.
- Per descrivere la distribuzione dei dati vengono utilizzate due tipi di statistiche:
 - **Misure della Tendenza Centrale**
Media, Mediana, Moda
 - **Misure di Dispersione**
Deviazione Standard, Varianza
- Inoltre è importante costruire dei grafici. La visualizzazione grafica dei dati fornisce dei dettagli più immediati di quelli ricavabili tramite l'osservazione di una tabella di numeri o di statistiche.

Statistiche Descrittive, ecc.

Un istogramma è un modo molto semplice ed efficace di riassumere l'informazione contenuta in una colonna o variabile.

Si possono determinare in modo rapido il range (min e max), la media, la mediana, la moda, e la varianza della variabile.

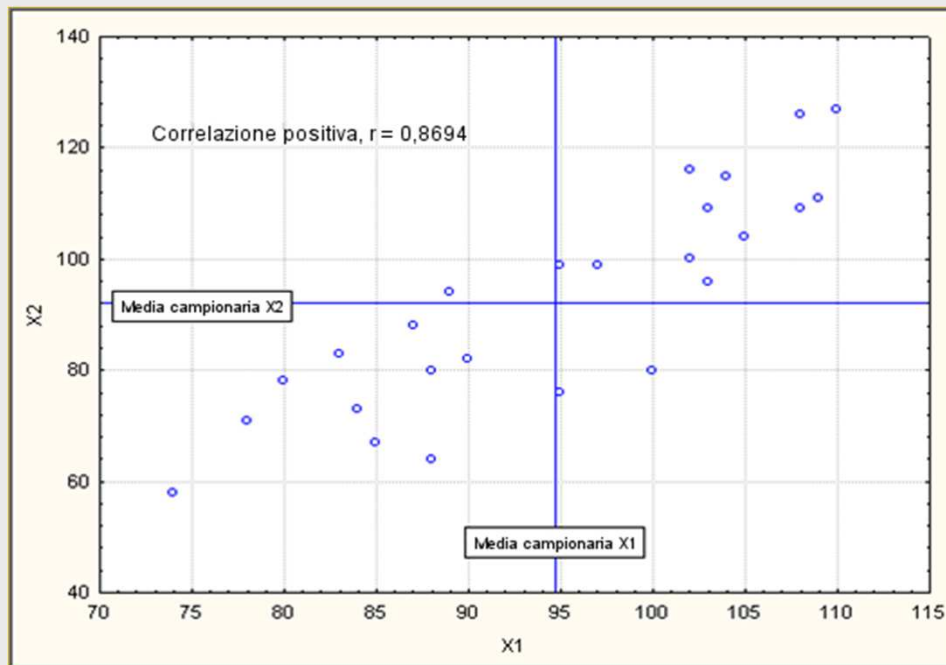


Variabile	Statistiche Descrittive (Spreadsheet1)						
	N Validi	Media	Mediana	Minimo	Massimo	Varianza	Dev.Std.
Var1	10000	3.992616	3.677125	3.000038	11.46572	0.995422	0.997708

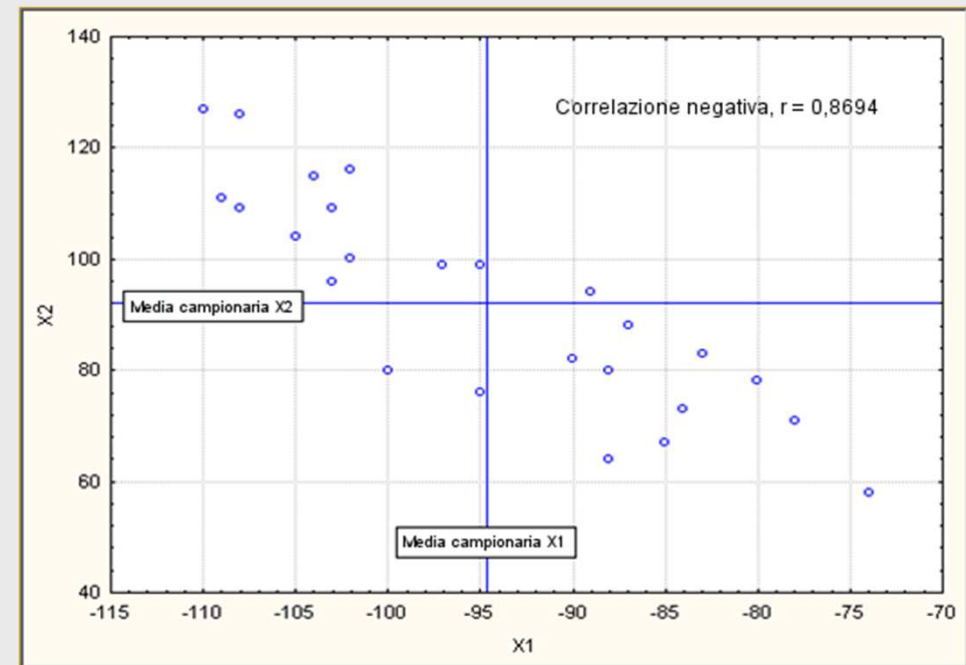
Statistiche Descrittive, ecc.

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Correlazione positiva



Correlazione Negativa



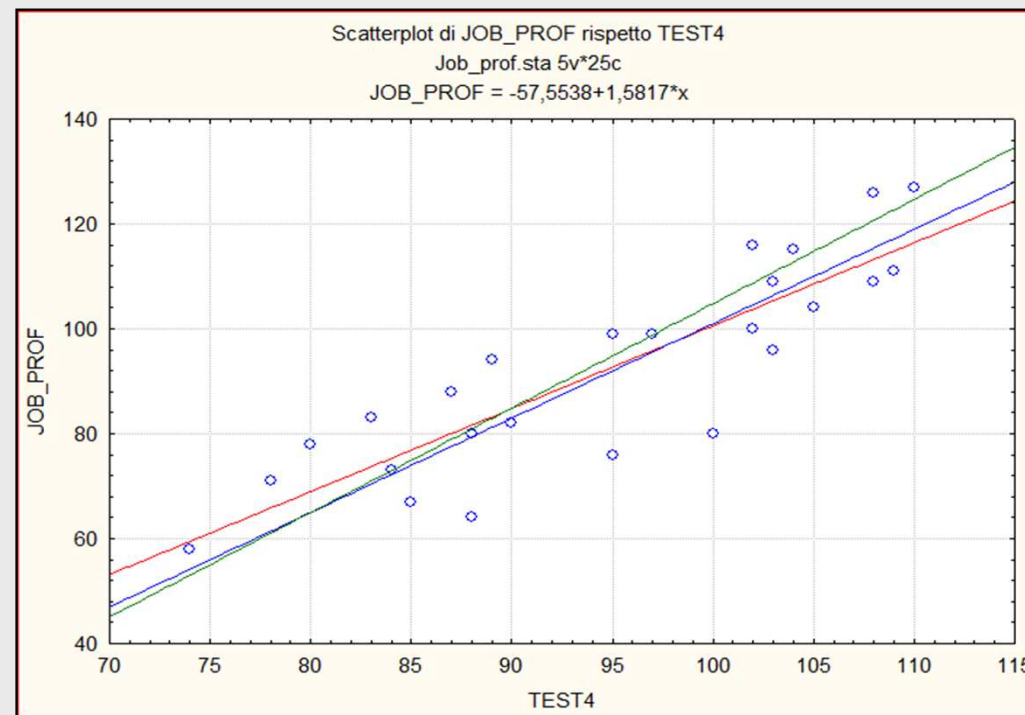
Regressione Lineare e Logistica

- La regressione è una metodologia statistica che utilizza la relazione tra due o più variabili quantitative, in modo da prevedere una variabile risposta tramite i valori delle altre.
- Si consideri in primis la regressione lineare, dove la variabile risposta è continua. Nella regressione logistica la risposta è invece dicotomica.
- Alcuni esempi:
 - L'ammontare delle vendite di un prodotto può essere prevista utilizzando la relazione esistente con gli investimenti in pubblicità.
 - La performance lavorativa di un impiegato può essere prevista utilizzando un insieme di test attitudinali.
 - La dimensione del vocabolario di un bambino può essere prevista utilizzando l'età del bimbo oppure il livello di istruzione dei genitori.

Regressione Lineare e Logistica

La forma più semplice di regressione contiene un predittore ed una risposta. $Y = \beta_0 + \beta_1 X$

I termini “coefficiente angolare” ed “intercetta” vengono determinati minimizzando la somma delle deviazioni quadratiche dalla retta. Questo metodo è noto come **minimi quadrati**.



Regressione Lineare e Logistica

- I responsabili del personale possono utilizzare le procedure di regressione multipla per determinare il giusto compenso. Gli analisti solitamente conducono sondaggi nelle aziende simili, registrando l'ammontare dei salari e le rispettive posizioni lavorative. Queste informazioni potranno essere utilizzate in un'analisi di regressione multipla per costruire un'equazione nella forma:

$$\text{Salario} = .5 * (\text{Responsabilità}) + .8 * (\text{Numero di Sottoposti})$$

- Cosa accade se il trend dei dati NON è lineare?
 - Si possono utilizzare più predittori
 - È possibile ricorrere a trasformazioni
 - È possibile aggiungere interazioni e termini polinomiali di ordine superiore (perché la regressione non è lineare nei predittori, ma lineare nei parametri), ossia si potrebbe costruire una curva di regressione invece della retta, quando le relazioni non sono lineari.
 - Usando questo approccio si possono sviluppare modelli complessi. Tuttavia, esso richiede notevole esperienza da parte dell'analista sia per quanto riguarda l'applicazione in uso che la metodologia. Le tecniche illustrate in seguito possono effettuare questo tipo di lavoro in modo automatico...

Regressione Lineare e Logistica

- Cosa accade se la risposta è dicotomica?
 - Potremmo usare la regressione lineare per prevedere la risposta 0/1. Tuttavia, nascono dei nuovi problemi, come ad esempio:
 - Se si utilizza la regressione lineare, i valori previsti potranno fuoriuscire dall'intervallo (0,1) nel caso di gravi scostamenti sull'asse X. Questi valori sarebbero teoricamente inammissibili.
 - Uno degli assunti della regressione è che la varianza di Y sia costante lungo i valori di X (omoschedasticità). Ciò potrebbe non verificarsi in presenza di una variabile binaria, in quanto la varianza è data da $P*(1-P)$. Quando il 50 per cento di persone corrisponde al valore 1, la varianza sarà pari a 0,25, il suo massimo valore. Via via che ci si muove verso i valori estremi, la varianza diminuirà. Quando $P=0,10$, la varianza sarà pari a $0,1*0,9 = 0,09$. Quindi più P tende ad 1 o a 0, più la varianza tende a 0.
 - Il test sulla significatività dei coefficienti b si basa sull'assunto che gli errori di previsione $(Y-Y')$ si distribuiscano come una normale. Dato che Y accetta solo i valori 0 e 1, questo assunto risulta piuttosto difficile da giustificare, anche approssimativamente. Pertanto, tali test diventano poco sicuri quando si utilizza una variabile risposta dicotomica.

Regressione Lineare e Logistica

- Diverse applicazioni della regressione dispongono di una variabile risposta con solo due possibili valori qualitativi (maschio/femmina, predefinito/non-predefinito, successo/insuccesso).
 - In uno studio sulla partecipazione alle forze lavoro delle donne sposate come funzione dell'età, del numero di figli e del reddito del marito, la variabile risposta sarà definita da due modalità: moglie con lavoro, moglie senza lavoro.
 - In uno studio longitudinale sulle malattie coronariche, sono stati raccolti i valori di età, sesso, fumatore/non fumatore, livello di colesterolo e pressione sanguigna; la variabile risposta avrà le seguenti due possibili modalità: la persona presenta ha sviluppato una malattia coronarica, la persona NON ha sviluppato una malattia coronarica durante lo studio.

ANOVA

- L'Analisi della Varianza (ANOVA) si utilizza per verificare le differenze tra le medie di più gruppi.
- Le variabili esplicative in un'ANOVA sono generalmente di tipo qualitativo (sesso, regione di residenza, ecc.)
- Se i predittori sono quantitativi, allora non saranno fatti degli assunti sulla natura della funzione di regressione tra la risposta e i predittori.

- Alcuni esempi includono:
 - Un esperimento per studiare gli effetti di cinque differenti marche di benzina sull'efficienza operativa dell'automobile (km/l).
 - Un esperimento per valutare gli effetti di differenti quantità di una particolare droga psichedelica sulla destrezza manuale.

Analisi Discriminante

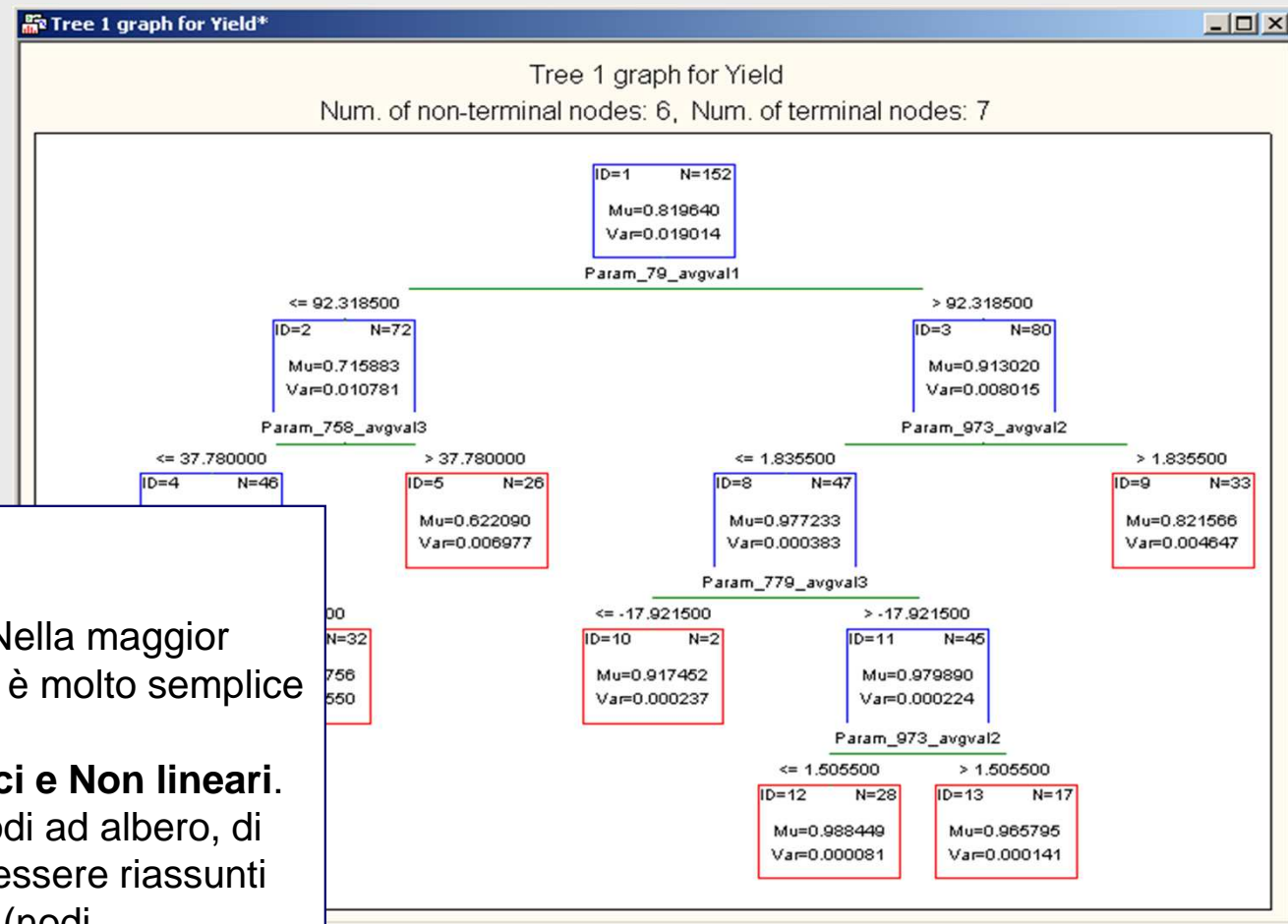
- L'analisi delle funzioni discriminanti si utilizza per determinare quali variabili discriminano tra due o più gruppi distinguibili.
- Per esempio, un educatore potrebbe voler stabilire quali variabili discriminano meglio tra studenti diplomati che decidono di
 - (1) andare all'università,
 - (2) iscriversi ad una scuola professionale,
 - (3) non proseguire con gli studi.

L'educatore potrebbe quindi raccogliere i dati relativi a diverse variabili, prima che gli studenti ottengano il diploma. Dopo il suo conseguimento, la maggioranza degli studenti cadrà naturalmente in una delle tre categorie. L'Analisi Discriminante potrebbe quindi venire utilizzata per determinare quali variabili siano i migliori predittori della scelta successiva al diploma.

- Un ricercatore medico potrebbe registrare diverse variabili riguardanti i propri pazienti, in modo da stabilire quali predicano meglio se una persona sarà
 - (1.) ricoverata stabilmente (gruppo 1),
 - (2.) ricoverata temporaneamente (gruppo 2),
 - (3.) non ricoverata (gruppo 3).

Alberi Decisionali

- Gli alberi sono modelli predittivi semplici da costruire ed interpretare.
- I modelli possono essere realizzati per prevedere risposte categoriali o continue.
- Un albero decisionale è una sequenza di condizioni se/allora.



Alcuni Vantaggi dei Modelli ad Albero

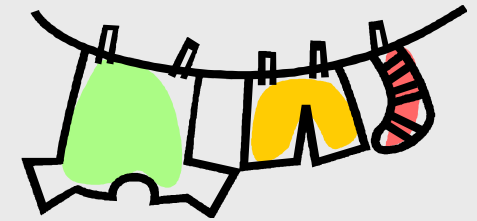
Semplice Interpretazione dei Risultati. Nella maggior parte dei casi, l'interpretazione dei risultati è molto semplice ed intuitiva.

I modelli ad albero sono Non parametrici e Non lineari.

I risultati finali derivanti dall'utilizzo di metodi ad albero, di classificazione o di regressione, possono essere riassunti in una serie di condizioni logiche se-allora (nodi dell'albero). Quindi, non si presuppone la linearità delle relazioni esistenti tra i predittori e la variabile dipendente, oppure che le funzioni legame siano non lineari o monotone.

Analisi dei Gruppi

- L'analisi dei gruppi è il metodo che consente di raggruppare elementi simili tra loro.
- Un semplice esempio di analisi dei gruppi è rappresentato dalla suddivisione dei capi da lavare – dividerli in capi delicati, bianchi, colorati, ecc.
- In questo caso, dove si colloca una maglietta bianca a righe rosse? Con cosa si può lavare?



Analisi dei Gruppi: Esempio di Marketing

Un tipico esempio di analisi dei gruppi può essere dato da una ricerca di marketing. Questo tipo di studio richiede il monitoraggio di un grande campione con molte variabili, legate al comportamento della clientela, in modo da determinare i, cosiddetti, “segmenti di mercato”, cioè dei gruppi di clienti in qualche modo simili tra loro e diversi da quelli appartenenti agli altri gruppi. Oltre ad individuare tali gruppi, di solito è di pari interesse la determinazione delle modalità di differenziazione, ossia identificare le variabili per le quali i gruppi si differenziano tra loro.

Analisi dei Gruppi

Le tecniche di Analisi dei Gruppi vengono utilizzate per...

- Identificare le caratteristiche delle persone appartenenti ai diversi gruppi (reddito, età, stato civile, ecc.).

Possibile applicazione dei risultati...

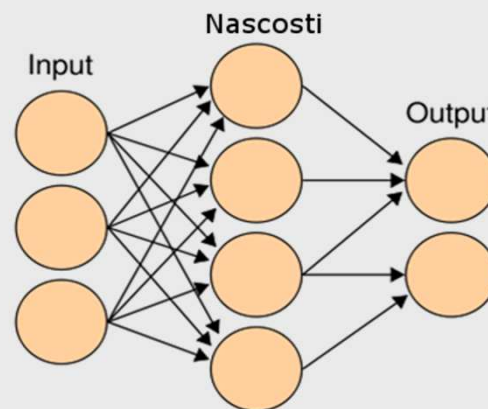
- Sviluppo di speciali campagne di marketing o raccomandazione di particolari punti vendita, in base alle caratteristiche individuate.
- Organizzazione delle giacenze di magazzino in base ai diversi gusti degli acquirenti.
- Miglioramento dell'appeal e della qualità di tutta l'esperienza di acquisto....

Raggruppamento K Means: Il classico algoritmo k-Means è diventato popolare grazie a Hartigan (1975; si veda anche Hartigan and Wong, 1978). Le operazioni alla base di questo algoritmo sono relativamente semplici: dato un numero fissato di k gruppi (desiderato o ipotizzato), le osservazioni saranno assegnate a tali gruppi, in modo che le medie di tutte le variabili siano quanto più differenti possibili tra loro.

Raggruppamento EM: L'algoritmo EM è descritto in dettaglio in Witten and Frank (2001). L'obiettivo del raggruppamento EM è stimare le medie e le deviazioni standard per ogni gruppo, in modo da massimizzare la verosimiglianza dei dati osservati (distribuzione). In altre parole, l'algoritmo EM prova ad approssimare le distribuzioni osservate sulla base delle misture delle differenti distribuzioni nei diversi gruppi.

Reti Neurali

Come la maggior parte dei metodi statistici, le reti neurali sono in grado di eseguire le tre principali operazioni, tra cui la *regressione* e la *classificazione*. La regressione comporta l'analisi delle relazioni esistenti tra diverse variabili x di input ed un insieme di t output continui (variabili target). Al contrario, la classificazione assegna la classe ad una variabile target categoriale, tramite un insieme di valori di input.



Analisi di Associazioni e Collegamenti

- Trova gli elementi che si presentano più spesso insieme (**Regole Associative**)
- Un'associazione è un'espressione nella forma:

Corpo \Rightarrow Testa (Supporto, Confidenza)

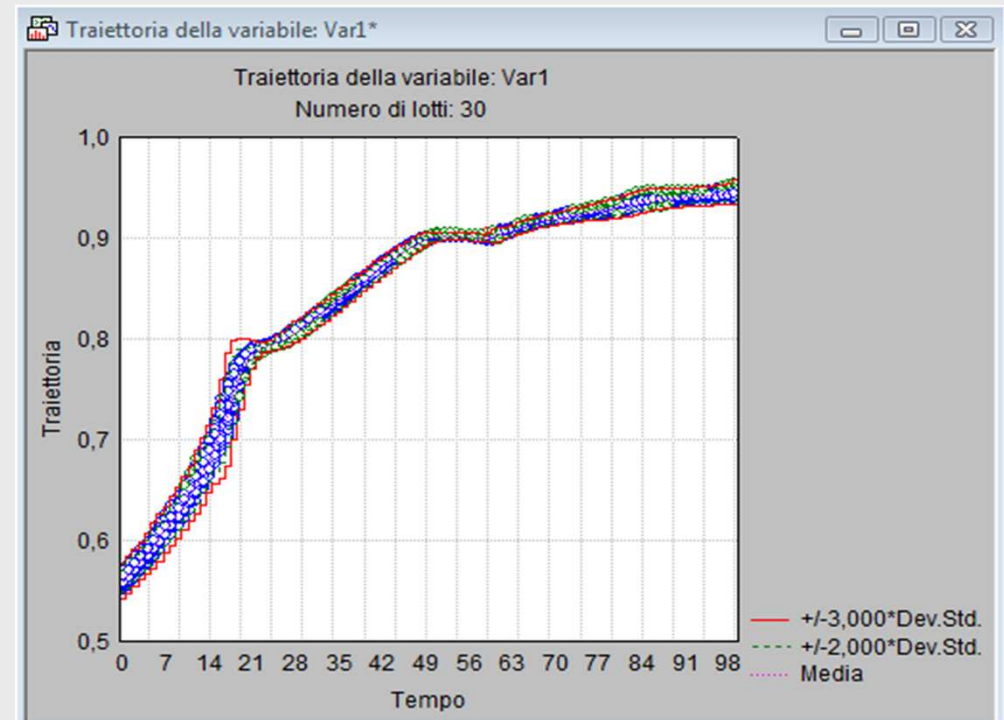
Se compra (x, "lampadine") \Rightarrow compra anche (x, "batterie") (250, 89%)

Dati: Riepilogo regole associative (Fastfood.sta)

Riepilogo regole associative (Fastfood.sta)
 Min: supporto = 40,0%, confidenza = 10,0%
 Max.dimensione di un insieme di item = 10

	Corpo	==>	Testa	Supporto(%)	Confidenza(%)	Lift
1	Gender==Male	==>	Pizza	57,50000	70,12195	1,016260
2	Pizza	==>	Gender==Male	57,50000	83,33333	1,016260
3	Hamburger	==>	Gender==Male	47,00000	82,45614	1,005563
4	Gender==Male	==>	Hamburger	47,00000	57,31707	1,005563

- Realizzato sulle capacità delle tecniche *PCA* e *PLS*, **MSPC** è una selezione di metodi progettati per il monitoraggio dei processi ed il controllo della qualità nelle elaborazioni di lotti industriali.
- In molti settori dell'industria, i processi a lotti sono di importanza considerevole nella realizzazione di prodotti caratterizzati da standard e specifiche, come ad esempio per i polimeri, le vernici, i fertilizzanti, i farmaci, i cementi, i prodotti petroliferi, i profumi e i semiconduttori.
- Gli obiettivi di queste tecniche sono legati al profitto raggiunto tramite la riduzione della variabilità del prodotto e l'aumento della qualità dello stesso.
- Dal punto di vista della qualità, questi processi possono essere divisi in lotti normali e lotti anomali. Più in generale, un lotto normale porta ad un prodotto con specifiche e standard desiderati, che non avviene invece con i lotti anomali, per i quali ci si aspetta che il prodotto finito sia di scarsa qualità.



- Un altro motivo per utilizzare questa tecnica è dato dalla necessità di conformarsi a normative regolamentari. Spesso è necessario che i prodotti industriali mantengano la traccia completa (storico) dell'andamento dei processi a lotti, in modo da consentire la verifica della modalità di controllo della qualità. *MSPC* aiuta a costruire un sistema efficiente per monitorare l'andamento dei lotti e per prevedere la qualità del prodotto finale.

Punti da Ricordare...

- Il data mining è uno strumento, non una bacchetta magica.
- Il data mining non scoprirà automaticamente delle soluzioni se non viene adeguatamente guidato.
- Le relazioni predittive trovate tramite il data mining non sono necessariamente di tipo causale.
- Per assicurare risultati significativi, è vitale comprendere la natura dei propri dati.
- **Questione centrale nel data mining:** Trovare tendenze significative con strumenti che consentono di evitare il sovra-adattamento.