UNIVERSITÀ DEGLI STUDI DI TERAMO

CL in BIOTECNOLOGIE

Anno Accademico 2022/2023

CHIMICA ANALITICA

Elaborazione dei dati

Di che cosa si occupa la Statistica?

- Fisica: fenomeni naturali
- Sociologia: fenomeni sociali
- Geologia: fenomeni che riguardano la crosta terrestre
- Astronomia: fenomeni celesti
- Biologia: fenomeni della vita (biologici)
- Medicina: fenomeni che riguardano lo stato di salute
- Economia: fenomeni di gestione delle risorse
- Chimica: fenomeni sulla composizione e trasformazioni della materia
- . . . . . .
- La Statistica si occupa di fenomeni reali!
  Si presta dunque a tutte le altre discipline.
  La Statistica studia i dati.

- Il punto di partenza di una indagine statistica è costituito da un insieme (che chiamiamo popolazione di riferimento), disomogeneo all'interno (ovvero non tutti gli elementi sono uguali tra di loro) e che costituisce la parte del mondo che ci interessa.
- Gli elementi di questo insieme (che di volta in volta nei problemi concreti saranno persone, animali, batteri, immagini raccolte da un satellite,...) vengono convenzionalmente indicati come unità statistiche.
- In genere, i ricercatori studiano un sottoinsieme della popolazione relativamente piccolo (campione) e desiderano trarre conclusioni circa l'intera popolazione.
- Inferenza: come utilizzare le informazioni nel campione per trarre conclusioni sulla distribuzione delle variabili di interesse nella popolazione.
- È importante anche poter associare alle analisi condotte su un campione una valutazione dell'affidabilità dei risultati.

- La prima fase di ogni analisi statistica è rappresentata dall'organizzazione e dalla sintesi dei DATI, le informazioni raccolte sulle UNITÀ STATISTICHE che compongono il CAMPIONE.
- Concetti e strumenti fondamentali dell'analisi esplorativa sono:
  - Variabili e tipi di variabili (qualitative sconnesse o ordinali, quantitative discrete o continue).
  - Frequenze (assolute, relative, percentuali, cumulate) e tabelle.
  - Grafici (a torta, a barre, istogramma).
  - Misure di posizione (media, mediana, moda, quantili).
  - Misure di variabilità (varianza, scarto interqurtile, campio di variazione).

- I DATI sono una raccolta di informazioni (espresse in forma numerica).
- Le entità (individui, ore del giorno, ...) che vengono osservate nello studio sono dette UNITÀ STATISTICHE (casi).
- L'insieme di tutte le unità statistiche di interesse per lo studio è detto POPOLAZIONE di riferimento.
- Invece, un sottoinsieme di unità statistiche selezionate (spesso casualmente) da una popolazione è detto CAMPIONE. La dimensione del campione può variare da poche unità a molte migliaia di osservazioni.
- Una quantità di interesse nella popolazione è detta parametro, mentre la quantità calcolata sul campione è detta statistica.

ESEMPIO: La popolazione oggetto di studio è l'insieme di tutti i pazienti affetti da patologia simile, anche in futuro (si tratta di una popolazione virtuale).
Il campione è costituito dai $n = 47$ pazienti che sono entrati nell'esperimento.

DEF: Una VARIABILE (o CARATTERE) è una caratteristica di interesse rilevata sulle unità statistiche (ad esempio, età, peso, trattamento, ...).

Il termine 'variabile' evidenzia che la caratteristica di interesse può assumere una pluralità di valori. L'insieme dei valori possibili si può pensare noto, ma prima di fare l'osservazione su una unità statistica, non sappiamo quale valore si osserverà.

DEF: I valori distinti assunti da una variabile sono detti MODALITÀ della variabile. Le modalità si presumono note preliminarmente.

Esempio: nello studio sul trattamento con la realtà virtuale, la variabile *FIM* può assumere valori nell'intervallo (0, 130). Le modalità sono dunque tutti i numeri reali appartenenti a questo intervallo.

Esempio: in uno studio sulla biodiversità, si può osservare la variabile *numero di esemplari di lupo* avvistati in una settimana da un certo punto di osservazione. Le modalità sono i valori $0, 1, 2, 3, \ldots$ (i numeri naturali), anche se difficilmente si osserveranno valori grandi.

Una variabile può essere:

• QUALITATIVA o CATEGORIALE quando le sue modalità sono espresse in forma verbale (*sesso*, *livello di istruzione*, *trattamento*, ...).

A sua volta una variabile qualitativa può essere:
– SCONNESSA o NOMINALE se non esiste nessun ordinamento tra le modalità.
Esempi:
la variabile *sesso* con modalità M e F;

la variabile *modo di somministrazione* con modalità ORALE, ENDOVENA, ...

– ORDINALE se è possibile individuare un ordinamento naturale delle modalità.
Esempi:
la variabile *livello di istruzione* con modalità ELEMENTARE, MEDIA INFERIORE, MEDIA SUPERIORE, ...;

la variabile *giudizio* con modalità INSUFFICIENTE, SUFFICIENTE, DISCRETO, OTTIMO.

• Se le modalità sono solo due si parla di variabili DICOTOMICHE o BINARIE (*sesso*, *presenza*, . . .). A volte le due modalità sono espresse con valori numerici (0,1, oppure 1,2,. . .), ma il valore del numero non vuol dire assolutamente nulla!!

Oppure, una variabile può essere:

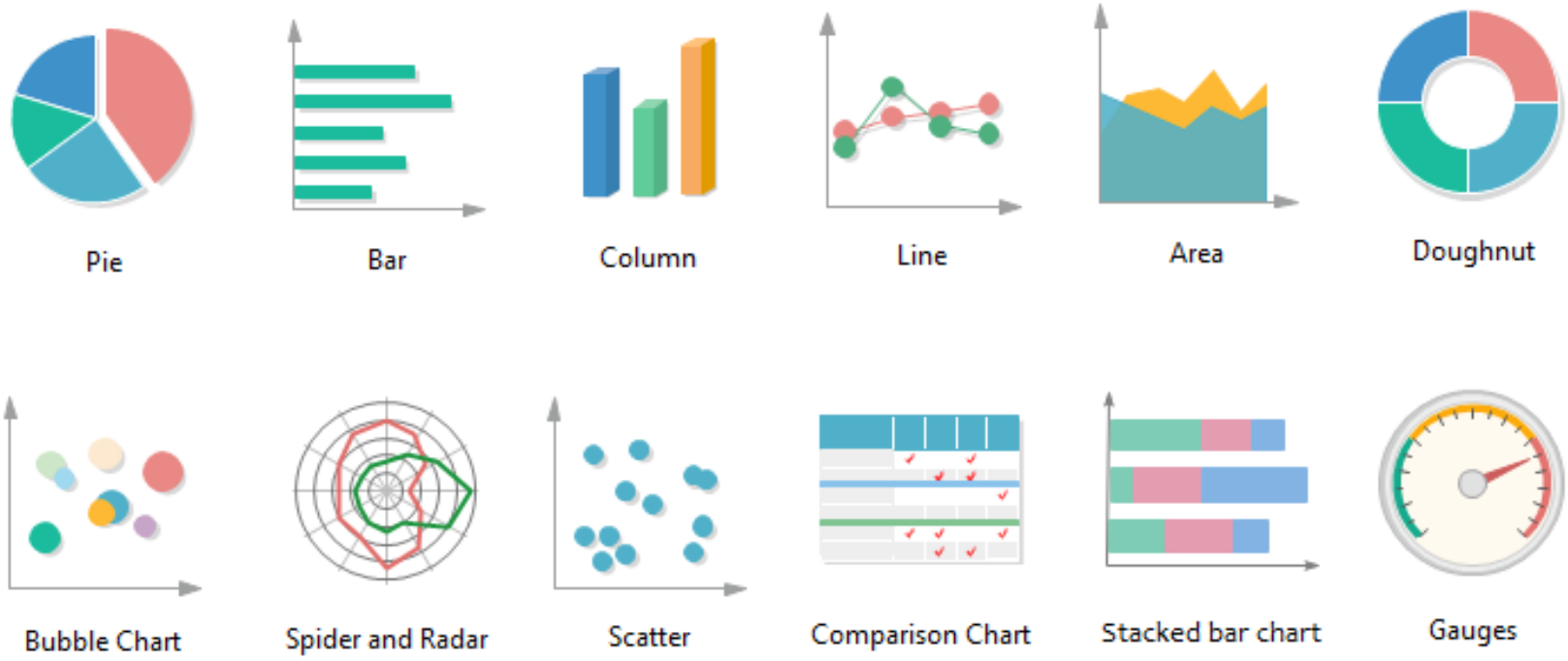
• QUANTITATIVA (o NUMERICA) quando le modalità sono espresse da numeri (età, peso,. . .). A sua volta una variabile quantitativa può essere:

– DISCRETA quando l'insieme delle modalità è finito o numerabile (stessa cardinalità dell'insieme dei naturali). Esempi:
- la variabile *numero di 'teste' in 10 lanci di una moneta*, con modalità 0,1, . . . , 10;
- le variabili *numero di sedute*, *numero di figli*, . . . con modalità 0, 1, 2, . . . ;

– CONTINUA quando l'insieme delle modalità è un intervallo, ossia un sottoinsieme, eventualmente illimitato, dei numeri reali. Esempi:
- la variabile *peso* (in kg) che ha come modalità possibili tutti i valori positivi,
- la variabile *dose* di un dato farmaco (in mg) con modalità da zero a 1000mg.
- eventuale suddivisione in classi.

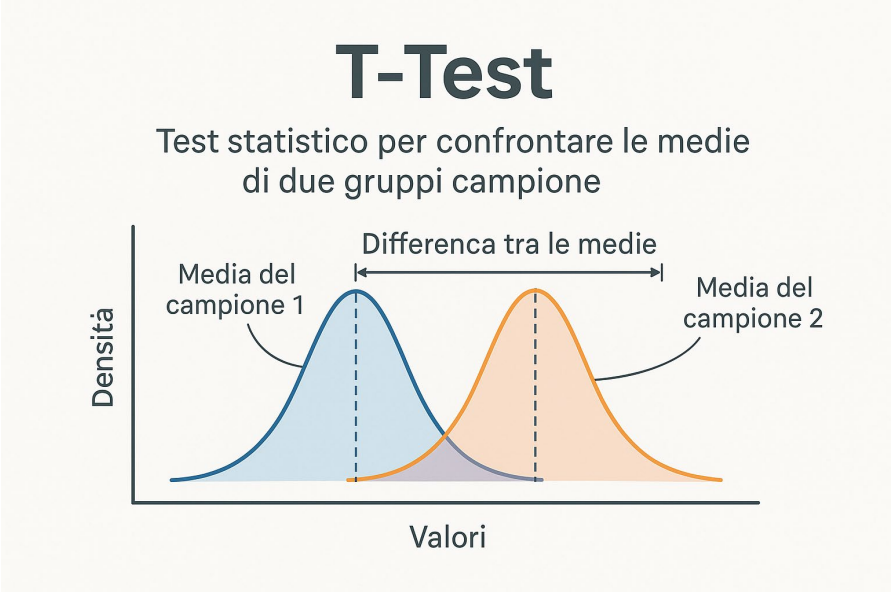
- VARIABILI QUALITATIVE vs QUANTITATIVE
  - A seconda del tipo di variabili osservate, sono possibili diverse analisi statistiche.
  - Ci sono degli strumenti statistici appositi per studiare tipi diversi di variabili.
  - Tra le varie tipologie di dati è implicita una gerarchia (le variabili quantitative possono essere discretizzate, le variabili quantitative discrete possono essere tradotte in variabili qualitative ordinali, quelle ordinali possono essere considerate nominali). Le analisi statistiche sono più ricche, per così dire, ascendendo la gerarchia.
- DATI UNIVARIATI vs MULTIVARIATI
  - Le analisi univariate considerano una sola variabile rilevata sulle unità.
  - Nello studio congiunto di due variabili si parla di analisi bivariata.
  - Lo studio congiunto di due o più variabili è detto analisi multivariata (ovviamente il multivariato include il bivariato).

Tipologie di grafici



T-test

Il test t (o t test) è un test parametrico di significatività statistica che utilizza la distribuzione t di Student per valutare la stima di una variabile o di un valore.
La distribuzione dei valori t ha come suoi parametri la media e l'errore standard della media, e, per campioni grandi, non è sensibile agli scostamenti dalla normalità della forma della distribuzione.
Inoltre, sempre per campioni grandi (maggiori di circa 30 casi), la distribuzione t tende a coincidere con la distribuzione normale standard (dei valori Z): come abbiamo visto e più volte ripetuto, infatti, all'aumentare delle dimensioni del campione, l'errore standard del campione tende a coincidere con la deviazione standard della popolazione.

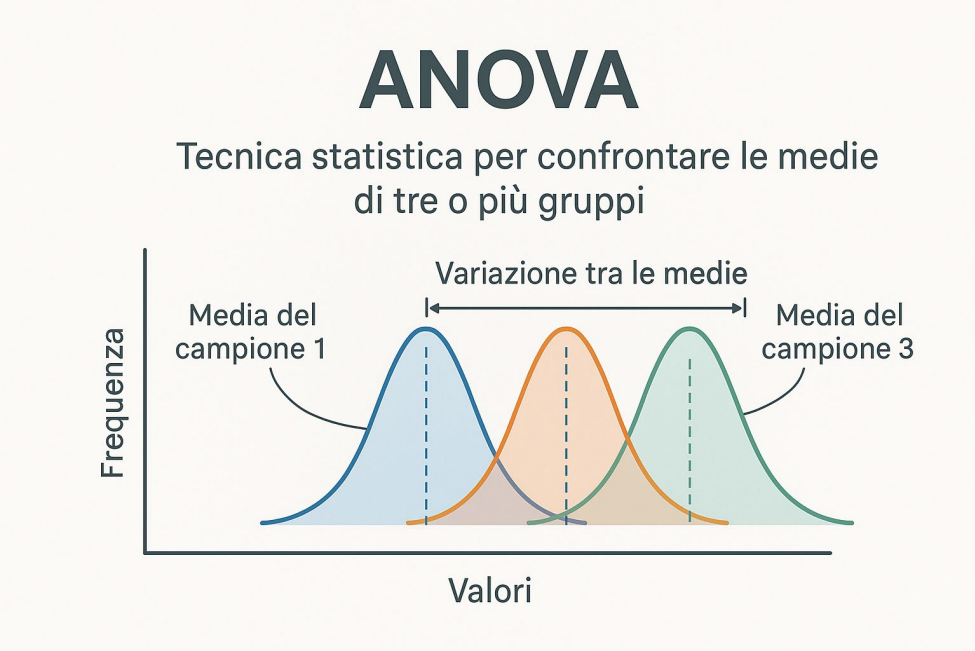


ANOVA: l’analisi della varianza spiegata semplice

L’analisi della varianza (ANOVA, dall’inglese Analysis of Variance) comprende una serie di test statistici che rientrano nell’ambito della statistica inferenziale. Scopri in questo articolo quando si può usare, quale test scegliere e come si interpretano i risultati.

L’ANOVA è una generalizzazione del test t. Entrambe le tecniche si utilizzano infatti per il confronto di valori medi. La differenza è che:

il test t permette di confrontare solo due gruppi
l’ANOVA permette di confrontare un numero qualsiasi di gruppi

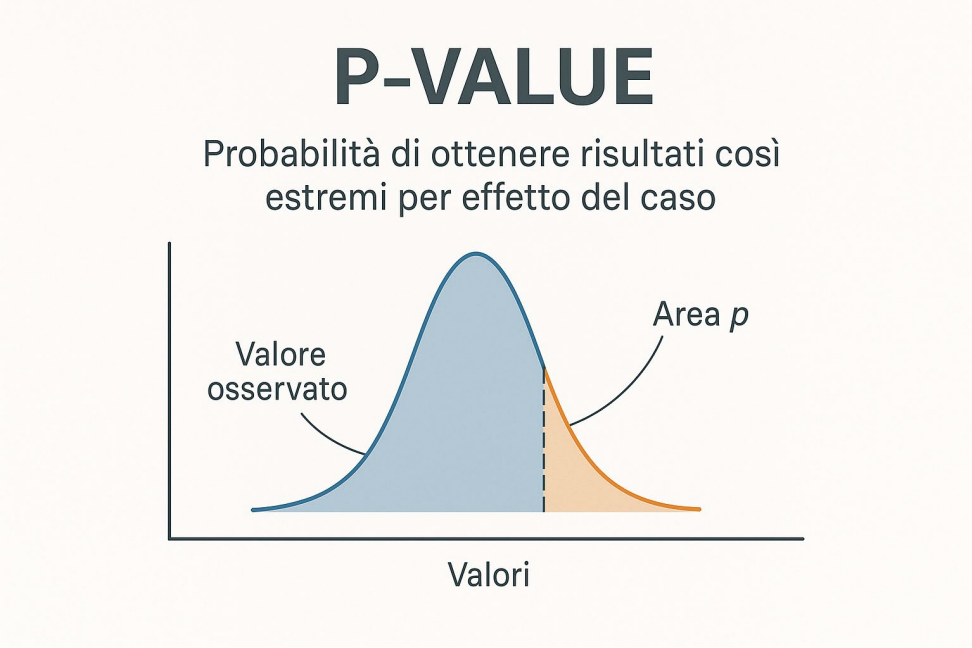


P-Value

Il p-value è la probabilità di ottenere un risultato almeno così estremo di quello osservato se l'ipotesi nulla è vera.In altre parole: indica quanto è compatibile il risultato con il caso.

p piccolo (es. p < 0.05)
→ È improbabile che la differenza osservata sia dovuta al caso.
→ *Si rifiuta l'ipotesi nulla*.

p grande (es. p > 0.05)
→ Il risultato è compatibile con la variabilità naturale.
→ *Non si rifiuta l'ipotesi nulla*.



Nel t-test (confronto tra due gruppi)

Il p-value indica quanto è probabile ottenere una differenza tra le **due medie** almeno così grande **se in realtà non c'è alcuna differenza reale** (ipotesi nulla: $\mu_1 = \mu_2$).

✔ **p-value basso (p < 0.05)**

→ La differenza osservata tra le due medie è **difficile da spiegare col caso**.

→ Concludiamo che **i due gruppi sono significativamente diversi**.

→ Esempio in Biotecnologie: due trattamenti cellulari rispondono diversamente.

✔ **p-value alto (p > 0.05)**

→ La differenza osservata potrebbe essere **solo rumore**.

→ Concludiamo che **non c'è evidenza sufficiente per dire che le medie siano divers**

Nell'ANOVA (confronto tra 3 o più gruppi)

Il p-value misura quanto è probabile osservare **variazioni tra le medie dei gruppi** più grandi di quelle attese **solo per variabilità casuale** (ipotesi nulla: $\mu_1 = \mu_2 = \mu_3 = \ldots$).

✔ **p-value basso (p < 0.05)**

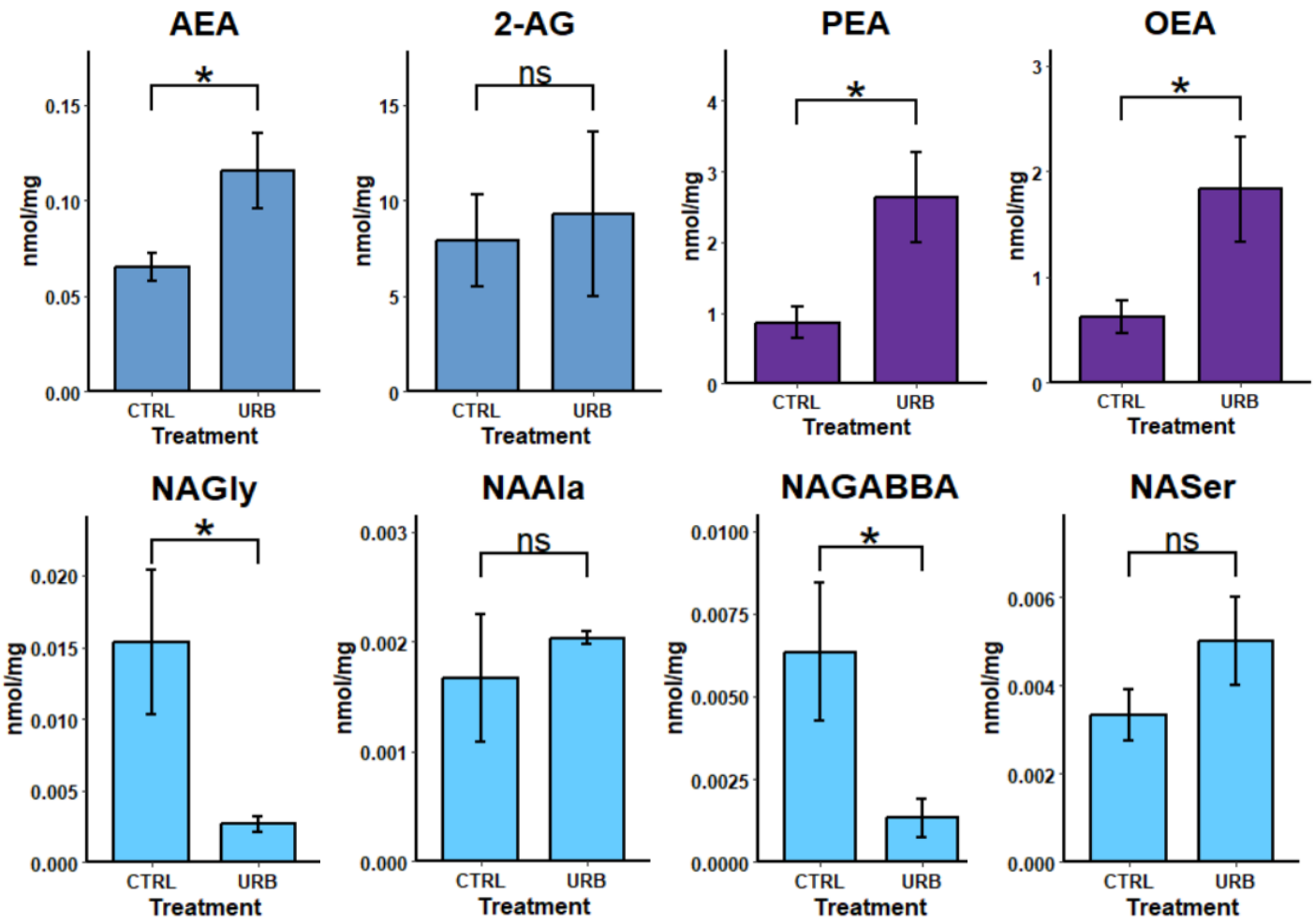
→ Le differenze tra i gruppi sono **troppo grandi per essere casuali**.

→ **Conclusione:** almeno **un gruppo** differisce dagli altri.

✔ **p-value alto (p > 0.05)**

→ Le medie dei gruppi sono **compatibili** con essere uguali.

→ Nessuna evidenza statistica di differenze.


AEA
*
0.15
0.10
0.05
0.00
nmol/mg
CTRL
URB
Treatment
2-AG
ns
15
10
5
0
nmol/mg
CTRL
URB
Treatment
PEA
*
4
3
2
1
0
nmol/mg
CTRL
URB
Treatment
OEA
*
3
2
1
0
nmol/mg
CTRL
URB
Treatment
NAGly
*
0.020
0.015
0.010
0.005
0.000
nmol/mg
CTRL
URB
Treatment
NAAla
ns
0.003
0.002
0.001
0.000
nmol/mg
CTRL
URB
Treatment
NAGABBA
*
0.0100
0.0075
0.0050
0.0025
0.0000
nmol/mg
CTRL
URB
Treatment
NASer
ns
0.006
0.004
0.002
0.000
nmol/mg
CTRL
URB
Treatment

L'Analisi fattoriale e la PCA

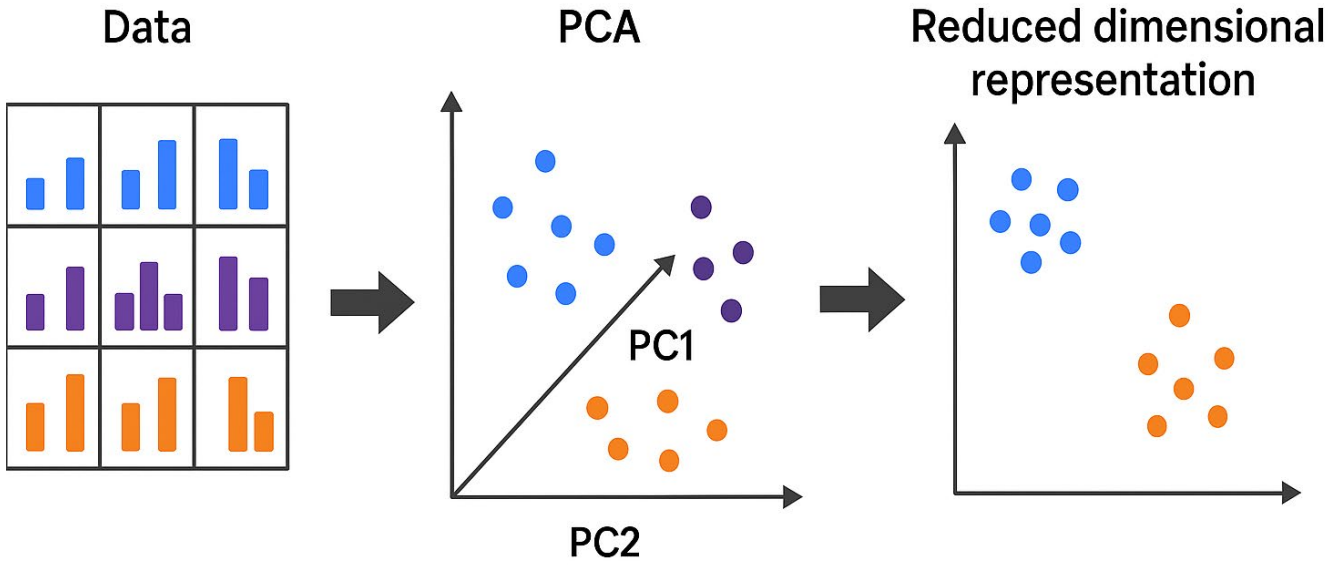
L'analisi fattoriale consiste in un insieme di tecniche statistiche che permettono di ottenere una **riduzione della complessità del numero di fattori che spiegano un fenomeno.**

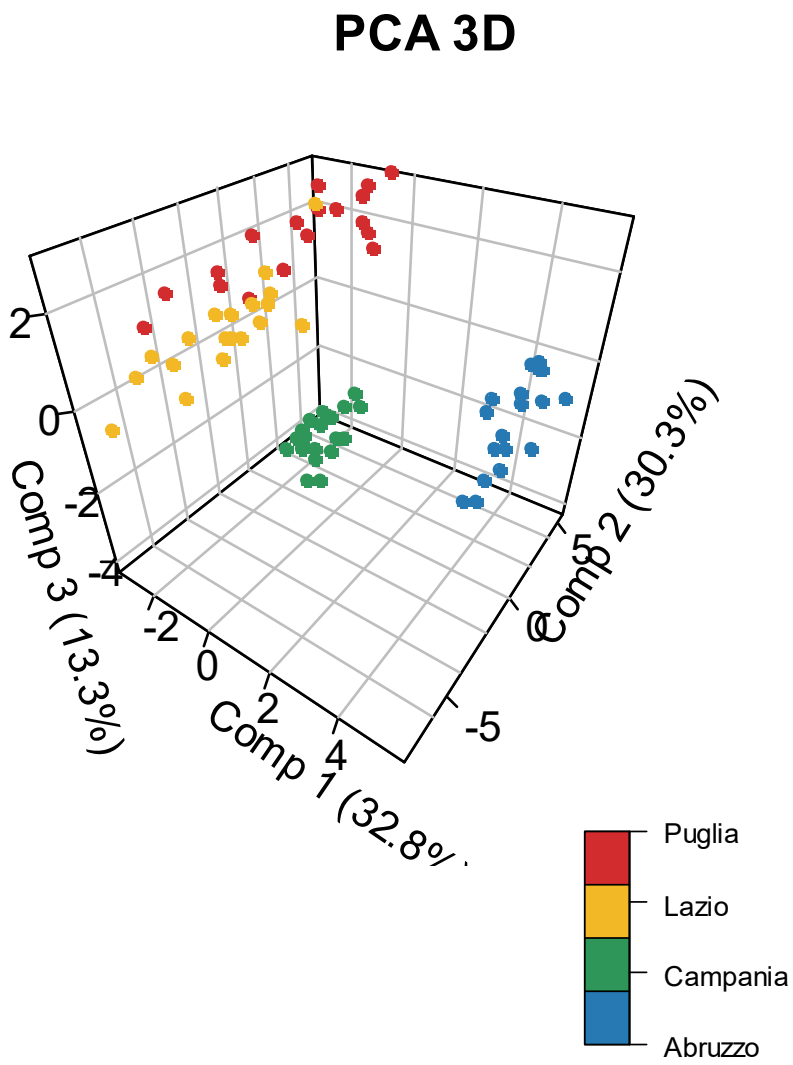
Si propone quindi di **determinare un certo numero di variabili "latenti"** (fattori non direttamente misurabili nella realtà) più ristretto e riassuntivo rispetto al numero di variabili di partenza.

La PCA, o *Principal Component Analysis*, è uno strumento che ci aiuta proprio a questo: prende un insieme di dati complesso e lo **semplifica**, mantenendo però la maggior parte delle informazioni importanti.

La PCA cerca le direzioni in cui questa nuvola "si allunga" di più. Queste direzioni sono chiamate **Componenti Principali**, e rappresentano gli assi lungo cui i dati mostrano la massima variabilità.

La PCA è una tecnica che **riduce la complessità** dei dati senza perdere il significato biologico. Ti aiuta a vedere pattern che sarebbero impossibili da cogliere guardando una variabile alla volta. È una lente che condensa centinaia di dimensioni in un'immagine comprensibile, rivelando relazioni, gruppi e differenze in modo immediato.



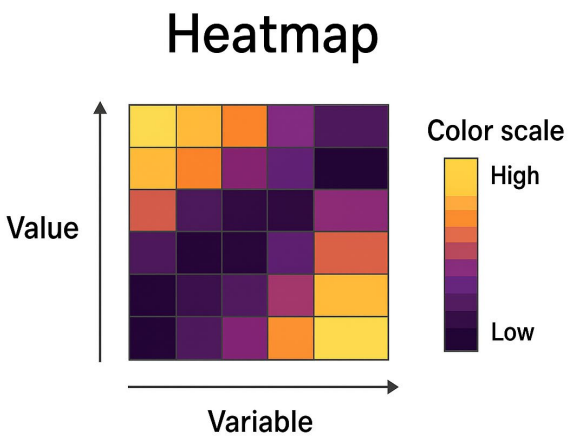


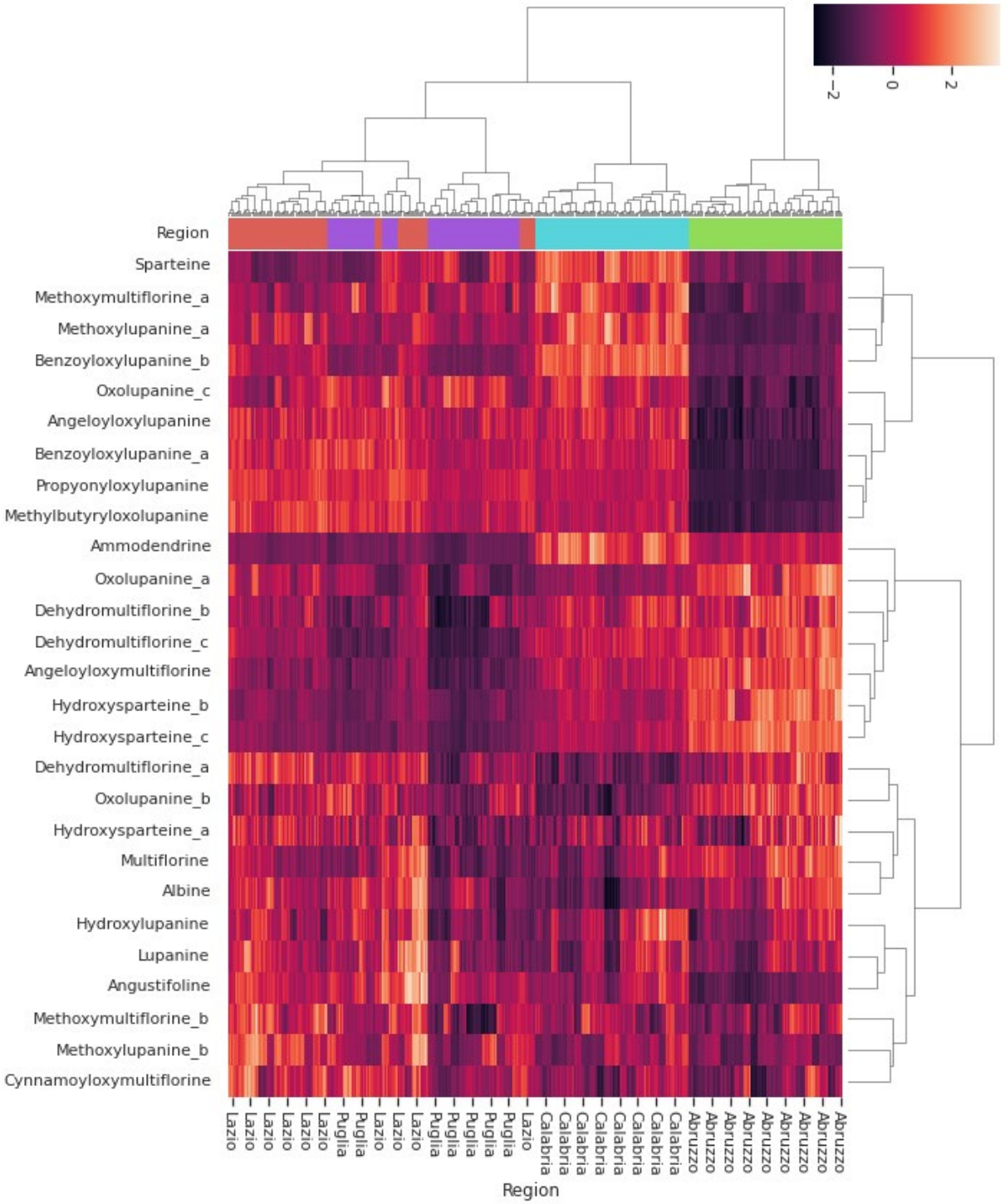
Cosa ci permette di vedere una PCA?
Se due campioni finiscono vicini sul grafico, significa che hanno profili molto simili. Se invece finiscono lontani, vuol dire che differiscono per molte variabili allo stesso tempo. In questo modo possiamo vedere, per esempio, se un trattamento separa chiaramente i campioni dal controllo, se due gruppi biologici hanno caratteristiche diverse o se c'è un outlier sospetto.

Una **heatmap** è una rappresentazione grafica che usa i **colori** per mostrare l'intensità di un valore.
Nelle analisi biologiche e biotecnologiche è uno degli strumenti più utili per visualizzare **grandi matrici di dati**, come profili metabolomici, proteomici o di espressione genica.
L'idea è molto intuitiva:

- ogni **riga** rappresenta un campione o una variabile (per esempio un metabolita);
- ogni **colonna** rappresenta un'altra dimensione del dataset (per esempio altri metaboliti, geni o condizioni sperimentali);
- il **colore** della cella indica l'intensità del valore (basso, medio, alto).

Spesso la heatmap è accompagnata da un **clustering gerarchico** (i dendrogrammi ai lati), che raggruppano in automatico campioni o variabili con comportamenti simili.
Questo permette di individuare **cluster biologicamente rilevanti**, come trattamenti che generano profili simili o metaboliti che rispondono alla stessa condizione.




−2
0
2
Region
Sparteine
Methoxymultiflorine_a
Methoxylupanine_a
Benzoyloxylupanine_b
Oxolupanine_c
Angeloyloxylupanine
Benzoyloxylupanine_a
Propyonyloxylupanine
Methylbutyryloxolupanine
Ammodendrine
Oxolupanine_a
Dehydromultiflorine_b
Dehydromultiflorine_c
Angeloyloxymultiflorine
Hydroxysparteine_b
Hydroxysparteine_c
Dehydromultiflorine_a
Oxolupanine_b
Hydroxysparteine_a
Multiflorine
Albine
Hydroxylupanine
Lupanine
Angustifoline
Methoxymultiflorine_b
Methoxylupanine_b
Cynnamoyloxymultiflorine
Lazio
Puglia
Calabria
Abruzzo
Region

La **PLS-DA** è una tecnica di classificazione multivariata molto usata in metabolomica, proteomica e dati "omics".
È simile alla PCA, ma con una differenza fondamentale:

- **la PCA è non supervisionata**, mentre
- **la PLS-DA è supervisionata**.

Cosa significa?
La PCA cerca di trovare pattern nei dati **senza sapere** quali campioni appartengono a quale gruppo.
La PLS-DA, invece, conosce già le classi (es. *controllo* vs *trattamento*, *malato* vs *sano*) e costruisce un modello per **massimizzare la separazione** tra questi gruppi.

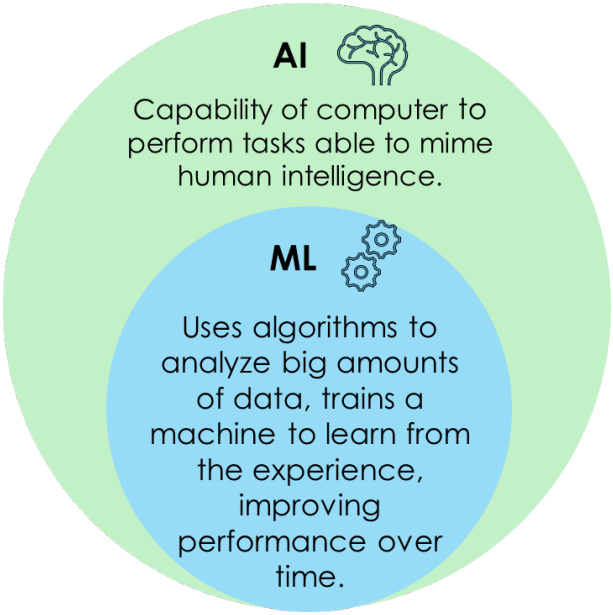
Come lavora la PLS-DA:

1. Prende il dataset multivariato (es. migliaia di metaboliti).
2. Cerca combinazioni lineari delle variabili che **spiegano la variazione tra i gruppi**.
3. Costruisce componenti latenti simili alle PC della PCA, ma orientate a **massimizzare la discriminazione**.
4. Produce grafici in cui i gruppi appaiono ben separati.

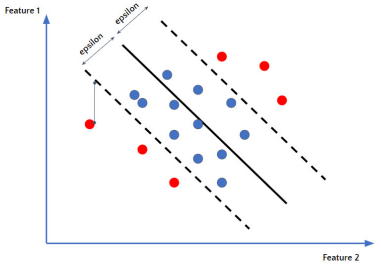
Il **machine learning** è un insieme di tecniche che permettono a un computer di imparare dai dati.

In biotecnologie è sempre più usato per:

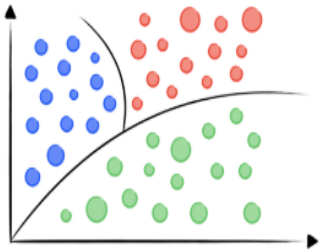
- classificare campioni (malato vs sano)
- predire risposte a trattamenti
- identificare biomarcatori
- analizzare dati omici ad alta dimensionalità



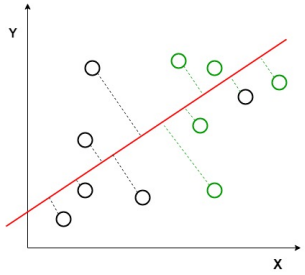
LDA



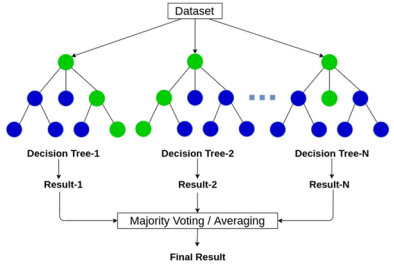
NB



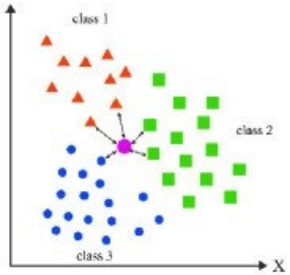
SVM



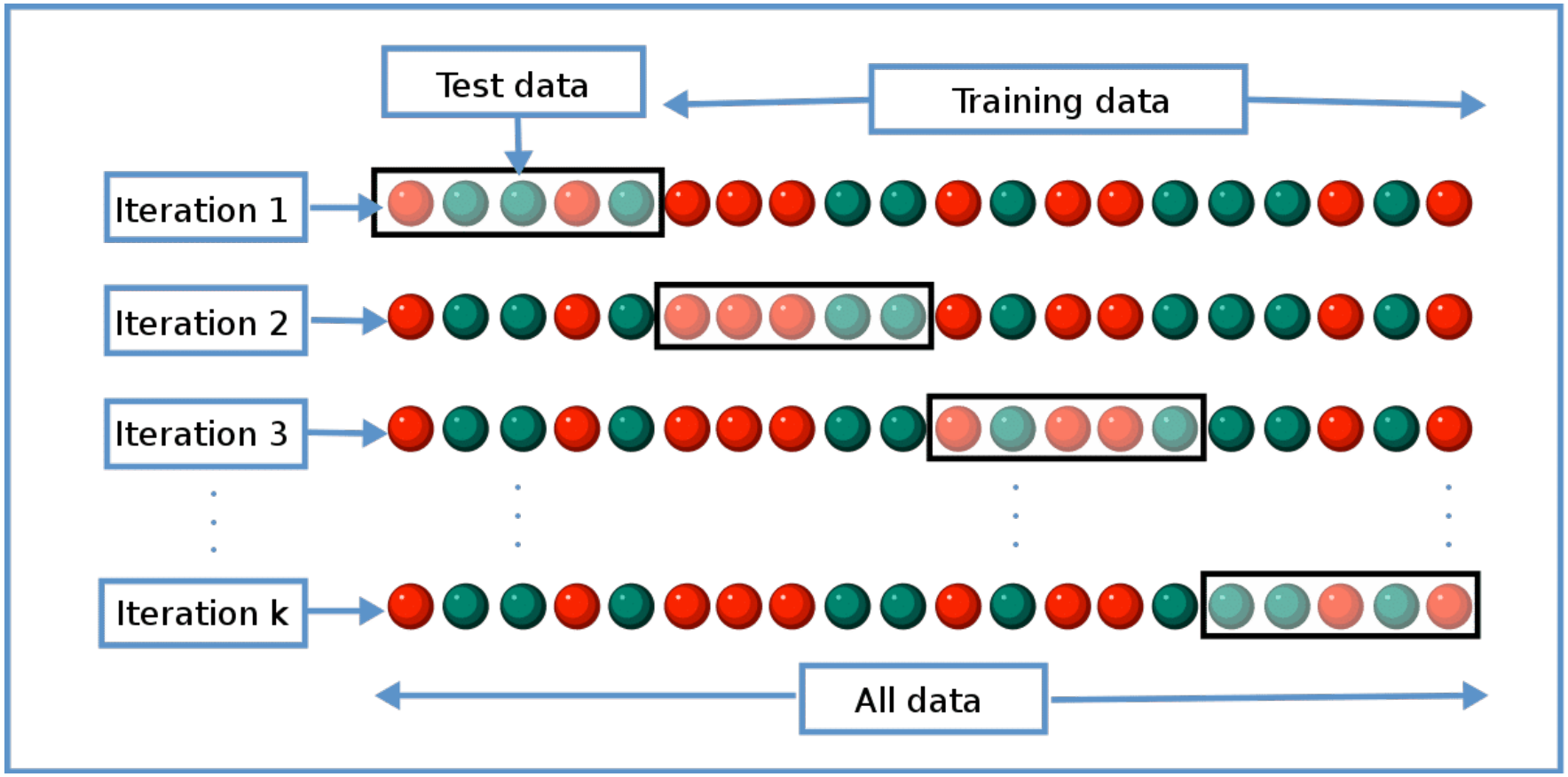
RF

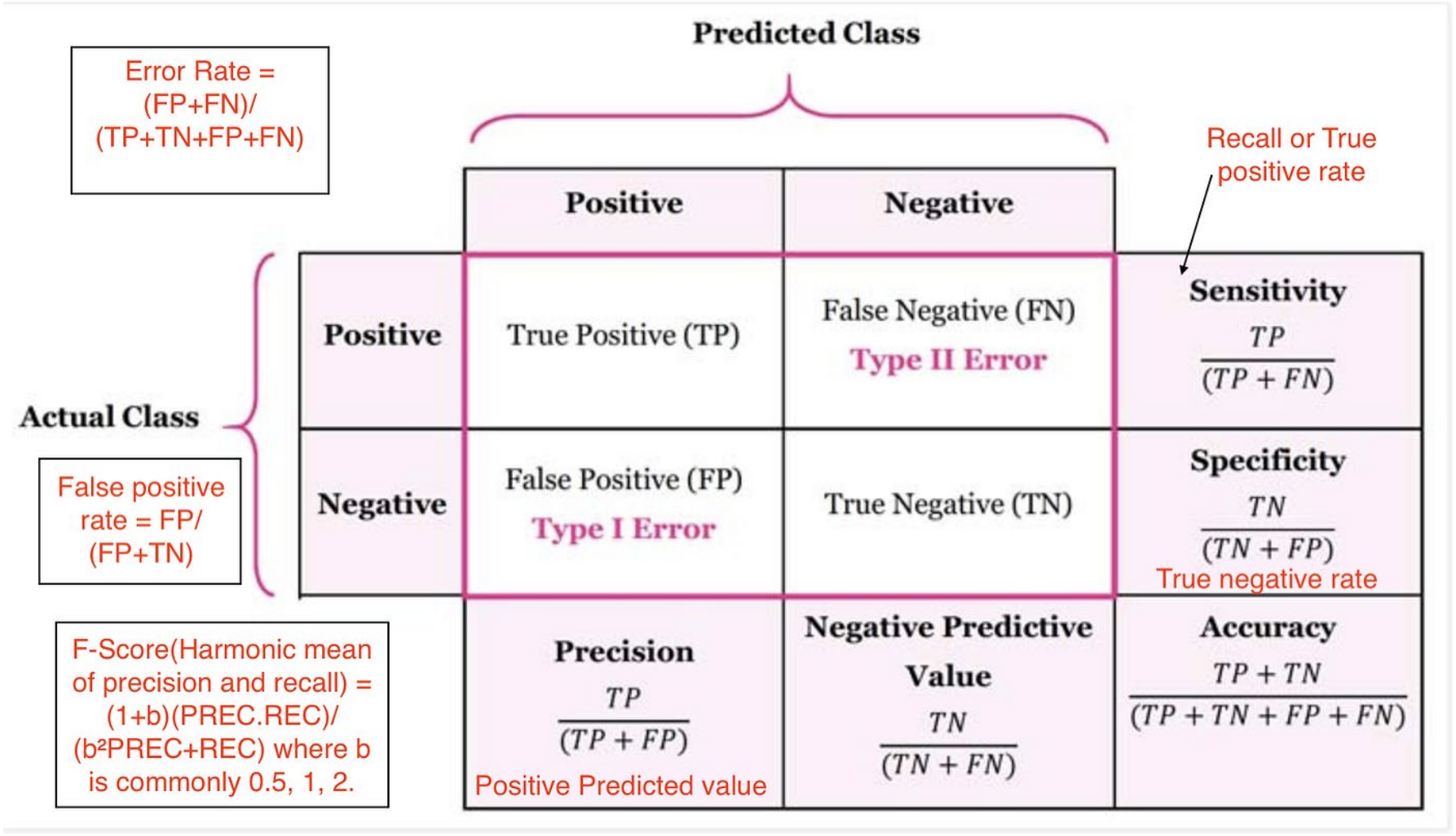


KNN



È uno strumento potente, ma va usato con attenzione, perché rischia di andare incontro ad **overfitting,** cioè imparare troppo bene il dataset e fallire su dati nuovi. Per questo si usano tecniche come **cross-validation**, **permutation test**, **VIP scores** e **plots di errore** per valutare la robustezza del modello.




Error Rate = (FP+FN)/ (TP+TN+FP+FN)
Predicted Class
Positive
Negative
Recall or True positive rate
Actual Class
Positive
True Positive (TP)
False Negative (FN)
Type II Error
Sensitivity
$\frac{TP}{(TP+FN)}$
False positive rate = FP/ (FP+TN)
Negative
False Positive (FP)
Type I Error
True Negative (TN)
Specificity
$\frac{TN}{(TN+FP)}$
True negative rate
F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/ (b²PREC+REC) where b is commonly 0.5, 1, 2.
Precision
$\frac{TP}{(TP+FP)}$
Positive Predicted value
Negative Predictive Value
$\frac{TN}{(TN+FN)}$
Accuracy
$\frac{TP+TN}{(TP+TN+FP+FN)}$

PLS- Discriminant Analysis Cross-Validation (10-Fold Summary)

Semi-Target 27 alkaloids

	Abruzzo	Campania	Lazio	Puglia
Sensitivity	1.00	0.96	0.91	0.97
Specificity	1.00	1.00	0.99	0.96
Pos Pred Value	1.00	1.00	0.96	0.90
Neg Pred Value	1.00	0.99	0.97	0.99
Balanced Accuracy	1.00	0.98	0.95	0.97