La regressione: uno strumento per capire e prevedere i fenomeni

DI MASSIMO VAILATI

08/06/2026

INTELLIGENZA ARTIFICIALE | MACHINE LEARNING | REGRESSIONE

LICEO , SCIENTIFICO S.A. , TECNICO , TT INFORMATICA

Contenuto: l'articolo introduce la regressione come metodo statistico per analizzare relazioni tra variabili e fare previsioni. Vengono spiegati anche concetti chiave come metodo dei minimi quadrati, coefficiente R^2, outlier e overfitting, con applicazioni nel machine learning.

Attività pratiche: applicazione della regressione lineare a dati reali, con previsione, analisi di outlier e valutazione del modello; analisi di regressione non lineare; riflessione sul problema dell'overfitting confrontando modelli di diversa complessità.

Nell'analisi dei dati uno degli obiettivi più importanti è capire **come una variabile influenza un'altra** e, possibilmente, **prevedere valori futuri**.
Per fare questo si utilizza una famiglia di metodi statistici chiamata **regressione**.
La regressione è una tecnica che permette di **modellare la relazione tra una variabile**

dipendente (quella che vogliamo spiegare o prevedere) **e una o più variabili indipendenti** (i fattori che possono influenzarla).
In termini semplici, la regressione risponde a domande come:

- Se aumenta la pubblicità, **quanto aumentano le vendite?**
- Se cresce il reddito medio, **come cambia il consumo?**
- Se aumenta la temperatura, **come varia la domanda di energia?**

Grazie alla regressione possiamo quindi:

- **capire relazioni tra variabili**
- **prevedere valori futuri o mancanti**
- **supportare decisioni economiche o aziendali**

LA REGRESSIONE COME MODELLO DI MACHINE LEARNING E INTELLIGENZA ARTIFICIALE

Negli ultimi anni la regressione ha assunto un ruolo fondamentale anche nel campo del **Machine Learning** e della **Intelligenza Artificiale**.
Nel machine learning la regressione è considerata un **metodo di apprendimento supervisionato**. In questo tipo di apprendimento il modello viene addestrato utilizzando dati storici in cui sono già noti:

- i valori delle variabili indipendenti (input)
- il valore della variabile da prevedere (output)

Il modello "impara" la relazione tra input e output e successivamente può **fare previsioni su nuovi dati**.
Molti algoritmi moderni di machine learning derivano proprio dai modelli di regressione statistica.
Nel contesto dell'intelligenza artificiale, la regressione viene utilizzata in numerosi sistemi predittivi, ad esempio per:

- previsione dei prezzi immobiliari
- stima della domanda di prodotti
- previsione dei consumi energetici
- analisi dei mercati finanziari

- previsione di variabili climatiche

In molti casi la regressione rappresenta **il punto di partenza dei modelli predittivi**, perché è semplice da interpretare e spesso sorprendentemente efficace.

LA REGRESSIONE LINEARE

La **regressione lineare** è il tipo di regressione più semplice e uno dei più utilizzati.
Essa descrive la relazione tra variabili attraverso **una linea retta**.
Nel caso più semplice (regressione lineare semplice) si studia la relazione tra:

- una variabile dipendente Y
- una variabile indipendente X

La relazione è descritta dalla formula:

$$Y = a + bX$$

dove:

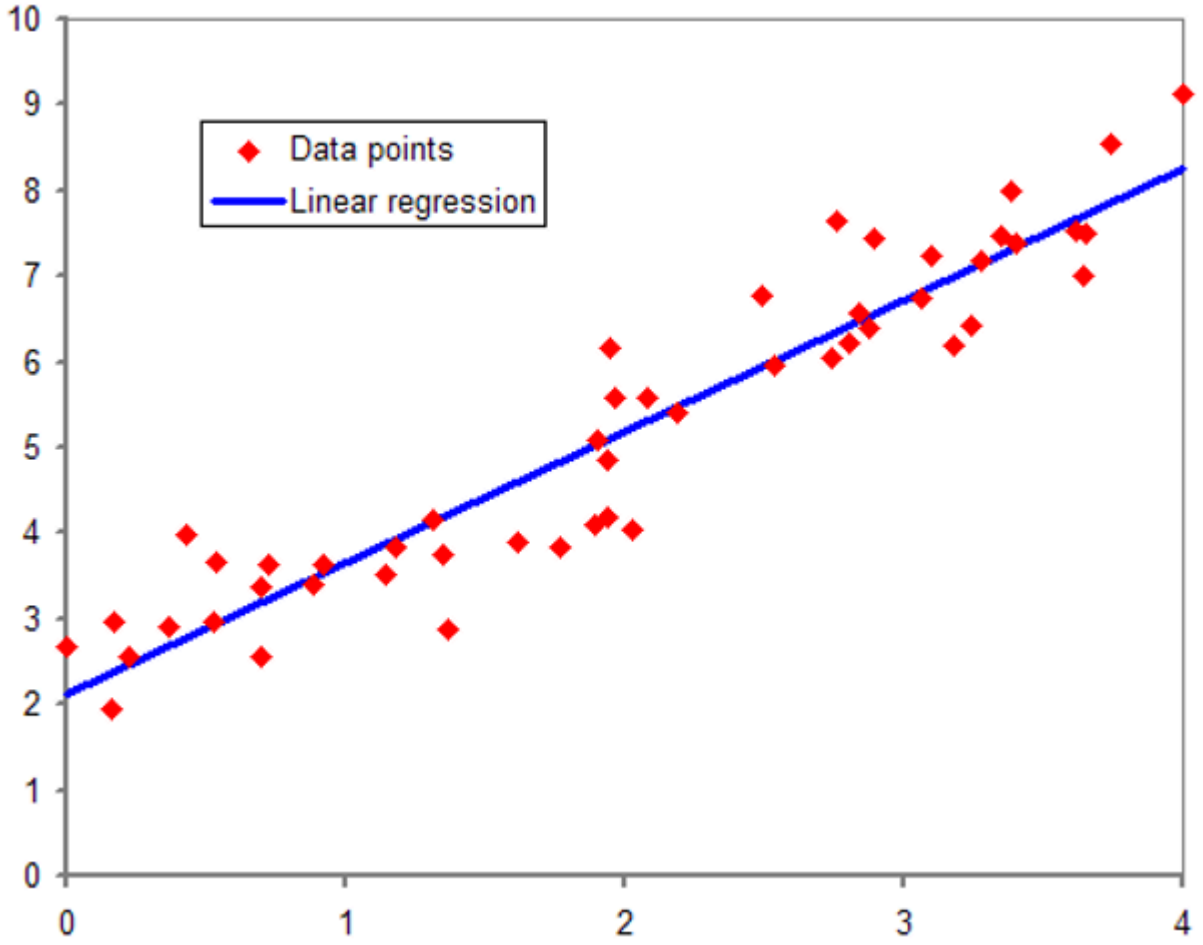
- **a** = intercetta (il valore di Y quando X = 0)
- **b** = coefficiente angolare o pendenza della retta
- **X** = variabile indipendente
- **Y** = variabile dipendente

Il coefficiente **b** indica **quanto cambia Y quando X aumenta di una unità**.
La retta di regressione non viene scelta casualmente.
L'obiettivo è trovare la retta che **minimizza l'errore tra i valori osservati e quelli stimati**.
Questo metodo è chiamato **metodo dei minimi quadrati**.

In pratica:

1. per ogni osservazione si calcola la differenza tra valore reale e valore previsto
2. queste differenze vengono elevate al quadrato
3. si sceglie la retta che rende **minima la somma di questi quadrati**

Il risultato è la **retta che meglio approssima l'andamento dei dati**.



TIPI PRINCIPALI DI REGRESSIONE

Oltre alla regressione lineare esistono molti altri modelli:

Regressione lineare semplice
Una variabile dipendente e una indipendente.
Esempio: vendite in funzione della spesa pubblicitaria.

Regressione lineare multipla
Una variabile dipendente e **più variabili indipendenti**.
Esempio: prezzo di una casa in funzione di: superficie, posizione, numero di stanze.

Regressione non lineare
La relazione tra variabili **non è una retta** ma una curva.
Esempi: regressione polinomiale, esponenziale, logaritmica.

LIMITI DELLA REGRESSIONE

Nonostante sia molto utile, la regressione presenta alcuni limiti:

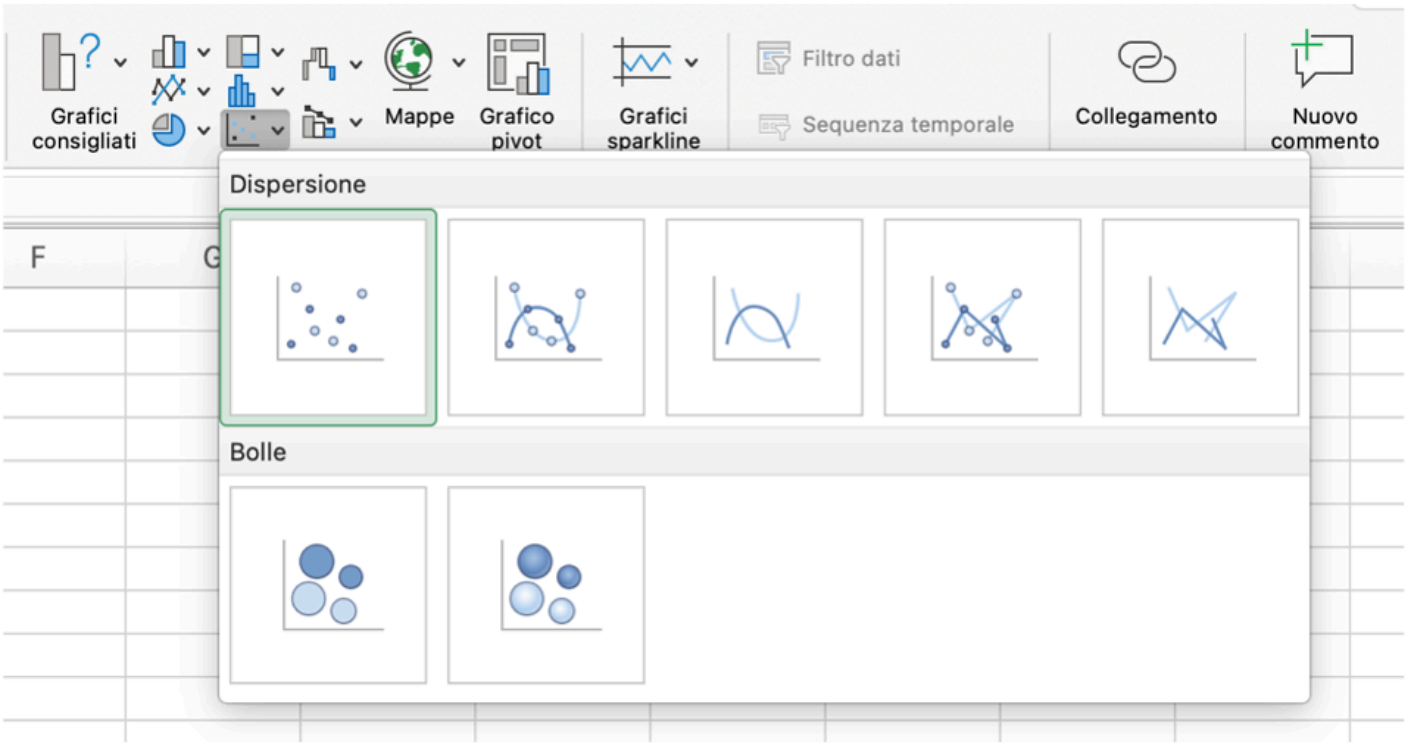
- è sensibile ai **valori anomali (outlier)**
  - Un outlier è un dato che si discosta molto dagli altri valori del dataset. La regressione, cercando di adattarsi a tutti i dati, può inclinare la retta in modo sbagliato per tener conto di questo punto isolato.
- la correlazione **non implica causalità**
  - La regressione evidenzia una relazione statistica, ma non dimostra automaticamente una relazione di causa-effetto tra le variabili.
- rischio di **sovradattamento (overfitting)**
  - L'overfitting si verifica quando un modello si adatta troppo bene ai dati di addestramento, catturando non solo la relazione reale tra le variabili ma anche il rumore casuale presente nei dati. In altre parole, il modello diventa troppo complesso e finisce per "memorizzare" i dati osservati invece di imparare la struttura generale del fenomeno.

Per questo motivo è importante **interpretare i dati e i risultati con attenzione**.

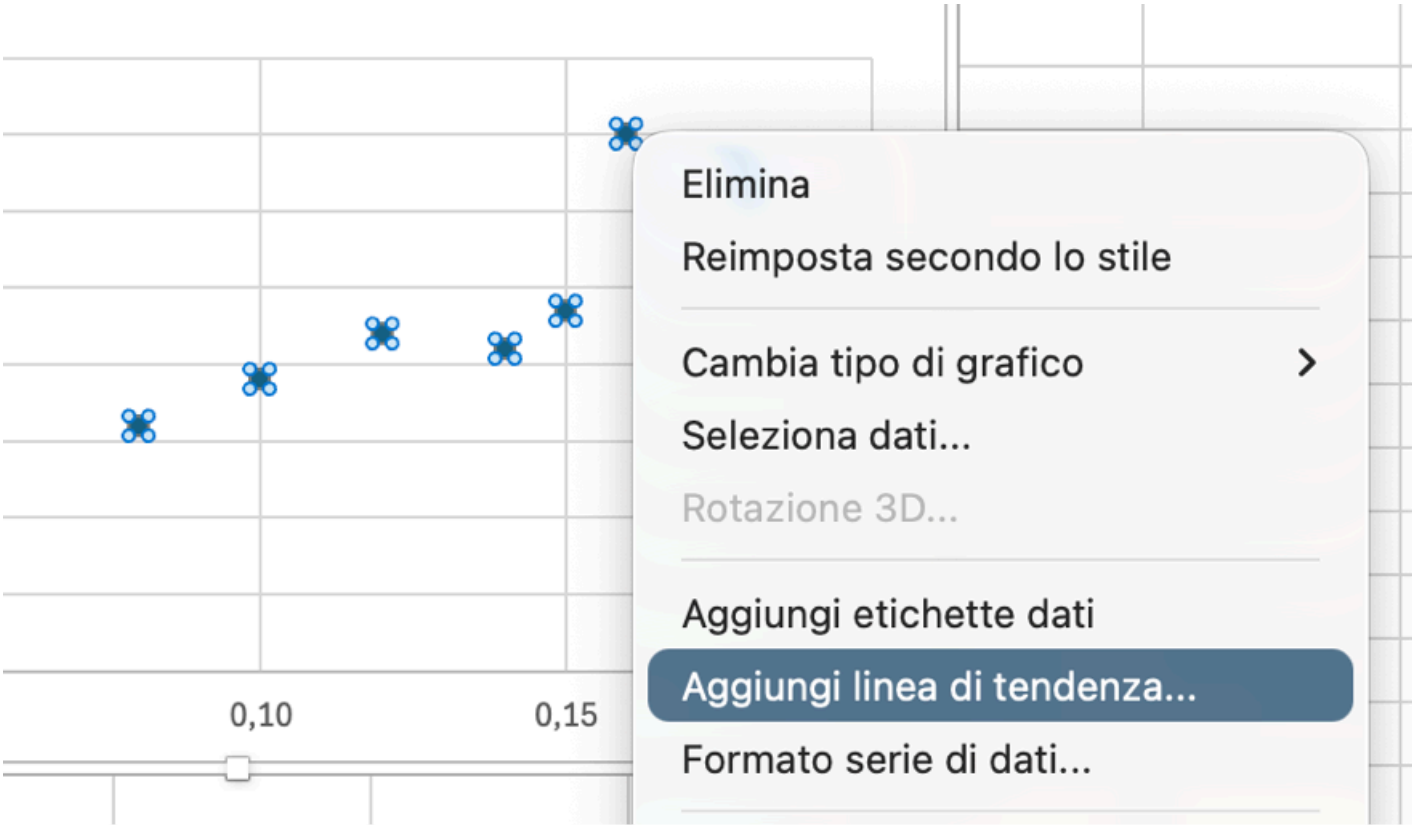
LA REGRESSIONE IN EXCEL: LE LINEE DI TENDENZA

Un modo semplice per applicare la regressione lineare è utilizzare **Microsoft Excel**.
Excel mette a disposizione una funzione chiamata **linea di tendenza**, che consente di stimare rapidamente una regressione.

Il procedimento è semplice:
1. inserire i dati in due colonne (X e Y)
2. creare un **grafico a dispersione**



3. selezionare i punti del grafico
4. scegliere **Aggiungi linea di tendenza**



Excel consente di scegliere diversi tipi di regressione:

- **lineare**
- esponenziale
- logaritmica
- polinomiale
- media mobile

È anche possibile:

- visualizzare **l'equazione della curva**
- visualizzare il **coefficiente di determinazione** R^2

Il coefficiente **R^2** indica **quanto bene il modello spiega i dati**:

- R^2 vicino a **1** → modello molto buono
- R^2 vicino a **0** → relazione debole

ATTIVITÀ PRATICA N.1

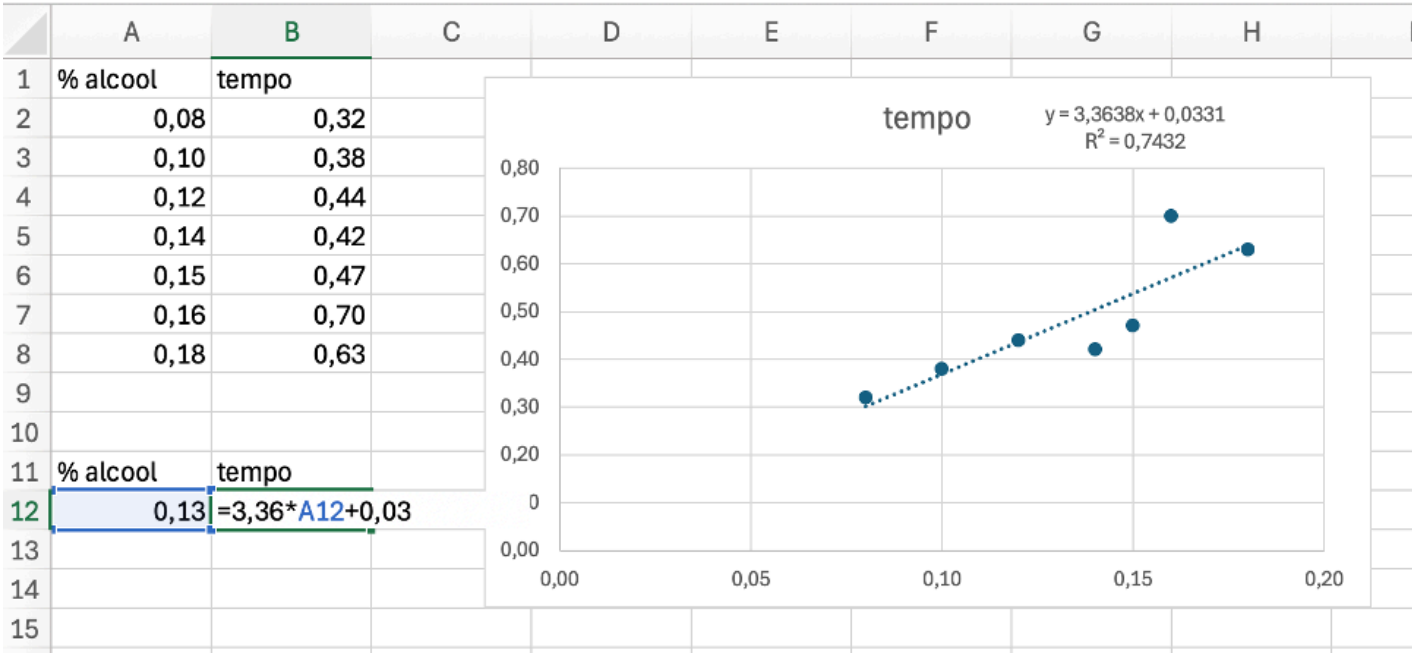
Regressione lineare

Si ritiene che più alcool c'è in circolo, più lento sia il tempo di reazione di una persona. Per verificare questa affermazione, 7 volontari assumono ciascuno una diversa quantità di alcool. La concentrazione di alcool nel sangue viene determinata come percentuale del peso corporeo. In seguito viene misurato il tempo di reazione di ciascuno a un certo stimolo, ottenendo i seguenti dati.

concentrazione di alcool nel sangue (%)	tempo di reazione (secondi)
0.08	0.32
0.10	0.38
0.12	0.44
0.14	0.42
0.15	0.47
0.16	0.70
0.18	0.63

1. Predire il tempo di reazione di un individuo con una concentrazione di alcool nel sangue di x=0,13 %
2. Nella tabella si può osservare un valore anomalo?
3. La regressione lineare è un modello valido?

Scriviamo nelle colonne A e B i dati raccolti, creiamo un grafico a dispersione e aggiungiamo una linea di tendenza di tipo lineare mostrando anche l'equazione della retta e il coefficiente R^2.



1. La retta di regressione è:

$$y = 3{,}3638x + 0{,}0331$$

Sostituiamo x = 0,13:

$$y = 3{,}3638 \cdot 0{,}13 + 0{,}0331$$
$$y = 0{,}4373 + 0{,}0331 \approx 0{,}4704$$

Il tempo di reazione previsto per x=0,13 è ≈ 0,47 secondi.

Nella riga 12 del foglio abbiamo riportato la formula per calcolare facilmente altre previsioni.

2. Si nota che:

- fino a **0,15** il tempo cresce abbastanza regolarmente
- a **0,16** salta improvvisamente a **0,70**

Il punto **(0.16 , 0.70)** è molto più alto degli altri e rompe il trend.
Quindi **può essere considerato un possibile valore anomalo (outlier)**.

3. Il coefficiente di determinazinoe è: $R^2 = 0{,}74$

Questo significa una correlazione **positiva abbastanza forte** ma **non perfetta.**

Il modello lineare è ragionevolmente valido, perché mostra una relazione positiva abbastanza chiara. Tuttavia il campione è **molto piccolo (7 dati)** e c'è **un possibile outlier (0.16, 0.70)** che può influenzare la retta.

ATTIVITÀ PRATICA N.2

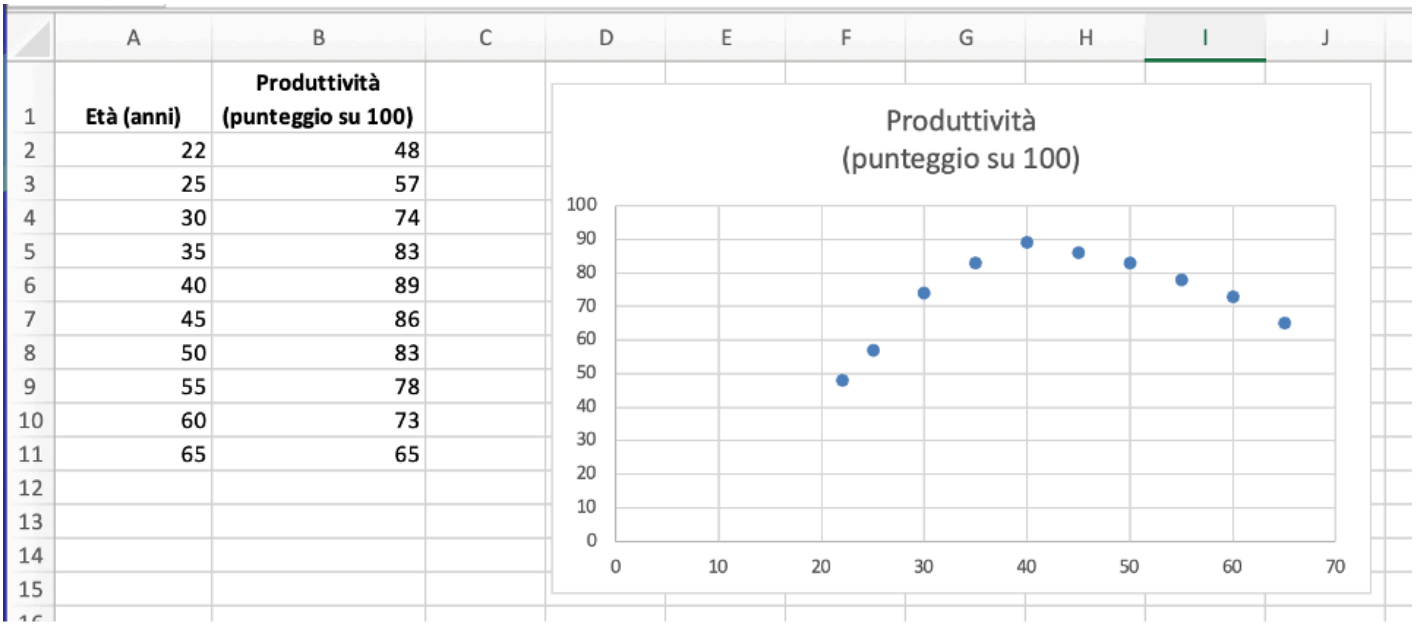
Regressione non lineare

Un'azienda vuole analizzare come varia la produttività lavorativa in base all'età dei dipendenti. Viene misurata la produttività (espressa in punteggio da 0 a 100) di un campione di lavoratori di diverse età.

I dati raccolti sono riportati nella tabella seguente:

Età (anni)	Produttività (punteggio su 100)
22	48
25	57
35	83
40	89
45	86
50	83
55	78
60	73
65	65

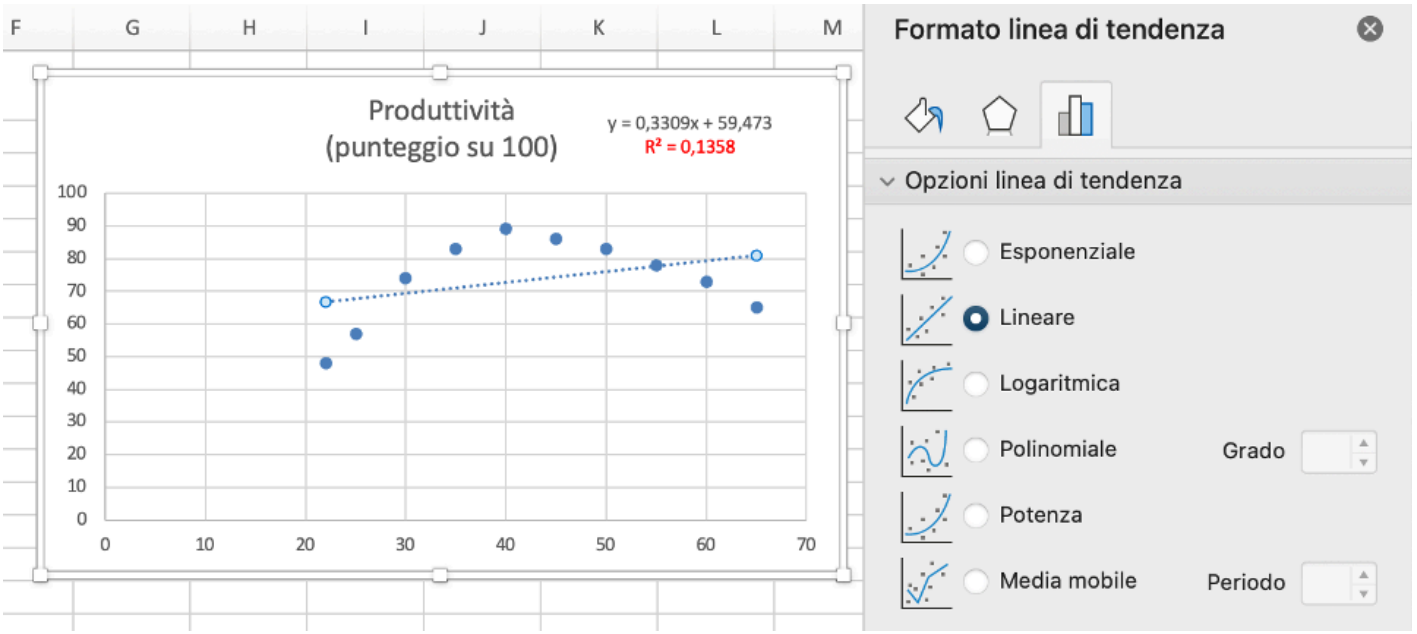
1. A quale età si raggiunge la massima produttività?
2. Qual è l'impatto dell'età sulla produttività dopo il picco?
3. Quale modello è ben adattato ai dati? (R^2 e significatività della regressione).



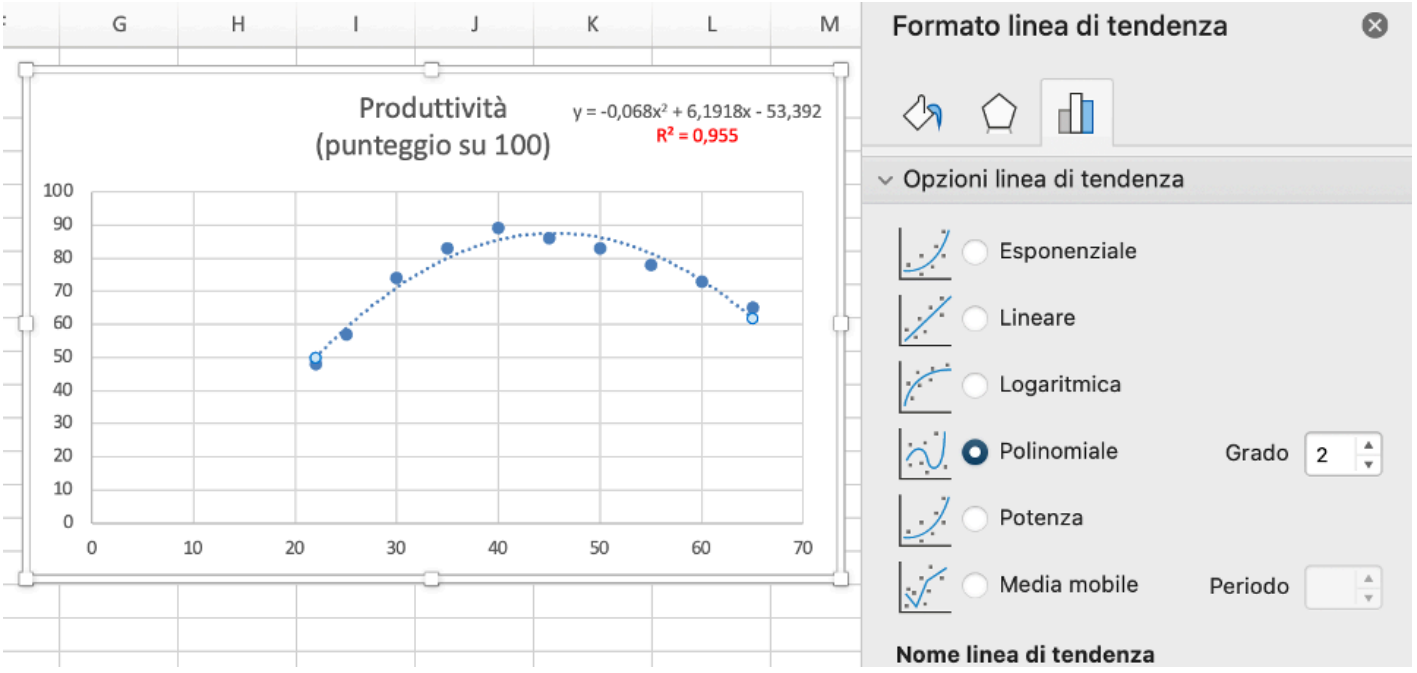
Osservando il grafico appare evidente che la produttività di un lavoratore in relazione all'età non è lineare:

- Aumenta nei primi anni con l'esperienza.
- Raggiunge un massimo a una certa età.
- Poi inizia a diminuire per l'invecchiamento.

Se aggiungiamo una curva di regressione lineare otteniamo un coefficiente di determinazione molto basso, poiché la correlazione è evidentemente non lineare:



Se scegliamo una curva di tipo polinomiale di secondo grado (una parabola) si ottiene un modello migliore:



L'analisi dei dati mostra che la **produttività raggiunge il suo massimo intorno ai 40 - 45 anni**. Dopo questo picco, **tende a diminuire gradualmente con l'aumentare dell'età**. Il modello di regressione utilizzato descrive bene i dati: il coefficiente di determinazione **R^2**

= 0,955 indica infatti **una correlazione molto elevata**, spiegando circa il **95,5% della variabilità della produttività**.

ATTIVITÀ PRATICA N. 3

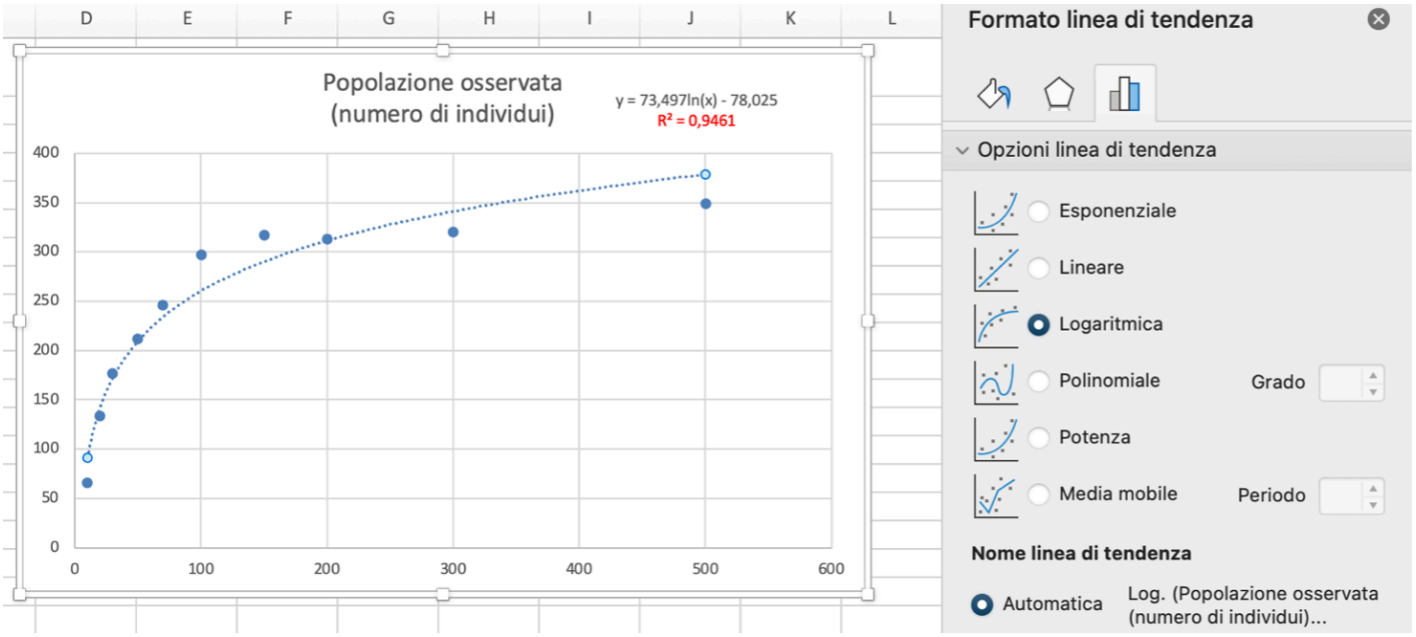
Regressione non lineare

Un team di ecologi sta studiando come la crescita della popolazione di una specie animale dipenda dalle risorse disponibili in un ecosistema. Si ipotizza che all'aumentare delle risorse, la popolazione cresca rapidamente all'inizio, ma con un rallentamento progressivo dovuto alla saturazione dell'habitat.

I dati raccolti sono riportati nella tabella seguente.

Risorse disponibili (tonnellate di cibo)	Popolazione osservata (numero di individui)
10	66
20	134
30	177
50	212
70	246
100	297
150	317
200	313
300	320
500	349

1. Qual è il tasso di crescita della popolazione rispetto alle risorse?
2. Dopo quante risorse la crescita della popolazione diventa marginale?
3. Quale modello si adatta bene ai dati?



L'analisi mostra che **la crescita della popolazione dipende dalla disponibilità di risorse**. Quando le risorse sono abbondanti, la popolazione cresce rapidamente. Tuttavia, con l'aumento del numero di individui, **le risorse disponibili devono essere condivise tra più soggetti**, quindi **la quantità di risorse pro capite diminuisce**. Questa riduzione porta progressivamente a **un rallentamento della crescita della popolazione**, fino a una situazione di quasi stabilizzazione.

Dai dati si osserva inoltre che **oltre circa 100 - 150 unità di risorse la crescita diventa marginale**, cioè la popolazione aumenta sempre meno. Per descrivere questo andamento, **il modello logaritmico risulta quello che si adatta meglio ai dati**, perché rappresenta bene una crescita inizialmente rapida che poi tende a rallentare.

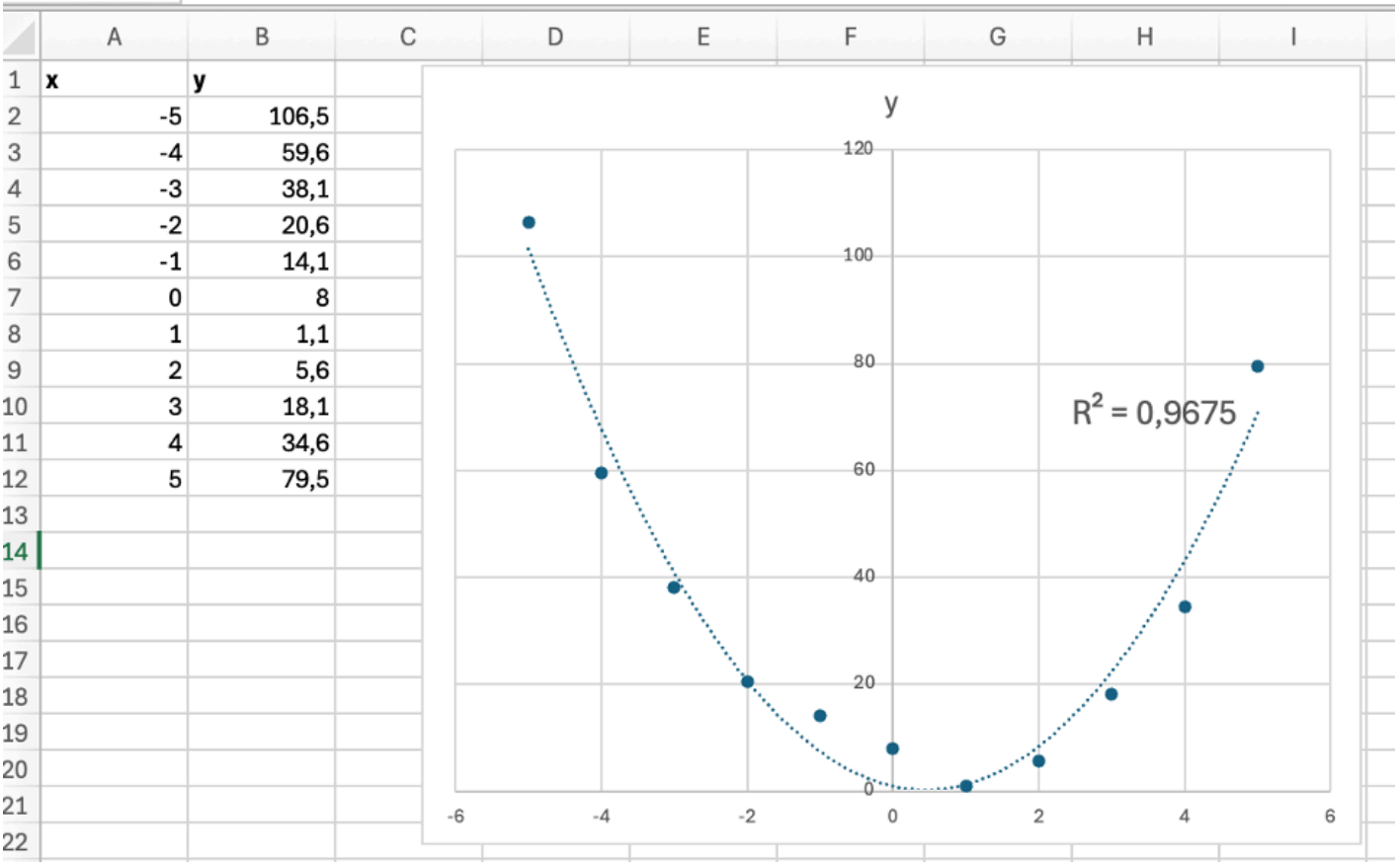
IL PROBLEMA DELL'OVERFITTING

L'overfitting è un fenomeno che si verifica quando un modello statistico si adatta eccessivamente, includendo anche il rumore o le fluttuazioni casuali che non rappresentano una tendenza generale.

Il modello può diventare troppo complesso (ad esempio, usando polinomi di grado molto elevato o troppi parametri), adattandosi perfettamente ai dati, ma risultando incapace di generalizzare bene dati di previsione.

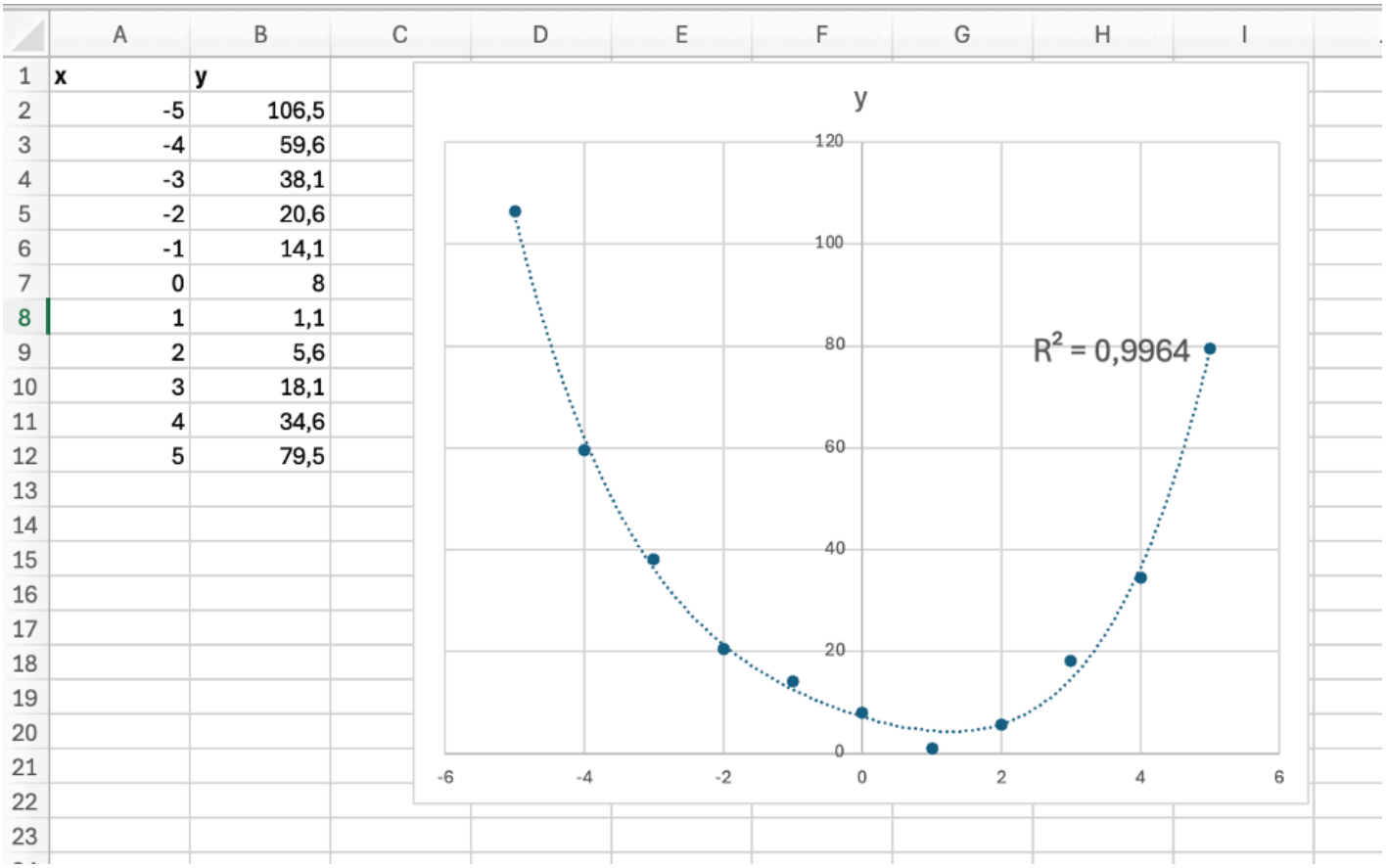
Immagina di cercare di prevedere l'andamento di una serie temporale. Se usi un modello molto complesso (come un polinomio di alto grado), il modello potrebbe "tracciare" perfettamente ogni piccola fluttuazione nei dati storici, ma non sarebbe in grado di prevedere correttamente i dati futuri, che potrebbero avere un andamento diverso.

Esempio: regressione polinomiale di grado 2



	A	B
1	**x**	**y**
2	-5	106,5
3	-4	59,6
4	-3	38,1
5	-2	20,6
6	-1	14,1
7	0	8
8	1	1,1
9	2	5,6
10	3	18,1
11	4	34,6
12	5	79,5

Esempio: regressione polinomiale di grado 4 sugli stessi dati



x	y
-5	106,5
-4	59,6
-3	38,1
-2	20,6
-1	14,1
0	8
1	1,1
2	5,6
3	18,1
4	34,6
5	79,5

Sebbene il coefficiente di determinazione del modello di grado 4 sia superiore (0,9964 rispetto a 0,9675) l'equazione della curva è più complessa: $y = 0,08x^4 + 0,06x^3 + 1,25x^2 - 4,13x + 7,19$ rispetto a $y = 3,4x^2 - 3,0x + 0,98$.

In generale i modelli più semplici, come la regressione lineare o polinomiale di basso grado, sono generalmente più facili da comprendere e interpretare. Questo è cruciale in molti settori dove l'interpretazione dei risultati è altrettanto importante quanto la previsione, come in medicina, finanza, e scienze sociali. Come già detto all'inizio di questo articolo, un modello troppo complesso può infatti adattarsi molto bene ai dati osservati ma **rischia di catturare anche il rumore o le variazioni casuali presenti nel campione**, perdendo così la capacità di generalizzare a nuovi dati. In questi casi il modello sembra molto accurato sui dati utilizzati per costruirlo, ma le sue previsioni risultano meno affidabili quando viene applicato a situazioni diverse.

Il problema dell'overfitting **non riguarda solo i modelli statistici tradizionali**, ma è presente anche nei **modelli di intelligenza artificiale e di machine learning**, che spesso sono molto complessi e basati su grandi quantità di parametri. Per questo motivo, nello sviluppo di tali modelli è importante trovare un equilibrio tra complessità e capacità di generalizzazione, in modo da ottenere previsioni accurate senza compromettere l'affidabilità del modello su nuovi dati.