

**WELCOME TO THE DARK SIDE OF  
SCIENCE**

**DATA SCIENCE**

# Machine Learning

```
graph LR; ML((Machine Learning)) --- UL((Unsupervised Learning)); ML --- SL((Supervised Learning)); ML --- RL((Reinforcement Learning)); UL --- DR((Dimensionality Reduction)); UL --- CL((Clustering)); DR --- MC(Meaningful Compression); DR --- SD(Structure Discovery); DR --- FE(Feature Elicitation); DR --- BDV(Big data Visualisation); CL --- RS(Recommender Systems); CL --- TM(Targetted Marketing); CL --- CS(Customer Segmentation); SL --- C((Classification)); SL --- R((Regression)); C --- IC(Image Classification); C --- IFD(Identity Fraud Detection); C --- CR(Customer Retention); C --- D(Diagnostics); R --- AP(Advertising Popularity Prediction); R --- WF(Weather Forecasting); R --- PG(Population Growth Prediction); R --- MF(Market Forecasting); R --- EL(Estimating life expectancy); RL --- RTD(Real-time decisions); RL --- GA(Game AI); RL --- RN(Robot Navigation); RL --- SA(Skill Acquisition); RL --- LT(Learning Tasks)
```

## Unsupervised Learning

### Dimensionality Reduction

- Meaningful Compression
- Structure Discovery
- Feature Elicitation
- Big data Visualisation

### Clustering

- Recommender Systems
- Targetted Marketing
- Customer Segmentation

## Supervised Learning

### Classification

- Image Classification
- Customer Retention
- Diagnostics
- Identity Fraud Detection

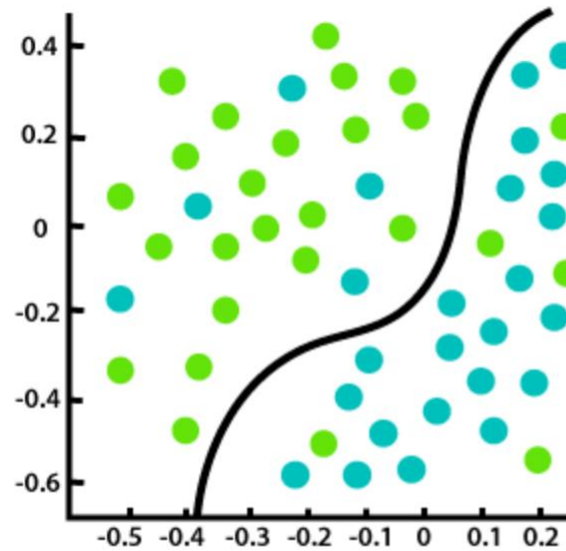
### Regression

- Advertising Popularity Prediction
- Weather Forecasting
- Market Forecasting
- Estimating life expectancy
- Population Growth Prediction

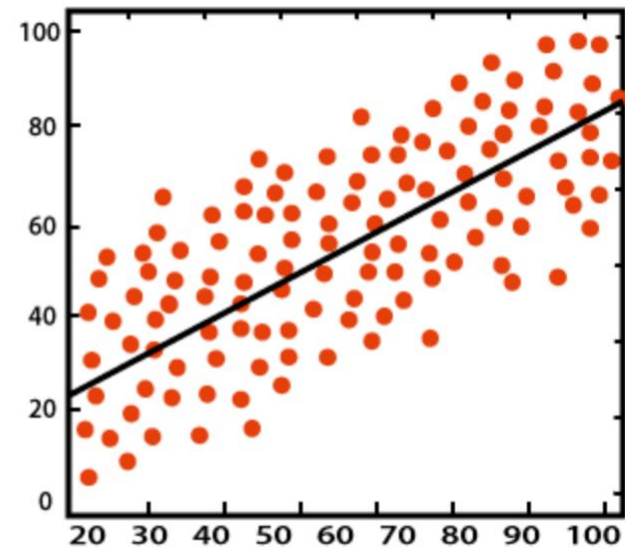
## Reinforcement Learning

- Real-time decisions
- Game AI
- Skill Acquisition
- Learning Tasks
- Robot Navigation

# SUPERVISIONATO

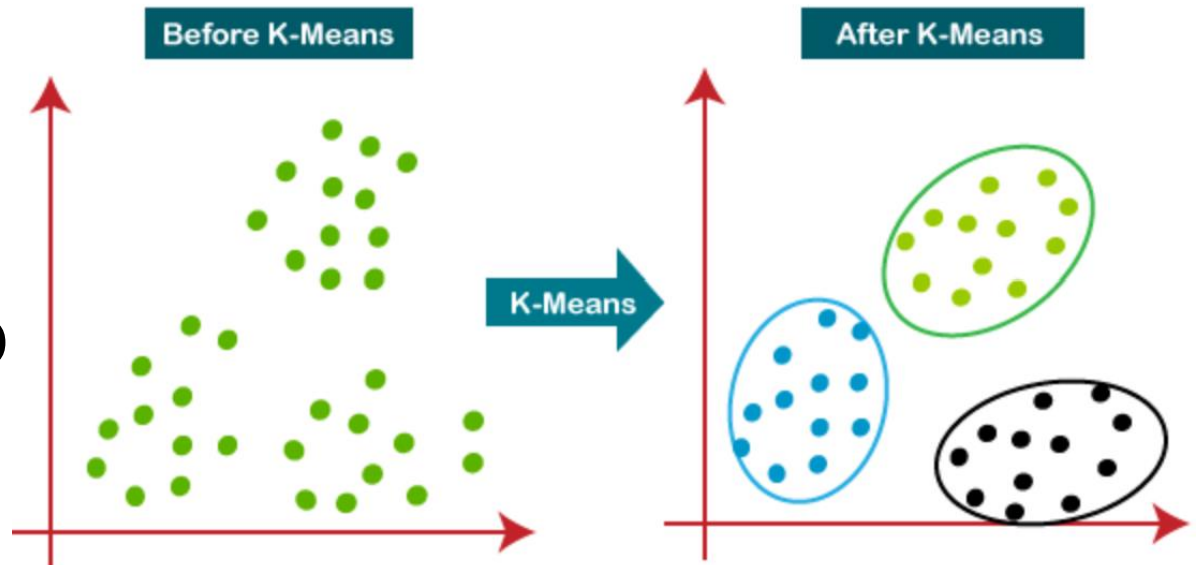


Classification

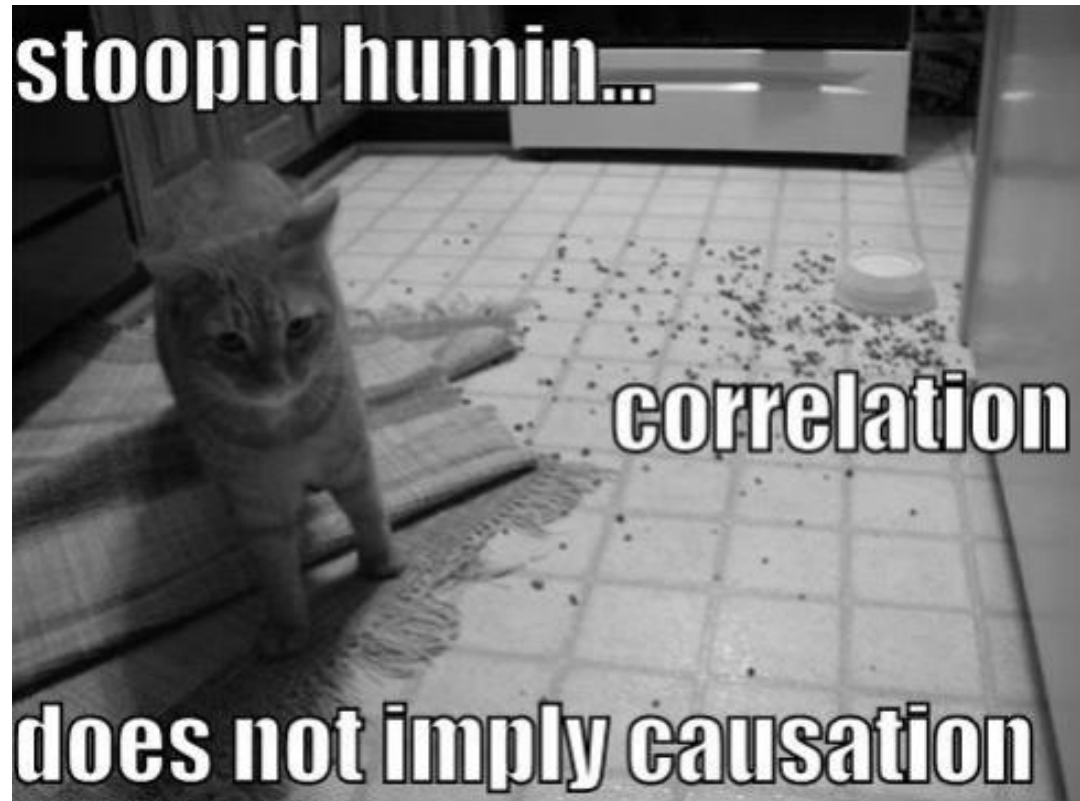


Regression

# NON SUPERVISIONATO



# LA REGRESSIONE LINEARE





Quando tra due variabili c'è una relazione di dipendenza, si può cercare di prevedere il valore di una variabile in funzione del valore assunto dall'altra.

Questo ha significato in senso stretto quando si ipotizza una relazione di causalità tra la variabile **indipendente**, su cui si agisce, e quella **dipendente**, su cui si vuole produrre un effetto.

Volendo costruire un **modello statistico** per prevedere  $Y$  in funzione di  $X$ , si pone la questione di quale relazione funzionale ipotizzare tra la variabile indipendente  $X$  e la variabile dipendente  $Y$ .

Il modello più semplice di relazione tra due variabili è quello lineare di primo grado, rappresentato da una retta, la cui equazione è :

$$y = a + b x$$

Una volta determinata la retta, il modello permetterà di stimare il valore della variabile  $Y$  sulla base del valore assunto dalla  $X$

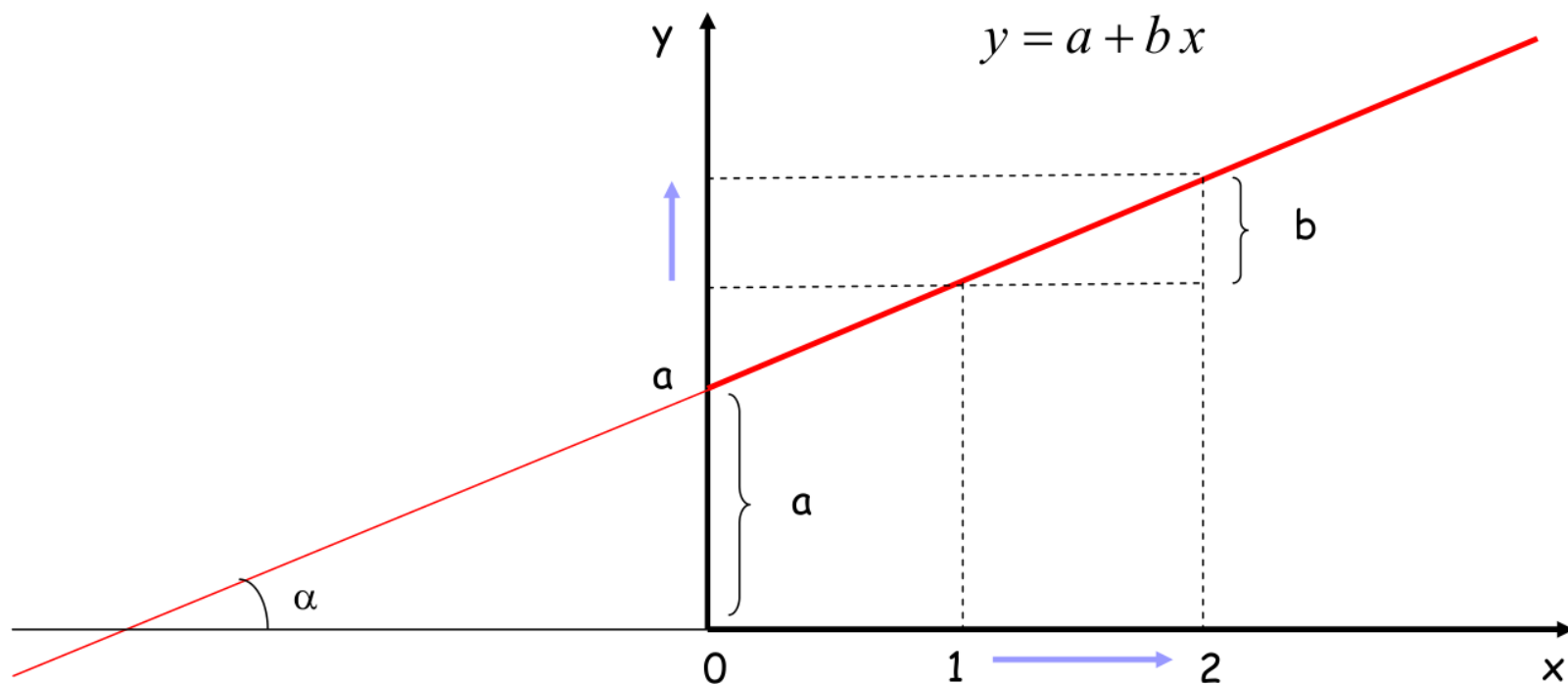
Per ottenere un buon modello, e quindi delle buone previsioni, occorre determinare la retta che meglio *descrive* i punti osservati: in pratica, si tratta di determinare i due coefficienti  $a$  e  $b$  che compaiono nell'equazione della retta:  $y = a + b x$

Ripassiamo un po' di geometria (e di trigonometria ... ):

- il parametro **a** è l'**intercetta** della retta con l'asse delle ordinate
- il parametro **b** è il **coefficiente angolare**: misura l'inclinazione della retta (è la tangente dell'angolo  $\alpha$  formato dalla retta con l'asse delle ascisse)

In pratica:

- a ci dice quanto vale Y quando X vale 0
- b ci dice di quanto aumenta Y all'aumentare di una unità di X



- Se i punti fossero solo due, o fossero tutti allineati, determinare la retta interpolante sarebbe facile (per due punti passa una sola retta), ma sfortunatamente nella realtà i dati osservativi non sono mai esattamente allineati
- Per determinare la retta di regressione che meglio descrive i dati osservati è allora necessario stabilire un criterio *statistico* di valutazione della bontà del modello:

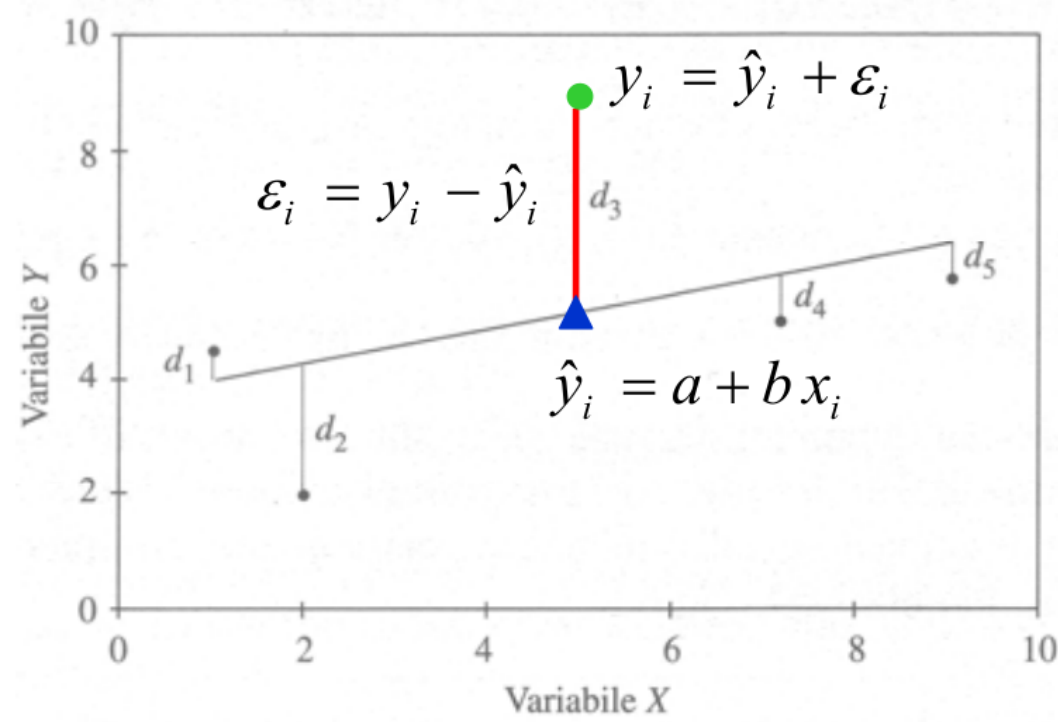
$$\hat{y}_i = a + b x_i \quad \forall i$$

- L'errore che si commette prevedendo ciascun Y osservato con il modello, può essere misurato come differenza tra il dato reale e quello previsto:

$$y_i = \hat{y}_i + \varepsilon_i$$

cioè, per ciascuna osservazione i, si commette un errore

$$\varepsilon_i = y_i - \hat{y}_i$$



**COME DETERMINARE  
LA SOLUZIONE MIGLIORE?**



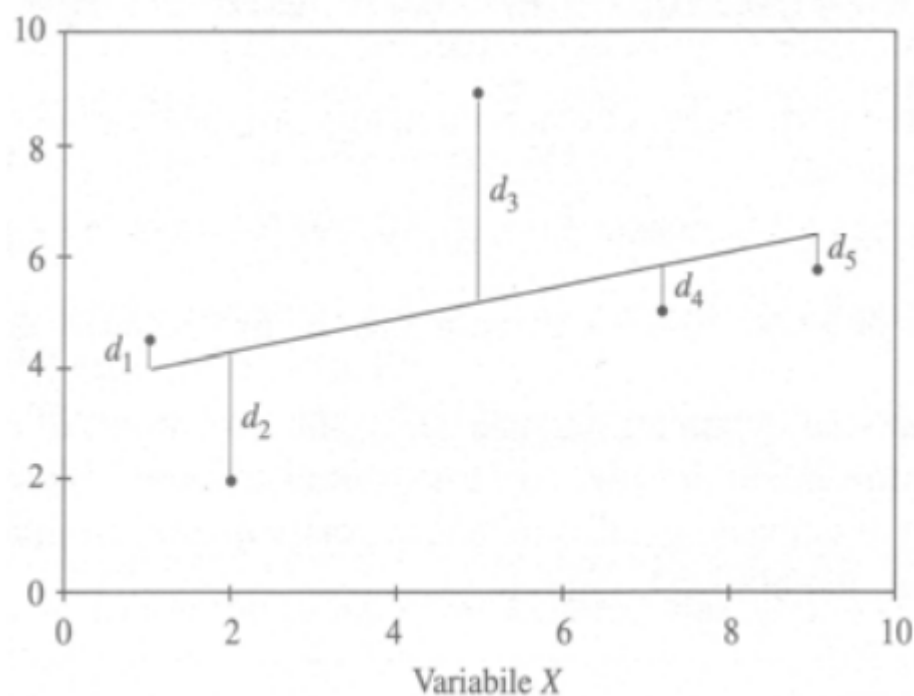
Il criterio detto dei **minimi quadrati** prevede di valutare la bontà del modello sulla base della somma dei quadrati di tutti errori di stima commessi:

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - b x_i)^2 = \min$$

La retta migliore, secondo questo criterio, è quella che minimizza la somma dei quadrati degli scarti dei valori stimati da quelli osservati, detti anche **residui** della regressione.

Perché proprio il quadrato dei residui ?

- per evitare che residui positivi e negativi si compensino
- il valore assoluto è matematicamente più scomodo da gestire e non sempre porta ad una soluzione univoca
- il quadrato dà peso maggiore agli scarti più grandi, che sono anche quelli che ci disturbano di più: è meglio fare tanti piccoli errori che non un errore molto grosso



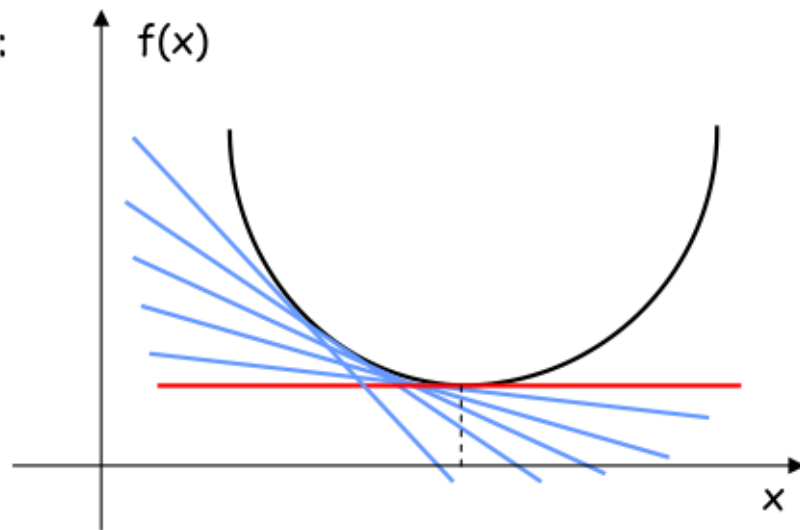
## ■ Ricerca del minimo di una funzione

- Il problema è determinare i coefficienti  $a$  e  $b$  del modello in modo da minimizzare la quantità :

$$\sum_{i=1}^n (y_i - a - b x_i)^2 = \min$$

- La soluzione di questo problema in matematica è semplice: si tratta di trovare il minimo di una funzione  $f(x)$
- La soluzione ad un problema di ricerca del minimo di una funzione si determina trovando i punti "di svolta" della funzione, in cui la curva cambia andamento (concavità) : in un punto di svolta, la retta tangente alla curva risulta orizzontale
- L'inclinazione della retta tangente ad una curva in un punto è data dalla derivata prima della  $f(x)$  in quel punto, indicata come:  $f'(x)$
- Quindi nel punto di svolta la derivata prima  $f'(x)$  della funzione  $f(x)$  si annulla
- Allora per determinare i punti di minimo di una funzione  $f(x)$  è necessario imporre che valga zero la sua derivata prima rispetto a  $x$ , in questo modo si determina il punto  $x$  in cui vale:

$$f'(x) = \frac{\partial f(x)}{\partial x} = 0$$



## Stima dei parametri ai minimi quadrati

Nel nostro caso vogliamo minimizzare una funzione  $f(a,b)$  rispetto ad  $a$  e  $b$ : dovremo imporre che siano nulle le *derivate prime parziali* della funzione rispetto ad  $a$  e  $b$ :

$$\sum_{i=1}^n (y_i - a - b x_i)^2 = \min$$

Derivando la funzione prima rispetto ad  $a$  e poi rispetto a  $b$ , e ponendo a zero tali derivate, si ottiene un sistema di due equazioni di primo grado in due incognite ( $a$  e  $b$ ):

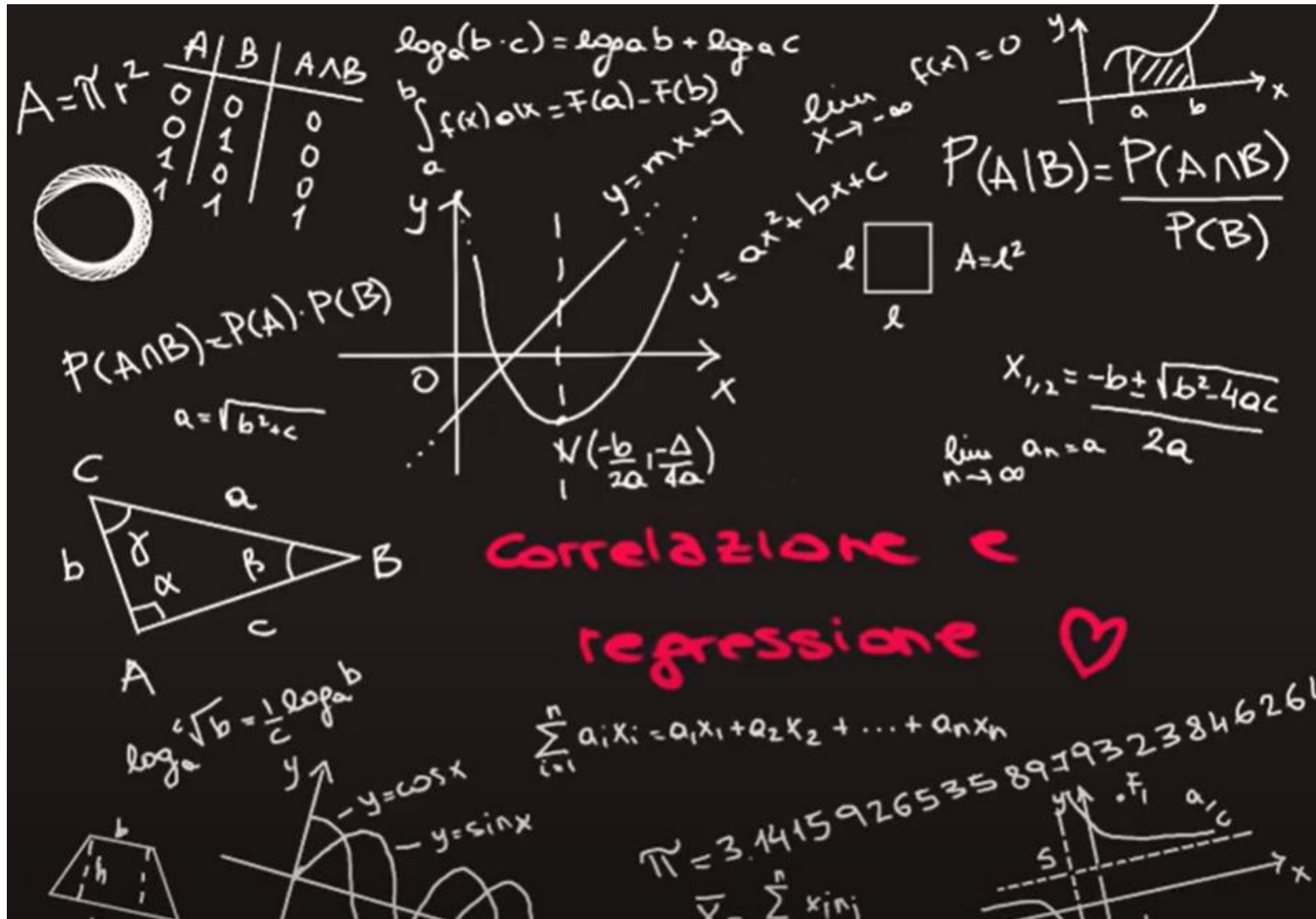
$$\begin{cases} \frac{\partial \sum (y_i - a - b x_i)^2}{\partial a} = 0 \\ \frac{\partial \sum (y_i - a - b x_i)^2}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} -2 \sum (y_i - a - b x_i) = 0 \\ -2 \sum (y_i - a - b x_i) x_i = 0 \end{cases} \Rightarrow$$

$$\begin{cases} \sum (y_i - a - b x_i) = 0 \\ \sum (y_i - a - b x_i) x_i = 0 \end{cases}$$

**Sistema di equazioni normali  
di regressione**

Risolvendo il sistema rispetto ad  $a$  e  $b$ , si determinano le stime ai minimi quadrati dei parametri della retta di regressione

# ES1: RISOLVERE IL SISTEMA DI EQUAZIONI – LET'S GO

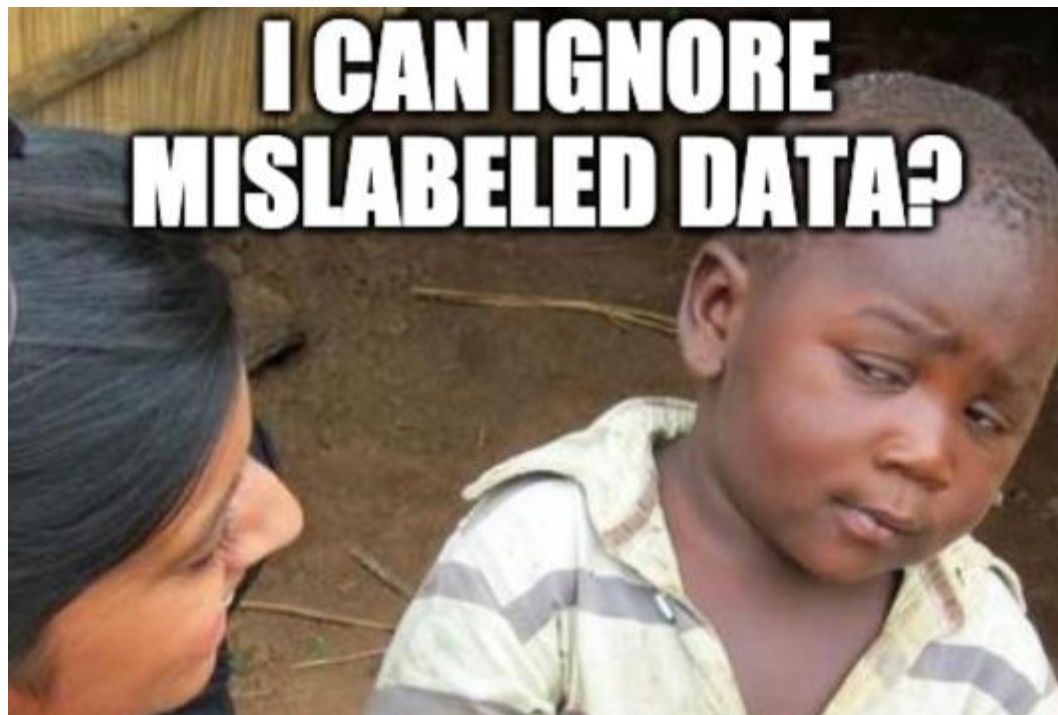




Machine Learning with Scikit-Learn

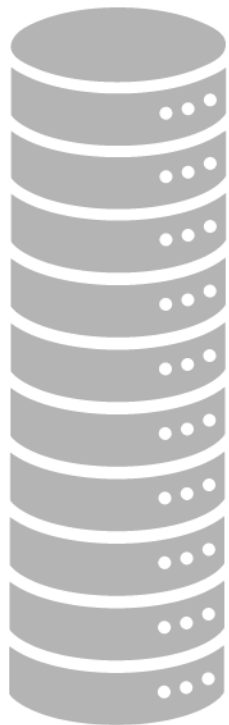
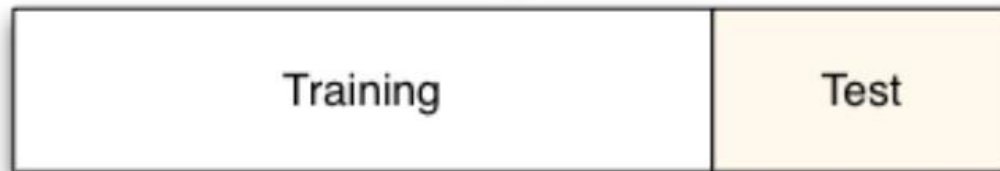


COME FACCIAMO A CAPIRE SE LE  
PERFORMANCE DEL MIO MODELLO  
SARANNO OTTIMALI ANCHE CON  
NUOVI DATI SENZA LABEL?





Data



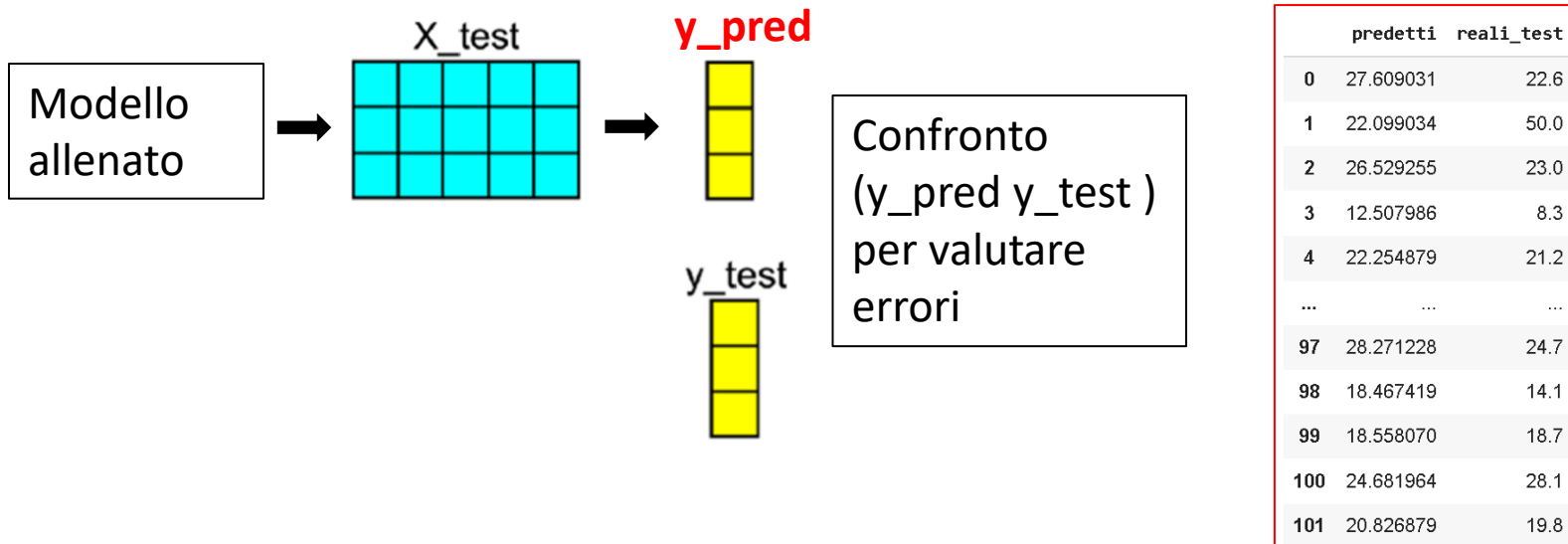
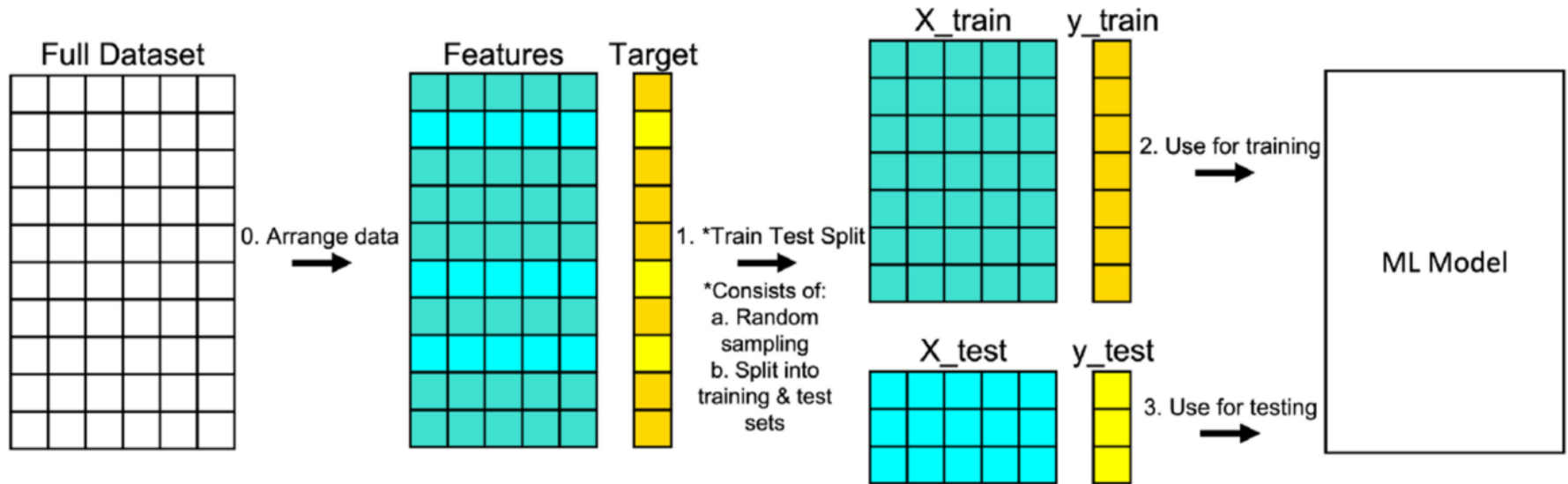
Testing set



Training set

# INPUT = FEATURES

# OUTPUT = TARGET



WHAT IS MY PURPOSE?

learn

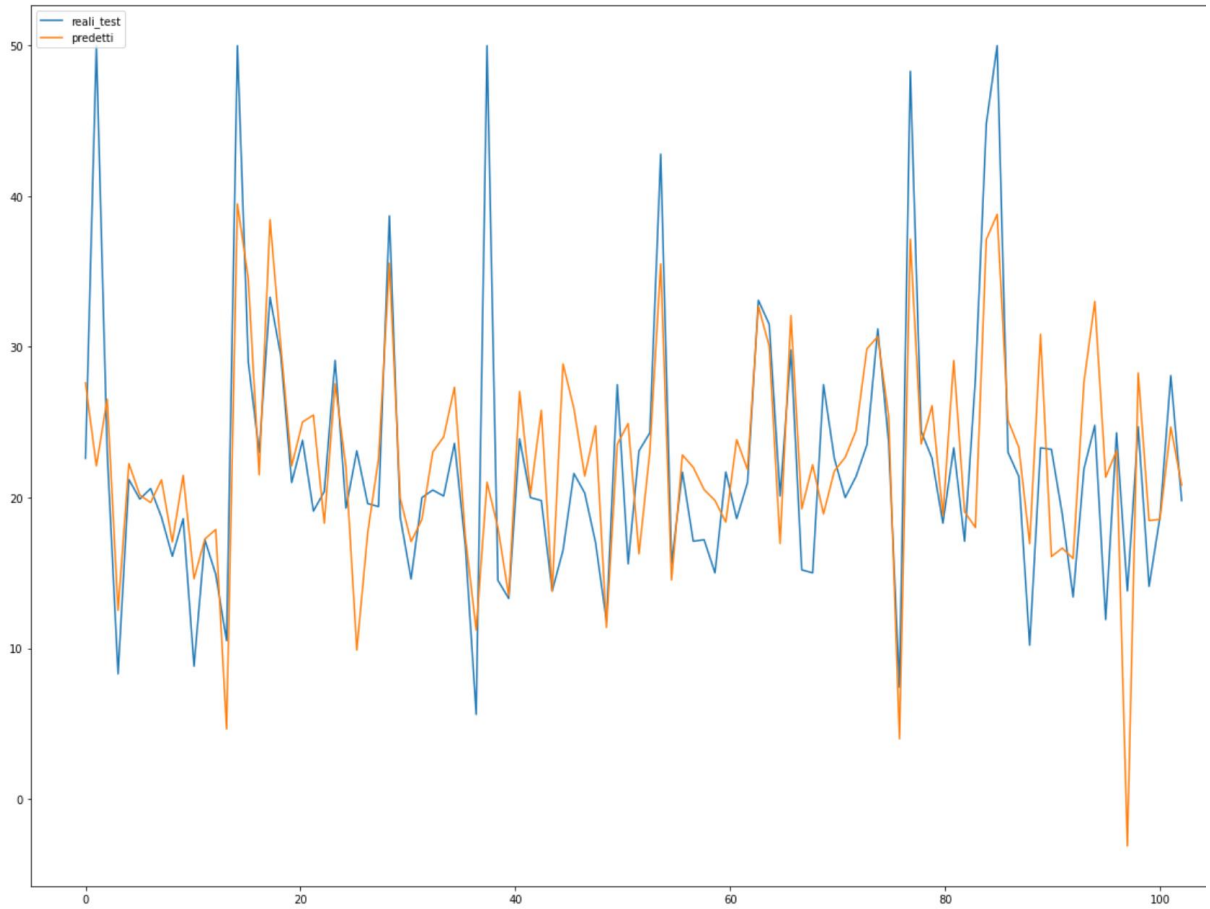
YOU SPLIT THE DATA

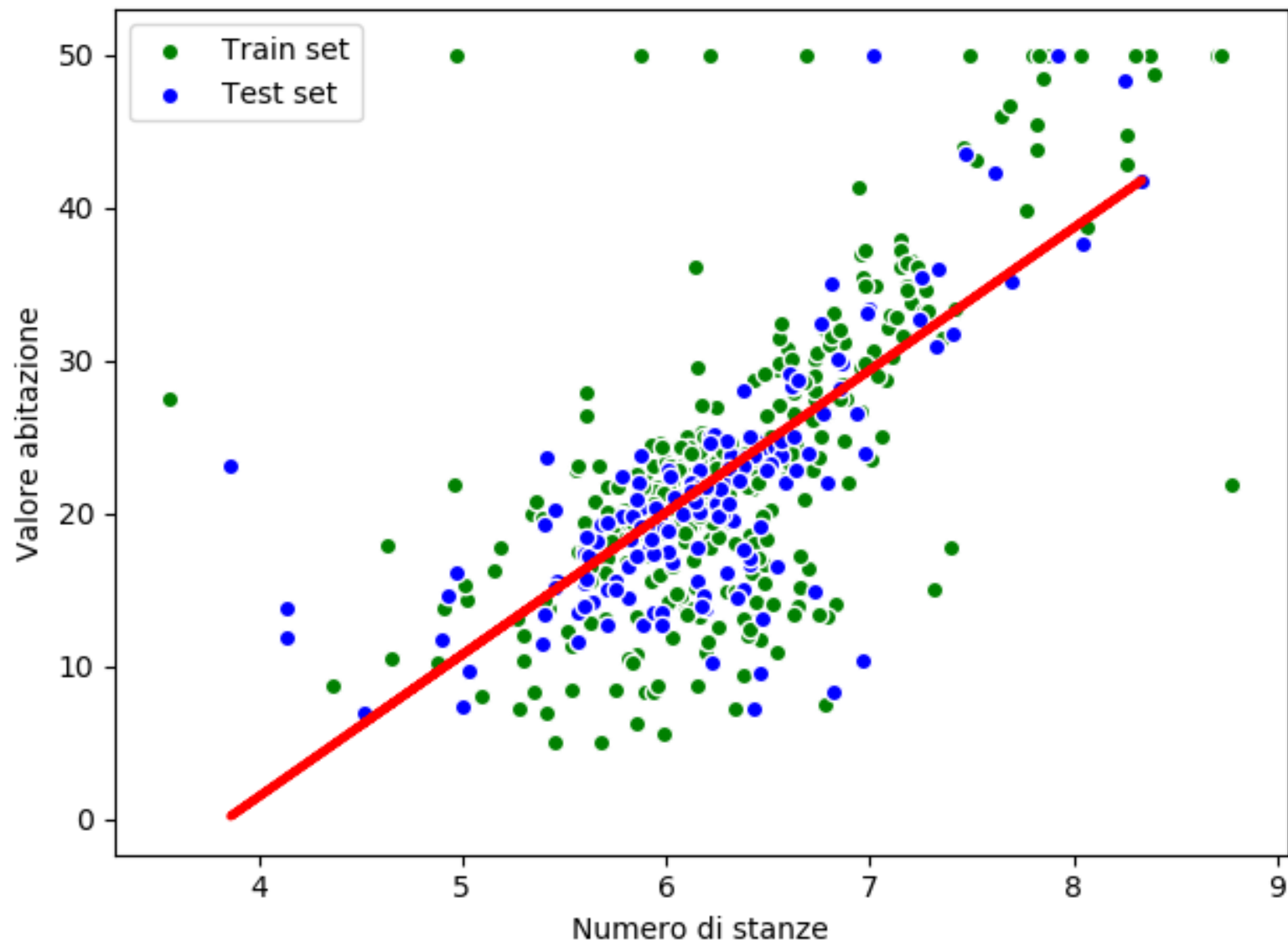
learn

OH MY GOD.

learn

	predetti	reali_test
0	27.609031	22.6
1	22.099034	50.0
2	26.529255	23.0
3	12.507986	8.3
4	22.254879	21.2
...	...	...
97	28.271228	24.7
98	18.467419	14.1
99	18.558070	18.7
100	24.681964	28.1
101	20.826879	19.8





Introducendo opportune assunzioni si ottiene il **modello di regressione lineare semplice**.

**Assunzione 1:**

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{per ogni osservazione } i=1, \dots, n$$

**Assunzione 2:**

**Le  $\varepsilon_i$  sono variabili casuali indipendenti con valore atteso  $E(\varepsilon_i) = 0$  e varianza costante  $V(\varepsilon_i) = \sigma^2$  per ogni  $i=1, \dots, n$**

**Assunzione 3:**

**I valori  $x_i$  della variabile esplicativa X sono noti senza errore**



**LEVEL 01**

**MISSION COMPLETE**

