



## **ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA – EDIZIONE 2**

**Operazione Rif. PA 2019-11596/RER “Anticipare la crescita con le nuove competenze sui Big Data”, approvata dalla Regione Emilia-Romagna con DGR n° 789 del 20 maggio 2019 e co-finanziata dal Fondo Sociale Europeo PO 2014-2020**



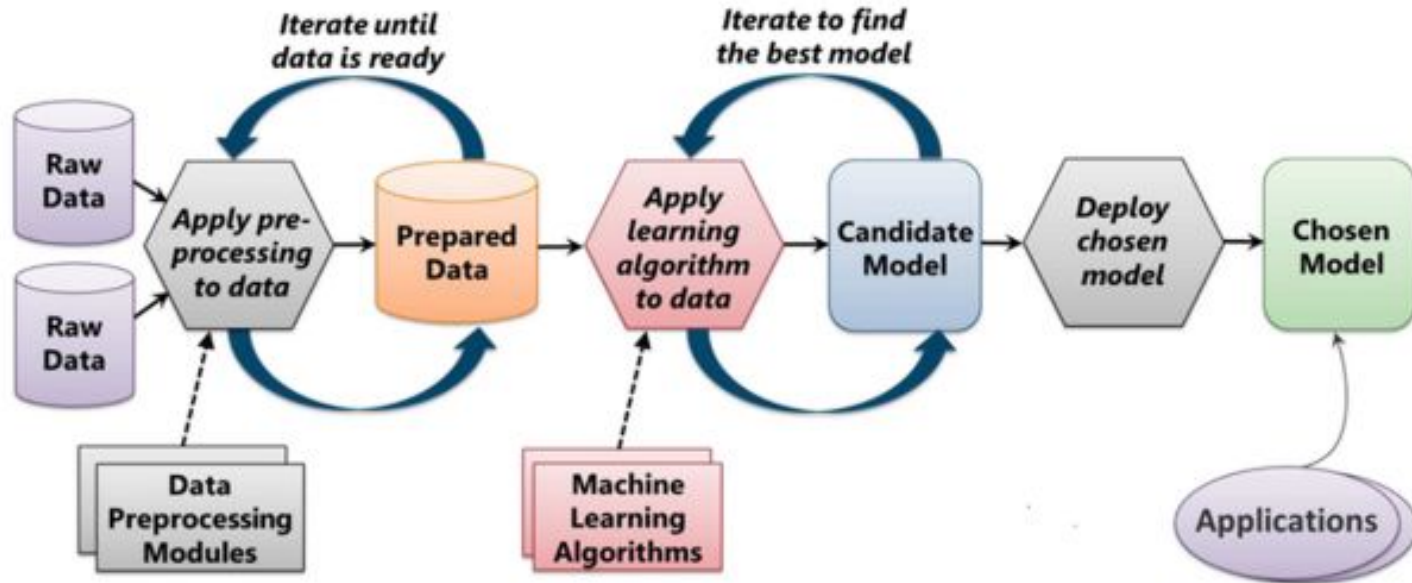
ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA  
CENTRO NAZIONALE DI RICERCA E INNOVAZIONE



# Programma della lezione

- Tecniche di pre-processing del dataset
- Come sostituire i missing
- Come codificare le variabili categoriche
- Metodi per normalizzare le variabile numeriche
- Come gestire un dataset sbilanciato

# Steps in progetto di Data Science



# Cos'è la Data Pre-Processing?

- Consiste in un insieme di tecniche per trasformare i dati grezzi in un formato comprensibile per i modelli
- Ogni modello di machine ha diverse necessità:
  - Esempio: KNN, SVM, K-means **hanno bisogno di standardizzare** i dati
  - Esempio: Random Forest e Gradient non hanno di standardizzare i dati

# Gestire i valori mancanti:

1. Media/Mediana/Moda
2. bfill, ffill
3. interpolate
4. Eliminare

# 1. Media/Moda/Mediana

Se la colonna è:

- Numerica: si sostituisce il valore mancante con la sua media/mediaiana
- Categorica: si sostituisce il valore mancante con la sua moda

Price		Price
100	Mean = 86.66 Median = 90 ➔	100
90		90
50		50
40		40
20		20
100		100
		86.66
60		60
120		120
		86.66
200		200

Ma come facciamo con dati temporali?

Day	Temp
Day 1	33
Day 2	33
Day 3	35
Day 4	NaN
Day 5	38
Day 6	37
Day 7	39

## 2. bfill

Backward fill (Riempimento all'indietro) - propaga all'indietro il primo valore non nullo osservato

	Day	Temp
0	Day 1	33.0
1	Day 2	33.0
2	Day 3	35.0
3	Day 4	NaN
4	Day 5	38.0
5	Day 6	37.0
6	Day 7	39.0

Having null values

	Day	Temp
0	Day 1	33.0
1	Day 2	33.0
2	Day 3	35.0
3	Day 4	38.0
4	Day 5	38.0
5	Day 6	37.0
6	Day 7	39.0

bfill





## 2. ffill

Forward fill (Riempimento in avanti) - propaga in avanti l'ultimo valore non nullo osservato

	Day	Temp
0	Day 1	33.0
1	Day 2	33.0
2	Day 3	35.0
3	Day 4	NaN
4	Day 5	38.0
5	Day 6	37.0
6	Day 7	39.0

Having null values

	Day	Temp
0	Day 1	33.0
1	Day 2	33.0
2	Day 3	35.0
3	Day 4	35.0
4	Day 5	38.0
5	Day 6	37.0
6	Day 7	39.0

ffill



### 3. Interpolate

Tecnica statistica che stima i valori sconosciuti basandosi sui valori esistenti

	Day	Temp
0	Day 1	34.0
1	Day 2	33.0
2	Day 3	35.0
3	Day 4	NaN
4	Day 5	NaN
5	Day 6	NaN
6	Day 7	38.0
7	Day 8	37.0
8	Day 9	39.0

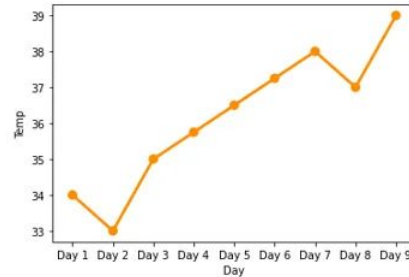
Having null values

	Day	Temp
0	Day 1	34.00
1	Day 2	33.00
2	Day 3	35.00
3	Day 4	35.75
4	Day 5	36.50
5	Day 6	37.25
6	Day 7	38.00
7	Day 8	37.00
8	Day 9	39.00

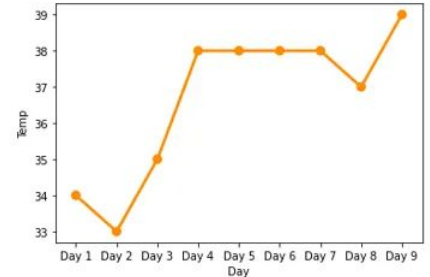
interpolate

	Day	Temp
0	Day 1	34.0
1	Day 2	33.0
2	Day 3	35.0
3	Day 4	38.0
4	Day 5	38.0
5	Day 6	38.0
6	Day 7	38.0
7	Day 8	37.0
8	Day 9	39.0

bfill



Filling missing values by interpolate method



Filling missing values by bfill method  
[Same value is filled for Day 5, Day 6, Day 7]

### 3. Interpolate

Ci sono due tipi principali:

- Lineare: Interpola i punti tracciando una linea retta tra i punti di dati circostanti.
- Polinomiale: Si costruisce un polinomio che passa tra i punti
  - Esempio di Polinomio di secondo grado:  $P(x) = a_0 + a_1x + a_2x^2$

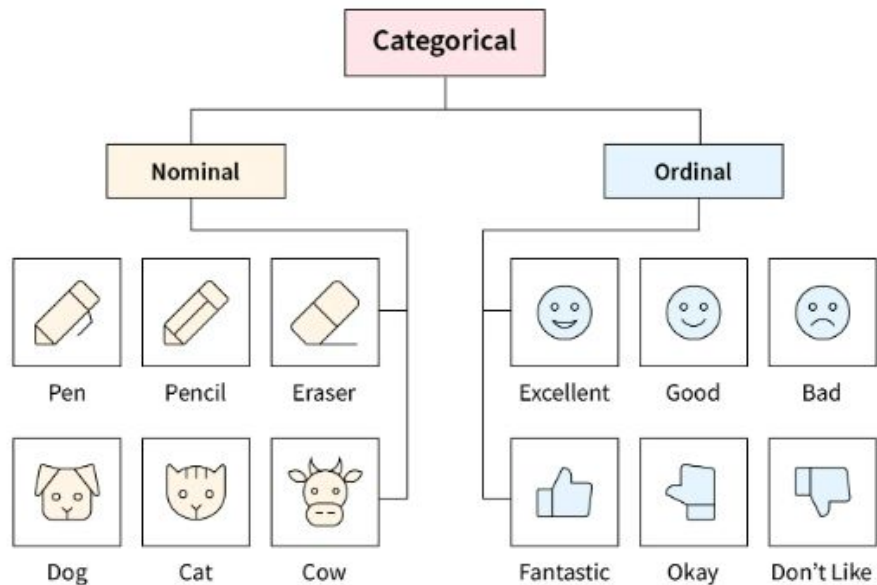
## 4. Eliminare i valori mancanti

- Se la colonna contiene più del 50% valori missing → diventa più rumorosa che informativa
- Bisogna eliminarla

# Come gestire le variabili categoriche

1. One-hot-encoding
2. Ordinal encoding

# Tipi di variabile categorica



# One-hot-encoding

id	color
1	red
2	blue
3	green
4	blue



id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Si usa con le variabili non-ordinali

# Ordinal encoding

Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

Si usa con le variabili ordinali



# Metodi per normalizzare

- Molti modelli di machine learning richiedono di avere le variabili numeriche normalizzate
  - Esempi: Regressione Lineare, SVM, KNN, K-means, DBscan
- Aiuta ad avere le variabili nello stesso range senza cambiare la distribuzione

# Normalizzazione

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Normalization



Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	0	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1

Range 0-1

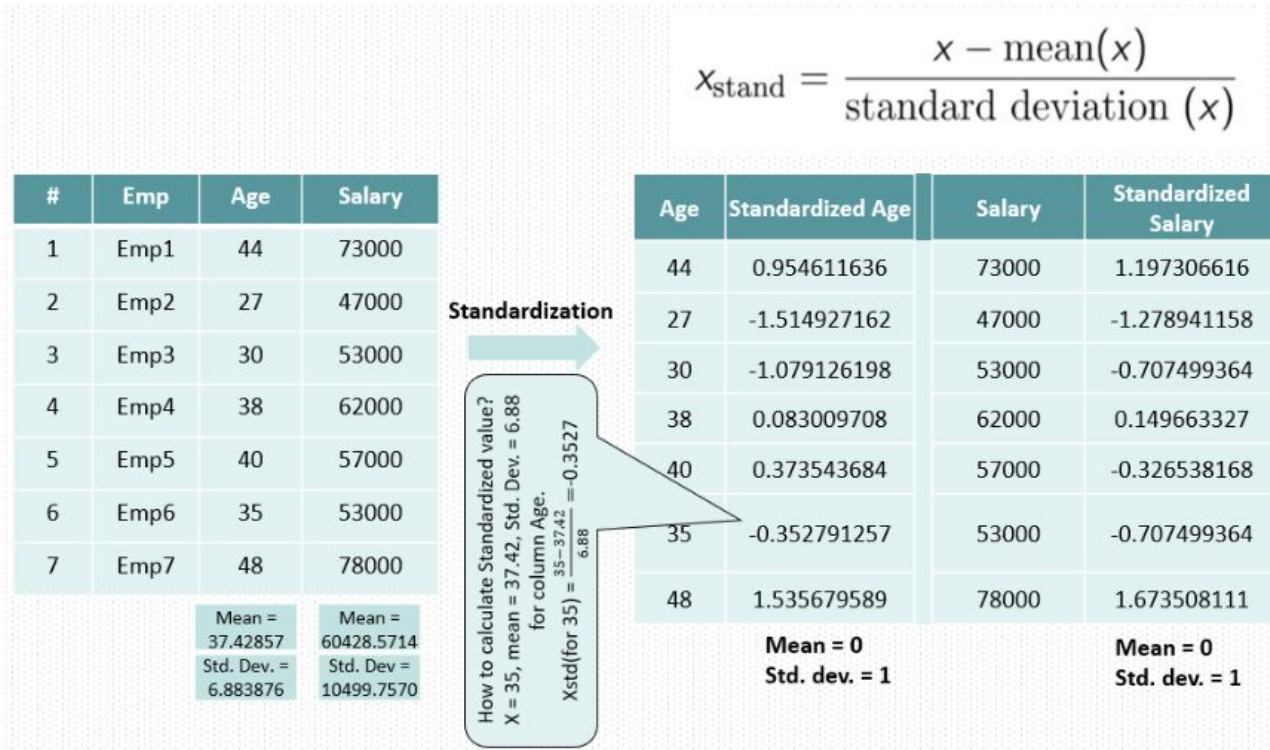
Range 0-1

How to calculate Normalized value?

X = 35, min = 27, max = 48 for column Age.

$$X_{\text{norm}}(\text{for } 35) = \frac{35 - 27}{48 - 27} = 0.3809$$

# Standardizzazione



# Robust Scaling

Si chiama “robusto” perché permette di evitare il problema della varianza alta dovuta agli outlier

The diagram illustrates the formula for Robust Scaling, showing how each part of the equation is derived from the data. The formula is  $x' = \frac{x - \text{median}(x)}{(Q3 - Q1)}$ . Annotations with arrows point to each component: 'Robust Standardised Value' points to  $x'$ ; 'Original Value' points to  $x$ ; 'Sample Median' points to  $\text{median}(x)$ ; and 'Interquartile Range =  $Q3 - Q1$ ' points to the denominator  $(Q3 - Q1)$ .

$$x' = \frac{x - \text{median}(x)}{(Q3 - Q1)}$$

Robust Standardised Value

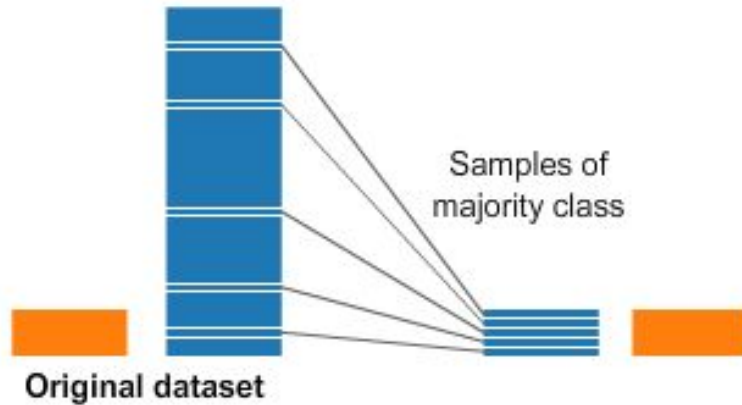
Original Value

Sample Median

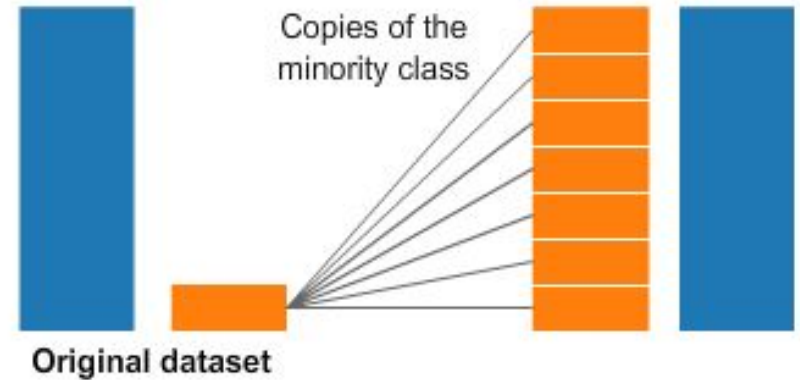
Interquartile Range =  $Q3 - Q1$

# Come gestire un dataset sbilanciato

## Undersampling



## Oversampling



# Next Steps

- Sporcatevi le mani con dati reali
- Provare a fare progetti personali di Data Science
- Creare un Portfolio con i vostri progetti