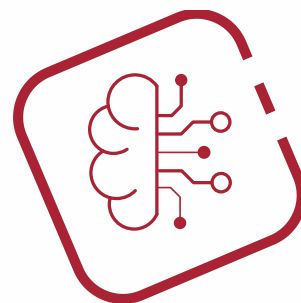
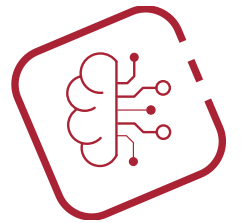


# Algoritmi per il Machine Learning

Ing Andrea Colleoni





# Pre-processing dei dati



Diagnose dirty data

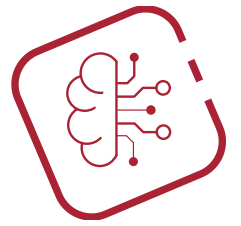


Side effects of dirty data



Clean data

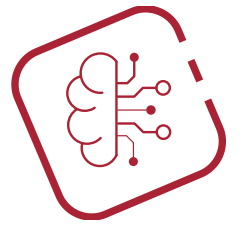
- Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.
- It is the first step marking the initiation of the process. Typically, real-world data is incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends.



# Pre-processing dei dati: tipi di dati

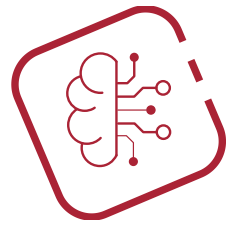
## Data type constraints

Datatype	Example	Python data type
Text data	First name, last name, address ...	<code>str</code>
Integers	# Subscribers, # products sold ...	<code>int</code>
Decimals	Temperature, \$ exchange rates ...	<code>float</code>
Binary	Is married, new customer, yes/no, ...	<code>bool</code>
Dates	Order dates, ship dates ...	<code>datetime</code>
Categories	Marriage status, gender ...	<code>category</code>

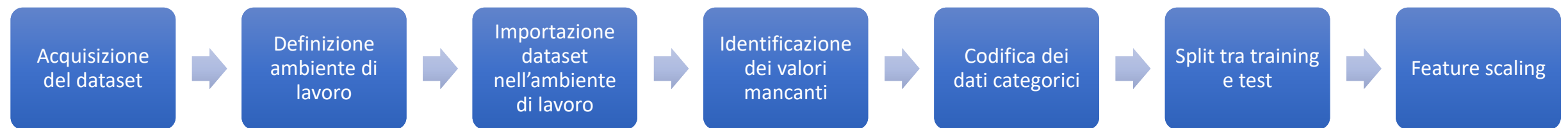


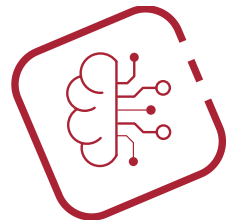
# Trasformazioni semplici

- Verifica preliminare dei tipi di dati
- Rimozione di informazioni superflue  
es: (€ 100,00 => 100,0)
- Convalida dei dati  
es: appartenenza ad un insieme, appartenenza ad un range, rispetto di un limite, minimo o massimo
- Dati non validi: come rimediare?
  - Eliminazione (solo se sono pochi)
  - Impostazione a un valore di default
  - Modifica dei vincoli



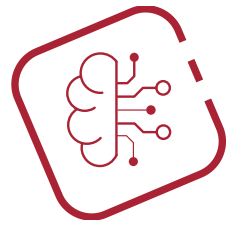
# Pre-processing dei dati





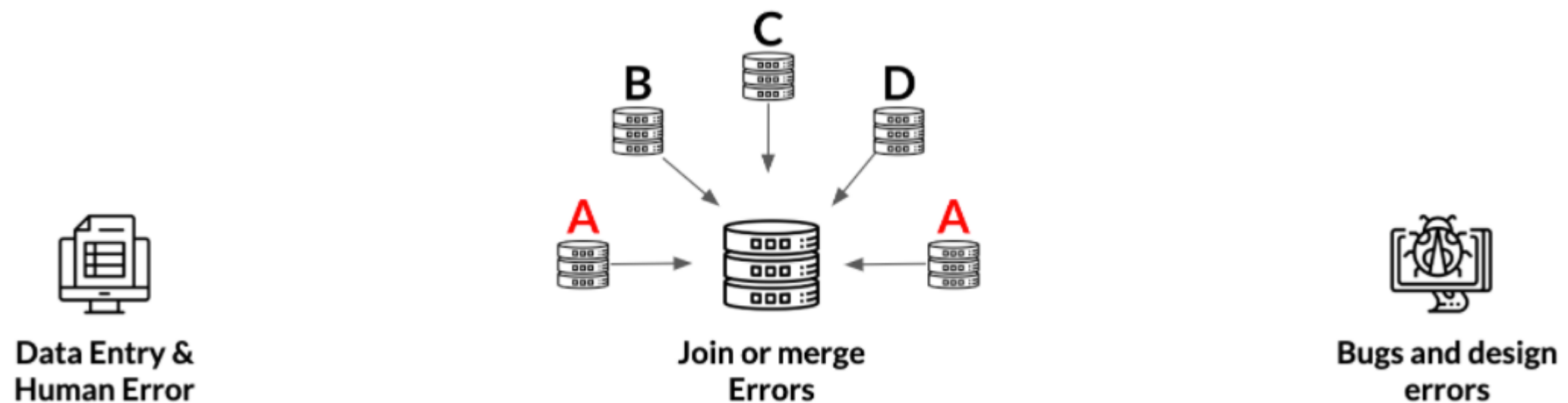
# Input mancanti

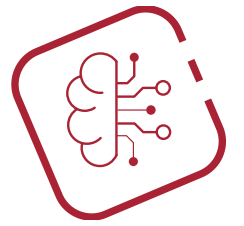
- **Deleting a particular row** – In this method, you remove a specific row that has a null value for a feature or a particular column where more than 75% of the values are missing. However, this method is not 100% efficient, and it is recommended that you use it only when the dataset has adequate samples.
- **Calculating the mean** – This method is useful for features having numeric data like age, salary, year, etc. Here, you can calculate the mean, median, or mode of a particular feature or column or row that contains a missing value and replace the result for the missing value.



# Valori duplicati

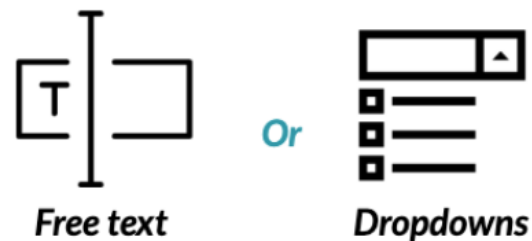
- I valori duplicati possono essere originati da molti fattori: errore umano o da errori di design o bug nel software.





# Input categorici

- Finora abbiamo sempre assunto che il vettore degli attributi (ingresso) sia sempre composto di quantità reali. Può accadere però che alcune osservazioni non riguardino grandezze continue, ma qualità discrete e non ordinate (colore, sesso, specie).
- L'algoritmo KNN, per esempio, si basa però sul calcolo di una distanza in uno spazio, quindi è necessario convertire le grandezze categoriche in una o più grandezze reali.

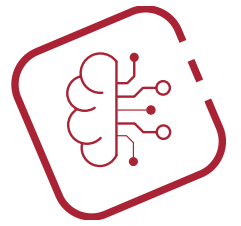


**Data Entry Errors**



**Parsing Errors**





# Membership e consistency

- Le categorie possono appartenere ad un dominio
  - Esempio del JOIN in algebra relazionale
- Errori di appartenenza
- Errori di «case»

**Capitalization:** 'married', 'Married', 'UNMARRIED', 'unmarried' ..

- Spazi in eccesso

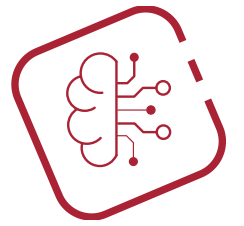
**Trailing spaces:** 'married ', 'married', 'unmarried', ' unmarried' ..

- Remapping di molte categorie in un numero

**Map categories to fewer ones:** reducing categories in categorical column.

operating\_system column is: 'Microsoft', 'MacOS', 'IOS', 'Android', 'Linux'

operating\_system column should become: 'DesktopOS', 'MobileOS'

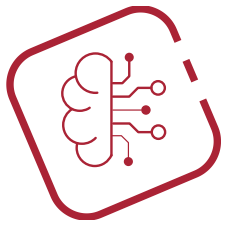


# Uniformità e cross field validation

- Temperature su varie scale?
- Date in vari formati?
- Grandezze in diverse unità di misura?
- I dati sono legati tra loro (se non sono 1NF, 2NF,..)

	flight_number	economy_class	business_class	first_class	total_passengers
0	DL140	100	60	40	200
1	BA248	130	100	70	300
2	MEA124	100	50	50	200
3			70	90	300
4			100	20	250

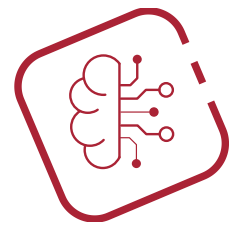
	user_id	Age	Birthday
0	32985	22	1998-03-02
1	94387	27	1993-12-04
2	34236	42	1978-11-24
3	12551	31	1989-01-03
4	55212	18	2002-07-02



# Encoding categorical

The techniques that you'll cover are the following:

- Replacing values
- Encoding labels
- One-Hot encoding
- Binary encoding
- Backward difference encoding
- Miscellaneous features

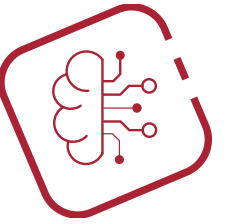


# Replace

	carrier	tailnum	origin	dest
0	AS	N508AS	PDX	ANC
1	US	N195UW	SEA	CLT
2	UA	N37422	PDX	IAH
3	US	N547UW	PDX	CLT
4	AS	N762AS	SEA	ANC



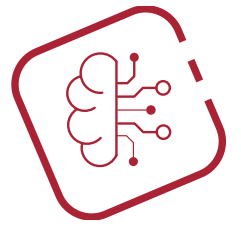
```
replace_map = {'carrier': {'AA': 1, 'AS': 2, 'B6': 3, 'DL': 4,  
                           'F9': 5, 'HA': 6, 'OO': 7, 'UA': 8, 'US': 9, 'VX': 10, 'WN': 11}}
```



# Encoding labels

- Invece che avere una mappa, ad ogni nuovo valore della categoria, genero un nuovo valore numerico dell'etichetta
- Hanno un peso questi valori numerici??





# Esempio

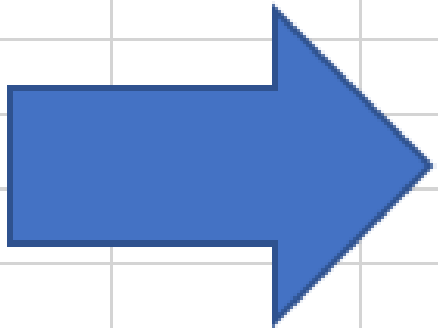
Come mappare con numeri reali i colori dei fiori?

	A	B	C	D
1	lung petalo ▾	colore ▾	altezza ▾	spine ▾
2	2,1	rosso	5	no
3	3,2	azzurro	5,4	no
4	2,6	rosso	7,2	sì
5	1,9	giallo	4,4	no
6	3	azzurro	5,6	sì
7	2,5	giallo	6	no

## ntazione unaria

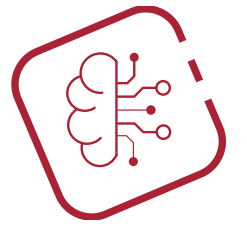
zata in tre colonne numeriche;  
ale1 se l'attributo categorico è  
me eccezione, se l'attributo è  
na sola colonna numerica,  
possibili valori categorici.

D	E	F	G	H	I	J	K	L	M	
e				lung petalo	rosso	azzurro	giallo	altezza	spine	
				2,1	1	0	0	5	0	
				3,2	0	1	0	5,4	0	
				2,6	1	0	0	7,2	1	
				1,9	0	0	1	4,4	0	
				3	0	1	0	5,6	1	
				2,5	0	0	1	6	0	



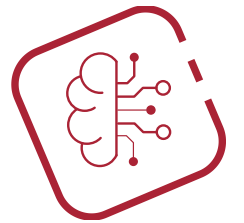






# Normalizzazione

- Con **normalizzazione** (o ridimensionamento) intendiamo la serie di operazioni che permettono di valutare il dataset semplificandolo e impedendo ai valori sproporzionati o fuori scala di influenzare col proprio peso il resto dei dati che ha valori “normali”.
- Il problema si risolve applicando il **Feature Scaling**.

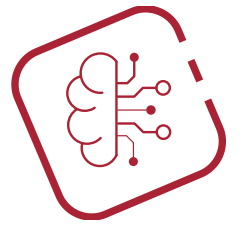


# Normalizzazione MIN-MAX

- è il metodo più semplice, i dati vengono **ridimensionati e scalati** su un intervallo fisso, in genere **[0, 1]**. Questa normalizzazione migliora l'accuratezza dell'analisi grazie alla migliore distribuzione dei dati.

$$m_j = \min_{i=1,\dots,m} x_{ij}; \quad M_j = \max_{i=1,\dots,m} x_{ij} \quad x'_{ij} = \frac{x_{ij} - m_j}{M_j - m_j}.$$

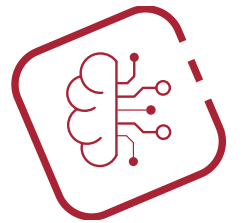
- In alcuni casi, il valore 0, anche se non compare mai nella matrice, ha un significato particolare e si desidera mantenerlo. Allora, è sufficiente porre  $m_j = 0$ .



# Normalizzazione MEDIA

- In alcune situazioni si può preferire mappare i dati su un intervallo  $[-1, 1]$  e utilizzando la media dei valori osservati

$$z = \frac{x - \text{media}(x)}{\max(x) - \min(x)}$$

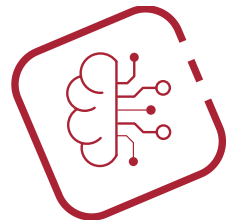


# Standardizzazione Z-SCORE

- Ridimensiona gli attributi in modo che il valore medio sia 0 e la deviazione standard 1.

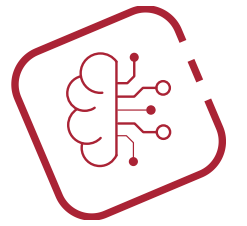
$$z = \frac{x - \mu}{\sigma}$$

- Dove  $\mu$  è la **media** dei campioni,  $\sigma$  è la **deviazione standard** dei dati di addestramento e  $x$  è il valore che si vuole standardizzare.



# Difetti della Normalizzazione

- L'effetto indesiderato più importante dell'uso indiscriminato di questa tecnica è la perdita di alcuni dati.  
Infatti tenendo conto del fatto che i dati vengono compressi in piccoli intervalli, viene ridotta la variazione standard e viene imposto un "peso" uguale per tutte le caratteristiche, capiamo che non è una cosa da usare con leggerezza.
- L'effetto è detto anche **sensibilità agli "outlier"**: se nella colonna è presente un valore molto più grande degli altri, questo viene mappato sul valore normalizzato 1, mentre tutti gli altri valori sono mappati vicino allo zero.



# Insiemi di validazione

- se il dataset è suddiviso fra insiemi di addestramento e validazione, bisogna assicurarsi che i minimi e i massimi siano calcolati soltanto sulle righe di addestramento
- Gli stessi minimi e massimi andranno poi utilizzati anche nella normalizzazione dei valori di validazione, che potrebbero quindi uscire dall'intervallo  $[0,1]$
- In generale, una volta ottenuti i parametri per la normalizzazione o per la standardizzazione, è fondamentale ricordarli in modo da applicare le stesse trasformazioni anche a nuovi insiemi di dati